APPLICATION OF SOMATIC VARIANT ANALYSIS IN CANCER EXOMES

YU WILLIE SHUN SHING

NATIONAL UNIVERSITY OF SINGAPORE

2015

APPLICATION OF SOMATIC VARIANT ANALYSIS IN CANCER EXOMES

YU WILLIE SHUN SHING

(B.Sc., UNIVERSITY OF CALIFORNIA, BERKELEY

M.Sc., BOSTON UNIVERSITY)

A THESIS SUBMITTED FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL OF INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2015

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

_____

YU Willie Shun Shing

28 December, 2014

Acknowledgements

First of all, I like to thank my father and mother for their unwavering love, support and patience over the years; it has been a long journey and I have finally made it.

I like to thank my uncle Michael, aunt Irene, Bernie, Li-Ann and Bebo for making me feel welcome in Singapore and helped make this country like a second home for me.

Thank you to my supervisors, Prof. Patrick Tan and Prof. Teh Bin Tean, for giving me the once-in-a-lifetime opportunity to do research at and to witness firsthand the birth of the cancer genomics era.

Thank you to Prof. Steve Rozen for your constructive advice on the computational aspects of cancer genomics. I look forward to working with you in the future.

Thank you Lian Dee for being there for me over the years; talking to you everyday has pushed me to keep in touch with experimental biology and made me realize it is an important partner to bioinformatics.

Finally, thank you Singapore for creating the environment where genomics research is not only possible but thriving. Happy 50^{th} birthday.

Two Quotes for Scientific Investigators

"*The fact that the scientific investigator works 50 percent of his time by non-rational means is, it seems, quite insufficiently recognized.*

Intuition, like a flash of lightning, lasts only for a second. It generally comes when one is tormented by a difficult decipherment and when one reviews in his mind the fruitless experiments already tried. Suddenly the light breaks through and one finds after a few minutes what previous days of labor were unable to reveal.

And, Randy's favorite,

As to luck, there is the old miners' proverb: 'Gold is where you find it.' "

Neal Stephenson, Cryptonomicon

"*TWO roads diverged in a yellow wood,*
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference. "

Robert Frost, The Road Not Taken

Table of Contents

Acknowledgements ……. i

Two Quotes for Scientific Investigators ……..... ii

Table of Contents .……. iii

Summary .…….. vi

List of Figures ...……...……. viii

List of Tables ……...…….. x

List of Abbreviations ..……...…… xii

Chapter One: Introduction ……...……..... 1

1.1 Somatic theory of evolution and the central role of the genome in cancer development ...…….. 2

1.2 Development of technologies to catalog and understand somatic mutations in cancer ..……..... 3

1.3 Description of general variant discovery pipeline used in analysis of next generation whole-exome sequencing data ….....……..... 8 - 16

1.3.1 Sequenced DNA data in FASTQ format ...……...…….…….. 9

1.3.2 Alignment of DNA fragments to the reference genome ...……...…… 10

1.3.3 PCR-duplicate removal ...……...……... 10

1.3.4 Variant calling and separation of somatic, germline and SNP variants ……. 11

1.3.5 Visualization and estimation of copy number and loss of heterozygosity changes ...……...…… 13

1.3.6 Inferring mutational processes in a tumour ……...……..... 15

1.4 Application of variant discovery pipeline …..……...…… 17 - 20

1.4.1 Summary of chapter two ...……...……. 17

1.4.2 Summary of chapter three ...……...….... 18

1.4.3 Summary of chapter four ..…… 19

Chapter Two: First Somatic Mutation of E2F1 in a Critical DNA Binding Residue Discovered in Well- Differentiated Papillary Mesothelioma of the Peritoneum ……..……..… 24

2.1 Introduction ...…...…… 25

2.2 Results ...…...…. 27 - 31

2.2.1 WDPMP whole-exome sequencing: mutation landscape changes big and small ...…...…..… 27

2.2.2 E2F1 R166H mutation affects critical DNA binding residue ..……. 28

2.2.3 R166H mutation is detrimental to E2F1's DNA binding ability and negatively affects downstream target gene expression ...…….. 30

2.2.4 Cells over expressing E2F1 R166H mutant show massive protein accumulation and increased protein stability ...…….. 31

2.2.5 Over expression of E2F1 R166H mutant does not adversely affect cell proliferation ...…… 32

2.3 Discussion ...……...……...…… 33 - 38

Chapter Three: Exome Sequencing of Liver Fluke-associated Cholangiocarcinoma …… 52

3.1 Introduction ...…..……. 53

3.2 Results ...…...…. 55 - 58

3.2.1 Clinical samples and information ...…….…….. 55

3.2.2 CCA whole-exome analysis ...…….…….. 55

3.2.3 Mutational analysis of CCA discovery set ..……. 56

3.2.4 Prevalence analysis of somatic mutations found in CCA discovery set ……. 56

3.2.5 Mutational landscape comparison between *O. Viverrini*-associated cholangiocarcinoma, pancreatic ductal adenocarcinoma and hepatitis C virus-associated hepatocarcinoma ...…….... 58

3.3 Discussion ...…….....…… 59 - 67

Chapter Four: Whole-exome sequencing studies of parathyroid carcinomas reveal novel *PRUNE2* mutations, distinctive mutational spectra related to APOBEC-catalyzed DNA mutagenesis and mutational enrichment in kinases associated with cell migration and invasion ...…...…… 93

4.1 Introduction ...…..…….…… 94

4.2 Results ...…… 95 - 99

4.2.1 Clinical samples and information ...…….. 95

4.2.2 PC whole-exome analysis 96

4.2.3 *CDC73* mutational status and its effect on the PC exome 97

4.2.4 Novel recurrent mutations of *PRUNE2* in PC 97

4.2.5 Kinase family is recurrently mutated in PC independent of *CDC73* mutation status 98

4.2.6 APOBEC mutational signature in PC 99

4.3 Discussion ..... 100 - 106

Chapter Five: General Discussion and Future Work 148

5.1 General discussion ..... 149 - 155

5.2 Hypothetical research proposal ..... 156 - 162

5.2.1 Title 156

5.2.2 Introduction 156

5.2.3 Conjecture 158

5.2.4 Proposed mechanism 158

5.2.5 Proposed milestones 159

5.2.6 Proposed experiments 159

5.2.7 Conclusion 161

References 163- 184

Summary

Whole-exome sequencing has revolutionized cancer research to accelerate the exploration and cataloging of somatic variants across multiple cancer samples. As the use of whole-exome sequencing is becoming increasingly prevalent, two natural questions arises: One is how to process and analyze the ever growing volume of sequencing data generated and the other is how to apply the results of the analysis to cancer research.

To start to answer the former, a general single nucleotide variant discovery pipeline is proposed to process and analyze whole-exome data; the results from this pipeline will be the starting points for downstream analysis such as functional analysis and cataloging of mutations, estimating copy number and loss of heterozygosity, and inferring mutational processes.

To start answering the latter question, three published studies will illustrate three possible applications of whole-exome sequencing.

The first study is whole-exome sequencing of well differentiated papillary mesothelioma of the peritoneum. The first *E2F1* somatic mutation was found and predicted to result in a R166H change to the protein product. R166 position is highly conserved and protein homology modeling indicates the position is a critical DNA contact point for binding. Downstream experimentation confirmed loss of DNA binding for E2F1 R166H mutant and also discovered that E2F1 mutant is much more stable than its wild type counterpart. This study highlights a collaborative application of bioinformatics with experimental biology where bioinformatics quickly predicts

the functional consequences of a mutation and presents high confidence hypothesis for experimental biologists to consider.

The second study is whole-exome sequencing of *Opisthorhis viverrini* (OV) - related cholangiocarcinoma (CCA); a malignant bile duct cancer that is endemic in northeastern Thailand due to OV infestation as a result of local dietary habits. In addition to finding recurrently mutated cancer-related genes such as *TP53* (44.4% mutation rate), *KRAS* (16.7%) and *SMAD4* (16.7%), another 10 novel recurrently mutated genes were cataloged such as *MLL3* (14.8%), *ROBO2* (9.3%), *RNF43* (9.3%), *PEG3* (5.6%) and *GNAS* oncogene (9.3%). Similarities in mutated genes and base substitution spectra between OV-related CCA, pancreatic ductal adenocarcinoma (PDAC) suggests therapies effective for PDAC may also be effective in OV-related CCA. Minnelide and LGK974, two therapeutics showing effectiveness against pancreatic cancer with *KRAS/TP53* mutations or *RNF43* mutations respectively, were suggested to be effective in treating CCAs with similar mutational background. This study highlights the medical translational application of whole-exome sequencing and analysis.

The third study outlines the mutational landscape of parathyroid carcinoma (PC) through PC whole-exome sequencing. *PRUNE2* is revealed to be the novel second recurrently mutated gene in PC with germline and somatic mutations clustered around an evolutionary conserved region of the protein. In addition, mutations to members of the kinase family related to cell migration and invasion were found to be enriched. APOBEC mediated mutagenesis was implicated for the first time in a subset of PC patients with high mutational burden and early age onset of disease. This study highlights the application of whole-exome analysis in opening new avenues of research not previously considered under hypothesis-driven approaches.

List of Figures

Figure 1.1: The ten hallmarks of cancer as defined by Hanahan and Weinberg 21

Figure 1.2: General variant discovery and analysis pipeline used for whole-exome sequencing data sets 22

Figure 1.3: FASTQ example and quality score encoding 23

Figure 2.1: Cumulative WDPMP exome coverage for tumor, normal and purified tumor cells 39

Figure 2.2: Compact representation of WDPMP exome using Hilbert plot 40

Figure 2.3: Sequencing coverage at *CDKN2A*, *RASSF1* and *NF2* 41

Figure 2.4: Sanger sequencing validation of somatic single nucleotide variants found in *E2F1*, *PPFIBP2* and *TRAF7*..... 42

Figure 2.5: Location and conservation analysis of E2F1 R166H 43

Figure 2.6: Visualization of p.Arg166His mutation location in E2F1 44

Figure 2.7: Homology modelling of wild type and mutant E2F1 around R166 residue 45

Figure 2.8: E2F1 R166 mutation affects binding efficiency on to promoter targets 46

Figure 2.9: Accumulation of mutant E2F1 protein in cells due to increased stability of E2F1 R166 mutation 47

Figure 2.10: Relative expression of E2F1 wild type or E2F1 mutant after co-transfection with EGFP in MSTO-211H and NCI-H28 48

Figure 2.11: Over expression of E2F1 R166H mutant in two mesothelial cell lines 49

Figure 3.1: Mutational landscape of OV-associated CCA 68

Figure 3.2: Proportion comparisons of mutational spectra in OV-associated CCA, PDAC and HCV-associated HCC 69

Figure 4.1: Mutational landscape of PC 107

Figure 4.2: Copy number estimation of chromosome 1 for each whole-exome sequenced PC sample using ASCAT 2.0 ..... 108- 112

Figure 4.3: Predicted LOH of chromosome 9 for sample 4 using ASCAT 2.0 113

Figure 4.4: Twenty eight mammalian species conservation analysis of PRUNE2 residue positions (Ser450, Val452, Gly455) corresponding to the three non-synonymous mutations (c.1349G>A, c.1354G>A, c.1364G>A) found in PC 114

Figure 4.5: Distribution of base substitutions in PC 115

Figure 4.6: Mutational signatures found by Emu 116

Figure 5.1: Life cycle of LINE-1 retrotransposon 162

List of Tables

Table 2.1: Overall WDPMP Exome Sequencing Summary ...…… 50

Table 2.2: Putative somatic nonsynonymous mutations found using the single nucleotide variant discovery pipeline ...…….. 51

Table 3.1a: Clinical information of the discovery set consisting of 8 patients diagnosed OV-associated CCA ...……... 70

Table 3.1b: Clinical information of the prevalence set consisting of 46 patients diagnosed OV-associated CCA ...…… 71 - 72

Table 3.2: Whole-exome sequencing summary of 8 matched pairs of OV-associated CCAs ...……... 73

Table 3.3: Nonsynonymous somatic mutations identified and validated in the discovery set ...……. 74 - 85

Table 3.4: Recurrently mutated genes as well as known recurrently mutated genes found in 54 OV-associated CCAs ...…… 86 - 90

Table 3.5: Frequency of recurrently mutated genes in OV-associated CCA, PDAC and HCV-associated HCC ...…….. 91

Table 3.6: Mutation spectra in OV-associated CCA, PDAC and HCV-associated HCC ……. 92

Table 4.1: Patient information for PC discovery set ...……. 117

Table 4.2: Sample information for PC validation set ...…… 118

Table 4.3: PC whole-exome sequencing summary ...……... 119

Table 4.4: Exome dbSNP concordance of whole-exome sequenced PC samples ... 120

Table 4.5: Validated single nucleotide variants for whole-exome sequenced PC samples ……..……. 121 - 136

Table 4.6: Zygosity summary of validated somatic mutations for whole-exome sequenced PC samples ……. 137

Table 4.7: Recurrent mutations in *CDC73* and *PRUNE2* for whole-exome sequenced PC ...…….. 138

Table 4.8: Mutated genes related to DNA damage repair in sample 7b ...……... 139

Table 4.9: Gene classification analysis of validated somatic mutations in PC ….. 140 - 143

Table 4.10: Kinase mutations in PC ...…….. 144

Table 4.11: Gene classification analysis of validated somatic mutations in PC excluding sample 7b …...……. 145 - 147

List of Abbreviations

A	Adenine
A or Ala	Alanine
ABL1	Abelson murine leukemia viral oncogene homolog 1
ANOLEA	Atomic Non-Local Environment Assessment
AP1	Activating protein-1
APAF1	Apoptotic peptidase activating factor 1
APC	adenomatous polyposis coli
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
APOBEC3C	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3C
APOBEC3D	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3D
APOBEC3G	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G
ARID2	AT rich interactive domain 2 (ARID, RFX-like)
ASCAT	Allele-Specific Copy number Analysis of Tumors
BAF	B allele frequency
BAP1	BRCA1-associated protein 1
BCH	BNIP-2 and Cdc42GAP Homology
BCR	Breakpoint Cluster Region
BGI	Beijing Genome Institute
BMCC1	Bcl2-/adenovirus E1B nineteen kDa-interacting protein 2 (BNIP-2) and Cdc42GAP homology BCH motif-containing molecule at the carboxyl terminal region 1)
BRAF	serine/threonine-protein kinase B-Raf
C	Cytosine
C or Cys	Cysteine

CASR	Calcium sensing receptor
CCA	Cholangiocarcinoma
CCNE1	Cyclin E1
CDC42BPA	CDC42 binding protein kinase alpha (DMPK-like)
CDC73	Cell division cycle 73
CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)
CDK6	Cyclin dependent kinase 6
CDKN2A	Cyclin-dependent kinase inhibitor 2A
CGP	Cancer Genome Project
CHEK2	Checkpoint kinase 2
ChIP	Chromatin immunoprecipitation
CI	Confidence interval
COSMIC	Catalogue of somatic mutations in Cancer
CTNNB1	Catenin (cadherin-associated protein), beta 1, 88kDa
D or Asp	Aspartic Acid
DAVID	Database for Annotation, Visualization and Integrated Discovery
dbSNP	Single nucleotide polymorphism database
ddNTPs	di-deoxynucleotidetriphosphates
DMXL1	Dmx-like 1
DNA	deoxyribonucleic acid
dNTPs	deoxynucleosidetriphosphates
E2F1	E2F transcription factor 1
E2F4	E2F transcription factor 4, p107/p130-binding
EGFP	Enhanced green fluorescent protein
Emu	Expectation maximization

FFPE	Formalin fixed paraffin embedded
G	Guanine
G or Gly	Glycine
GATK	Genome analyzer toolkit
GNAS	GNAS complex locus
GROMOS	Groningen Molecular Simulation
HCC	Hepatocarcinoma
HCV	Hepatitis C virus
HDAC2	Histone deacetylase 2
HDAC4	Histone deacetylase 4
His	Histidine
HPT	Primary hyperthyroidism
HPT-JT	Hyperthyroidism-jaw tumor syndrome
HRAS	Harvey rat sarcoma viral oncogene homolog
I or Iso	Isoleucine
IDH1	Isocitrate dehydrogenase 1
IL17RA	Interleukin 17 receptor A
JAK1	Janus kinase 1
KAP1	KRAB-associated protein-1
kDa	Kilo Daltons
KRAB	Krueppel-associated box
KRAS	V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
L or Leu	Leucine
Lbc	A kinase (PRKA) anchor protein 13
LIMK2	Lim kinase domain 2

LINE-1	Long interspersed nuclear elements-1
LOH	Loss of heterozygosity
LTK	Leukocyte receptor tyrosine kinase
M or Met	Methionine
MAP3K11	Mitogen-activated protein kinase kinase kinase 11
MEKK3	Mitogen-activated protein kinase kinase kinase 3
MEN1	Multiple endocrine neoplasia type 1
MEN2A	Multiple endocrine neoplasia type 2A
MH2	Mad homology domain 2
MLL3	Lysine (K)-specific methyltransferase 2C
MPM	Malignant peritoneal mesothelioma
N or Asn	Asparagine
NDC80	NDC80 kinetochore complex component
NEDL1	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1
NF2	Neurofibromatosis type 2
NGF	Neuronal growth factor
NLRP1	NLR family, pyrin domain containing 1
NMF	Nonnegative matrix factorization
ODZ3	Teneurin transmembrane protein 3
OR	Odds ratio
ORF1	Open reading frame 1
ORF2	Open reading frame 2
OV	Opisthorhis viverrini
P or Pro	Proline
PA	Parathyroid adenoma

PARP1	poly (ADP-ribose) polymerase 1
PC	Parathyroid carcinoma
PCDHA13	Protocadherin alpha 13
PCM1	Pericentriolar material 1
PCR	Polymerase chain reaction
PDAC	Pancreatic ductal adenocarcinoma
PEG3	Paternally expressed 3
POLH	Polymerase (DNA directed), eta
PORCN	Porcupine homolog (Drosophila)
PPFIBP2	PTPRF interacting protein, binding protein 2 (liprin beta 2)
PRMT6	Protein arginine methyltransferase 6
PRUNE2	Prune homolog 2 [Drosophila]
PTEN	Phosphatase and tensin homolog
PTH	Parathyroid hormone
PTPRM	Protein tyrosine phosphatase, receptor type, M
Q or GLN	Glutamine
R or Arg	Arginine
RADIL	Ras association and DIL domains
RASSF1A	Ras association domain family 1 isoform A
RB1	Retinoblastoma 1
RhoA	Ras homolog family member A
RIOK3	RIO kinase 3
RNA	Ribonucleic acid
RNF43	Ring finger protein 43
ROBO2	Roundabout, axon guidance receptor, homolog 2 (Drosophila)

rtTA	recombinant tetracycline controlled transcription factor
S or Ser	Serine
SAD	SMAD4 activation domain
SHANK3	SH3 and multiple ankyrin repeat domains 3
SIAH1A	Siah E3 ubiquitin protein ligase 1A
SIRT1	Sirtuin 1
SMAD4	SMAD family member 4
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV40	Simian vacuolating virus 40
T	Thymine
Tet	Tetracycline-Controlled Transcription Activation
TFDP1	Transcription factor, Dp1
TIE1	Tyrosine kinase with immunoglobulin-like and EGF-like domains 1
TP53	Tumor protein p53
TRAF7	TNF receptor-associated factor 7, E3 ubiquitin protein ligase
TRE	Tetracycline responsive element
UTR	Untranslated region
V or Val	Valine
WD40	Beta-transducin repeat 40
WDPMP	Well differentiated papillary mesothelioma of the peritoneum
XIRP2	Xin actin-binding repeat containing 2
Y or Tyr	Tyrosine

Chapter One: Introduction

1.1 Somatic theory of evolution and the central role of the genome in cancer development

Majority of cells within an organism has only limited replicative potential; the replicative trajectory of these cells inevitably leads to a state of senescence, where the cell can no longer divide but still alive and metabolically active, and finally to apoptosis, the process of programmed cell death. During these cells' limited lifetime, they can accumulate changes to its genome. Some of the earliest observations of these genomic changes were observed through microscopy in studies by Hansemann and Boveri (1,2). By observing the characteristics of cancer cells undergoing cell divisions, they noticed the chromosomes of cancer cells looked markedly different from the chromosomes of normal cells. This led to the conjecture that cancer cells are caused by genomic abnormalities. Following the elucidation of deoxyribonucleic acid (DNA) structure as well as its role as the vehicle of inheritance, studies showed that genomic DNA changes or somatic mutations can come about due to endogenous processes, such as mistakes in DNA replication during cell division, or exogenous processes, such as radiation or chemical insults (3,4,5). The key study demonstrating the importance of abnormal genes in the development of cancer is the identification of a naturally occurring sequence change in the form a guanine to thymine single base substitution that results in a glycine to valine amino acid change in codon 12 of the Harvey rat sarcoma viral oncogene homolog (HRAS) protein; insertion of total genomic DNA containing this genetic mutation into NIH3T3 cells, a phenotypically normal primary mouse embryonic fibroblast cells, resulted in conversion to cancer cells (6).

Some somatic mutations confer increased survival and proliferation capabilities in cells that acquired these mutations when compared with cells without

these mutations in the context of the local tissue environment. A classic example is the development of chronic myeloid leukemia through a specific genomic translocation event between chromosome 9 and chromosome 22 creating the chromosomal anomaly known as the Philadelphia chromosome (7). This key transformation event results in the creation of a fusion gene between the breakpoint cluster region (*BCR*) gene and the Abelson murine leukemia viral oncogene homolog 1 (*ABL1*) gene where the resulting fusion protein product drives unregulated cell division (8). Cells acquiring through mutations the ability to escape the normal cell fate of senescence and apoptosis will hold a tremendous evolutionary advantage over non-mutated cells in propagating their genetic material; therefore, cancer is a group of mutated cells with advantageous mutations that sweeps through a cell population, pushing aside cells lacking these mutations, to become the dominant cell type within the context of its environment. In evolutionary terms, the development of cancer is due to the process of positive selection or selection of adaptive traits that overcome the replicative or growth limitations imposed on a cell. These adaptive traits displayed by a cancer cell were classified into ten distinct categories by Hanahan and Weinberg in two seminal review articles (9,10) (Figure 1.1).

1.2 Development of technologies to catalog and understand somatic mutations in cancer

The key study demonstrating a single somatic base substitution to *HRAS* is sufficient for cancerous transformation led to the continuous search for and cataloging of gene mutations that is still ongoing. There are two critical technologies that first enabled and subsequently accelerated our ability to discover these genetic mutations.

The first technology is DNA sequencing or the capability to generate single base resolution of a DNA molecule. There are two methods of DNA sequencing developed during the 1970's. The Maxam-Gilbert method employs chemical treatment of radiolabelled DNA in four reactions to generate breaks at one or two of the four nucleotides. Size separations of the chemically treated fragment were performed using acrylamide gels with visualization through gel exposure to X-ray film (11). The Sanger method employs the use of modified di-deoxynucleotidetriphosphates (ddNTPs) to introduce premature terminations of DNA elongation at specific nucleotides where normal deoxynucleosidetriphosphates (dNTPs) are substituted for ddNTPs (12,13). There are four separate sequencing reactions where each reaction contains one of the four possible ddNTPs, that is radio or fluorescent labeled, as well as a mixture of the four normal nucleotides, the DNA template of interest, primer oligonucleotides and DNA polymerases. After several rounds of DNA template extension of each reaction mixture will result in DNA fragments of various sizes ending at the site of ddNTP insertion; size separation using acrylamide gels of the four reactions will enable the DNA sequence information to be deduced. Due to the relative ease of use and lower use of radioactive and toxic chemicals, the Sanger method became the dominant method of DNA sequencing that is still in use today.

The second technology is polymerase chain reaction (PCR) or the ability to amplify small quantities of DNA fragments by several orders of magnitude. First proposed by Kary Mullis in 1983, the method employ the heat-stable DNA polymerases to replicate DNA and selective amplification is achieved by use of oligonucleotides or a “primer” complementary to nearby DNA region of interest (14,15). This method effectively eliminated the experimental biology bottle neck of

limited DNA availability and enabled much greater latitudes of experimental manipulations.

Subsequent improvements and automation to the above two discoveries enabled the application extension from examination of DNA sequences at a gene level to the total DNA examination of an organism. In 1990, the publicly funded Human Genome Project was started with the goal of sequencing and identifying the over three billion nucleotides present in the human genome. In competition with the privately funded Celera Genomics, who started sequencing the human genome in 1998, both sides announced their sequencing draft of the human genome in February 2001 and published their findings detailing methods used in production and analysis of the draft sequence (16,17).

The availability of a human reference genome accelerated the study and cataloging of genetic alterations in human cancer genomes in two ways. One, the reference genome provide a single template for PCR primer design. This enables an efficient, systematic design of primers with sufficient coverage to amplify larger and larger portions of the protein coding regions in the human genome. In combination with automated DNA-sequencing instruments based on the Sanger method, these technologies enables a broader simultaneous sampling of the cancer genome through sequencing of gene families, such as kinomes, to eventually sequencing most coding exons of the genomes, now commonly called exomes (18,19).

Two, the reference human genome is a template where all subsequently sequenced human DNA samples can be computationally mapped and compared against. There is no longer a necessity to de-novo assemble each new sequenced human genome of interest resulting in a tremendous saving in computational time;

with the substantial savings in computational time, genomic studies of a large part or even the whole of the protein coding regions across a cohort of samples became possible. Such genomic studies ranged from targeted screenings of hundreds of genes in hundreds of cancer samples to entire exome screens (~22,000 protein coding genes) in a targeted cancer class of 10-20 samples (20,21). While these studies were successful in finding single nucleotide mutations in numerous cancer genes, there are two point mutation discoveries in two separate genes that became the standard bearers for advocates of systematic mutational screens as the discovery of both mutations eventually led to development of targeted therapeutics approved for medical use or currently undergoing clinical trials.

The first point mutation was found to occur in over 80% of melanomas that resulted in a valine to glutamic acid change in position 600 of the serine/threonine-protein kinase B-Raf (BRAF) protein (22); Vemurafenib, a targeted inhibitor specific for BRAF with V600E mutation, was developed in 2006, only 4 years after the mutation's initial report, and received government approval for melanoma treatment in 2011 (23,24,25). The second point mutation was found in the isocitrate dehydrogenase 1 (*IDH1*) gene resulting in the arginine residue changing to a histidine residue at position 132 of the protein product; this gene was found to be recurrently mutated using exome screening of 22 glioblastoma multiforme samples in 2008 initially and with subsequent studies revealing this gene to be also recurrently mutated in acute myeloid leukemia and cholangiocarcinoma (21,26,27). A targeted inhibitor of IDH1 with R132H mutation was first reported in 2013 with the inhibitor currently undergoing Phase I clinical trials as of December 2014 (28,29,30).

While there are significant knowledge to be gained from large scale systematic sequencing, more ambitious whole-exome or even whole-genome screening through a

large cohort of involving hundreds of cancer samples remained out of reach due the low throughput and high costs associated in using automated Sanger type capillary sequencing technology. The introduction of massively parallel sequencing technologies or next generation sequencing by companies such as Roche, Illumina and Applied Biosystems, resulted the great leap forward in increased throughput and lowered cost that allowed large scale screenings across large sample numbers to become a reality. The common principle uniting these novel technologies is the concept of shotgun sequencing: the random fragmentation of a genome followed by sequencing a short stretch of DNA, called a read, for large numbers of these DNA fragments such that each base in the reference human genome is covered several times. This "shotgun sequencing" paradigm was first employed by The Institute for Genomics Research to sequence the *Haemophilus influenzae* genome then by Celera Genomics in the sequencing of *Drosophila melanogaster* and *Homo sapiens* genome (17,31,32). As a proof of concept demonstrating the ability of this new sequencing technology to overcome barriers in both the throughput and cost associated with whole genome sequencing, the human genome project was repeated, using this massively parallel sequencing technology, to sequence the genome of Dr. James Watson (33). This project, published in 2009, was completed in only two months at approximately 1% of the cost associated with the first Human Genome Project. With next generation sequencing in combination with DNA capturing technology capable of extracting just the DNA fragments corresponding to the protein coding regions of the human genome, the capability to rapidly and inexpensively performed whole-exome type sequencing across large numbers of samples became a reality. In 2010, the first application of this novel next generation whole-exome sequencing technology to the study of human cancer was the screening of 31 uveal melanoma samples

revealing recurrent inactivating mutations to the gene encoding the BRCA1-associated protein 1 (BAP1) (34). In 2009, there were recurrent somatic mutations identified in 350 protein-coding genes in the human genome representing a quarter century of cancer research (35). A mere 5 years later, the number of protein-coding genes implicated in cancer has grown to 547, a greater than 50% growth highlighting how next generation sequencing technology increased the effectiveness of systematic cancer sequencing studies.

1.3 Description of general variant discovery pipeline used in analysis of next generation whole-exome sequencing data

In parallel to the rapid development of next generation sequencing, there is an increasing need for bioinformatics to develop a systematic method or pipeline in order to analysis the ever growing volume of sequenced DNA data. The computational pipeline described below (Figure 1.2) outlines the basic steps required to align short reads data generated by Illumina sequencing technology to a reference genome and generate a list of high confidence variants. Downstream use of these variants will be to catalog somatic mutations, to estimate copy number/loss of heterozygosity (LOH) changes and to infer signatures of mutational processes. Due to the need to differentiate between somatic and germline variants, DNA extracted from non-cancer tissues or blood is also sequenced along with tumor DNA extracted from the same patient to form a matched pair for comparison. Computationally, the steps taken to generate high confidence variants remains the same between normal and tumor DNA data; the cost of sequencing, computational analysis and data storage as well as the time need to generate and analyze the data due to the need for matched pair DNA sequencing should be taken into account during the project planning stages.

1.3.1 Sequenced DNA data in FASTQ format

The basic starting point for this pipeline is a flat text file containing information about the sequenced DNA fragments or short reads from a single sample, tumor or normal. There is a general format in which the sequenced DNA data is presented; this format is called FASTQ and is the dominant data format used to present sequenced DNA data in all public databases.

The FASTQ data format uses four lines to present information from a single read as shown in figure 1.3A:

Line1: '@' character is used to start the first line followed by information concerning the sequence or the machine where the DNA was sequenced.

Line2: The DNA sequence of the short read described in Line1

Line3: '+' character is used to start the third line and may display the information presented in Line1 or be left blank.

Line4: The number of characters must equal to the number of characters in Line2; each character is a quality score, encoded in ASCII format, of the corresponding sequenced base in Line2.

The ASCII characters used to encode the quality scores ranging from 0 – 93 are shown in Figure 1.3B for reference. The quality score (Q) is an integer mapping of the probability (p) that the corresponding base is sequenced incorrectly. The conversion equation between quality score and probability is shown below.

$$Q = -10*\log_{10}(p)$$

In addition to the Sanger format of quality score encoding, there are three legacy quality score formats proposed by Solexa/Illumina: Solexa, Illumina 1.3+ and Illumina 1.5+ (Figure 1.3B). There are two main differences between Sanger and Solexa/Illumina formats; one is the narrower range of possible quality scores from Solexa/Illumina formats and the other is a shift to the higher range of ASCII

encoding. As of March 2011, Illumina quality score for its fastq output returned to the Sanger format.

1.3.2 Alignment of DNA fragments to the reference genome:

BOWTIE2, BWA and SOAP3-dp represents a popular family of short-read sequence aligners designed specifically for mapping short read sequencing data produced by next generation sequencing technology (36,37,38). All three programs employs the use of Burrows-Wheeler transform to create a compressed reusable index of the human reference genome to reduce the memory requirements for high speed mapping of short reads. BWA is used the aligner for this pipeline, all three alignment programs are essentially equivalent in terms of performance, requirements, and output format and can be substituted in a modular manner (37).

1.3.3 PCR-duplicate removal:

After alignment to a reference genome, PCR duplicates present in the aligned data set must be removed; PCR duplicates of short reads arise when two or more copies of the same DNA fragment is sequenced; this phenomenon is created due to the necessity of using PCR to amplify the original DNA molecules to ensure adequate quantities will be available not only for sequencing but subsequent downstream experimentation. Higher number of amplification cycles needed to compensate for low starting amounts of DNA will increase the amount of PCR duplicates; large variance in DNA fragments due to non-optimized DNA shattering protocol will also result in PCR duplicates as PCR reaction is biased towards amplifying shorter DNA fragments. Not filtering for PCR duplicates will result in an increase in false positive variant calls due to PCR errors that are amplified or false calls in copy number

alterations due to preferential PCR amplifications. There are two popular open-source toolkits currently available with utilities to process the aligned output from aligners described above and remove PCR duplicates: SAMtools' rmdup function and PICARD's MarkDuplicates function (39,40). SAMtools' rmdup function is markedly faster and consumes significantly less memory intensive than PICARD's MarkDuplicates function; however, MarkDuplicates is able to remove interchromosomal duplicates whereas rmdup do not have this capability.

1.3.4 Variant calling and separation of somatic, germline and SNP variants:

To detect single nucleotide variants (SNVs), a suite of programs, collectively known as the Genome Analyzer Toolkit (GATK), is employed using the aligned, PCR duplicates removed data set as the starting point (41). As a pre-processing step, aligned reads predicted to contain small insertion/deletion events (micro-indels), between 3bp – 10bps, undergo base quality recalibration followed by realignment to the reference genome; the purpose of this pre-processing step is to ensure a better local alignment in reads containing micro-indels to reduce false positive variant calls. The realigned data file is filtered such that only well-mapped reads with a mapping quality score greater than 30 and less than three mismatches within a 40 bp window were used as input to the GATK Unified Genotyper; this program performs the consensus calling in order to identify SNVs. These SNVs are compared against common polymorphisms listed in Single Nucleotide Polymorphism Database (dbSNP) and in the 1000 genomes database, and any SNVs present in either database will be discarded (42,43). However, some somatic mutations implicated in cancer, such as variants leading to glycine mutations in codon 12 of V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog (KRAS), were also found in dbSNP; an additional

comparison is made to the Catalogue of Somatic Mutations in Cancer (COSMIC) database and SNVs present in COSMIC database will be retained (44). An explanation for the presence of known oncogenic variants may be due to the relaxed submission requirements practiced by dbSNP. According to dbSNP, submissions may be polymorphisms, common variations, AND mutations, rare allele variations. In addition, even if a variant submission might be somatic but cannot be determined due to lack of matched normal DNA, dbSNP will still accept the submission as long as the submitted method states the submitter has no way of determining if the submission is somatic/germline. Such relaxed submission requirements may account for the presence of oncogenic variants within dbSNP. All SNVs remaining after this step will be considered to be "novel" and will be placed in a novel variant file; the filtered SNVs, considered to be common polymorphisms, will be stored in a separate SNP file. Several gene transcript annotation databases (CCDS, RefSeq, Ensembl, UCSC) will be used for transcript identification and for determining the amino acid change. Only SNVs in exons or in canonical splice sites will be annotated with amino acid changes annotated according to the largest transcript of the gene.

The steps described thus far will be performed twice; once for the sequenced data from the tumour sample and once for the sequenced data from the corresponding normal sample resulting in four separate variant files: tumour novel variant file, normal novel variant file, tumour SNP file and normal SNP file. The intersection of SNVs between the tumour novel variants and the normal novel variants will produce a list of germline variants or inherited mutations or mutations unique to an individual; this list is useful in locating mutations that predispose an individual to develop certain cancers. SNVs that are present in the tumour novel variants list but not in the normal variants list will produce a list of somatic mutations or mutations acquired during the

development of cancer; these predicted somatic mutations will be verified using Sanger capillary sequencing. As the number of validations can be high, a high throughput primer design software, Primer3Plus (45), is employed to design the forward and reverse DNA primers. The DNA primer sequences for each predicted somatic mutation is included as part of the final analysis report. In addition, nonsynonymous mutations or mutations that will result in a corresponding amino acid change in the gene's protein product are submitted to PolyPhen2 for functional prediction (46). If the protein crystal structure corresponding to a gene of interest is available in the RCSB Protein Data Bank (PDB), the protein structure containing the mutation can be modeled using SWISS-MODEL, an online fully automated protein structure homology-modelling server; the predicted mutated protein structure output by SWISS-MODEL as well as the original protein structure can be viewed using Deepview, a freely available program linked to SWISS-MODEL that allows for visualization, analysis and comparison of several protein structures simultaneously (47,48,49). Functional and, where possible, structural prediction of novel somatic mutations using computational tools represents a critical first step in the identification of gene alterations contributing to the development of cancer.

1.3.5 Visualization and estimation of copy number and loss of heterozygosity changes

Hilbert plot is an early method to visualize copy number changes across the entire sequenced exome in a compact graphical manner (50); instead of linearly plotting the sequencing depth versus the chromosomal position, Hilbert plot computationally wraps the chromosomal positions, essentially a DNA string, in a fractal manner onto a two dimensional grid of pre-determined size and presents the

sequencing depth via a heat map. By comparing the tumor and normal Hilbert plots, copy number changes of the tumor, if present, will reveal itself through color intensity changes; when compared to normal, intensity changes will reveal regions of the plot where copy number change occurs as well as systemic targeted DNA capturing and sequencing bias. While this visualization method is useful in quickly establishing gross changes in copy number, it is difficult to estimate, from a glance, which chromosome or where on the chromosome the copy number change is occurring due to the two color display limit of the program and the non-intuitive fractal mapping of a one dimensional string on a two dimensional surface.

The above method of copy number estimation has been superseded by ASCAT (Allele-Specific Copy number Analysis of Tumors) which offers, in addition to copy number analysis, loss of heterozygosity and ploidy analysis (51); originally designed for analysis of SNP arrays, the analog input parameters of total signal intensity, Log R, and allele contrast, B allele frequency (BAF), are equivalently represented in a genomic sequencing context. Only heterozygous variants in the sequenced DNA of the normal sample, corresponding to SNPs or germline mutations, will be considered in the ASCAT analysis as homozygous variants are uninformative in copy number estimation. Log R parameter is equivalent to the Log of the ratio between tumor and normal total sequencing depth at the position of a heterozygous variant. BAF parameter is equivalent to the ratio between the number of reads calling for the variant and the total sequencing depth for the tumor sample. A log R value around zero means there are no copy number changes between tumor and normal samples while A BAF value around 0.5 means the number of paternal and maternal alleles are balanced; significant deviation these values represents copy number changes and/or LOH events in the tumor. The usage of ASCAT, through the use of normally

discarded or neglected SNPs and germline variants, enabled another parallel level of exome analysis in addition to the search for somatic nonsynonymous mutations and highlights the inherent richness of the exome data.

1.3.6 Inferring mutational processes in a tumor

The list of somatic SNVs obtained in variant analysis can be viewed as the end result of X mutational processes operating during the development of the cancer tumor. These mutational processes may be distinguished from one another through nucleotide context preferences in mutating the genome resulting in different mutational signatures. One example is Aristolochic acid, a known carcinogen, is shown to have a characteristic genome wide mutational signature corresponding to adenine to thymine substitution pattern due to the carcinogen's preferentially forming adducts with the adenine base (52,53). Another example is the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) mediated mutagenesis, resulting in a C > G or T mutation in a TpCpA or TpCpT trinucleotide context, that is found to be operative in a number of different cancer types (54,55,56).

There are currently two different methods of inferring mutational signatures and their contributions given somatic SNVs obtained from a sequenced DNA set of cancer samples; they are nonnegative matrix factorization (NMF) and expectation-maximization (EMu) (57,58). The differences in algorithmic approaches of the two methods highlight the different philosophical approaches between the two methods. NMF is, at its heart, a neutral mathematical construct designed to find two matrices corresponding to X mutational signatures and their corresponding Y contributions to each cancer sample such that when X and Y are multiplied together, the original mutation list will be recovered as closely as possible. The NMF algorithm used in

deconvoluting mutational signatures can be applied as is to divergent applications such as gene expression analysis, facial recognition, text mining and spectral data analysis of space debris (58-61). EMu, on the other hand, seeks to take advantage of available biological information such as the differences in trinucleotide context distributions and copy number changes unique to each cancer tumor; this *a priori* information is used to generate a probabilistic method that not only takes into account the differences in mutational opportunity but also account for noisy data inherent in the stochastic nature of mutational processes. Both methods showed similar results in locating defined mutational signatures associated with known mutational processes such as the signature of APOBEC mediated mutagenesis or the signature of spontaneous deamination of 5-methylcytosine to thymine (57,58). However, without experimental evidence of a one to one correspondence between signature and process, it is not possible to determine which algorithm is more accurate in the location and assignment of novel mutational signatures. The EMu algorithm was selected to infer mutational signatures for this thesis as it requires substantially less hardware requirements, no specialized proprietary software and orders of magnitude faster than NMF in results generation.

1.4 Application of variant discovery pipeline

This thesis seeks to apply the methodologies discussed above in three areas of cancer research presented in three chapters below.

1.4.1 Summary of chapter two

In this study, fresh well-differentiated papillary mesothelioma of the peritoneum samples as well as matching blood from a single patient were obtained. Fresh tumor samples enabled the culturing of the tumor cells to purify its tumor content. Using the DNA extracted from the primary tumor, its purified tumor cells and blood, we performed whole-exome sequencing. The use of Hilbert plot to compactly display the sequenced exomes displayed no gross chromosomal anomalies. Somatic variant detection followed by validation revealed only three somatic single nucleotide mutations present. One of the mutations is predicted to alter the arginine (Arg) 166 codon to histadine (His) of E2F transcription factor 1 (E2F1), a gene implicated in cancer but was never found to be mutated in cancer thus far. Conservation analysis across paralogues and orthologues of E2F1 indicated the Arg166 position is completely conserved suggesting the position's functional importance. Protein homology modeling revealed the Arg166 to be a critical DNA contact point for E2F1 and modeling of Arg166His alteration suggested a functional loss of DNA binding for E2F1. Chromatin immunoprecipation as well as real-time PCR on E2F1 targets revealed Arg166His alteration abrogated the DNA binding ability of E2F1 and negatively affected the gene expression of E2F1 binding targets. Massive accumulation of mutant E2F1 protein was observed in transfected cells when compared with cells transfected with wild-type E2F1. By comparing the protein quantities of wild-type and mutant E2F1 in transfected cells dosed with

cycloheximide, a potent protein synthesis inhibitor, at different time intervals, mutant E2F1 were observed to be resistant to degradation when compared with wild-type E2F1. Interaction between E2F1 and RB1 constitutes a critical process in controlling a cell's entry from G1 to S phase. RB1 binds and inhibits E2F members which are responsible for initiating S phase and the cell's commitment to division. As long as E2F members are bound to RB1, the cell is stalled at the G1 phase of cell cycle. A conjecture was proposed that mutant E2F1, resistant to degradation and accumulating in much larger quantities than its wild-type counterpart, was more likely by chance to bind to Retinoblastoma 1 (RB1) and thus leaving behind a small pool of unbound wild-type E2F1 that was able to bypass the G1/S checkpoint to drive aberrant cell division.

This study highlights the ability of computational analysis to quickly narrow the field of possible functional consequences of a mutation and present high confidence hypothesis for experimental biologists to consider. In addition, this study also demonstrates the synergy between computational and wet lab studies.

1.4.2 Summary of chapter three

This study outlines the mutational landscape of *Opisthorhis viverrini*-related (OV-related) cholangiocarcinoma (CCA), a malignant cancer of the bile duct prevalent in northeastern Thailand and Laos. A discovery set of eight OV-related tumors and matched normal tissue were selected for whole-exome sequencing with 46 additional CCA matched samples constituting the prevalence set. In addition to somatic mutations in cancer related genes tumor protein p53 (*TP53*) (44.4% mutation rate), *KRAS* (16.7%) and SMAD family member 4 (*SMAD4*) (16.7%), another 10 novel recurrently mutated genes were identified: These include inactivating mutations

in lysine (K)-specific methyltransferase 2C (*MLL3*) (14.8%), roundabout, axon guidance receptor, homolog 2 (Drosophila) (*ROBO2*) (9.3%), Ring finger protein 43 (*RNF43*) (9.3%), paternally expressed 3 (*PEG3*) (5.6%) and activating mutations of GNAS complex locus (*GNAS*) oncogene (9.3%).

Minnelide, a water-soluble form of the plant extract Triptolide, has been shown to be effective for in-vitro and in-vivo models of pancreatic cancer with a background of *KRAS* and *TP53* mutations. The naturally occurring Triptolide has been shown to be effective against CCA suggesting Minnelide may also be effective in treating the subset of CCAs with *KRAS* and/or *TP53* mutations. Recurrent mutations to *TP53, RNF43* and *PEG3* points to aberrant Wnt signaling activation suggesting the use of O-acyltransferase Porcupine inhibitor (LGK974), shown to be effective in *RNF43* inactivated pancreatic cancer cell lines, as a targeted therapeutic in treating *RNF43* inactivated CCAs.

Comparison of OV-related CCA, pancreatic ductal adenocarcinoma (PDAC) and hepatitis C virus (HCV)-associated hepatocarcinoma (HCC) revealed a distinctive grouping, at both recurrently mutated genes and base substitution spectra level, with OV-related CCA/PDAC in one group and HCV-associated HCC in a separate group. As endogenous and exogenous mutational processes drives the observed mutational spectra, a conjecture was made that individual stochastic mutational processes may be driving the emerging recurrent gene mutational patterns observed in different cancers.

1.4.3 Summary of chapter four

This study outlines the whole-exome mutational landscape of parathyroid carcinoma (PC) and attempts to characterize the mutational processes involved in PC. Recurrent inactivating mutations in known PC associated gene *Cell division cycle 73*

(*CDC73*) were verified and loss of heterozygosity (LOH) accompanied by recurrent amplifications of mutant *CDC73* allele were computationally predicted. Whole-exome analysis identified *prune homolog 2* [Drosophila] (*PRUNE2*) to be the second recurrently mutated gene in PC with germline and somatic mutations clustered around a functionally unknown but evolutionary conserved region of the protein. Members of the kinase family related to cell migration and invasion were also found to be mutated in PC. APOBEC mutational signature was found to be dominant in a subset of PC patients with high mutational burden and early age onset of disease with APOBEC mediated mutagenesis implicated for the first time in parathyroid carcinoma. This study highlights the ability of mutational screening studies to open new avenues of research not previously considered under hypothesis-driven approaches.

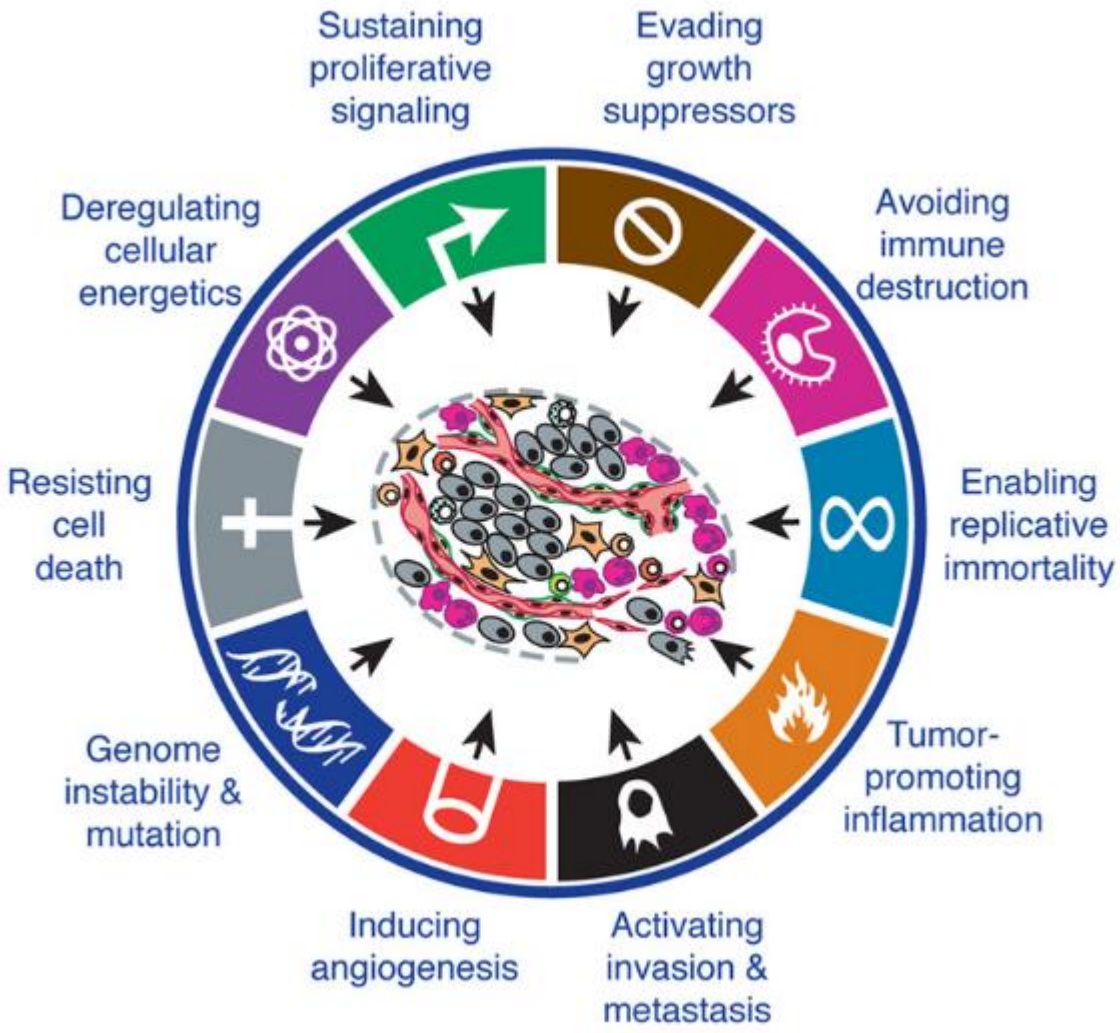


Figure 1.1: The ten hallmarks of cancer as defined by Hanahan and Weinberg. Figure extracted and modified from figure 6 of Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell* 2011, 144:646-674. License to reproduce has been obtained from Elsevier Limited and can be produced upon request.

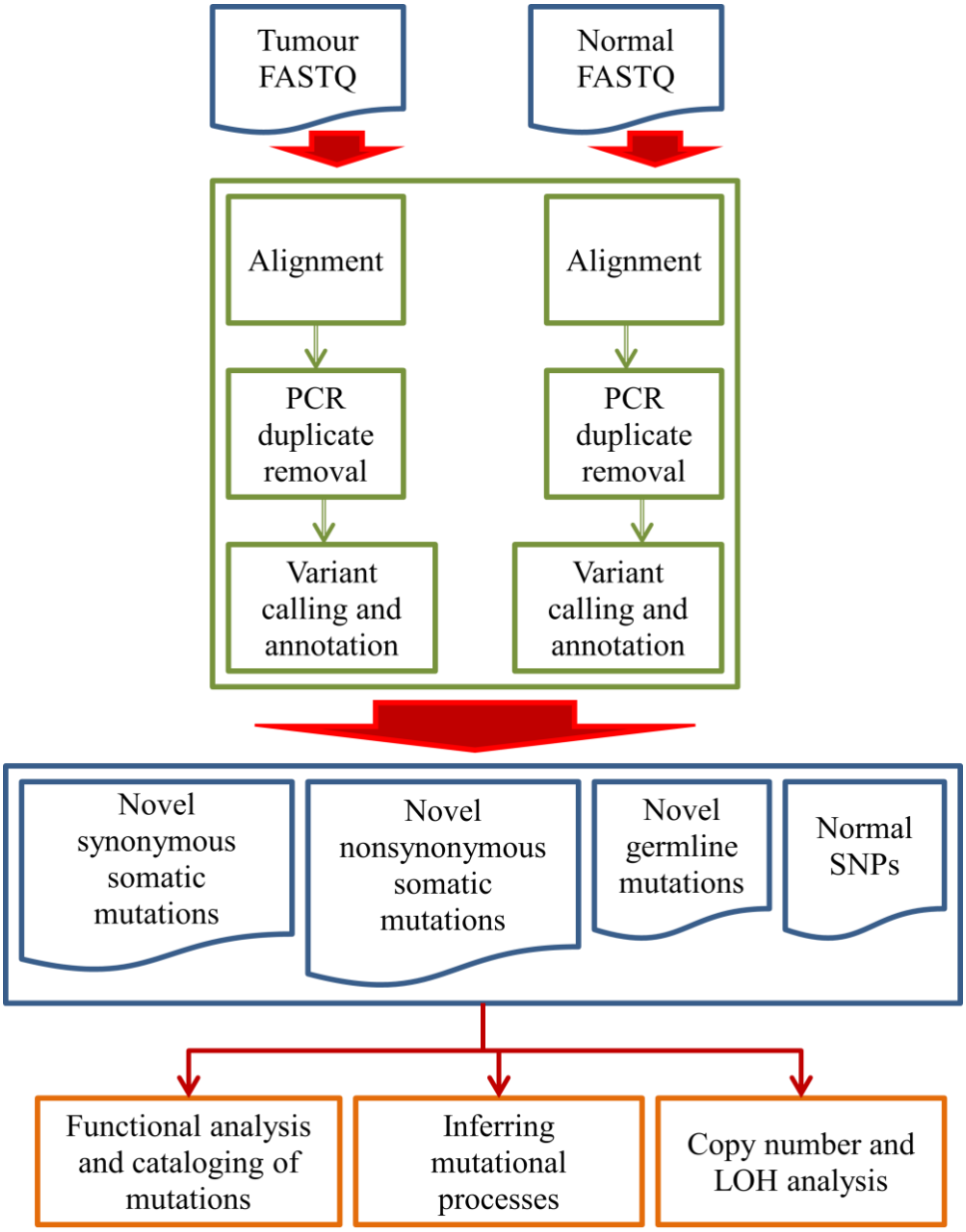


Figure 1.2: General variant discovery and analysis pipeline used for whole-exome sequencing data sets. Blue color boxes represent input or output files required or generated by the variant pipeline. Green boxes represent key steps in the variant calling pipeline. Orange boxes represent different downstream analysis that can be performed based of the output files generated by the variant calling pipeline.

A

```
@HWUSI-EAS300R_0005_FC62TL2AAXX:8:30:18447:12115#0/1
CGTAGCTGTGTGTACAAGGCCCGGGAACGTATTCACCGTG
+HWUSI-EAS300R_0005_FC62TL2AAXX:8:30:18447:12115#0/1
acdd^aa_Z^d^ddc`^_Q_aaa`_ddc\dfdffff\fff
```

B

!	**"**	**#**	**$**	**%**	**&**	**'**	**(**	**)**	*****	**+**	**,**	**-**	**.**	**/**	**0**	**1**	**2**	**3**	**4**
33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

5	**6**	**7**	**8**	**9**	**:**	**;**	**<**	**=**	**>**	**?**	**@**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
						-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
											0	1	2	3	4	5	6	7	8
														3	4	5	6	7	8

I	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**	**[**	****
73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92
40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28

]	**^**	**_**	**`**	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**
93	94	95	96	97	98	99	100	101	102	103	104
60	61	62	63	64	65	66	67	68	69	70	71
29	30	31	32	33	34	35	36	37	38	39	40
29	30	31	32	33	34	35	36	37	38	39	40
29	30	31	32	33	34	35	36	37	38	39	40

Figure 1.3: FASTQ example and quality score encoding. A) An example of a single sequenced read in FASTQ format. B) ASCII characters (bold) used to encode typical quality scores from different FASTQ formats. Green: Sanger, Blue: Solexa, Orange: Illumina v1.3+, Red: Illumina v1.5+.

Chapter Two: First Somatic Mutation of E2F1 in a Critical DNA Binding Residue Discovered in Well- Differentiated Papillary Mesothelioma of the Peritoneum.

Part of the findings in this Chapter was published in Yu et al. (2011), *Genome Biol*; 12(9):R96 (pp 25-51 of this thesis).

The downstream bench-top studies were performed by Dr. Waraporn Chan-On and I have indicated clearly the sections where the work was performed by her.

2.1: Introduction

Mesothelioma is an uncommon neoplasm that develops from the mesothelium, a protective lining covering the majority of the body's internal organs, and is divided into four subtypes: pleural, peritoneum, pericardium and tunica vaginalis (62). The malignant pleural subtype of this cancer captured the world's attention through its association with asbestos exposures (63-66). Malignant peritoneal mesothelioma (MPM) has since been also shown to afflict asbestos exposed males in the age range of 50-60 years old (66). Unlike its more infamous siblings, well-differentiated papillary mesothelioma of the peritoneum (WDPMP) is an extremely rare subtype of mesothelioma that was first discovered incidentally, in 1958, in a 41 year old female undergoing surgery to repair a cystocele, bladder herniation into the vagina, and uterine prolapse (62,68). Since its initial discovery, there are fewer than 60 WDPMP cases described in the literature (69) with most tumors being discovered incidentally and only in rare cases can the tumor be associated with symptoms (62,70,71).

Distinguishing features of WDPMP are no invasive activity into surrounding structures or tissues possesses well defined papillary/tubular structures and lined by cuboidal mesothelial cells with low or absent mitotic activity. Consensus recommendation for treatment is surgical resection followed by routine observation; chemo or radiotherapy is not recommended due the tumor's benign nature and potential significant side-effects due to treatment (69-72).

Recurrent WDPMP is extremely rare with only a single case reported in literature at the time of publication (71); the particular case indicated the recurrent WDPMP tumor was discovered incidentally during another surgical procedure almost 4 years after the initial WDPMP resection. Whether WDPMP can, in time, progress to

malignancy is a subject of debate at the time of this study's publication; there was a single report of WDPMP progressing to malignant mesothelioma but was criticized by Malpica et al. for lacking in pathologic examination (71,73).

Overall, the general consensus is that WDPMP is a tumor of low malignant potential found predominately in young women with no definitive exposure to asbestos (63,69,70,71). While much scientific research has been done on asbestos related malignant mesothelioma (74-77), the rarity of WDPMP coupled with its good prognosis relegated its research to case reports and reviews by medical oncologists concentrating in the area of diagnosis, prognosis and treatment options.

Second generation sequencing technologies coupled with newly developed whole exome capturing technologies (78) allow for rapid, relatively inexpensive approach to obtain an overview of large complex genomes concentrating on the critical coding areas of the genome. In December of 2010, Harbour et al's discovery of *BAP1* mutations in metastasizing uveal melanomas demonstrated the application of whole-exome capture followed by massively parallel sequencing to accelerate detection of novel recurrently mutated genes in cancer (34). Through the use of whole-exome sequencing technology to rapidly catalogue somatic mutations in WDPMP, we can begin to address the long standing question of whether the benign WDPMP have the potential to progress to malignancy. This will have major implications in the clinical treatment of this disease. In September 2011, we report the use of this new sequencing paradigm on a matched pair of WDPMP tumor and its purified tumor cells to discover the first somatic single nucleotide mutation of *E2F1*, a critical player in the initiation of cell division.

2.2: Results

2.2.1: WDPMP whole-exome sequencing: mutation landscape changes big and small

Whole-exome captured sample libraries comprising of DNA from WDPMP tumor, DNA from patient's blood, and DNA from purified tumor cells were sequenced using Illumina GAIIx 76bp Pair-End sequencing technology; in brief, cells extracted from fresh tumor were plated according to protocol for the initial deposition of tumor cells onto the appropriate dish. After seven passages, the tumor cells were collected for DNA extraction and exome capturing. Table 2.1 shows the summary of the sequenced exome data; in total, ~34 Gbases of sequence data were obtained in which >92% of the reads successfully mapped back to the hg18 reference genome using BWA short read aligner (37). After removal of low quality reads and PCR duplicate reads using SAMtools (39), ~24.3 Gbases of sequence data remained. Of the remaining sequence data, ~64% or ~15.5 Gbases fell within the exon regions with the average exome coverage per sample being 152x depth; Figure 2.1 shows the breakdown of coverage vs sequencing depth, the key statistics being 97% of the exome were covered by at least a single good quality read, ~92% of the exome were covered at least 10 good quality reads and 82-86% of the exome were covered by at least 20 reads indicating the overall exome capturing and sequencing were successful with large amounts of good quality data.

A novel way to visualize large copy number changes using exome sequencing data is the use of HilbertVis (50), an R statistical package, to plot exome sequencing depth versus chromosomal position in a compact graphical manner. Copy number changes, if present, will reveal itself through color intensity changes in regions of the

plot where copy number change occurs when comparing between tumor/purified tumor cells versus normal. Figure 2.2 shows the Hilbert plots of the sequenced tumor, normal and purified tumor cells exome revealing some systemic capturing biases but no deletion/amplification events detected with particular attention paid to known somatic deletions of 3p21, 9p13~21 and 22q associated with loss of Ras association domain family 1 isoform A (*RASSF1A*), cyclin-dependent kinase inhibitor 2A (*CDKN2A*) and neufibromatosis type 2 (*NF2*) genes respectively in malignant mesothelioma (79). Sequencing depth was also adequate for the regions of exon capture for these genes (Figure 2.3) indicating these genes were truly not somatically mutated and lack of mutations detected were not due to a lack of coverage.

Since the Hilbert plots showed no gross anomalies, we turned our attention to mining the whole-exome data for somatic single nucleotide mutations. The single nucleotide variant discovery pipeline, described in the Methods section, was performed using GATK (41) for tumor, normal and purified tumor cells exomes. Filtering was set to accept candidate SNV's with quality/depth score of greater than three and were present in both tumor and purified tumor cells and not in normal. 19 potential somatic mutations remain and validation was attempted using Sanger sequencing (Table 2.2); putative mutations in *E2F1*, PTPRF interacting protein, binding protein 2 (liprin beta 2) (*PPFIBP2*) and TNF receptor-associated factor 7, E3 ubiquitin protein ligase (*TRAF7*) were validated to be true somatic mutations (Figure 2.4).

2.2.2: E2F1 R166H mutation affects critical DNA binding residue

E2F1 R166H somatic mutation is of particular interest as there was no reported mutation of this gene in cancer prior to this study's publication. Figure 3 top

shows the genomic location of E2F1 as well as the specific location of the mutation. Sanger sequencing around the mutated nucleotide for the tumor, purified tumor cells and normal revealed the mutation to be heterozygous (Figure 2.4, top). A check of UniProt for E2F1 [UniProtKB: Q01094] showed the mutation to be located in the DNA binding domain of the protein. To study the evolutionary conservation of the R166 residue, a CLUSTALW analysis was performed on paralogues of the human E2F family and single nucleotide polymorphism (SNP) analysis, using SNPS3D (80,81), was performed across orthologues of E2F1 (Figure 2.5). Figure 5 bottom shows the results of the paralogues and orthologues conservation analysis respectively. In addition, three functional prediction programs, Polyphen2, SIFT and CADD, were employed to predict the impact of altering the R166 residue to histidine (46,82,83). In summary, all three programs predict the mutation to be damaging with CADD predicting the mutation to be among the 1% of the most deleterious substitution that can occur in the human genome. The conclusion drawn is E2F1 is never observed to be mutated; its R166 is conserved in evolution with histidine alteration predicted to functionally deleterious.

Since there is no E2F1 crystal structure containing the R166 residue, E2F transcription factor 4, p107/p130-binding (E2F4) X-ray crystal structure [PDB: 1CF7] was used to determine the mutation location and its role in DNA binding using Swiss-PDB viewer (49). The E2F4 DNA binding structure was used as an adequate representation of the E2F1 counterpart due to the conserved status of the R165-R166 residues across the E2F paralogues (Figure 2.5, bottom right) as well as the affected residue being a part of the transcription factor e2f/dimerization partner domain observed across all E2F family of transcription factors (84). The arginine residues of E2F4 and its transcription factor, Dp1 (TFDP1) binding partner responsible for DNA

binding (Figure 2.6, top) and the analysis clearly shows R166 as one of four Arginine residues contacting the DNA target (Figure 2.6, bottom).

Since the crystal structure for the DNA binding domain of E2F4 was available, computational modeling of the mutation was amenable to homology-modeling using SWISS-MODEL (48). Figure 7 top shows the modeling of E2F1 mutant and wild-type DNA binding domain; Calculation of individual residue energy using Atomic Non-Local Environment Assessment (ANOLEA) and Groningen Molecular Simulation (GROMOS) indicated the mutant histidine‘s predicted position and conformation was still favorable as indicated by the predicted negative energy value (Figure 2.7, bottom). While there is a difference in the size and charge between the mutant histidine and wild-type arginine residue coupled with a conformational shift at the mutated position, the overall 3-D structure of the domain appears minimally affected by the mutation. Even though the mutation effect on DNA binding is inconclusive computationally, these results did pinpoint structural location and functional importance of the R166 residue thus pointing the way for the functional experiments below.

NOTE: The experiments leading to the results described from this point on were performed by Dr. Waraporn Chan-On.

2.2.3: R166H mutation is detrimental to E2F1's DNA binding ability and negatively affects downstream target gene expression

In order to conclusively show the R166H mutation effect on DNA binding, chromatin immunoprecipitation (ChIP) assays were used on the promoters of two known transcriptional targets of E2F1, SIRT1 and APAF1, using MSTO-211H cells over-expressing either wild type or mutant E2F1 (85,86). The mutant E2F1 (Figure 2.8a lane 7) showed significantly decreased quantities of *APAF1* (top) and *SIRT1*

promoter DNA binding (bottom) when compared with WT E2F1 (Figure 2.8a lane 6) although the amount of input DNA for E2F1 mutant was greater than E2F1 wild type (Figure 2.8a lane 2 and 3 respectively). The ChIP result indicates the R166H mutation has a detrimental effect on the E2F1's DNA binding ability.

To show the R166H mutant's reduced DNA binding affinity affected the expression of E2F1 target genes, expression of *SIRT1*, *APAF1* and *cyclin E1* (*CCNE1*) were examined by real-time PCR in MSTO-211H and NCI-H28 that were transfected with the E2F1 mutant or wild-type. Interestingly, over-expression of E2F1 R166H could not up-regulate expression of *SIRT1* and *APAF1* as high as E2F1-WT over-expression in both cell lines (Figure 2.8b and c). In particular, levels of *SIRT1* and *APAF1* in MSTO-211H observed in E2F1-R166H were significantly lower than the levels in E2F1 wild-type ($p = 0.032$ for *SIRT1* and $p = 0.005$ for *APAF1*). However, the expression of *CCNE1*, a well-known target of E2F1 (87), was minimally affected in the over-expression context which may be indicative of compensatory effect by other members of the E2F family. The observed SIRT1 and APAF1 transcription differences between MSTO-211H and NCI-H28 may be due to compensatory effects of other transcriptional activators and repressors such as c-MYC, p53, HIC1 and other members of the E2F family (86,88-90).

2.2.4: Cells over expressing E2F1 R166H mutant show massive protein accumulation and increased protein stability

To study cellular phenotypes that might be affected by the R166H mutation, we initially over-expressed the mutant and wild type in the cells. Surprisingly, an obvious difference in E2F1 protein levels between wild-type and mutant was observed in both cell lines as determined by western blot (Fig. 2.9a). In order to ensure the protein differences were not due to differences in transfection efficiency,

the two cell lines; MSTO-211H and NCI-H28, were co-transfected with E2F1 and Enhanced green fluorescent protein (EGFP) vectors simultaneously with protein lysate obtained at 48 hr time point for western blot analysis. Clearly, expressions of E2F1 wild type and mutant normalized by EGFP levels were similar (Figure 2.10) indicating that the transfection efficiency of R166H is not different from wild type. This suggests that the large increase in the level of mutant E2F1 protein might be caused by other mechanisms such as increased protein stability.

To monitor E2F1 protein stability, we over-expressed E2F1 wild type and mutant in MSTO-211H before treating the cells with 25µg/ml cyclohexamide to block newly synthesized protein in half hour intervals. As shown in figure 2.9b, the protein levels of E2F1 mutant remained almost constant throughout the 3 hour period of the experiment while the E2F1 wild type protein level was decreasing in a time-dependent manner. This result suggests that the mutant protein is more stable and resistant to degradation than the wild type and an increased stability of R166H is the cause of its accumulation within the mutant over expressing cells.

2.2.5: Over expression of E2F1 R166H mutant does not adversely affect cell proliferation

Since the R166H mutant is demonstrated to have exceptional stability and accumulates heavily in mutant over expressing cells, it would be instructive to observe what effect if any does this mutant have on cell proliferation. Proliferation assay was performed on the transiently transfected cell lines. The result showed that high expression of E2F1 wild type decreased the growth rate of the cells whereas the mutant showed a increased growth rate although both results were not statistically significant (Figure 2.11a and b). Although E2F1 R166H mutation does not show

significant effect on regulating cell proliferation, it is possible that the mutation is advantageous to cancer cells as it does not inhibit cell growth when the mutant is highly expressed in cells.

2.3: Discussion

For this study we have performed whole-exome sequencing using DNA obtained from a matched pair of WDPMP along with its purified tumor cells. A barrier to accurate somatic mutations prediction is the amount of normal cells present in the tumor tissue. Proportional increase in normal cell content will result in equivalent decrease in amount of tumor DNA sequenced; this will result in increased false positive and false negative somatic mutation predictions due to decreased amount of tumor DNA sequenced requiring additional sequencing to increase the tumor resolution. One method to increase the tumor content is to treat the cells from tumor tissue as a cell line and processing them for several passages to increase the tumor cell content. We have shown the sequencing amount of purified tumor cells is only 2/3 of the amount for whole tumor sequencing (Table 2.1) with the true somatic mutations being recovered by both purified and whole tumor sequencing; subsequent Sanger sequencing validation showed greater clarity of the mutant peak for the purified tumor cells.

Analysis of the exomes revealed the tumor contained none of the chromosomal aberrations or focal gene deletions commonly associated with asbestos-related malignant mesothelioma. We were able to verify somatic mutations in *PPFIBP2, TRAF7* and *E2F1*.

TRAF7 is an E3 ubiquitin ligase shown to be involved in mitogen-activated protein kinase kinase kinase 3 (MEKK3) signaling and apoptosis (91). The mutation Y621D occurs in the beta-transducin repeat 40 (WD40) repeat domain and the domain was shown to be involved in MEKK3–induced activating protein-1 (AP1) activation (92). Since AP1 in turn controls a large number of cellular processes involved in differentiation, proliferation and apoptosis (93), mutation in TRAF7's WD40 repeat domain may de-regulate MEKK3's control over AP1 activation which may contribute to WDPMP transformation. Since the publication of this study, two additional studies reported recurrent *TRAF7* mutations clustering around its WD40 repeat domains in secretory as well as in non-*NF2* mutated meningiomas; they also reported a more aggressive clinical course in tumors harboring both *TRAF7* and Kruppel-like factor 4 (gut) (*KLF4*) mutations (94,95). With these two studies in mind, the seemingly random *TRAF7* mutation found in a seemingly benign and indolent WDPMP takes on a more sinister meaning; the conjecture can be made that the tumor is accumulating mutations that not only partially transforms the affected cells into a benign tumor but also mutations that may synergize with other gene mutations to accelerate cancerous progression. Due to WDPMP's non-symptomatic nature, the tumor may not be detected for significant period of time allowing the tumor a greater chance to mutate the "correct" combination of genes to transform fully.

PPFIBP2 is a member of the LAR protein-tyrosine-phosphatase-interacting protein (liprin) family (96). While there are no functional studies published on PPFIBP2, it was reported as a potential biomarker for endometrial carcinomas (97). However, the Q791H mutation itself is predicted by Polyphen to be benign and COSMIC did not show this particular mutation to recur in other cancers thus this mutation is likely to be of a passenger variety.

At the time of this study's publication, there was no reported somatic mutation observed for *E2F1* despite its critical roles in cell cycle control, apoptosis and DNA repair (87,98,99). Since then, over 52 unique somatic alterations have been found for *E2F1* according to COSMIC database (44). Of these 52 mutations found, 39 were nonsynonymous and the rest were synonymous. Of interest are 10 mutations found within the transcription factor e2f/dimerization partner domain of *E2F1* and all were nonsynonymous; four of these 10 mutations are clustered around the ultra-conserved arginine165 and arginine166 codons, the two DNA contact points of E2F1. Using various bioinformatics tools, the *E2F1* mutation found in this study was identified to mutate an arginine residue into a histidine residue thus altering a critical evolutionary conserved DNA contact point responsible for DNA binding and motif recognition.

Since computational modeling is sufficient to pinpoint the mutation's structural location but is inconclusive in showing the mutation's functional effect on DNA binding, ChIP assay was performed showing the R166H mutation abrogates E2F1 DNA binding. Gene expression study on selected E2F1 target genes in over expression system showed inability of E2F1 mutant to adequately up-regulate expression of *SIRT1* and *APAF1* when compared with E2F1 wild type. Of interest is the lack of expression change in *CCNE1*, a known target of E2F1 and an important component in starting S-phase of cell cycle. A possible explanation is the functional redundancy of the E2F family to ensure the cell's replication machinery is operational as mice studies have shown E2F1 -/- mice can be grown to maturity (100,101).

Our study has also shown R166H mutant is much more stable than its wild type counterpart enabling massive accumulation within the cell. Previous study have shown over-expression of E2F1 results in apoptosis induction (99) which is in line with our observation of a drop in proliferation when cells were over-expressing wild

type E2F1; curiously over expressing mutant E2F1 protein did not lead to any noticeable effect on cellular proliferation even though mutant protein levels were many folds higher than its wild type counterpart in equivalent transfection conditions. One explanation for this phenomenon is inactivation of E2F1 decrease apoptosis and its abrogated cell cycle role is compensated by other members of its family. E2F1 -/- mice can grow to maturity and reproduce normally but display a predisposition to develop various cancers (101) indicating the greater importance of tumor suppressive function of E2F1 rather than its cell cycle genes activation function.

An alternative but not mutually exclusive explanation is stable and numerous E2F1 R166H mutants behave functionally like Simian vacuolating virus 40 (SV40) Large T antigens, taking up the lion's share of Rb interaction but with no gene activation ability resulting in free wild type E2F1 to drive cell cycle. While R166H mutation crippled E2F1's DNA binding ability, its other interaction domains including the Rb interaction domain are still active. The mutant's stability and large quantities will favor its preferential binding to Rb due to its sheer numbers and the heterozygous nature of the mutation in the WDPMP tumor would ensure active copies of wild type E2F1 were present to drive cell cycle. This theory is supported by Cress et al. and Halaban et al.; Cress et al. created an E2F1-E132 mutant that is artificially mutated in position 132 within E2F1's DNA binding domain and the mutant is demonstrated to have loss of DNA binding capacity (102) like our R166H mutant. Halaban et al. demonstrated expression of E2F1-E132 mutant can induce a partially transformed phenotype by conferring growth factor independent cell cycle progression in mice melanocytes (103). One possible reason proliferation of E2F1 mutant over expressing cells was not greater than control cells is both mesothelial cell lines used in this study already have a homozygous deletion of *CDKN2A* gene

resulting in p16 null cells. A key part of G1/S checkpoint of cell cycle is p16 deactivation of cyclin dependent kinase 6 (CDK6) which keeps Rb hypophosphorylated thus keeping E2F1 sequestered (104). A p16 null cell already lost its G1/S checkpoint control thus introducing another mutation that will cause the same checkpoint loss will not cause noticeable growth differences.

Given that WDPMP is a rare sub-type of mesothelioma, it is of interest to extrapolate E2F1's role to the more prevalent malignant pleural mesothelioma (MPM). Given *CDKN2A* homozygous deletion is prevalent in MPM with up to 72% of tumors affected (105), G1/S checkpoint is already broken in *CDKN2A* deleted tumors thus in terms of proliferation it is unlikely that an additional E2F1 R166H mutation will be useful as the mutation will be redundant in this context; on the other hand, E2F1 also plays an important role in the activation of apoptosis pathways (99); and the R166H mutation, with its abrogated DNA binding, may contribute to the survival of the cancer cell harboring this mutation. It would be worth checking the remaining 28% of MPMs without *CDKN2A* deletion for possible mutations in *E2F1* and other related genes. It is interesting to note that BAP1, a nuclear deubiquitinase affecting E2F and Polycomb target genes, was recently shown to be inactivated by somatic mutations in 23% of MPMs suggesting that the genes within the E2F pathways might play an important role in mesothelioma in general (106).

Since the publication of this study in 2011, two studies have been published addressing two important questions surrounding this mysterious disease: Is it possible for WDPMP to progress to malignant mesothelioma? Is there a hereditary component to this rare cancer? Both are affirmative in answer. Ribeiro et al. addressed the hereditary question in a study reporting two sisters, the elder developing WDPMP and the younger developing both WDPMP and uveal melanoma, in a Portuguese family

harboring germline *BAP1* mutation (107). As germline *BAP1* mutations has been shown to predispose affected individuals to malignant mesothelioma, it is not surprising that the benign variant, though much rarer, will also develop at an increased frequency in affected individuals (108). Nemoto et al. addressed the progression question in the affirmative in a report describing a Japanese woman diagnosed with WDPMP that progressed to malignancy in 54 months (109). SNP analysis revealed a loss of heterozygosity for the *NF2* gene as well as the neighboring interleukin 17 receptor A (*IL17RA*), checkpoint kinase 2 *(CHEK2*) and SH3 and multiple ankyrin repeat domains 3 (*SHANK3*) genes. LOH of *NF2* was found in the early stage of WDPMP suggesting the LOH event was an early alteration that enabled WDPMP to develop into malignancy. While WDPMP is generally a benign cancer with low chance of recurrence, the above two studies highlight again that given the correct germline or somatic alterations, the chance of developing WDPMP or of WDPMP progressing to full blown mesothelioma can increase substantially. In light of this additional evidence, additional molecular diagnostic tests may be needed to accompany the standard histopathology method of diagnosis to determine the chance of progression and the question of chemo- or radiotherapy in additional to surgical resection will need to be revisited.

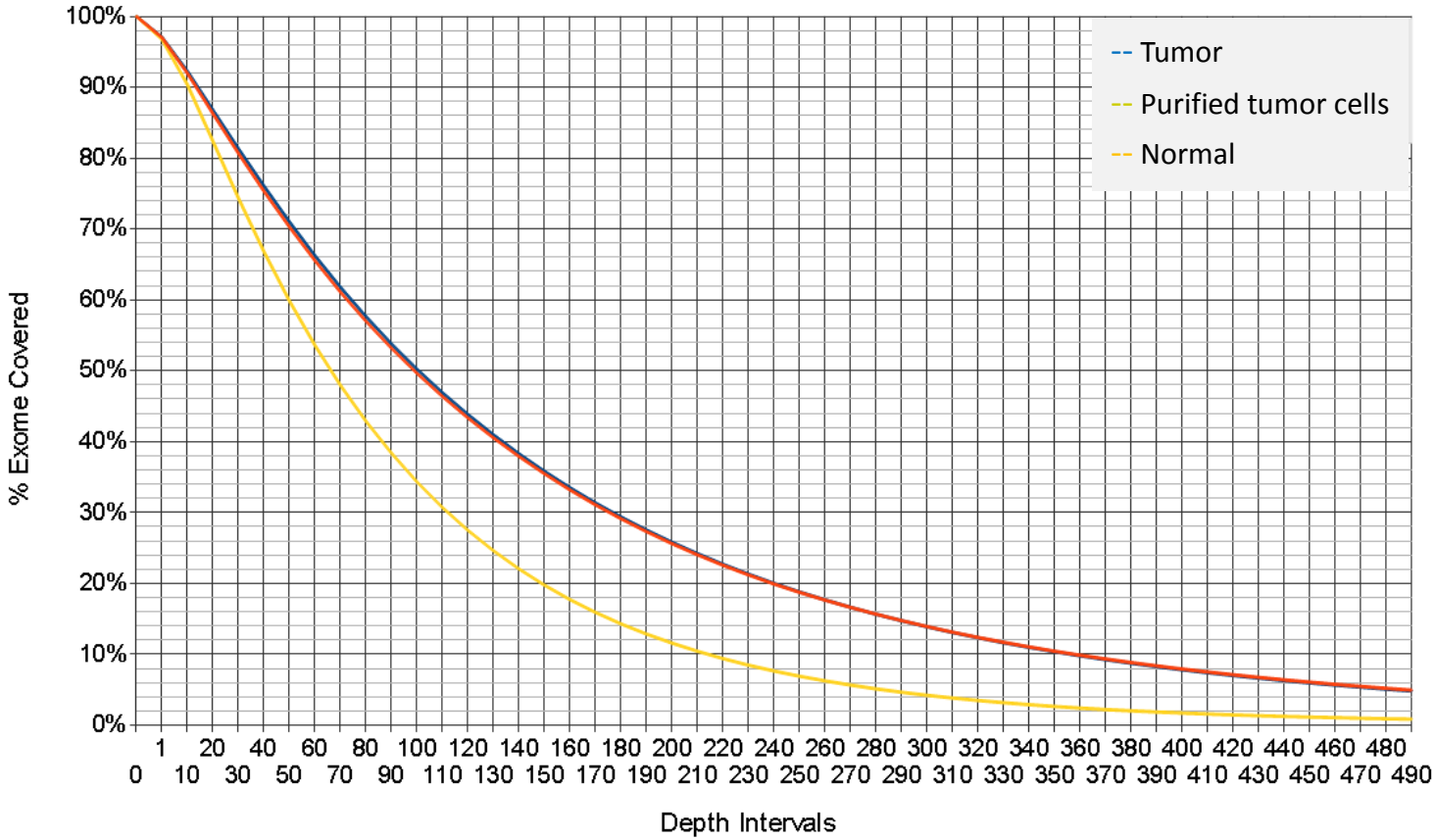


Figure 2.1: **Cumulative WDPMP exome coverage for tumor, normal and purified tumor cells**. Cumulative exome coverage curve for tumor (blue), normal (orange) and purified tumor cells (yellow) is generated by plotting the percentage of the exome represented by different read depths where read depth is defined as number of individual 75bp sequenced read mapped to a particular exome position.

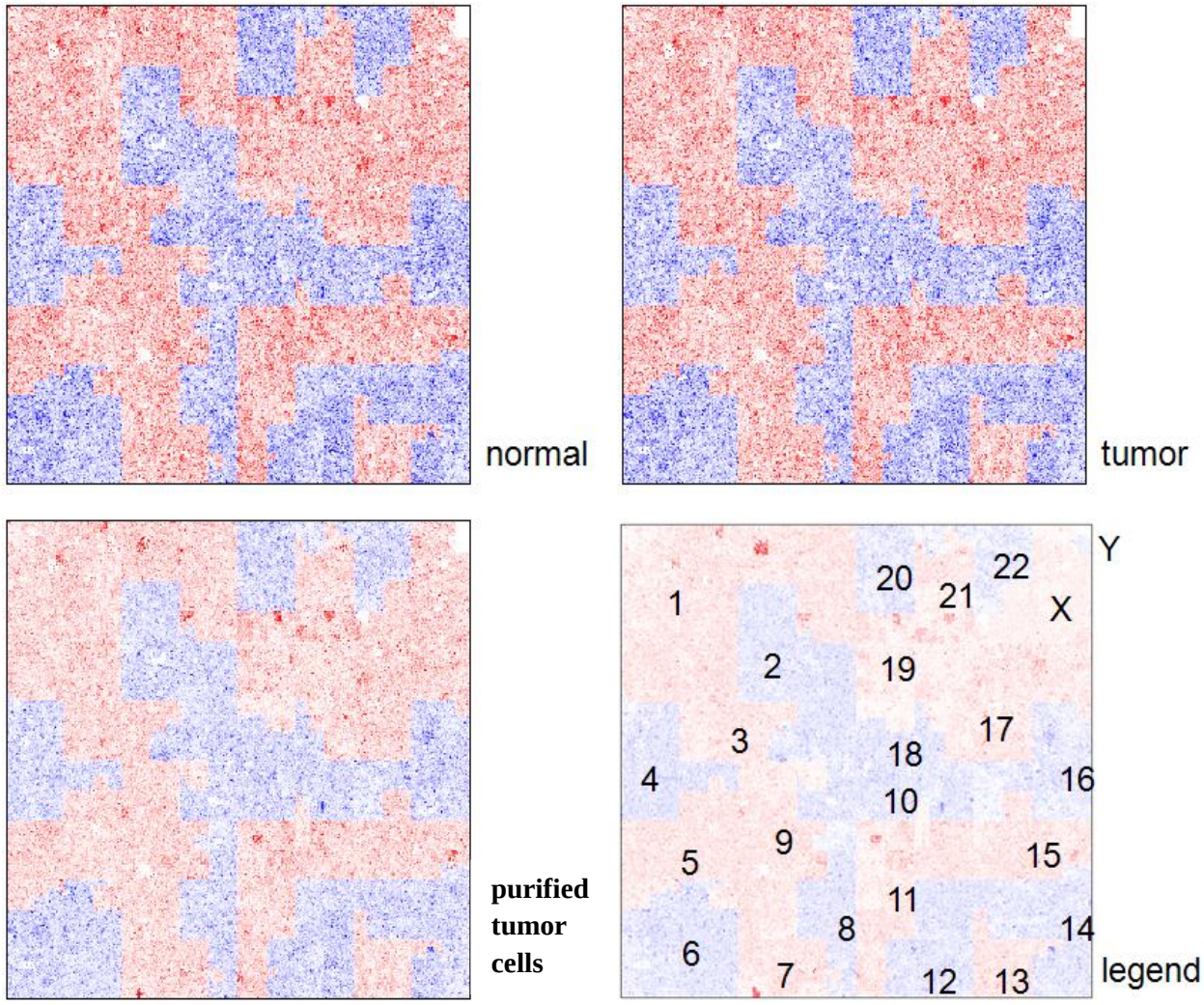


Figure 2.2: Compact representation of WDPMP exome using Hilbert plot. Red and blue color heat mapping is used to demarcate the borders of each chromosome. Increasing color intensity corresponds to increasing sequencing read depth.

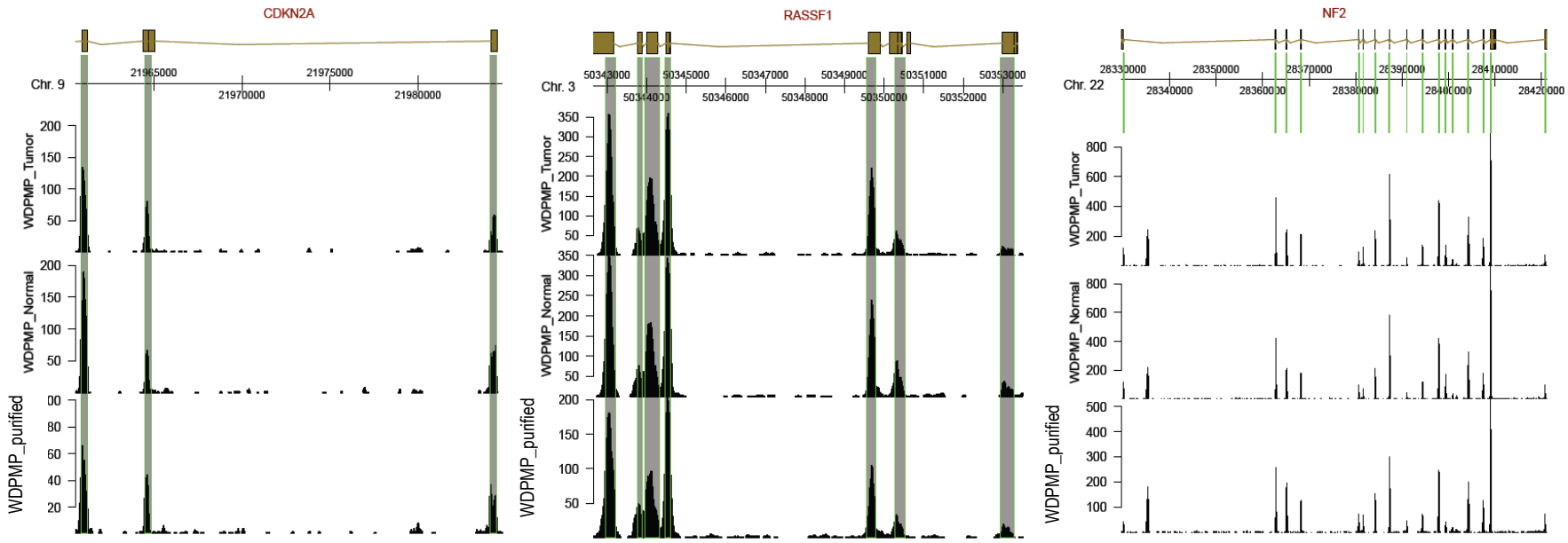


Figure 2.3: **Sequencing coverage at *CDKN2A, RASSF1* and *NF2*.** Each graph shows the exons (brown box) and introns (brown line) as defined by ENSEMBL, the chromosome and chromosomal coordinates of the gene, the actual capture region as defined by Agilent SureSelect Human All Exon Kit v1.01 (gray box with green outlines or green lines if capture region is very small relative to distance between exons), and three plots showing sequencing depth versus chromosomal coordinates for tumor, normal and purified tumor cells.

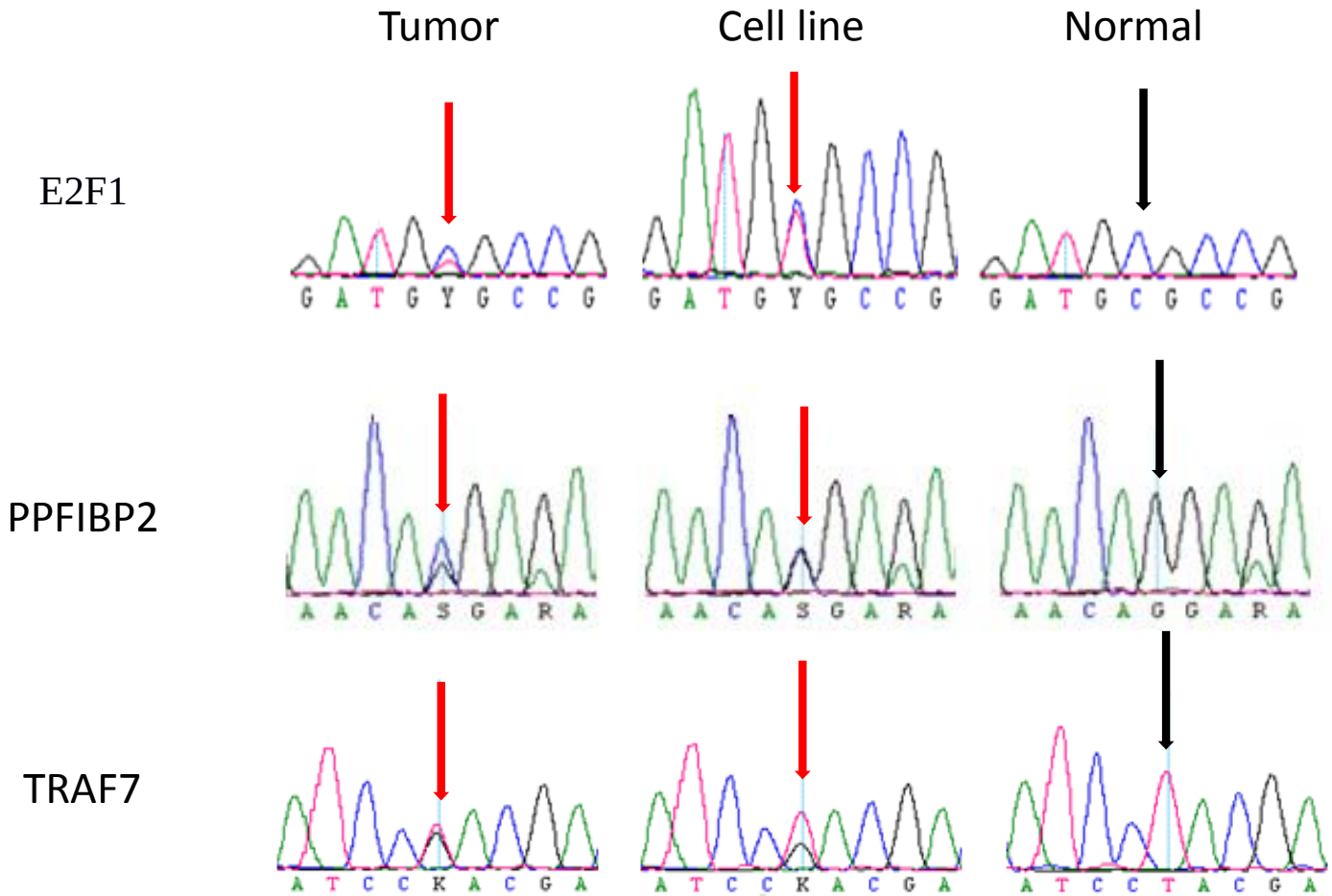


Figure 2.4: Sanger sequencing validation of somatic single nucleotide variants found in *E2F1*, *PPFIBP2* and *TRAF7*.

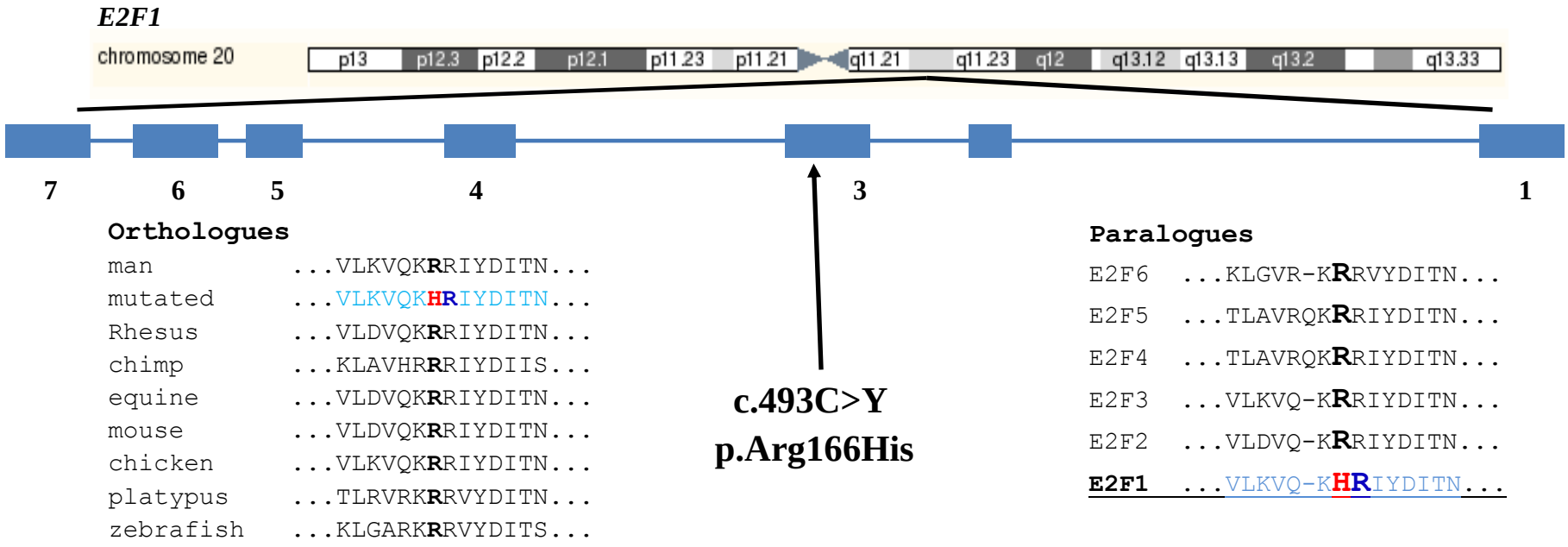


Figure 2.5: Location and conservation analysis of E2F1 R166H. The top part of the panel shows the chromosomal location of E2F1 and then focusing on the location of its exons. E2F1 orthologues conservation analysis with the E2F1 mutated protein sequence is shown in light blue (bottom left). The Arginine-Arginine conservation across diverse species is shown with the Histidine mutation highlighted in red and its Arginine partner highlighted in blue. E2F1 paralogues conservation analysis (bottom right) is shown with E2F1 mutated protein sequence in underlined light blue with the Histidine mutation shown in red and its partner Arginine shown in blue.

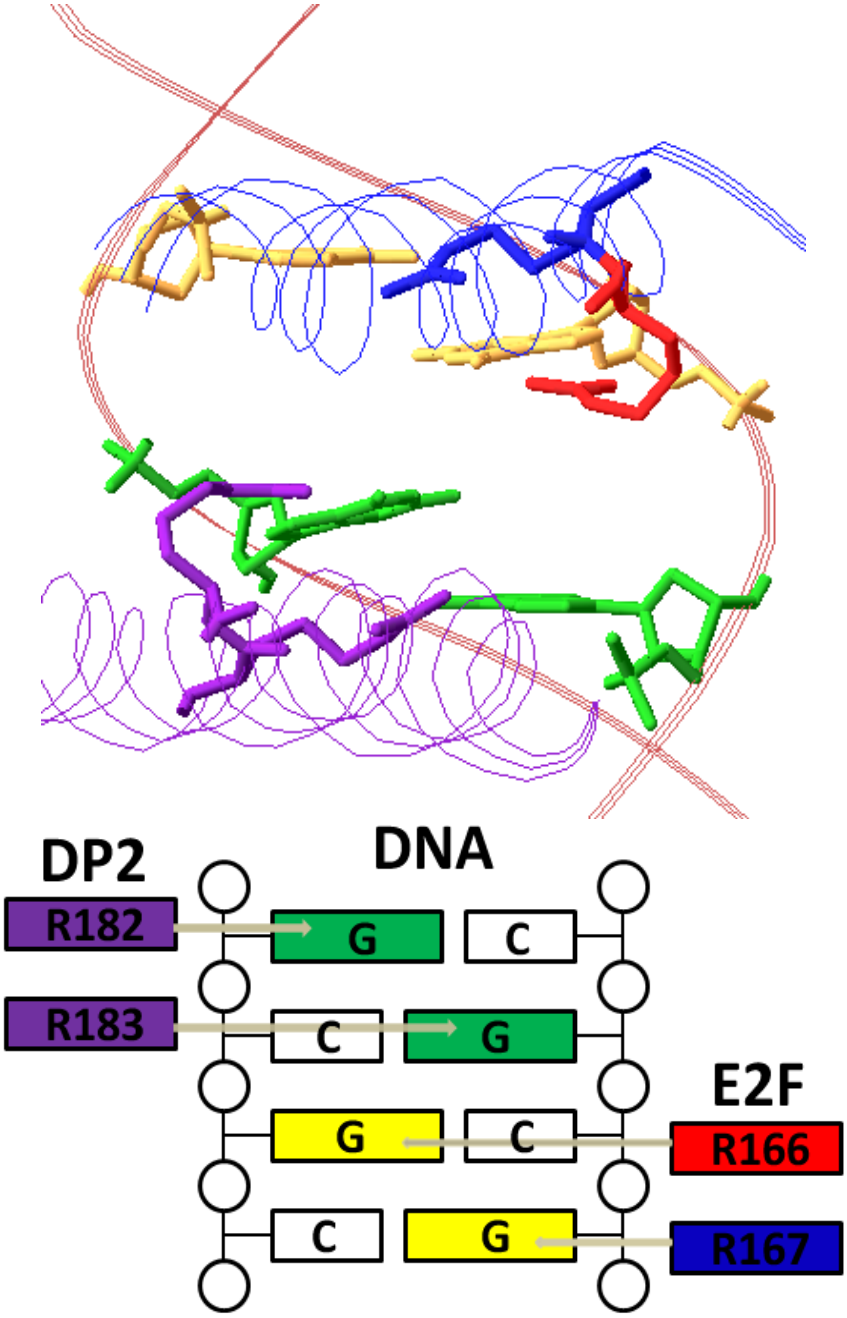


Figure 2.6: Visualization of p.Arg166His mutation location in E2F1. Top panel presents the E2F4 crystal structure [PDB ID: 1CF7] for visualizing the location of the p.Arg166His mutation while the bottom presents a schematic to clearly indicate the residue to nucleotide binding sites The brown double helix is the DNA binding motif with green colored Guanine nucleotides representing binding targets of Arg182-Arg183 of DP2 protein and yellow colored Guanine nucleotides representing binding targets of Arg166-Arg165 of E2F protein. The blue ribbon represents the DNA binding region of E2F with the Arg166 mutation target in red and Arg165 in blue while the purple ribbon represents the DNA binding region of DP2 with the Arg182 and Arg183 in purple.

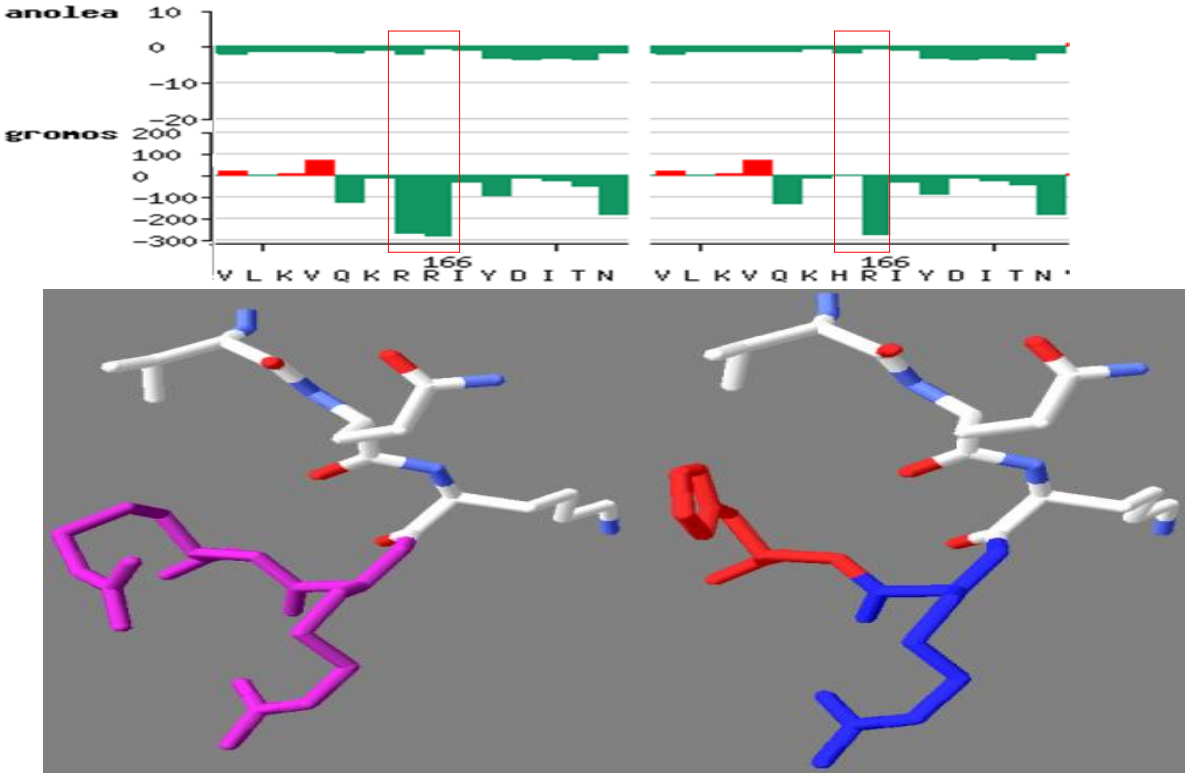


Figure 2.7: Homology modelling of wild type and mutant E2F1 around R166 residue. ANOLEA and GROMOS are used by SWISS-MODEL to assess the quality of the model structure of E2F1 WT and E2F1 R166H mutant DNA binding domain (top). Y-axis represents the energy for each amino acid of the protein with negative energy values (in green) representing favorable energy environment while positive energy values (in red) represents unfavorable energy environments for the amino acid in question. The predicted three dimensional structure of the residues VQK(R/H)R (bottom) is shown with the wild type Arg-Arg residues shown in purple (bottom left), the mutated histidine residue shown in red and its arginine neighbor shown in blue (bottom right). The side-chain of the histidine mutation is clearly predicted to be oriented ~90 degrees counter clockwise compared to the side-chains of its wild type arginine counterpart.

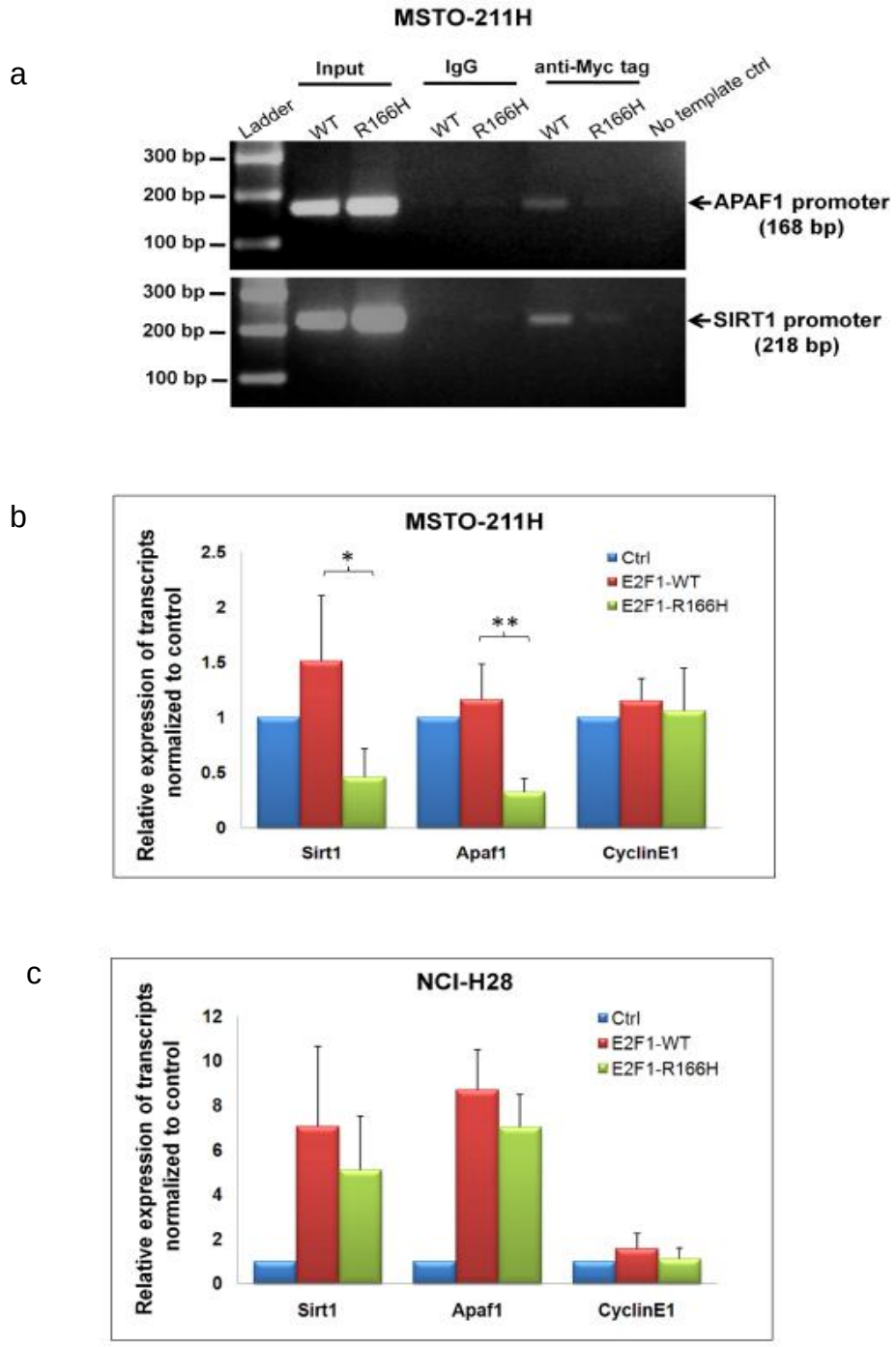


Figure 2.8: E2F1 R166 mutation affects binding efficiency onto promoter targets. a) Chip assay on MSTO-211H transiently transfected with E2F1-WT or E2F1-R166H for 48 hours using anti-Myc antibody. Amplification levels of APAF1 (top) and SIRT1 promoter (bottom) were determined by PCR. Anti-IgG antibody was used as negative control. b and c) Expression levels of E2F1 targets; SIRT1, APAF1, and CCNE1 in MSTO-211H and NCI-H28 that were transfected with indicated plasmids. Each bar represents means ± s.d (n = 3, single asterisk indicates $P < 0.05$, double asterisk indicates $P < 0.01$). Ctrl; Empty vector. Figure reproduced from Yu et al. (2011), *Genome Biol*; 12(9):R96 originally published by BioMed Central.

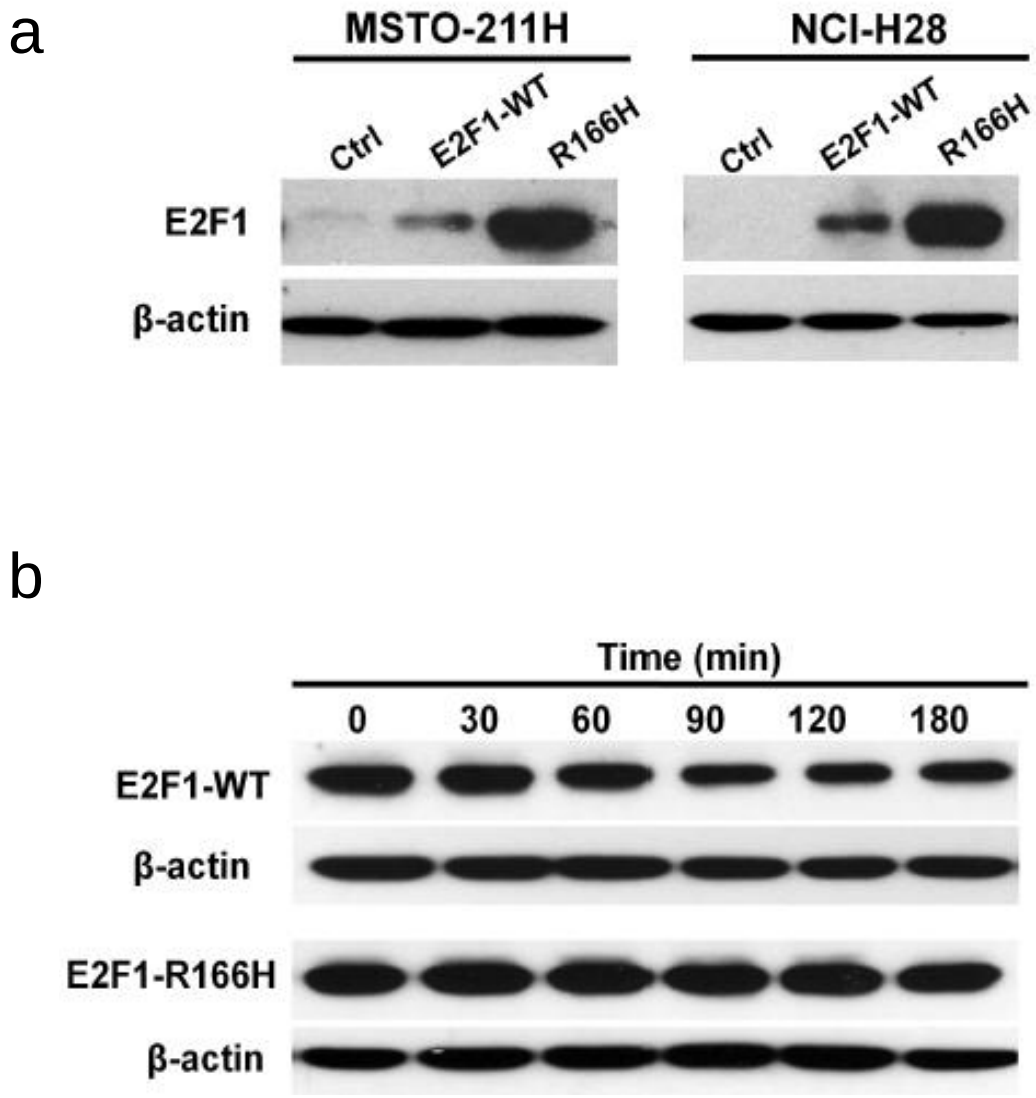


Figure 2.9: Accumulation of mutant E2F1 protein in cells due to increased stability of E2F1 R166 mutation. a) E2F1 protein levels detected by anti-E2F1 Ab (KH95) 48 hours-post transfection. b) Degradation assay performed in MSTO-211H over-expressing E2F1 treated with 25µg/ml cycloheximide. Levels of E2F1 protein were monitored every 30 min up to 3 hours using anti-E2F1 Ab. Figures reproduced from Yu et al. (2011), *Genome Biol*; 12(9):R96 originally published by BioMed Central.

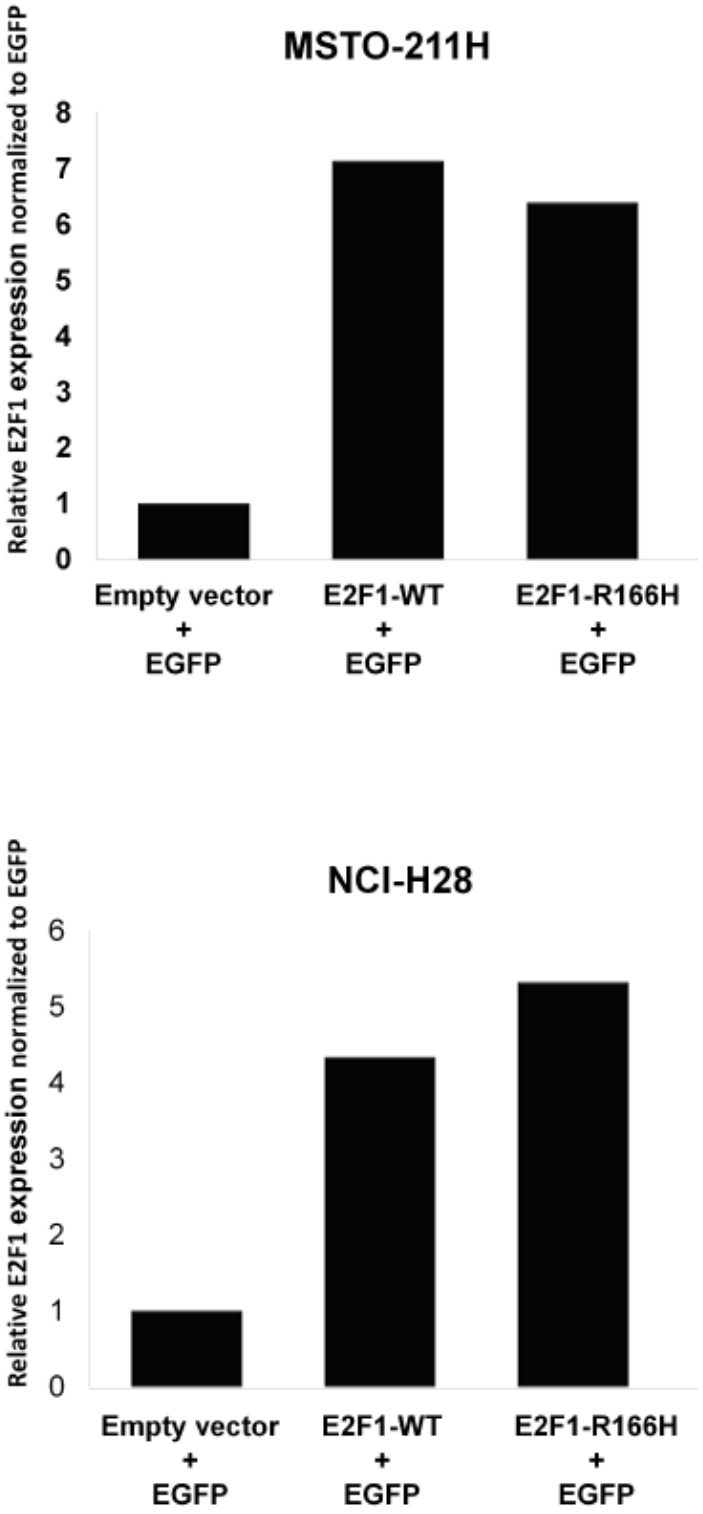


Figure 2.10: Relative expression of E2F1 wild type or E2F1 mutant after co-transfection with EGFP in MSTO-211H and NCI-H28. Figure reproduced from Yu et al. (2011), *Genome Biol*; 12(9):R96 originally published by BioMed Central.

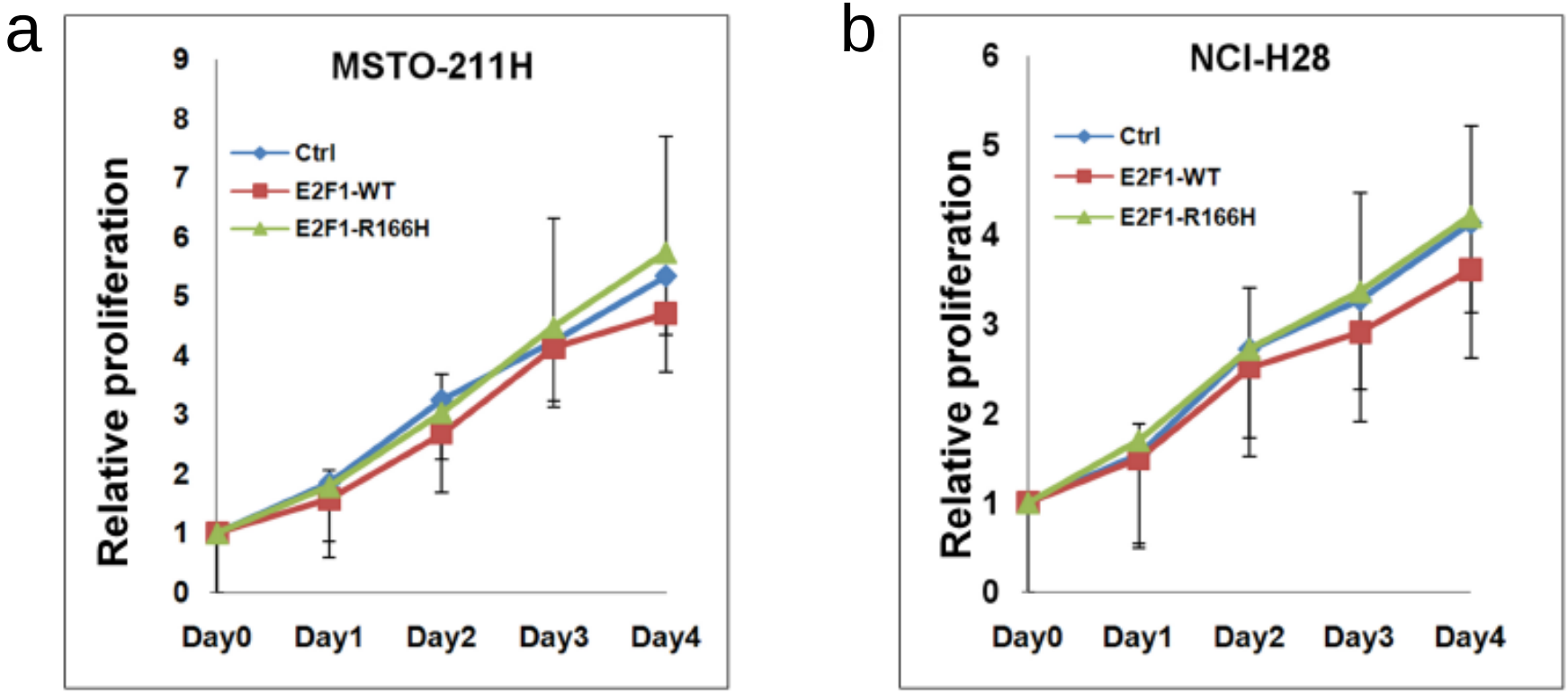


Figure 2.11: Over expression of E2F1 R166H mutant in two mesothelial cell lines. a and b) Proliferation assay after over-expressing E2F1 in MSTO-211H and NCI-H28. Cells were transfected with indicated plasmids for 48 hours. Data are means ± s.d (n = 3). Ctrl; Empty vector. Figure reproduced from Yu et al. (2011), *Genome Biol*; 12(9):R96 originally published by BioMed Central.

Sample ID	Raw Reads	Unalignable Reads	Aligned Reads	% alignable
Tumor	187,023,594	14,717,058	172,306,536	92.13%
Purified tumor cells	119,030,552	3,778,934	115,251,618	96.83%
Normal	190,772,020	14,190,612	176,581,408	92.56%

Sample ID	Reads passing filter	% remaining after filter	# of PCR duplicates	% PCR duplicates
Tumor	129,919,859	69.47%	8,498,978	6.54%
Purified tumor cells	92,512,679	77.72%	15,903,654	17.19%
Normal	137,134,828	71.88%	6,070,718	4.43%

Sample ID	Reads within exome	% reads overall in exome	% exome covered @ 1x	% exome covered @ 20x
Tumor	80,042,870	61.61%	97.10%	86.80%
Purified tumor cells	49,312,008	53.30%	96.80%	82.40%
Normal	80,869,734	58.97%	97.00%	86.30%

Table 2.1: Overall WDPMP Exome Sequencing Summary.

Chr	Position	Reference base	Consensus base called	Gene Symbol	Change Type	AA Change
chr7	2718841	G	A	AMZ1	Nonsyn	V434M
chr6	88180328	A	G	C6orf165	Nonsyn	T92A
chr20	60920345	C	C/A	COL9A3	Nonsyn	P60T
chr20	31731297	C	C/T	E2F1	Nonsyn	R166H
chr16	88408515	C	C/T	FANCA	Nonsyn	G66D
chr7	72381368	G	T	FKBP6	Nonsyn	R82L
chr12	150548	A	A/G	IQSEC3	Nonsyn	E722G
chr1	165225280	C	T	MAEL	Nonsyn	R23W
chr7	151513605	C	C/A	MLL3	Nonsyn	A1685S
chr1	1236253	C	C/G	PUSL1	Nonsyn	H267Q
chr11	9629588	G	G/C	PPFIBP2	Nonsyn	Q791H
chr7	5747327	C	A	RNF216	Nonsyn	G283W
chr16	30687806	G	A	RNF40	Syn_or_UTR,possible_5_prime_splice	-
chr3	43364013	G	G/A	SNRK	Nonsyn	A420T
chr16	21666165	T	T/G	TRAF7	Nonsyn	Y621D
chr6	41234432	T	A	TREM2	Nonsyn	S213C
chr1	169954974	A	A/G	VAMP4	Nonsyn	S42P
chr1	85324174	A	A/G	WDR63	Nonsyn	S205G
chr11	64640552	G	A	ZNHIT2	Nonsyn	R384W

Table 2.2: Putative somatic nonsynonymous mutations found using the single nucleotide variant discovery pipeline.

Chapter Three: Exome Sequencing of Liver Fluke-associated Cholangiocarcinoma

Part of the findings in this Chapter was published in Ong et al. (2012), *Nat Genet*; 44(6):690-693 (pp 53-92 of this thesis).

3.1: Introduction:

Cholangiocarcinoma (CCA) is a rare malignant cancer of the biliary tract classified as an adenocarcinoma: carcinoma derived from glandular tissue or in which the tumor cells form recognizable glandular structures. Other than CCA, biliary tract cancers also include cancer of the pancreas, gall bladder and ampulla of Vater. There are several risk factors associated with the development of CCA such as primary sclerosing cholangitis, *Opisthorhis viverrini* (OV) parasitic infection and hepatolithiasis, the presence of gallstones in the biliary ducts (110). Incidences of CCA in Thailand (30/100,000) are much higher than the rest of the world (1.36/100,000) (111); in particular, CCA incidences in northeastern Thailand, where OV infestation are also particularly high, double when compared with the rest of the country (112). Presence of OV eggs in bile or liver biopsy correlate with increased risk of CCA (odds ratio [OR] 10.8; 95% confidence interval [CI] 1.1-108.4) with the highest risk of CCA observed when the number of eggs per gram of feces exceeds 6000 (OR 14.1; 95% CI 1.7-119) (113,114).

OV life cycle involves two intermediate hosts, snails and fresh water fish, as well as humans as its definitive host. As fresh water fish is a major indigenous food source for northeastern Thai population, the main vector of infestation is via consumption of raw, fermented and/or undercooked OV infested fishes (115). OV infestations are mainly found in the bile ducts and acute inflammation of large intrahepatic bile ducts and portal connective tissues are signs of early human host response (108). Long term OV infestation results in periductal fibrosis and scarring of the bile duct epithelium (116,117). The standard treatment for OV infestations is praziquantel; while the treatment is effective, some existing periductal fibrosis persists even after parasite removal indicating the damage wrought by long term OV

habitation may be irreversible (118). Due to regional dietary and culinary traditions, the cycle of re-infestations followed by praziquantel treatment frequently occurs and treatment frequency is a risk factor for CCA (119).

CCA is a lethal malignancy since complete surgical resection of primary tumor and any metastases is currently the only potential curative option and the majority of CCA cases being inoperable (110,120). The 5-year survival rate for patients diagnosed with CCA is <5% with survival rate falling to zero for patients with CCA that has metastasized to distal lymph nodes (121,122). There is no standard chemotherapy treatment for CCA; pancreatic cancer treatments, such as gemcitabine plus cisplatin, are often adopted as a starting point although with only marginal improvement to survival length (123).

Given the dismal survival length for patients diagnosed with CCA and the lack of effective therapeutic options in the treatment of this disease, exploring and cataloguing the somatic alterations present in CCA represent a first step towards understanding the genetic basis for this cancer. Recent advances in the ability to selectively capture DNA representing the coding regions of the human genome as well as the ability to perform massively parallel sequencing presents a timely and cost-effective way to explore and catalogue the somatic alterations of multiple cancer samples. For this study, whole-exome DNA sequencing was performed on a discovery set of eight OV-associated CCAs as well as their matched normal tissue. Recurrently mutated genes found in the discovery set were validated in a separate prevalence set of 46 matched OV-associated CCA cases.

3.2: Results

3.2.1: Clinical samples and information

DNA from 8 patients diagnosed with CCA, consisting of 5 males and 3 females [mean age 57 years (range 48 to 66 years)] were obtained along with DNA from matched normal tissue from each patient, constitutes the discovery set (Table 3.1a). Whole-exome captured DNA libraries obtained from the discovery set were sequenced using Illumina GAIIx 76bp Pair-End sequencing technology. For the prevalence set, DNA of 46 CCAs along with matched normal tissue from 29 males and 17 females [mean age 56 years (range 37 – 73 years)], were selected (Table 3.1b). Variant verification for both sample sets was performed using Sanger sequencing.

3.2.2: CCA whole-exome analysis

Whole-exome data were aligned using BWA (37) against the hg18/NCBI36.1 reference genome build. Read quality filtering and PCR duplicate removal were performed using SAMtools (39). We obtained an average sequencing depth of 79.7 with >81% of the exome sequenced to 20x depth, enabling high confidence variant calling (Table 3.2). To detect single nucleotide variants and small indels, a discovery pipeline based the Genome Analyzer ToolKit (41) was employed. Details of this discovery pipeline are discussed in Chapter 1. Our discovery set had an average of 26 somatic non-synonymous variants/tumor (range 19-34) affecting 187 genes and of the 206 Sanger sequencing confirmed variants, 191 were due to somatic single nucleotide base substitutions and 15 were due to small indels of between 1 – 12 base pairs (Table 3.3). Of the 191 base substitutions, 165 were predicted missense type mutations, 16 were predicted nonsense type mutations and 9 were mutations predicted to disrupt the splice site of the affected gene. Of the 15 validated somatic indel events, six were

deletion events of between 1 – 12 base pairs and nine were insertion events of 1 – 4 base pairs.

3.2.3: Mutational analysis of CCA discovery set

In the discovery set, we found recurrent mutations, defined as somatic non-synonymous alterations in two or more samples, for 13 genes in total (Figure 3.1A). 75% of the samples (6/8) in the discovery set contain *TP53* and/or *SMAD4* mutations with over 50% of the mutations being of nonsense and frameshift mutations in keeping with their known tumor suppressive roles. 50% (4/8) of the discovery set contains *KRAS* and/or *GNAS* mutations corresponding to well-known mutational "hotspot" codon sites: p.Gly12Ala or Gly13Asp for KRAS and p.Arg201Cys for GNAS (Figure 3.1A, Table 3.4). There are a number of somatic nonsense and frameshift mutations to *RNF43* (1/4), *MLL3* (2/2), *ROBO2* (2/2) and NDC80 kinetochore complex component (*NDC80*) (1/2) suggesting these genes serve an important tumor suppressive function in the context of this cancer. Since phosphatase and tensin homolog (*PTEN*) and *CDKN2A* were known to be recurrently mutated for CCA (124,125), the mutation status of both genes are included for completeness and both are found to be singly mutated in the discovery set.

3.2.4: Prevalence analysis of somatic mutations found in CCA discovery set

In order to more accurately determine the mutation frequencies of the 15 recurrently mutated genes found in the discovery set or in previous studies, further whole-gene validations were performed using a prevalence set of 46 matched CCA tumor pairs. We verified that 44% (24/54) of the CCA samples contained a somatic mutation in the *TP53* gene, all of which have been reported in COSMIC database

(Table 3.4). Of these, 46% (11/24) are mutations of the nonsense or frameshift category predicted to produce a truncated protein.

KRAS activating mutations (codon 12 and 13) are identified in 16.7% (9/54) of cases (Figure 3.1C, Table 3.4). We confirmed a recurrent mutation rate of 16.7% (9/54) in *SMAD4* for CCA with 44% (4/9) of mutations being of nonsense or frameshift type predicted to produce a truncated protein lacking its Smad4 activation domain (SAD) and/or its Mad homology 2 domain (MH2). In addition, two different *SMAD4* missense mutations (p.R145Q, p.R261C), each mutation falling within either MH1 or MH2 domain of *SMAD4*, are detected in a single sample (B149) (Figure 3.1B, Table 3.4).

One of the novel recurrently mutated genes identified in this study is *MLL3*, a histone-lysine N-methyltransferase. Somatic mutations of *MLL3* are identified in 14.8% (8/54) of CCA tumors at a similar mutation frequency observed in *SMAD4* (16.7%) and *KRAS* (15%) (Figure 3.1C). Seventy five percent of the *MLL3* mutations (6/8) observed are in the mutational categories of nonsense, frameshift or splice-site type. The mutations found are distributed throughout the entirety of the gene suggesting *MLL3*'s function as a tumor suppressor gene (Table 3.4) with all truncating mutations predicted to result in protein products lacking the whole or part of the key methyltransferase domain.

Other genes found mutated in approximately 10% of CCA cases were *GNAS, ROBO2* and *RNF43.* The mutations to *GNAS* are "hotspot" in nature and are predicted to alter a single arginine codon (R201). Of the nine samples mutated in *RNF43* or *ROBO2*, over half (5/9) are of the nonsense or frameshift category. While xin actin-binding repeat containing 2 (*XIRP2*), *PEG3,* Ras association and DIL domains

(RADIL), *NDC80* and protocadherin alpha 13 (*PCDHA13*) are mutated at a lower frequency of 3.7% to 5.6%, only *XIRP2* and *PEG3* are found to have additional mutations in the prevalence set (Figure 3.1B).

3.2.5: Mutational landscape comparison between *O. Viverrini*-associated cholangiocarcinoma, pancreatic ductal adenocarcinoma and hepatitis C virus-associated hepatocarcinoma

With the availability of mutational data from PDAC and HCV-associated HCC studies (19,126), a comparison can be made between these two data sets and the OV-associated CCA data set at the level of mutated genes and also in the distribution of base substitutions. At the level of recurrently mutated genes, the three cancers display several interesting points of similarities and differences (Table 3.5). Overall, there appears to be two distinct genetic paths to cancer that differentiates OV-associated CCA and PDAC from HCV-associated HCC. Other than *TP53* being the commonly mutated gene across the three cancers, there is a clear separation of HCV-associated HCC from the other two cancers from a gene level perspective with HCV-associated HCC displaying a mutually exclusive set of recurrently mutated genes. *TP53, KRAS, SMAD4, CDKN2A and MLL3* are commonly mutated in both OV-associated CCA and PDAC, although the mutation frequency for the first four genes is lower in OV-associated CCA while the opposite is true for *MLL3* mutation frequency. Novel recurrently mutated genes in our study (*GNAS, RNF43, ROBO2* & *PEG3*) in CCA are not found to be mutated in PDAC nor HCV-associated HCC.

Next we categorize the somatic base substitutions for each cancer into eight substitution classes (Figure 3.2a). Again, two distinctive patterns of base substitutions separate CCA and PDAC from HCV-associated HCC. One pattern is the high proportion of C>T substitutions in the context of XpCpG for CCA and PDAC when

compared with HCV-associated HCC. The other pattern is the high proportion of thymine substitutions unique to HCV-associated HCC especially in the T>C context as well as proportionally higher C>A substitutions. The Chi-Square statistical test indicates OV-associated CCA has a mutational spectra distinct but related to PDAC ($P = 0.0099$), but has a very different mutational spectra when compared to HCV-associated HCC ($P < 1 \times 10^{-16}$) (Table 3.6, Figure 3.2b).

3.3: Discussion

CCA occurs at a much higher rate in northeastern Thailand and Laos due to consumption of *O. Viverrini* infested fish products. As surgical resection is the only potential curative option, CCA is an almost universally fatal disease due to late stage disease detection and diagnosis where surgery is not possible. Due to the relative rarity of this cancer outside of Thailand and Laos and the socio-economic status of the population afflicted with this disease, therapeutic options for CCA has not advanced beyond palliative measures employing guidelines used in the treatment of pancreatic cancers as a crude guide. Whole-exome capturing and sequencing presents a unique opportunity to not only begin to catalogue the somatic mutations of this neglected disease but also to start to elucidate and compare the mutational patterns of OV-associated CCA with that of more vigorously studied PDAC and HCV-associated HCC.

In total, 206 somatic non-synonymous alterations are identified and validated in the discovery set consisting of eight whole-exome sequenced OV-associated CCAs and their corresponding normal tissues. Fifteen recurrently mutated genes are identified from the discovery set and their mutational frequencies characterized through whole-gene Sanger sequencing in a prevalence set consisting of 46 matched

pairs of OV-associated CCA. This list of recurrently mutated genes are by no means exhaustive; this is evidenced by the necessity to include *PTEN* and *CDKN2A* in the mutational frequency assessment even though they are not found to be recurrently mutated in our discovery set but nonetheless were previously known recurrently mutated CCA-associated genes (124,125). With the decreasing cost of high throughput sequencing and advances in data analysis, subsequent follow-up studies involving larger numbers of whole-exomes and eventually whole-genomes will no doubt add to the mutational catalog and refine the mutational frequencies that are observed in this modest study.

The top three recurrently mutated genes in this study are *TP53* (44.4%), *KRAS* (16.7%) and *SMAD4* (16.7%). Almost one-half of the *TP53* mutations in this data set are predicted to produce a truncated protein; as *TP53* demonstrated haploinsufficiency in previous studies, a state where a single functional copy does not produce enough gene product for wild-type function, even a single inactivated *TP53* copy will be enough to abrogate normal TP53 functions (127-130). The remaining *TP53* mutations are missense mutations all occurring in the DNA binding domain and are likely to alter the proper function of the protein (131). One of TP53's many functional roles is to act as a regulator of Wnt signaling where TP53 exerts its influence through transcriptional activation of the microRNA miR-34. This microRNA directly targets *WNT* and *β-catenin* genes, potent activators of the Wnt pathway, with knockdown of TP53 leading to increased Wnt signaling (132,133). As the *TP53* mutations in our study are all predicted to interfere with the transcriptional activity of the protein or to produce a truncated version of the protein, one of the likely functional consequences of these mutations will be a reduction in miR-34 production leading to aberrant Wnt signaling.

RNF43 are found to be altered in 9.3% (5/54) in our study and its gene product is a RING-type E3 ubiquitin ligase, highly expressed in colon cancer, that interacts with HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1 (NEDL1) and p53, and suppresses p53-mediated apoptosis (134). Two out of five somatic *RNF43* mutations found in this study are of nonsense type, while the remaining missense mutations were predicted by Polyphen to be functionally damaging to the protein (Table 3.4) suggesting RNF43 is inactivated in OV-related CCA. Since the publication of this study, additional research has implicated *RNF43* inactivation to the aberrant activation of the Wnt/β-catenin signaling pathway (135,136). Jiang et al's study revealed RNF43 inhibits Wnt/β-catenin signaling by reducing the membrane level of the Frizzled protein in pancreatic cancer cells. Giannakis et al. revealed truncating mutations to *RNF43* segregate from inactivating adenomatous polyposis coli (*APC*) mutations in colorectal tumors suggesting a functional equivalency focused on loss of control over β-catenin degradation. Interestingly, pancreatic cell lines with *RNF43* inactivation appeared to be onco-addicted to the Wnt/β-catenin signaling pathway; this is evidenced by their sensitivity to LGK974, a small molecule inhibitor of porcupine homolog (Drosophila) (PORCN) whose function is as a mediator of Wnt signaling and the inhibitor is currently in Phase I clinical trial (137). Given the presence of *RNF43* inactivations in CCA and the mutational landscape similarities between CCA and pancreatic cancer, there is a high chance LGK974 may also be similarly effective in treating a subset of *RNF43* inactivated CCA.

PEG3 is a maternally imprinted gene and its encoded product interacts with siah E3 ubiquitin protein ligase 1A (SIAH1A), and inhibiting PEG3 activities blocks p53-induced apoptosis (138). Through the use of zebrafish model, PEG3 was shown to play a role in the Wnt signaling regulation through its binding to β-catenin and

promoting its degradation (139). In the context of cancer, promoter hypermethylation of PEG3 were found in primary human gliomas with loss of PEG3 aberrantly activating Wnt signaling leading to chromosomal instability (139,140). These above reports support a tumor suppressive role of PEG3 in CCA.

The mutational consequences of *TP53*, *RNF43* and *PEG3* point to abnormal activation of Wnt signaling through loss of control over Wnt ligand production and/or β-catenin degradation. The PORCN inhibitor, LGK974, has shown effectiveness in *RNF43* inactivated pancreatic cell lines and has since been approved for clinical trials. LGK974 may be equally effective in treating *RNF43* inactivated CCA and further studies showing the effectiveness of LGK974 in CCA context will lead to not only new therapy options for patients suffering from CCA but also a new line of approach in studying CCA in the context of the Wnt signaling pathway.

For this study, all mutations to *KRAS*, a member of the Ras superfamily of small GTPases, are known activating mutations altering a critical glycine residue in codon 12 of the protein; the mutation rate of *KRAS* is similar to a previous report showing *KRAS* mutations in 15% of CCA cases (141). Another gene showing similar "hotspot" type activating mutations is *GNAS*, a stimulatory G-protein alpha subunit involved in classical activation of adenyl cyclase that has previously been reported in thyroid carcinomas, adenocortical lesions, pituitary tumors, kidney, leydig cell tumors and colorectal cancer (142); the somatic mutations to this gene are predicted to alter the arginine residue of codon 201 resulting in aberrant activation of this protein and was shown to be associated with invasive progression in intraductal papillary mucinous neoplasm of the pancreas (143,144). Given the singular alteration to this gene, the wide spread occurrence of this alteration in different cancers and its

association with invasive progression, these lines of evidence present a compelling argument in support of *GNAS* mutations belonging to the driver mutations category as well as its potential for targeted small molecule inhibition.

SMAD4 is a known tumor suppressor gene and plays a pivotal role in TGF-β signaling pathway (145,146,147). In a previous study, *SMAD4* was shown to be mutated in 16% (5/32) of CCA samples (148). While the ethnic origin or OV association of the samples used in the study was not explicitly stated, the authors' affiliations with hospitals located in Germany suggest the samples were of European origins and most likely not associated with OV. In our larger cohort of OV-associated CCAs, there is a remarkable similarity in *SMAD4* mutation frequency, 16.7% (9/54) for our study compared with 16% (5/32) found by Hahn et al., pointing to not only the gene as a key player in CCA development but also suggesting a convergent tumor development independent of risk factors associated with CCA or perhaps the different risk factors converge to a common CCA mutational process. Interestingly, an in-vivo study demonstrated that hepatocyte and bile duct epithelial cell specific *SMAD4* knock-out mice were not observed to develop cancer thus demonstrating that *SMAD4* inactivation alone is insufficient to cause CCA (149). However, the same study also demonstrated double inactivation of *SMAD4* as well as *PTEN* is sufficient for development of CCA in mice highlighting *SMAD4*'s role as an enhancer to another activated oncogenic pathway on the road to tumorigenesis rather than as an initiator of cancer development. Interestingly, each of the two *PTEN* mutated tumors also harbored *SMAD4* mutations, highlighting the synergistic significance of both genes demonstrated in the double knockout model.

Among the novel mutated genes identified in this study is *MLL3*; this gene codes for a histone-lysine N-methyltransferase and implicated in numerous cancer types such as pancreatic cancer and medulloblastoma (19,150). In this study, *MLL3* is found to be mutated at a frequency of 14.8% comparable to the rate observed in *KRAS* and *SMAD4* but higher than the observed rates in pancreatic cancer (7.8%) and medulloblastoma (3.4%). Notably, most of the *MLL3* mutated tumors do not harbor *TP53, KRAS* and *SMAD4* mutation despite the fact that these three genes are mutated in 57% of CCA. Taken together, the data suggested that *MLL3* might play an important role in CCA by presenting an alternate mutational route to tumorigenesis. Furthermore, three other samples in our discovery set, excluding the two *MLL3* mutated tumors, harbor somatic mutations in genes encoding histone modifying enzymes such as histone deacetylase 2 (*HDAC2*), histone deacetylase 4 *(HDAC4*) and protein arginine methyltransferase 6 (*PRMT6*), indicating that histone modifying enzymes are involved in the tumorigenesis of CCA.

Cholangiocytes and hepatocytes are known to differentiate from the same hepatic progenitor cells located at the Canal of Herring in liver and these liver stem cells has been implicated in both CCA and HCC (151). Several studies indicate that the biliary tree contains stem cell compartments for liver, pancreas and the bile duct system and persist into adulthood (152). In another study, the biliary tract shows some potential for pancreatic differentiation and both biliary tract and pancreas has similar pathological features (153). The above studies highlight the intimate relationship and common origins between these three organs and lead us logically to compare and contrast our data with published mutations in PDAC and HCV-associated HCC (Table 1) (19,126). By comparing these three data sets at the gene level and also at the base substitutions level, there is a clear segregation, at both levels, with HCV-

associated HCC on one side and OV-associated CCA with PDAC on the other side. Other than *TP53*, the most commonly mutated gene across a wide spectrum of cancers, being the commonly mutated gene across all three data sets, *KRAS*, *SMAD4*, *CDKN2A* and *MLL3* are commonly mutated in both OV-associated CCA and PDAC but are not mutated in HCV-associated HCC; while recurrently mutated genes, catenin (cadherin-associated protein), beta 1, 88kDa (*CTNNB1*), AT rich interactive domain 2 (ARID, RFX-like) *(ARID2*), Dmx-like 1 *(DMXL1*) and NLR family, pyrin domain containing 1 (*NLRP1*), in HCV-associated HCC are not found to be mutated in OV-associated CCA or PDAC. As somatic mutations observed in a tumor are a result of one or more endogenous or exogenous mutational processes, differences observed in the distributions of base substitutions may point to different mutational processes at work. There appears to be two distinct mutational processes at work clearly separating OV-associated CCA and PDAC from HCV-associated HCC. For OV-associated CCA and PDAC, the observed mutational pattern of XpCpG at C>T base substitutions indicate the mutational process of 5-methylcytosine deamination to thymine (154). While for HCV-associated HCC, the abundance of thymine to cytosine or its complement adenine to guanine substitutions points to the mutational process of adenine deamination to hypoxanthine which pairs with cytosine in a selective manner resulting in a post-replicative transition mutation pattern observed (155). An interesting conjecture arising from these observations is whether mutational processes, fundamentally a stochastic process, can create emerging mutation patterns at a gene level; certain groups of genes might not be readily targeted by one mutational process due to its base mutation preferences: for example, mutational process creating A>G:T>C substitutions will not be able to create the activating *KRAS* G12 mutation thus *KRAS* is less likely to be mutated in an environment dominant in

A>G:T>C substitutions. However, that same gene group may be more amenable to deleterious mutations by another mutational process: for example, the same gene *KRAS* is much more likely to be mutated at its G12 codon under a dominant C>T mutational process.

We have shown that CCA and PDAC are mutationally similar in the gene and base substitution level with an interesting conjecture that the common mutational process driving both cancers may also be driving the appearances of commonly mutated genes. Due to these similarities, therapies developed for PDAC, a more prevalent and better studied disease, may have translational applications in treating CCA. There are several lines of evidence that suggest the drug Minnelide, recently approved for phase I clinical trials in treating pancreatic cancer, may also be effective in the treatment of CCA (156). Minnelide is a synthetic analog of triptolide, an extract from the plant *Trypterigium wifordii*. Triptolide has shown effectiveness in the CCA cell line study and also in in-vivo hamster model study before (157,158) but failed to advance to human trials due to its low solubility in water. Minnelide is rationally designed to be more water soluble while retaining the effectiveness of its naturally occurring predecessor. Chugh et al. demonstrated Minnelide's effectiveness in decreasing cell viabilities of *KRAS*/*TP53* mutated pancreatic cell lines: S2-013, MIA PaCa-2, S2-VP10, and Panc-1 (159). In addition, three independent complementary mice models (orthotopic, xenograft and KRasG12D/Trp53R172H/Pdx-1Cre spontaneous pancreatic cancer mouse model) were employed in the study to test the effectiveness of Minnelide and all three models showed the Minnelide was effective in reducing tumor growth and spread while improving survival (159). Given the evidence of Triptolide effectiveness for in-vitro and in-vivo CCA models and the

mutational similarities between pancreatic cancer and CCA, it is highly probable that Minnelide will also be effective in treating *KRAS/TP53* mutated CCA subset.

The short term implications of this study are two possible targeted therapy paths in the treatment of CCA where there were none before: One, the use of PORCN inhibitor LGK974 to target Wnt signaling dependent CCAs driven by *RNF43* and possibly *PEG3* inactivating mutations and two, the use of Minnelide, synthetic soluble version of triptolide, to treat *KRAS/TP53* mutated CCAs. Assuming the effectiveness of both LGK974 and Minnelide in the treatment of CCA, there will be a need for a companion biomarker panel in order for oncologists to personalize the therapy for each patient encountered. Based on current knowledge, a simple targeted DNA capture and sequencing of the genes *TP53, KRAS* and *RNF43* can be a starting point in streaming patients to the most effective therapy. This panel can be expanded as additional genetic mutations are implicated in the effectiveness or resistance to existing therapies and/or new therapies. This will provide a systematic approach in the treatment of CCA through the concept of personalized therapy. The longer term implication of this study is the idea that individual stochastic mutational processes may be driving the emerging recurrent gene mutational patterns that we see in different cancers. If this is true, there may only be a small number of gene pathways that can be effectively mutated by a particular mutational process and blocking these mutational processes may severely limit the number of developmental paths that a cancer can take.

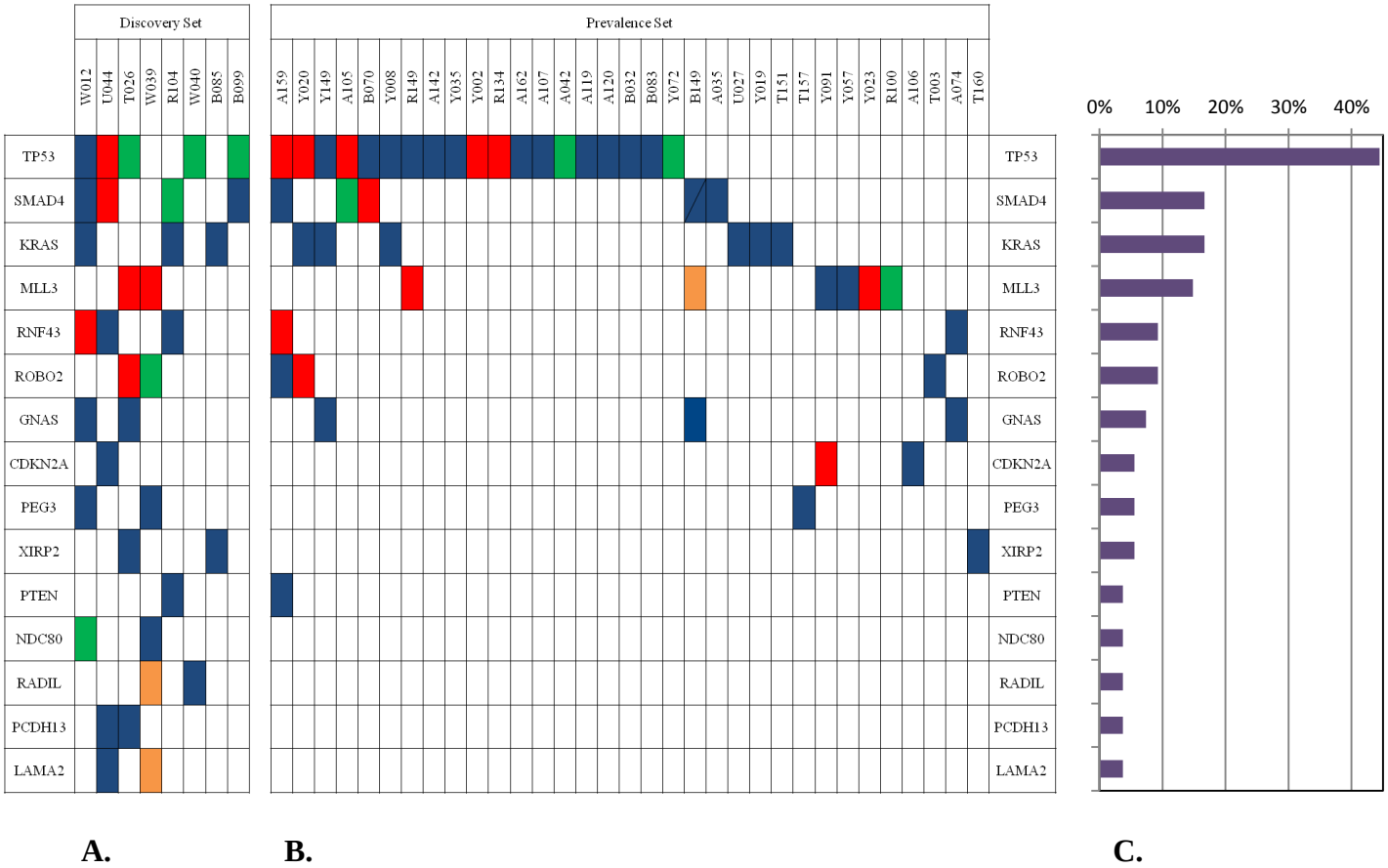


Figure 3.1: Mutational landscape of OV-associated CCA. A) Validated recurrent mutations in discovery set B) Validated recurrent mutations in prevalence set C) Total mutation rate of individual genes. Color legend: Blue = missense mutation, Red = nonsense mutation, Orange = splice site mutation, Green = frameshift mutation.

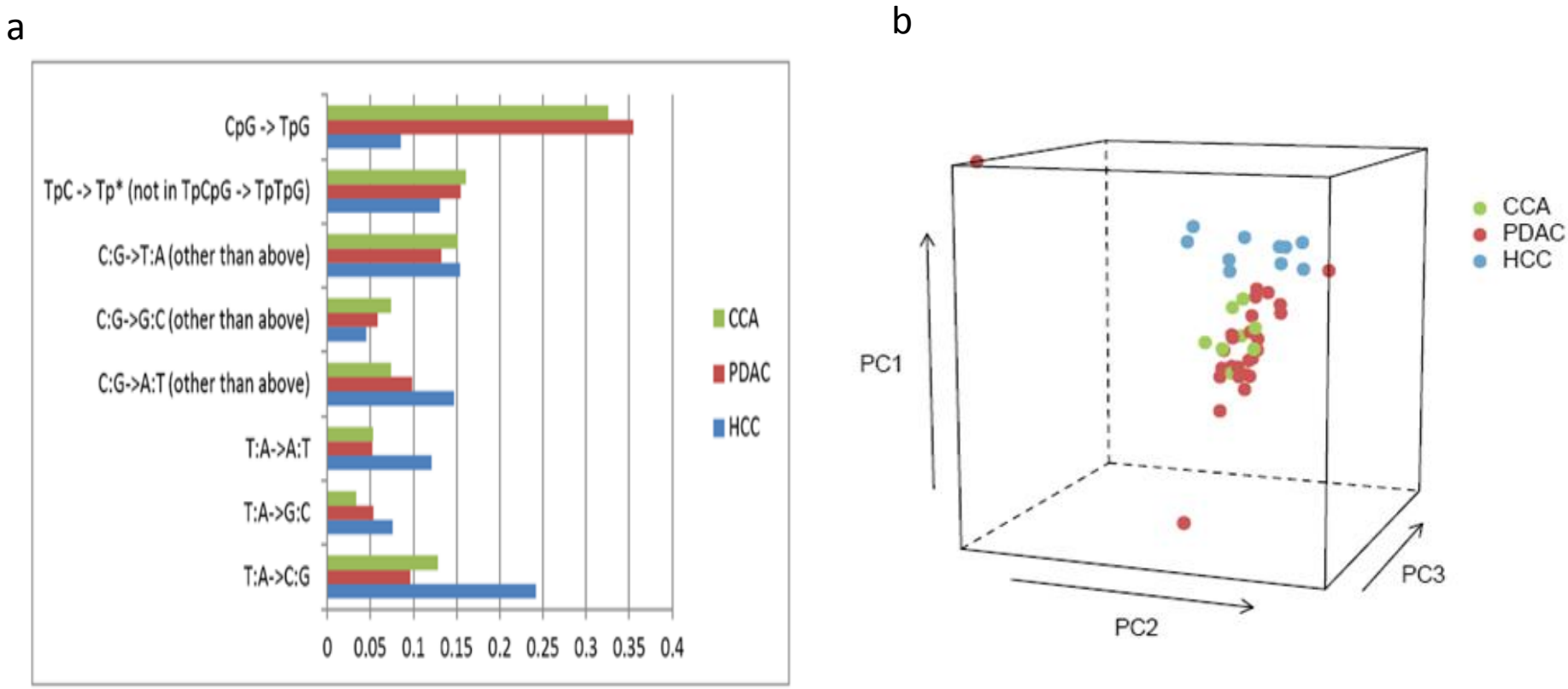


Figure 3.2: Comparisons of mutational spectra between OV-associated CCA, PDAC and HCV-associated HCC. a) Proportional comparisons between OV-associated CCA, PDAC and HCV-associated HCC. b) Principal components analysis between OV-associated CCA, PDAC and HCV-associated HCC. PDAC data extracted from Jones et al.[19]. HCV-associated HCC data extracted from Li et al.[126].

Code	Sex	Age	TMN	Staging	Histological types
B099	M	48	T3N0M0	III	WD tubular adenocarcinoma
T026	M	52	T3N1M0	IIIB	WD tubular adenocarcinoma
B083	F	53	T3N0M0	III	WD tubular adenocarcinoma
W040	M	56	T4N1M0	IVA	WD tubular adenocarcinoma
U044	M	57	T3N1M0	IIIB	Papillary carcinoma
W012	F	61	T3N0M0	III	WD tubular adenocarcinoma
R104	F	64	T3N0M0	III	WD tubular adenocarcinoma
W039	M	66	T3N0M0	III	WD tubular adenocarcinoma

Histological types: tumor differentiation, WD = Well differentiation
Staging: TMN stage according to TMN classification AJCC sixth edition.

Table 3.1a: Clinical information of the discovery set consisting of 8 patients diagnosed OV-associated CCA.

Table 3.1b: Clinical information of the prevalence set consisting of 46 patients diagnosed OV-associated CCA.

Sample	Sex	Age	TMN	Staging	Histological types
A142	F	38	T2bN1M0	IVA	WD tubular adenocarcinoma
Y074	F	40	T1N0M0	I	Papillary carcinoma
Y002	F	45	T3N1M0	IVA	Papillary carcinoma
A035	F	49	T3N0M0	III	WD tubular adenocarcinoma
R100	F	49	T2bN1M0	IIIB	WD tubular adenocarcinoma
A119	F	51	T3N0M0	III	Papillary carcinoma
B048	F	51	T2bN0M0	IVA	WD tubular adenocarcinoma
Y033	F	51	T4N1M0	IVA	WD tubular adenocarcinoma
Y091	F	51	T4N0M0	IVA	WD tubular adenocarcinoma
A042	F	52	T3N1M0	IVA	WD tubular adenocarcinoma
A120	F	53	T3N1M0	IVA	Papillary carcinoma
Y140	F	56	T2N0M0	II	Papillary carcinoma
A028	F	59	T3N1M0	IVA	WD tubular adenocarcinoma
Y035	F	64	T3N0M0	III	WD tubular adenocarcinoma
Y008	F	65	T3N0M0	III	Papillary carcinoma
U027	F	66	T2N1M0	IVA	Papillary carcinoma
A039	F	72	T3N1M0	IVA	PD tubular adenocarcinoma
R149	M	37	T4N1M0	IIIA	Papillary carcinoma
T157	M	42	T3N0M0	IIIa	WD tubular adenocarcinoma
A043	M	45	T3N0M0	III	Papillary carcinoma
T160	M	46	T3N0M0	IIIa	Papillary carcinoma
R134	M	50	T3N0M0	III	WD tubular adenocarcinoma
Y020	M	50	T3N1M0	IVA	Papillary carcinoma
A106	M	51	T3N1M0	IVA	WD tubular adenocarcinoma
T003	M	52	T4N0M0	IVA	MD tubular adenocarcinoma
A105	M	53	T3N1M0	IVA	WD tubular adenocarcinoma
B032	M	53	T3N0M0	III	MD tubular adenocarcinoma

Sample	Sex	Age	TMN	Staging	Histological types
A107	M	54	T3N0M0	III	WD tubular adenocarcinoma
A074	M	55	T4N0M0	IVA	WD tubular adenocarcinoma
A159	M	55	T3N0M0	IIIB	Papillary carcinoma
Y123	M	55	T4N0M0	IVA	WD tubular adenocarcinoma
B087	M	56	T2bN0M0	II	WD tubular adenocarcinoma
Y057	M	57	T3N1M0	IVA	WD tubular adenocarcinoma
A128	M	61	T3N1M0	IVA	Papillary carcinoma
B070	M	61	T3N0M0	IIIB	Papillary carcinoma
B113	M	61	T3N1M0	IVA	Papillary carcinoma
Y072	M	61	T3N0M0	III	Papillary carcinoma
Y023	M	62	T4N0M0	IVA	Papillary carcinoma
B149	M	63	T3N0M0	IIIA	WD tubular adenocarcinoma
Y065	M	63	T4N1M0	IVA	Papillary carcinoma
A162	M	68	T3N1M0	IVA	Papillary carcinoma
Y032	M	69	T3N1M0	IVA	Papillary carcinoma
Y149	M	69	T4N1M0	IVA	WD tubular adenocarcinoma
Y019	M	70	T4NXM0	IVA	WD tubular adenocarcinoma
B085	M	73	T3N0M0	III	WD tubular adenocarcinoma
T151	M	73	T2bN1M1	IVB	MD tubular adenocarcinoma

Histological types: tumor differentiation, WD = Well differentiation, MD = Moderate differentiation, PD = Poorly differentiation, PAP = Papillary differentiation

Staging: TMN stage according to TMN classification AJCC sixth edition.

		Bases in Target Region	Bases Mapped to Target Region	Ave. Depth Per Targeted Base	Targeted Bases with Depth at Least 1X	Targeted Bases with Depth at Least 20X	Somatic mutations identified
B085	Normal	37,806,033	3,905,980,377	103.3	97.7	87	
	Tumor	37,806,033	3,812,200,909	100.8	97.6	85.4	22
B099	Normal	37,806,033	2,930,117,037	77.5	97.7	85	
	Tumor	37,806,033	2,798,044,325	74	97.5	80.6	21
R104	Normal	37,806,033	1,903,794,100	50.4	96.5	71	
	Tumor	37,806,033	2,668,373,072	70.6	96.9	78.6	34
T026	Normal	37806033	3,668,218,669	97	97.5	86	
	Tumor	37,806,033	3,782,109,048	100	97.5	86.2	22
U044	Normal	37,806,033	2,526,282,000	66.8	96.9	79.4	
	Tumor	37,806,033	2,404,785,459	63.6	96.9	78.2	33
W012	Normal	37,806,033	2,969,234,364	78.5	97.3	81.4	
	Tumor	37,806,033	2,519,928,206	66.7	96.9	77.2	19
W039	Normal	37,806,033	2,428,397,663	64.2	97.1	77.2	
	Tumor	37,806,033	2,799,365,832	74.1	97.1	79.1	28
W040	Normal	37,806,033	3,409,974,611	90.2	97.6	86.1	
	Tumor	37,806,033	3,673,858,289	97.2	97.3	85.7	27
	AVERAGE	**37,806,033**	**3,012,541,498**	**79.7**	**97.2**	**81.5**	**26**

Table 3.2: Whole-exome sequencing summary of 8 matched pairs of OV-associated CCAs.

Table 3.3: Nonsynonymous somatic mutations identified and validated in the discovery set.

Gene Symbol	Sample	Transcript accession ID	Genomic location	cDNA location	Mutation type	Predicted residue change
ADAM12	B085	CCDS7653.1	g.chr10: 127721608 C>T	c.2104C>T	Missense	p.R702W
ALPP	B085	CCDS2490.1	g.chr2: 232953395 C>T	c.814C>T	Missense	p.R272C
ANO4	B085	CCDS31884.1	g.chr12: 99955057 G>T	c.790G>T	Nonsense	p.E264X
CHRD	B085	CCDS3266.1	g.chr3: 185585091 G>A	c.1513G>A	Missense	p.V505M
CRISPLD1	B085	CCDS6219.1	g.chr8: 76060995 G>A	c.218G>A	Missense	p.R73Q
DDX52	B085	CCDS11323.1	g.chr17: 33055103 G>A	c.1505G>A	Missense	p.R502Q
EIF2C1	B085	CCDS398.1	g.chr1: 36153669 G>A	c.1967G>A	Missense	p.R656H
FLNC	B085	CCDS43644.1	g.chr7: 128265344 G>A	c.1037G>A	Missense	p.G346E
GHSR	B085	CCDS3218.1	g.chr3: 173645905 G>A	c.841G>A	Missense	p.V281I
JAG1	B085	CCDS13112.1	g.chr20: 10573540 A>T	c.2315A>T	Missense	p.E772V
KCNH7	B085	CCDS2219.1	g.chr2: 163082641 G>A	c.737G>A	Missense	p.R246Q
KCNN4	B085	CCDS12630.1	g.chr19: 48963656 C>T	c.1163C>T	Missense	p.S388L
KRAS	B085	CCDS8703.1	g.chr12: 25289551 G>C	c.35G>C	Missense	p.G12A
LRP1B	B085	CCDS2182.1	g.chr2: 140847041 A>G	IVS69+6 A>G	Splice site	-
MAGEA11	B085	CCDS14690	g.chrX: 148576284 G>T	c.975G>T	Missense	p.E325D
NLGN3	B085	CCDS14407.1	g.chrX: 70284565 G>A	c.241G>A	Missense	p.E81K

PRMT6	B085	CCDS41360.1	g.chr1: 107401336 A>G	c.299A>G	Missense	p.Y100C
SEMA5A	B085	CCDS3875.1	g.chr5: 9119662 C>A	c.2170C>A	Missense	p.H724N
SYNJ2	B085	CCDS5254.1	g.chr6: 158396002 C>A	c.811C>A	Missense	p.L271M
TRIM28	B085	CCDS12985.1	g.chr19: 63752232 G>T	c.1475G>T	Missense	p.R492L
WNK3	B085	CCDS14357.1	g.chrX: 54241639 G>A	c.5246G>A	Missense	p.G1749E
XIRP2	B085	CCDS42769.1	g.chr2: 167816151 insT	c.9704 insT	Insertion	frameshift
CD109	B099	CCDS4982.1	g.chr6: 74589979 C>T	c.4240C>T	Missense	p.R1414C
DLGAP5	B099	CCDS9723.1	g.chr14: 54712389 G>C	c.1150G>C	Missense	p.G384R
FBN2	B099	CCDS34222.1	g.chr5: 127828512 T>A	c.630T>A	Missense	p.D210E
GIGYF2	B099	CCDS33401.1	g.chr2: 233364254 G>A	c.1136G>A	Missense	p.R379T
HCN4	B099	CCDS10248.1	g.chr15: 71422829 C>T	c.1159C>T	Missense	p.R387C
MASP2	B099	CCDS123.1	g.chr1: 11009590 G>C	c.2000G>C	Missense	p.G667A
P4HA3	B099	CCDS8230.1	g.chr11: 73691252 A>G	c.377A>G	Missense	p.D126G
PBX3	B099	CCDS6865.1	g.chr9: 127737667 A>T	c.803A>T	Missense	p.K268I
PHACTR4	B099	CCDS41294.1	g.chr1: 28688303 G>T	c.1825G>T	Nonsense	p.E609X
PRDM4	B099	CCDS9115.1	g.chr12: 106669331 T>C	c.1117T>C	Missense	p.F373L
RBM26	B099	CCDS9462.1	g.chr13: 78843081 G>C	c.634G>C	Nonsense	p.E212X
RNF133	B099	CCDS5784.1	g.chr7: 122125755 G>T	c.454G>T	Missense	p.V152L
SACS	B099	CCDS9300.1	g.chr13: 22809914 A>T	c.5660A>T	Missense	p.K1887I
SGSM2	B099	CCDS32526.1	g.chr17: 2223502 G>C	c.2045G>C	Missense	p.S682T

SMAD4	B099	CCDS11950.1	g.chr18: 46847530 T>C	c.1283T>C	Missense	p.K428T
SMARCA1	B099	CCDS14612.1	g.chrX: 128458440 C>A	c.2594C>A	Missense	p.L532M
SNAPC4	B099	CCDS6998.1	g.chr9: 138409599 C>T	c.443C>T	Missense	p.P148L
SPSB4	B099	CCDS3115.1	g.chr3: 142268228 G>T	c.592G>T	Missense	p.G198C
SYNE1	B099	CCDS5236.1	g.chr6: 152694560 C>T	c.12953C>T	Missense	p.T4318M
TFAP2D	B099	CCDS4933.1	g.chr6: 50848327 C>T	c.1150C>T	Missense	p.H384Y
TRPC5	B099	CCDS14561.1	g.chrX: 110977036 G>T	c.1662G>T	Missense	p.K554N
ASAP3	R104	CCDS235.1	g.chr1: 23638263 G>A	c.992G>A	Missense	p.C331Y
BCAS2	R104	CCDS874	g.chr1: 114925731delC	c.4 delG	Deletion	frameshift
CALN1	R104	CCDS5541.1	g.chr7: 70913409 A>G	c.380A>G	Missense	p.D127G
CCDC97	R104	CCDS12578.1	g.chr19: 46517546 G>T	c.730G>T	Nonsense	p.E244X
CHD5	R104	CCDS57.1	g.chr1: 6107811 G>A	c.4330G>A	Nonsense	p.Q1444X
EHBP1	R104	CCDS1872.1	g.chr2: 62787877 G>A	c.47G>A	Missense	p.S16F
ENTPD8	R104	CCDS43913.1	g.chr9: 139452306 C>T	c.178C>T	Nonsense	p.Q60X
EPHA2	R104	CCDS169.1	g.chr1: 16337356 G>A	IVS1+1 G>A	Splice site	-
EXOSC8	R104	CCDS31958.1	g.chr13: 36479133 G>A	c.412G>A	Missense	p.D138N
GHRHR	R104	CCDS5432	g.chr7: 30981934 T>C	c.346T>C	Missense	p.S116P
HECW2	R104	CCDS33354.1	g.chr2: 196892592 A>G	c.1267A>G	Missense	p.I423V
HOOK1	R104	CCDS612.1	g.chr1: 60060136 G>T	c.82G>T	Missense	p.A28S
IP6K2	R104	CCDS2777.1	g.chr3: 48707692 C>T	c.37C>T	Missense	p.R13C

KCNQ2	R104	CCDS13520.1	g.chr20: 61535669 C>T	c.1055C>T	Missense	p.S352L
KRAS	R104	CCDS8703.1	g.chr12: 25289548 C>T	c.358C>T	Missense	p.G13D
LRRC15	R104	CCDS3306.1	g.chr3: 195561906 G>A	c.1162G>A	Missense	p.V388I
MAP3K13	R104	CCDS3270.1	g.chr3: 186644068 G>A	c.801G>A	Missense	p.M267I
MARK4	R104	CCDS12658.1	g.chr19: 50458430 T>A	c.320T>A	Missense	p.V107D
MLYCD	R104	CCDS42206.1	g.chr16: 82499236 G>C	c.646G>C	Missense	p.E216Q
NUP160	R104	CCDS31484.1	g.chr11: 47813884 A>G	c.996A>G	Missense	p.K332N
NUP160	R104	CCDS31484.1	g.chr11: 47813886 A>T	c.994A>T	Missense	p.K332E
OR6Q1	R104	CCDS31541.1	g.chr11: 57555662 C>A	c.662C>A	Missense	p.S221Y
PCDHGA7	R104	ENST00000518325	g.chr5: 140744757 G>A	c.2107G>A	Missense	p.V703I
PTEN	R104	CCDS31238.1	g.chr10: 89682960 A>G	c.464A>G	Missense	p.Y155C
RAB11FIP5	R104	CCDS1923.1	g.chr2: 73169134 C>T	c.1120C>T	Missense	p.R374W
RGS3	R104	CCDS35114.1	g.chr9: 115396282 C>T	c.262C>T	Missense	p.A88T
RMI1	R104	CCDS6669.1	g.chr9: 85806275 A>G	c.554A>G	Missense	p.E185G
RNF43	R104	CCDS11607.1	g.chr17: 53803291 A>G	c.355A>G	Missense	p.C119R
SCN11A	R104	CCDS33737.1	g.chr3: 38888713 C>T	c.3470C>T	Missense	p.A1157V
SMAD4	R104	CCDS11950.1	g.chr18: 46857143delAGTA	c.1447 to 1450 delAGTA	Deletion	frameshift
STT3B	R104	CCDS2650.1	g.chr3: 31592975 C>T	c.398C>T	Missense	p.P133L
TMEM222	R104	CCDS297.2	g.chr1: 27533346 G>A	c.526G>A	Missense	p.G176R

VCAN	R104	CCDS4060.1	g.chr5: 82911578 G>A	c.9904G>A	Missense	p.V3302I
VEGFC	R104	CCDS43285.1	g.chr4: 177887807 A>G	c.235A>G	Missense	p.K79E
ANK2	T026	CCDS3702.1	g.chr4: 114476537 C>T	c.3466C>T	Missense	p.R1156C
ARID1A	T026	CCDS285.1	g.chr1: 26962263 C>T	c.2632C>T	Nonsense	p.Q878X
CALML5	T026	CCDS7068.1	g.chr10: 5531395 C>T	c.7C>T	Missense	p.G3S
FKBP3	T026	CCDS9683.1	g.chr14: 44668830 A>G	c.235A>G	Missense	p.S79G
GLI3	T026	CCDS5465.1	g.chr7: 42054630 C>A	c.664C>A	Missense	p.L222M
GNAS	T026	CCDS13472.1	g.chr20: 56917815 C>T	c.601C>T	Missense	p.R201C
IL1RAPL1	T026	CCDS14218.1	g.chrX: 29211170 G>A	c.277G>A	Missense	p.G93R
INHBA	T026	CCDS5464.1	g.chr7: 41696138 C>T	c.916C>T	Missense	p.R306C
ITGA2B	T026	CCDS32665.1	g.chr17: 39812940 G>A	c.1708G>A	Missense	p.G570R
MLL3	T026	CCDS5931.1	g.chr7: 151466816 G>T	c.14641G>T	Nonsense	p.E4881X
OR51B2	T026	CCDS31377.1	g.chr11: 5302021 C>T	c.83C>T	Missense	p.P28L
PAPD4	T026	CCDS4048.1	g.chr5: 78954967 C>T	c.364C>T	Missense	p.H122Y
PCDHA13	T026	CCDS4240.1	g.chr5: 140244022 C>T	c.1985C>T	Missense	p.T662M
PPM1E	T026	CCDS11613.1	g.chr17: 54412237 A>G	c.1331A>G	Missense	p.D444G
ROBO2	T026	CCDS43109.1	g.chr3: 77695073 C>T	c.1585CC>T	Nonsense	p.Q529X
SEMA6D	T026	CCDS32224.1	g.chr15: 45849983 C>T	IVS16-3 C>T	Splice site	-
STK36	T026	CCDS2421.1	g.chr2: 219270018 A>G	c.2599A>G	Missense	p.S867G

TARS	T026	CCDS3899.1	g.chr5: 33484483 C>T	c.218C>T	Missense	p.A73V
TCEAL2	T026	CCDS14496.1	g.chrX: 101269027 C>T	c.569C>T	Missense	p.A190V
TP53	T026	CCDS11118.1	g.chr17: 7520195insG	c.216 insG	Insertion	frameshift
VWCE	T026	CCDS8002.1	g.chr11: 60805889 C>G	c.732C>G	Missense	p.F244L
XIRP2	T026	CCDS42769.1	g.chr2: 167814046 C>T	c.7898C>T	Missense	p.S2633L
ANKRD35	U044	CCDS919.1	g.chr1: 144274448 G>A	c.2779G>A	Missense	p.E927K
ARAF	U044	CCDS35232.1	g.chrX: 47311074 A>T	c.650A>T	Missense	p.N217I
ARL6IP5	U044	CCDS2912.1	g.chr3: 69233820 T>C	c.317T>C	Missense	p.M106T
CDH8	U044	CCDS10802.1	g.chr16: 60448611 G>A	c.580G>A	Missense	p.A194T
CDKN2A	U044	CCDS6510.1	g.chr9: 21961108 C>T	c.416C>T	Missense	p.R139Q
COL11A2	U044	CCDS43452.1	g.chr6: 33247518 C>T	c.2842C>T	Missense	p.R948C
DICER1	U044	CCDS9931.1	g.chr14: 94632359 G>A	c.4651G>A	Missense	p.E1551K
DSP	U044	CCDS4501.1	g.chr6: 7526185 C>T	c.4763C>T	Missense	p.S1588F
EIF3E	U044	CCDS6308.1	g.chr8: 109330028 C>G	c.80C>G	Missense	p.S27C
FOXR2	U044	ENST00000339140	g.chrX: 55667644 G>C	c.775G>C	Missense	p.D259H
GATM	U044	CCDS10122.1	g.chr15: 43447670 C>T	c.565C>T	Missense	p.R189C
HIST1H2AG	U044	CCDS4619.1	g.chr6: 27209193 G>C	c.364G>C	Missense	p.E122Q
HOXC11	U044	CCDS8867.1	g.chr12: 52653746 G>A	c.454G>A	Missense	p.D152N
IQCB1	U044	CCDS33837.1	g.chr3: 123027678 G>T	c.303G>T	Missense	p.E101D

IRX1	U044	CCDS34132.1	g.chr5: 3652548 G>C	c.486>C	Missense	p.M162I
KIAA1267	U044	CCDS11503.1	g.chr17: 41500732 G>A	IVS4+5 G>A	Splice site	-
LAMA2	U044	CCDS5138.1	g.chr6: 129422627 C>A	c.289C>A	Missense	p.H97N
LMX1A	U044	CCDS1247.1	g.chr1: 163589070 G>C	c.130G>C	Missense	p.D44H
MXRA5	U044	CCDS14124.1	g.chrX: 3250001 G>A	c.3725G>A	Missense	p.R1242Q
NPY5R	U044	CCDS3804.1	g.chr4: 164491125 G>T	c.250G>T	Missense	p.A84S
ODF3	U044	CCDS7688.1	g.chr11: 187641 T>G	c.190T>G	Missense	p.C64G
PCBP3	U044	CCDS42974.1	g.chr21: 46158300 C>T	c.512C>T	Nonsense	p.P171L
PCDH11X	U044	CCDS14461.1	g.chrX: 91760167 G>A	c.3616G>A	Missense	p.A1206T
PCDHA13	U044	CCDS4240.1	g.chr5: 140243338 C>T	c.1301C>T	Missense	p.S434L
PCDHA7	U044	CCDS34252.1	g.chr5: 140195506 G>A	c.1354G>A	Missense	p.A452T
PPP1R16B	U044	CCDS13309.1	g.chr20: 36980694 A>G	c.1675A>G	Missense	p.K559E
RNF43	U044	CCDS11607.1	g.chr17: 53795717 A>T	c.500A>T	Missense	p.N167I
SMAD4	U044	CCDS11950.1	g.chr18: 46835241 C>T	c.547C>T	Nonsense	p.Q183X
SPAG17	U044	CCDS899.1	g.chr1: 118529214 G>A	IVS1+3 G>A	Splice site	-
TP53	U044	CCDS11118.1	g.chr17: 7520102 G>A	c.310C>T	Nonsense	p.Q104X
TUBA3C	U044	CCDS9284.1	g.chr13: 18646011 G>T	c.1345G>T	Nonsense	p.E449X
ZFY	U044	CCDS14774.1	g.chrY: 2907586 C>T	c.1958C>T	Missense	p.T653M
ZNF790	U044	CCDS12496.1	g.chr19: 42001246 G>A	c.1840G>A	Missense	p.E614K

EXTL3	W012	CCDS6070.1	g.chr8: 28651097 A>G	c.2419A>G	Missense	p.K807E
FER	W012	CCDS4098.1	g.chr5: 108247045 A>T	c.973A>T	Missense	p.N325Y
GJB6	W012	CCDS9291.1	g.chr13: 19695240 G>A	c.380G>A	Missense	p.R127Q
GNAS	W012	CCDS13472.1	g.chr20: 56917815 C>T	c.601C>T	Missense	p.R201C
HDAC2	W012	CCDS43493.1	g.chr6: 114386557 C>T	c.514C>T	Missense	p.R172W
KRAS	W012	CCDS8703.1	g.chr12: 25289551 G>C	c.35G>C	Missense	p.G12A
LRP2	W012	CCDS2232.1	g.chr2: 169740065 G>A	c.10652G>A	Missense	p.R3551H
MEFV	W012	CCDS10498.1	g.chr16: 3233338 G>A	c.2150G>A	Missense	p.R717H
NDC80	W012	CCDS11827.1	g.chr18: 2577914delA	c.756 delA	Deletion	frameshift
P2RY13	W012	CCDS3158.1	g.chr3: 152528720 C>T	c.751C>T	Missense	p.H251Y
PCNX	W012	CCDS9806.1	g.chr14: 70584285 C>T	c.4169C>T	Missense	p.A1390V
PEG3	W012	CCDS12948.1	g.chr19: 62019086 A>G	c.2536A>G	Missense	p.S846R
PTPRS	W012	CCDS12140.1	g.chr19: 5225217 G>A	c.230G>A	Missense	p.R77H
RASAL2	W012	CCDS1321.1	g.chr1: 176708963 C>T	c.3776C>T	Missense	p.T1259M
RNF43	W012	CCDS11607.1	g.chr17: 53789969 C>T	c.21667C>T	Nonsense	p.Q723X
SIGLEC12	W012	CCDS12833.1	g.chr19: 56696603 C>T	c.197C>T	Missense	p.A66V
SMAD4	W012	CCDS11950.1	g.chr18: 46829198 G>T	c.394G>T	Missense	p.H132D
TP53	W012	CCDS11118.1	g.chr17: 7518264 G>A	c.742G>A	Missense	p.R248W
YSK4	W012	CCDS2176.2	g.chr2: 135462100 C>T	c.812C>T	Missense	p.S271L

DYNC1H1	W039	CCDS9966.1	g.chr14: 101530788 G>A	c.3182G>A	Nonsense	p.W1061X
FBLN1	W039	CCDS14067.1	g.chr22: 44349096 C>T	c.1739C>T	Missense	p.S580F
FH	W039	CCDS1617.1	g.chr1: 239735946 C>T	c.884C>T	Missense	p.A295V
FMNL1	W039	CCDS11497.1	g.chr17: 40678429 G>A	c.2755G>A	Missense	p.V919M
GPSM1	W039	CCDS48055	g.chr9: 138372433 G>A	c.441G>A	Missense	p.R75Q
ITPR2	W039	CCDS41764.1	g.chr12: 26644222 C>T	c.3766C>T	Missense	p.L1256F
KCNH5	W039	CCDS9756.1	g.chr14: 62486984 G>A	c.989G>A	Missense	p.R330Q
KLHL4	W039	CCDS14456.1	g.chrX: 86756148 C>T	c.646C>T	Missense	p.L216F
LAMA2	W039	CCDS5138.1	g.chr6: 129615138 T>A	IVS14+5 T>A	Splice site	-
LRRK2	W039	CCDS31774.1	g.chr12: 38920555 C>T	c.575C>T	Missense	p.S192L
MAP2K4	W039	CCDS11162.1	g.chr17: 11983883 T>G	c.1043T>G	Missense	p.L348R
MLL3	W039	CCDS5931.1	g.chr7: 151482155 G>T	c.12149G>T	Nonsense	p.S4050X
MYH2	W039	CCDS11156.1	g.chr17: 10383329 G>A	c.1334G>A	Missense	p.R445H
NDC80	W039	CCDS11827.1	g.chr18: 2579298 G>A	c.859G>A	Missense	p.E287K
PDGFD	W039	CCDS41703.1	g.chr11: 103302867 G>A	c.970G>A	Missense	p.V324M
PEG3	W039	CCDS12948.1	g.chr19: 62018675 G>A	c.2947G>A	Missense	p.D983N
PGBD5	W039	CCDS1583.1	g.chr1: 228539503 C>T	c.1139C>T	Missense	p.T380M
POLL	W039	CCDS7513.1	g.chr10: 103332547 G>A	c.1157G>A	Missense	p.R386H
PPP2R3A	W039	CCDS3088.1	g.chr3: 137224638 G>A	c.37G>A	Missense	p.D13N

PREX2	W039	CCDS6201.1	g.chr8: 69266560 G>C	c.4396G>C	Missense	p.A1466P
RADIL	W039	CCDS43544.1	g.chr7: 4828547 C>T	IVS6+4 C>T	Splice site	-
RASGRF2	W039	CCDS4052.1	g.chr5: 80399605 A>G	IVS2-2 A>G	Splice site	-
ROBO2	W039	CCDS43109.1	g.chr3: 77739727insC	c.3234 insC	Insertion	frameshift
SAMD7	W039	CCDS3209.1	g.chr3: 171120005 C>G	c.25C>G	Missense	p.P9A
THBS2	W039	CCDS34574.1	g.chr6: 169390871 C>T	c.175C>T	Missense	p.R59C
TNKS2	W039	CCDS7417.1	g.chr10: 93601053 G>A	c.2795G>A	Missense	p.R932K
TRIM48	W039	CCDS7947	g.chr11: 54794908 C>T	c.1048C>T	Nonsense	p.Q366X
UTP20	W039	CCDS9081.1	g.chr12: 100291384 T>A	IVS53+5 T>A	Splicesite	-
AFF1	W040	CCDS3616.1	g.chr4: 88255312 T>C	c.2282T>C	Missense	p.L761S
BBS4	W040	CCDS10246	g.chr15: 70803932delGTT	c.471 to 473 delGTT	Deletion	frameshift
CNBP	W040	CCDS3056.1	g.chr3: 130372666 C>T	c.362C>T	Missense	p.S121F
COL11A1	W040	CCDS779.1	g.chr1: 103269277 G>A	c.793G>A	Missense	p.E255K
CTNNA2	W040	CCDS42703.1	g.chr2: 80654802 T>A	c.1748T>A	Missense	p.M583K
FAM47A	W040	CCDS43926.1	g.chrX: 34059452 G>A	c.865G>A	Missense	p.E289K
GDPD3	W040	CCDS10671	g.chr16: 30023708insAGCT	c.938 insAGCT	Insertion	frameshift
HDAC4	W040	CCDS2529.1	g.chr2: 239701829 C>T	c.1633C>T	Missense	p.P545S
KCNJ9	W040	CCDS1194	g.chr: 158320685insT	c.241 insT	Insertion	frameshift

KIAA1984	W040	CCDS43906	g.chr9: 138814726delCC	c.504 to 505 delCC	Deletion	frameshift
NFKB1	W040	CCDS43906	g.chr4: 103724981insC	c.1032 insC	Insertion	frameshift
NFKBIL2	W040	CCDS34968.1	g.chr8: 145632580 C>T	c.1567C>T	Missense	p.P523S
NUP210	W040	CCDS33704.1	g.ch3r: 13347016 G>A	c.4054G>A	Missense	p.E1352K
POLDIP3	W040	CCDS14038	g.chr22: 41328941insC	c.227 insC	Insertion	frameshift
RADIL	W040	CCDS43544.1	g.chr7: 4883911 G>A	c.386G>A	Missense	p.R129Q
RBBP7	W040	CCDS14179	g.chrX: 16797171delGGGTCATAACCA	c.98 to 109 delGGGTCATAACCA	Deletion	frameshift
SEMA4F	W040	CCDS1955.1	g.chr2: 74760598 G>C	c.2067G>C	Missense	p.Q689H
SLC23A1	W040	CCDS4213.1	g.chr5: 138745561 A>T	c.227A>T	Missense	p.Q76L
SLC5A5	W040	CCDS12368.1	g.chr19: 17862725 G>A	c.1682G>A	Missense	p.G561E
SMC6	W040	CCDS1690.1	g.chr2: 17759681 A>G	c.1658A>G	Missense	p.Y553C
TMCO4	W040	CCDS198.1	g.chr1: 19970415 G>T	c.327G>T	Missense	p.L109F
TP53	W040	CCDS11118.1	g.chr17:7518289insT	c.716 insT	Insertion	frameshift
TROVE2	W040	CCDS1379	g.chr1: 191312721insTG	c.1007 insTG	Insertion	frameshift
UBAC1	W040	CCDS35177.1	g.chr9: 137977996 G>C	c.484G>C	Missense	p.V162M
VANGL2	W040	CCDS30915.1	g.chr1: 158655525 T>A	c.302T>A	Missense	p.L101Q
VAV1	W040	CCDS12174.1	g.chr19: 6805008 G>A	c.2383G>A	Missense	p.A795T

ZNF136	W040	CCDS32916.1	g.chr19: 12158762 A>T	c.569A>T	Missense	p.H190L

Table 3.4: Recurrently mutated genes as well as known recurrently mutated genes found in 54 OV-associated CCAs. Bold = somatic mutations identified in discovery set.

Sample	Gene Symbol	Transcript accession ID	Genomic location	cDNA location	Mutation type	Predicted residue change
A042	*TP53*	CCDS11118.1	g.chr17:7518992 delT	c.582 delT	Deletion	Frameshift
A105	*TP53*	CCDS11118.1	g.chr17:7517590 T>G	c.981 T>G	Nonsense	p.Y327X
A107	*TP53*	CCDS11118.1	g.chr17:7518996 A>G	c.578 A>G	Missense	p.H193R
A119	*TP53*	CCDS11118.1	g.chr17:7517819 C>T	c.844 C>T	Missense	p.R282W
A120	*TP53*	CCDS11118.1	g.chr17:7517845 G>A	c.818 G>A	Missense	p.R273H
A142	*TP53*	CCDS11118.1	g.chr17:7518334 G>A	c.673 G>A	Missense	p.V225I
A159	*TP53*	CCDS11118.1	g.chr17:7517747 C>T	c.916 C>T	Nonsense	p.R306X
A162	*TP53*	CCDS11118.1	g.chr17:7519260 A>G	c.395 A>G	Missense	p.K132R
B032	*TP53*	CCDS11118.1	g.chr17:7517837 C>T	c.832 C>T	Missense	p.P278S
B070	*TP53*	CCDS11118.1	g.chr17:7518948 A>G	c.707 A>G	Missense	p.Y236C
B083	*TP53*	CCDS11118.1	g.chr17:7519252 T>C	c.403 T>C	Missense	p.C135R
B099	***TP53***	**CCDS11118.1**	**g.chr17:7518320 G>T & insT**	**c.686 G>T & ins T**	**Insertion**	**Frameshift**
R134	*TP53*	CCDS11118.1	g.chr17:7514725 G>T	c.1027 G>T	Nonsense	p.E343X
R149	*TP53*	CCDS11118.1	g.chr17:7517810 G>A	c. 853 G>A	Missense	p.E285K
T026	***TP53***	**CCDS11118.1**	**g.chr17: 7520195insG**	**c.216 insG**	**Insertion**	**Frameshift**
U044	***TP53***	**CCDS11118.1**	**g.chr17: 7520102 G>A**	**c.310 C>T**	**Nonsense**	**p.Q104X**

W012	***TP53***	**CCDS11118.1**	**g.chr17: 7518264 G>A**	**c.742 C>T**	**Missense**	**p.R248W**
W040	***TP53***	**CCDS11118.1**	**g.chr17:7518289insT**	**c.716 ins T**	**Insertion**	**Frameshift**
Y002	*TP53*	CCDS11118.1	g.chr17:7518988 C>T	c.586 C>T	Nonsense	p.R196X
Y008	*TP53*	CCDS11118.1	g.chr17:7518930 G>A	c.644 G>A	Missense	p.S215N
Y020	*TP53*	CCDS11118.1	g.chr17:7519167 C>A	c.489 C>A	Nonsense	p.Y163X
Y035	*TP53*	CCDS11118.1	g.chr17:7517846 C>T	c.817 C>T	Missense	p.R273C
Y072	*TP53*	CCDS11118.1	g. chr17: 7577839__7578578 del	c. del 750bp	Deletion	Frameshift
Y149	*TP53*	CCDS11118.1	g.chr17:7517846 C>T	c.817 C>T	Missense	p.R273H
A035	*SMAD4*	CCDS11950.1	g.chr18:46858835 A>G	c.1659 A>G	Missense	p.X553W
A105	*SMAD4*	CCDS11950.1	g.chr18:46827503 delGA	c.90 delGA	Deletion	Frameshift
A159	*SMAD4*	CCDS11950.1	g.chr18:46845823 G>A	c.988 G>A	Missense	p.E330K
B070	*SMAD4*	CCDS11950.1	g.chr18:46829150 C>T	c.346 C>T	Nonsense	p.Q116X
B099	***SMAD4***	**CCDS11950.1**	**g.chr18: 46847530 T>C**	**c.428 A>C**	**Missense**	**p.K428T**
B149	*SMAD4*	CCDS11950.1	g.chr18:46829208 G>A	c.404 G>A	Missense	p.R135Q
B149	*SMAD4*	CCDS11950.1	g.chr18:46845916 C>T	c.1081 C>T	Missense	p.R361C
R104	***SMAD4***	**CCDS11950.1**	**g.chr18: 46857143 delAGTA**	**c.1447 delAGTA**	**Deletion**	**Frameshift**
U044	***SMAD4***	**CCDS11950.1**	**g.chr18: 46835241 C>T**	**c.547 C>T**	**Nonsense**	**p.Q183X**
W012	***SMAD4***	**CCDS11950.1**	**g.chr18: 46829198 G>T**	**c.394 C>G**	**Missense**	**p.H132D**
R104	***KRAS***	**CCDS8703.1**	**g.chr12: 25289548 C>T**	**c.38 C>T**	**Missense**	**p.G13D**
B085	***KRAS***	**CCDS8703.1**	**g.chr12: 25289551 G>C**	**c.35G>C**	**Missense**	**p.G12A**
T151	*KRAS*	CCDS8703.1	**g.chr12: 25289551 G>A**	c.35 G>A	Missense	p.G12D

U027	*KRAS*	CCDS8703.1	**g.chr12: 25289552 G>A**	c.34 G>A	Missense	p.G12S
W012	***KRAS***	**CCDS8703.1**	**g.chr12: 25289551 G>C**	**c.35 G>C**	**Missense**	**p.G12A**
Y008	*KRAS*	CCDS8703.1	g.chr12: 25289551 G>A	c.35 G>A	Missense	p.G12D
Y019	*KRAS*	CCDS8703.1	g.chr12: 25289552 G>T	c.34 G>T	Missense	p.G12C
Y020	*KRAS*	CCDS8703.1	g.chr12: 25289551 G>T	c.35 G>T	Missense	p.G12V
Y149	*KRAS*	CCDS8703.1	g.chr12: 25289551 G>T	c.35 G>T	Missense	p.G12V
B149	*MLL3*	CCDS5931.1	g.chr7:151686605 T>C	IVS2+2 T>C	Splice site	-
R100	*MLL3*	CCDS5931.1	g.chr7:151483981-151483982insTT	c.11908_11909 ins TT	Indel	Frameshift
R149	*MLL3*	CCDS5931.1	g.chr7:151491661 C>T	c.9934 C>T	Nonsense	p.Q3312X
T026	***MLL3***	**CCDS5931.1**	**g.chr7: 151466816 G>T**	**c.14641 G>T**	**Nonsense**	**p.E4881X**
W039	***MLL3***	**CCDS5931.1**	**g.chr7: 151482155 G>T**	**c.12149 G>T**	**Nonsense**	**p.S4050X**
Y023	*MLL3*	CCDS5931.1	g.chr7:151476865 G>A	c.13080 G>A	Nonsense	p.W4360X
Y057	*MLL3*	CCDS5931.1	g.chr7:151643305 A>T	c.411 A>T	Missense	p.Q147H
Y091	*MLL3*	CCDS5931.1	g.chr7: 151476365 T>C	c.13580 T>C	Missense	p.V4527A
A074	*RNF43*	CCDS11607.1	g.chr17:53794980 C>G	c.611 C>G	Missense	p.T204R
A159	*RNF43*	CCDS11607.1	g.chr17:53803309 C>T	c.337 C>T	Nonsense	p.R113X
R104	***RNF43***	**CCDS11607.1**	**g.chr17: 53803291 A>G**	**c.355 T>C**	**Missense**	**p.C119R**
U044	***RNF43***	**CCDS11607.1**	**g.chr17: 53795717 A>T**	**c.500 A>T**	**Missense**	**p.N167I**
W12	***RNF43***	**CCDS11607.1**	**g.chr17: 53789969 C>T**	**c.2167 C>T**	**Nonsense**	**p.Q723X**
A159	*ROBO2*	CCDS43109.1	g.chr3:77776576 T>A	c.3999 T>A	Missense	p.S1322R
T003	*ROBO2*	CCDS43109.1	g.chr3:77706407 C>T	c.2039 C>T	Missense	p.A680V

T026	***ROBO2***	**CCDS43109.1**	**g.chr3: 77695073 C>T**	**c.1585 C> T**	**Nonsense**	**p.Q529X**
W039	***ROBO2***	**CCDS43109.1**	**g.chr3: 77739727insC**	**c.3234 insC**	**Insertion**	**Frameshift**
Y020	*ROBO2*	CCDS43109.1	g.chr3:77625129 G>T	c.712 G>T	Nonsense	p.E238X
A074	*GNAS*	CCDS13472.1	g.chr20: 56917816 G>T	c.602 G>T	Missense	p.R201L
B149	*GNAS*	CCDS13472.1	g.chr20: 56917815 C>T	c.601 C>T	Missense	p.R201C
T026	***GNAS***	**CCDS13472.1**	**g.chr20: 56917815 C>T**	**c.601 C>T**	**Missense**	**p.R201C**
W012	***GNAS***	**CCDS13472.1**	**g.chr20: 56917815 C>T**	**c.601 C>T**	**Missense**	**p.R201C**
Y149	*GNAS*	CCDS13472.1	g.chr20: 56917816 G>A	c.602 G>A	Missense	p.R201H
A106	*CDKN2A*	CCDS6510.1	g.chr9:21961015 G>T	c.343 G>T	Missense	p.V115L
U044	***CDKN2A***	**CCDS6510.1**	**g.chr9: 21961108 C>T**	**c.416C>T**	**Missense**	**p.R139Q**
Y091	*CDKN2A*	CCDS6510.1	g.chr9:21961120 C>T	c.238 C>T	Nonsense	p.R80X
T157	*PEG3*	CCDS12948.1	g.chr19:62019610 C>T	c.2012 C>T	Missense	p.S671F
W012	***PEG3***	**CCDS12948.1**	**g.chr19: 62019086 A>G**	**c.2536 A>C**	**Missense**	**p.S846R**
W039	***PEG3***	**CCDS12948.1**	**g.chr19: 62018675 G>A**	**c.2947 G>A**	**Missense**	**p.D983N**
B085	***XIRP2***	**CCDS42769.1**	**g.chr2: 167816151 insT**	**c.9704 insT**	**Insertion**	**Frameshift**
T026	***XIRP2***	**CCDS42769.1**	**g.chr2: 167814046 C>T**	**c.7898C>T**	**Missense**	**p.S2633L**
T160	*XIRP2*	CCDS42769.1	g.chr2: 167760210 C>T	c.218 C>T	Missense	p.S73L
A159	*PTEN*	CCDS31238.1	g.chr10:89701871 G>A	c.509 G>A	Missense	P.S170N
R104	***PTEN***	**CCDS31238.1**	**g.chr10: 89682960 A>G**	**c.464A>G**	**Missense**	**p.Y155C**
W012	***NDC80***	**CCDS11827.1**	**g.chr18: 2577914delA**	**c.759 delA**	**Deletion**	**Frameshift**
W039	***NDC80***	**CCDS11827.1**	**g.chr18: 2579298 G>A**	**c.859 G>A**	**Missense**	**p.E287K**

W039	***RADIL***	**CCDS43544.1**	**g.chr7: 4828547 C>T**	**IVS6+4 C>T**	**Splice site**	**-**
W040	***RADIL***	**CCDS43544.1**	**g.chr7: 4883911 G>A**	**c.386 G>A**	**Missense**	**p.R129Q**
T026	***PCDHA13***	**CCDS4240.1**	**g.chr5: 140244022 C>T**	**c.1985 C>T**	**Missense**	**p.T662M**
U044	***PCDHA13***	**CCDS4240.1**	**g.chr5: 140243338 C>T**	**c.1301 C>T**	**Missense**	**p.S434L**
U044	***LAMA2***	**CCDS5138.1**	**g.chr6: 129422627 C>A**	**c.289C>A**	**Missense**	**p.H97N**
W039	***LAMA2***	**CCDS5138.1**	**g.chr6: 129615138 T>A**	**IVS14+5 T>A**	**Splice site**	**-**

Genes [a]	OV-associated CCA N = 54	PDAC[b] N = 114	HCV-associated HCC[c] N = 95
TP53	44.4 % (24)	85%	33.70%
KRAS	16.7 % (9)	100%	0/10
SMAD4	16.7 % (9)	27%	0/10
CDKN2A	5.6 % (3)	25%	0/10
MLL3	14.8 %(8)	7.90%	0/10
ROBO2	9.3 % (5)	0/24	0/10
GNAS	9.3 % (5)	0/24	0/10
RNF43	9.3 % (5)	0/24	0/10
PEG3	5.6 % (3)	1/24	0/10
PTEN	3.7 % (2)	0/24	0/10
RADIL	3.7 % (2)	0/24	0/10
NDC80	3.7 % (2)	0/24	0/10
PCDHA13	3.7 % (2)	0/24	0/10
CTNNB1	0/8	0/24	20%
ARID2	0/8	0/24	7.40%
DMXL1	0/8	0/24	4.20%
NLRP1	0/8	0/24	4.20%

Table 3.5: Frequency of recurrently mutated genes in OV-associated CCA, PDAC and HCV-associated HCC. [a]Including genes affected by point mutations, indels, and splice site mutations. [b]Data extracted from Jones et al.[19]. [c]Data extracted from Li et al.[126].

Mutation Category	OV-associated CCA	PDAC[a]	HCV-associated HCC[b]
CpG -> TpG	282	523	36
TpC -> Tp* (not in TpCpG -> TpTpG)	139	228	55
C:G->T:A (other than above)	131	195	65
C:G->G:C (other than above)	64	86	19
C:G->A:T (other than above)	64	145	62
T:A->A:T	46	77	51
T:A->G:C	29	79	32
T:A->C:G	111	142	102
Total	**866**	**1475**	**422**
P-values (chisq)[c]	-	**0.0099**	**2.20E-016**

Table 3.6: Mutation spectra in OV-associated CCA, PDAC and HCV-associated HCC. [a]Data extracted from Jones et al.[19]. [b]Data extracted from Li et al.[126]. [c]P-values for pair-wise comparison of mutation spectrum of OV-associated CCA with PDAC or HCV-associated HCC.

Chapter Four: Whole-exome sequencing studies of parathyroid carcinomas reveal novel *PRUNE2* mutations, distinctive mutational spectra related to APOBEC-catalyzed DNA mutagenesis and mutational enrichment in kinases associated with cell migration and invasion.

The findings in this Chapter were published in Yu et al. (2015), *J Clin Endocrinol Metab*, 100(2):E360-4 (pp 94-147 of this thesis).

4.1: Introduction

PC is a rare, malignant subset of parathyroid tumors associated with primary hyperparathyroidism (HPT). Most PCs have deregulation in the secretion of parathyroid hormone (PTH) leading to hypercalcemia with complications arising from this condition being the major cause of morbidity (160). While PC accounts for only 0.1-5% of primary HPT cases (161), the occurrence of PC rises to ~15% in a subset of primary HPT associated with hyperparathyroidism-jaw tumor syndrome (HPT-JT) (160). Genetic analysis of kindreds with HPT-JT syndrome and a subset of kindred with familial isolated primary HPT revealed frequent germline mutations of the *CDC73* gene (162). Additional investigations into *CDC73* mutation status in sporadic cases of PC reveal somatic mutations in 60% to 100% of cases (163,164). A majority of PC cases positive for *CDC73* disruption show two distinct mutations or a single mutation in combination with LOH supporting the two-hit tumor suppressor mechanism and pointing to *CDC73* as a major driving gene in PC (163,164). Furthermore, the mutation is highly specific in PC, rarely described in other tumor types.

Other than diseases associated with *CDC73* mutations, PC is only rarely linked to other genes implicated in other familial HPT syndromes; only a small number of PC cases have been reported in patients with multiple endocrine neoplasia type 1 (*MEN1*) (165) or multiple endocrine neoplasia type 2A (*MEN2A*) (166). While down-regulation of calcium sensing receptor (*CASR*) has been demonstrated to aid in PC diagnosis (167), somatic *CASR* mutations was not found in PC. While significant focus has been put into determining the functional roles of *CDC73* (168,169,170) as well as its use for diagnosis of PC (171), the question of whether there are additional mutational characteristics and gene disruptions that provides a more comprehensive

genetic view of PC is still unanswered. The rapid maturation of high throughput sequencing technology and targeted DNA capture of protein coding regions in the past few years has enabled scientists to explore whole exomes across multiple samples enabling a wider and deeper view into the genetics of a disease. While whole genome sequencing of a single PC case was recently described highlighting mutational events driving this particular case (172), a wider and deeper view into mutational events and processes driving PC as a whole is still lacking and can be achieved through multi-sample DNA sequencing. Here, we present the whole-exome sequencing analysis of DNA from seven matched pairs of PC, tumor and corresponding leukocyte normal, as well as one matched triplet consisting of DNA from double primary tumor and corresponding leukocyte normal.

4.2: Results

4.2.1: Clinical samples and information

DNA from 8 patients diagnosed with PC, consisting of 6 males and 2 females [mean age 42 years (range 14 to 69 years)] are obtained consisting of 7 single primary tumors and 1 double primary tumor (7a and 7b) removed from the same surgery (Table 4.1). DNA from the above tumors, along with DNA from matched normal leukocyte from each patient, constitutes the discovery set. Exon captured sample DNA libraries obtained from the discovery set are sequenced using Illumina Hi-Seq 76bp Pair-End sequencing technology. For a PC validation set, DNA of thirteen PCs from 7 males and 6 females [mean age 50 years (range 29 – 75 years)], are selected from a previously described PC cohort consisting of formalin fixed paraffin embedded (FFPE) PCs (Table 4.2) (167). For parathyroid adenoma (PA) validation set, DNA are obtained from 40 patients, consisting of 7 males and 33 females [mean

age 69.5 years (range 47 to 89 years)]; Sixteen patients were described previously by Newey et al. 2012 (177). Variants to be verified by either validation sets are performed using Sanger sequencing.

4.2.2: PC whole-exome analysis

Whole-exome data are aligned using BWA (37) against the hg19/GRCh37 reference genome build. Read quality filtering and PCR duplicate removal are performed using SAMtools (39). We obtained an average sequencing depth of 105 with >86% of the exome sequenced to 20x depth, enabling high confidence variant calling (Table 4.3). To detect single nucleotide variants and small indels, a discovery pipeline based the Genome Analyzer ToolKit (41) is employed. Details of this discovery pipeline are discussed in Chapter 1. An exome SNPs concordance analysis is performed for all sample pairs with a >93% average concordance rate indicating the samples were correctly paired (Table 4.4). In addition, the SNPs of sample 7a and 7b are compared and showed an average concordance of 94% indicating single patient origin. Our discovery set has an average of 51 somatic variants per tumour (range 3-176) (Figure 4.1A) and of the 459 Sanger sequencing confirmed variants, 390 are due to somatic single nucleotide base substitutions and 69 are due to LOH, where LOH is taken to be a heterozygous variant in the normal changing to a homozygous variant in tumor DNA. Of the 390 base substitutions, 384 are heterozygous and 6 are homozygous with 265 non-synonymous mutations and 125 synonymous mutations (Figure 4.1A, Table 4.5 and 4.6).

4.2.3: *CDC73* mutational status and its effect on the PC exome

The presence of a high number of germ line and somatic mutations in *CDC73* in PC are confirmed in our discovery set in 7/9 samples with one novel indel (sample 8, c.539_544insA, p.I182NfsX10) (Figure 4.1B, Table 4.7). Sample 7a and 7b, both tumors excised at the same time from the patient involved, showed mutually exclusive somatic mutations and 10x differences in mutation numbers despite its common origin (Figure 4.1A, Table 4.4 and 4.5). Sample 7a and 7b are observed to contain different somatic "second hit" to their remaining wildtype copy of *CDC73*; the former has a somatic SNV predicted to cause a Leu95Pro amino acid substitution while the latter has a LOH of the wild type allele (Figure 4.1B, Table 4.7). The presence of mutations in genes related to DNA damage repair such as poly (ADP-ribose) polymerase 1 (*PARP1*) (Table 4.8) for sample 7b may indicate inefficiency in repairing somatic mutations in this lesion. ASCAT 2.0 Copy number estimation (51) using exome sequencing data shows 4/6 *CDC73* mutated samples (1, 2, 6, 7b) with aberrant 1q LOH or whole *CDC73* gene deletion (sample 6) as well as three to five copy number gain of the 1q allele containing the inactivated CDC73 copy (Figure 4.2A-I).

4.2.4: Novel recurrent mutations of *PRUNE2* in PC

We identified a novel PC gene, *PRUNE2*, mutated at both the germ line and somatic level. A *PRUNE2* germ line missense mutation (c.1354G>A, Val452Met) is found in a *CDC73* wildtype sample (sample 4; Table 4.7) with a deduced LOH of chromosome 9, where *PRUNE2* is located (Figure 4.1B, Figure 4.3). Two non-sense somatic mutations (c.1609G>T and c.1420G>T) of *PRUNE2* are seen in a *CDC73* mutated sample (sample 6; Table 4.7). The nonsense mutations are within 100 amino

acids downstream of the reported missense mutations and predicted to produce a truncated PRUNE2 protein lacking its BNIP-2 and Cdc42GAP Homology (BCH) domain. As the mutations are localized to exon 8 of *PRUNE2*, we further screened this exon across our FFPE-PC validation set which revealed two other somatic missense mutations (c.1364G>A, p.Gly455Asp; c.1349G>A, p.Ser450Asn) in samples negative for *CDC73* or *MEN1* mutations (Table 4.2). The three missense mutations are clustered within 6 amino acids of one another (codon 450-455) with conservation analysis showing all three amino acids are conserved across 28 mammalian species (Figure 4.4); all three amino acid mutations are computationally predicted by HumVar-trained PolyPhen model (46) to be probably damaging, in keeping with a likely pathogenic role in disrupting the function of PRUNE2. In total, 4/22 (18%) of PCs carried *PRUNE2* mutations. Screening of exon 8 of *PRUNE2* through the PA validation set revealed a single rare missense polymorphism (p.Asp1677Asn) in 40 tumors. Other than the Val452Met, all other *PRUNE2* variants are not found in COSMIC, ENSEMBL, dbSNP, 1000genomes or exome variant server (release 6500).

4.2.5: Kinase family is recurrently mutated in PC independent of *CDC73* mutation status

We have taken the validated list of somatically mutated genes from all our sequenced samples and performed a gene functional classification analysis using DAVID v6.7 (178) (Database for Annotation, Visualization and Integrated Discovery). The dominant representation of kinase genes in the functional classification highlights the importance of the kinase family in PC (Table 4.9). Interestingly, mutation status of *CDC73* does not affect the distribution of mutated

kinases and samples harboring mutated kinase(s) contain at least one predicted deleterious kinase mutation (Table 4.10). As sample 7b, which is from the recurrent parathyroid carcinoma, has a much higher number of somatic mutations and may skew the gene classification analysis, we repeated the same analysis without the mutational contributions of sample 7b and the results shows close agreement with our original analysis (Table 4.11).

4.2.6: APOBEC mutational signatures in PC

Using six classes of base substitutions (C > G, C > T, C > A, T > C, T > A, T > G) to detect mutational patterns, whole-exome PCs show a prevalence of C > T and C > G base substitutions (Figure 4.1C). The prevalence of C > T and C > G persists when samples are analyzed according to their *CDC73* mutation status (Figure 4.5A-B). We further broke down the C > G and C > T base substitutions into sixteen distinct classes of trinucleotide sequences by tracking the bases immediately 5' and 3' for each substitution (Figure 4.1D). Looking at the C > (G|T) base substitutions, we observed a distinctive pattern of TpCpW trinucleotide context most clearly shown for sample 7b, 6 and 2.

Given the similarity of the above mutational spectra to the APOBEC mutational signature in literature (55,56), we investigated if the signature is indeed present in our PC samples and if so, in what proportion. Since the APOBEC signature was shown to have significant contribution to the mutational spectra observed in bladder cancer (25,26), EMu (58), a probabilistic method incorporating tumor-specific opportunity for different mutation types according to sequence composition, is employed to infer PC mutational spectra and contribution from the trinucleotide context data extracted from our PC whole-exomes as well as from 328 bladder cancer whole-exomes downloaded from The Cancer Genome Atlas (179) and Beijing

Genome Institute (BGI) (180). Using the combined PC and bladder cancer data sets as input for EMu, the result showed clearly the presence of the APOBEC signature (Figure 4.6) as well as its contribution to mutational landscape of PC (Figure 4.1E). APOBEC signature appears to be particularly strong in samples with higher mutational burden such as sample 7b, 6 and 2 with the mutational process predicted to contribute to 80% - 98% of the mutations.

4.3: Discussion

PC is a rare endocrine malignancy primarily associated with HPT-JT due to inactivating mutations in the *CDC73* gene (162). Research thus far has yet to find any additional recurrently mutated genes in PC and the mutational landscape of PC is still completely unknown. In this study, we performed whole PC exome sequencing to analyze the mutational status of *CDC73*, to explore the involvement of novel PC-related genes and finally to study the mutational signatures of PC. Supported by previous reports (162,163,164), our studies confirmed the most frequently mutated PC-related genes, both at germline or somatic level is the *CDC73* gene. All except one mutation (6/7) in *CDC73* mutations found in our discovery set are indel in nature with predicted truncation of the affected protein within 15 amino acids from site of mutation; the exception being a SNV event predicted to cause a L95P amino acid substitution. In addition, we reported one novel *CDC73* germline indel mutation (Sample 8: c.539_544insA, p.I182NfsX10) not previously reported in literature.

Interestingly, the two cases (sample 3 and 8), each harboring a single heterozygous *CDC73* indel, are also the samples without any detectable somatic non-synonymous mutations indicating that *CDC73* indeed play an important early role in driving PC tumorigenesis. Of particular interest is sample 7a (primary tumor) and 7b

(recurrent tumor) which showed mutually exclusive somatic mutation sets with the exception of the germline *CDC73* mutation (c.356delA; p.Gln119ArgfsX14) present in both samples; tumor 7a has a second hit to *CDC73* as a SNV (c.284T > C) predicted to cause a L95P amino acid change while tumor 7b, on the other hand, has a LOH of *CDC73* wild type allele as the second hit. The surgical removal of both tumors during the same surgery, as well as the mutual exclusivity of the somatic mutations, points to the parallel and independent development of both tumors. Functional annotation of the mutated genes in sample 7b revealed a group of genes related to DNA damage repair (Table 4.8). In particular, somatic mutations predicted to cause amino acid substitutions in PARP1 (p.Asp678His) and polymerase (DNA directed), eta (POLH) (p.Asp67Asn) are computationally predicted by both Polyphen (45) and SIFT (82) to be damaging. While the aging process will certainly contribute to the number of somatic mutations, the young age of the patient 7 (age = 14) at the time of diagnosis along with the low number of somatic mutations (n = 15) in the parallel tumor (sample 7a) strongly point to the impairment of DNA repair machinery in sample 7b as a major factor in the substantial increase in somatic mutations. The trinucleotide context of the SNVs in sample 7b shows a preference for C > (T|G) in the TpCpW context; computational inference of mutational spectra and contribution reveals the APOBEC mutational process to be the major mutational source for the SNVs (98%) found in tumor 7b. Thus the following conjecture can be made; activation of APOBEC family of proteins leads to C > (T|G) genomic mutations in TpCpW context and, through random chance, introduced damaging mutations to genes responsible for DNA damage repair. Impairing DNA repair enabled the APOBEC mutational process to be much more effective in introducing its signature C > (T|G) genomic mutations in TpCpW context onto the tumor genome.

Previous PC research showed prevalence for chromosome 1 aberrations in malignant parathyroid tumors (181,182) and we confirm the finding in our data set with 7/9 samples containing predicted aberrant chromosome 1 copy number status. As *CDC73* is located within the chromosome 1q arm, we match the gene's mutation status with 1q copy number status; we found 4/7 CDC73 mutated samples (sample 1, 2, 6,7b) has a predicted 3 – 5 copy number gain of the mutant allele. Three of the samples (1, 2, 7b) showed LOH of the 1q arm containing the wild type allele. Sample 6 has a germline whole-gene deletion of *CDC73* and somatic indel of its remaining copy (c.32delA, p.Tyr11SerfsX10) followed by amplification of the allele containing the somatic indel. The copy number gain of allele containing mutant *CDC73* as well as LOH or whole-gene loss of its wild type allele is intriguing as the evidence appeared to contradict *CDC73*'s role as a tumor suppressor (183,184,185). The amplification events encompassed large segments of the 1q arm and in addition, samples containing wild type CDC73 (samples 4 and 5) or single copy loss of CDC73 (samples 3 and 8) showed no chromosomal aberration in 1q. The evidence suggests the presence of hidden proto-oncogenes in 1q that are regulated and suppressed by CDC73, as part of the polymerase associated factor complex, in a haplosufficient manner as single copy loss of CDC73 did not lead to 1q alterations. The loss of the remaining functional CDC73 copy through LOH or other somatic alterations enabled these proto-oncogenes in 1q to be unregulated with subsequent multiple large segment amplifications of these genes contributing to cellular transformation.

Through analyzing the mutational landscape of PC, we identified a novel PC-specific cancer gene, *PRUNE2*, recurrently mutated at both the germline and somatic level in PC. Sequence analysis of *PRUNE2* in 40 parathyroid adenomas revealed one rare missense polymorphism (p.Asp1677Asn) suggesting *PRUNE2* to be a PC

specific tumor suppressor gene. *PRUNE2*, also known as *BMCC1*, is a large 350kDa protein containing a BCH domain in the C-terminal region. Research showed that *PRUNE2* is up-regulated during neural growth factor (NGF)-depletion-induced apoptosis and high expression of *PRUNE2* was found to correlate with favorable prognosis in neuroblastoma and leiomyosarcoma (186,187). A functional study of PRUNE2 revealed its BCH domain suppresses Ras homolog family member A (RhoA) activity through interference of binding between RhoA and A kinase (PRKA) anchor protein 13 (Lbc), a Rho-specific guanine exchange factor; this results in reduced stress fiber formation and suppression of oncogenic cellular transformation (188). The three missense mutations we found are clustered within 6 amino acids of one another (codon 450-455) within a highly conserved region (Figure 4.4) and computationally predicted by PolyPhen (46) to be probably damaging. In addition, the nonsense mutations of *PRUNE2* we reported are within 100 amino acids downstream of the reported missense mutations and predicted to produce a truncated protein lacking the BCH domain. Based on the clustering of mutations and sequence conservation around this region of PRUNE2, we propose this region is important to the overall function of PRUNE2 and mutations in this region may contribute to increased susceptibility to developing PC through loss of control over cellular transformation.

With patterns emerging from base substitution classifications, we wonder whether the mutations can be classified at the gene level. Using a set of genes that were validated to have somatic mutations and asking the question whether this set of mutated genes have any functional similarities, our analysis showed that mutations in the kinase family is significantly over-represented in PC. Strikingly, we found the majority of mutated kinase genes to be involved in controlling cell migration and

invasion properties. Janus Kinase 1 (JAK1), Lim domain kinase 2 (LIMK2), CDC42 binding protein kinase alpha (DMPK-like) (CDC42BPA), RIO Kinase 3 (RIOK3) has been shown to act through the Rho-kinase dependent signaling pathway causing changes in the cytoskeletal structure that allows for increased migration and invasion in a variety of cancers (189-192). Similarly, fyn-related Src family tyrosine kinase (FRK) has been shown to suppress cell migration and invasion in glioma cells through c-Jun signaling pathway (193), tyrosine kinase with immunoglobulin-like and EGF-like domains 1 (TIE1) suppression leads to endothelial-mesenchymal transition in human endothelial cells (194), leukocyte receptor tyrosine kinase (LTK) mutations leads to loss of contact inhibition and anchorage-independent growth in epithelial cells (195) and loss of mitogen-activated protein kinase kinase kinase 11 (MAP3K11) expression has been shown to increase the invasive properties of AGS cell line (196). This over-representation of kinase genes connected to control of cellular migration and invasion processes can begin to offer an explanation to why higher proportion of PCs are locally invasive.

Taking the validated list of somatic single nucleotide mutations for each sample and classifying them according to six classes of base substitutions (C > G, C > T, C > A, A > G, A > T, A > C) revealed a distinct prevalence of C > T and C > G base substitutions for PC with the overall mutational spectra matching closely with breast cancer reported by Greenman et al. (18). Grouping the base substitutions distribution in terms of *CDC73* mutated and *CDC73* wildtype samples, the prevalence for C > (G|T) substitutions remain invariant to *CDC73* status indicating the base substitution pattern is a characteristic of PC as a whole. This result is also suggesting while *CDC73* inactivation is important to the development of PC, there may be separate process or processes driving the somatic base substitution patterns. Recent

studies have shown that mutational processes can be gleaned by taking into consideration the trinucleotide context surrounding the base substitution (54,197,198). Overall, we do not see an over-representation of C > T substitutions at the XpCpG triplets indicating that the elevated C > T mutation rate in PC is not due to deamination of methylated cytosines to thymine, a well-known mutational mechanism prevalent at XpCpG triplets (154,199). However, there is a distinctive spectra of C > (T|G) substitutions at TpCpW context for samples with higher mutational burden (sample 7b, 6, 2; Figure 4.1a and 1d) and computational inference pointing strongly to APOBEC mutational process as the main culprit contributing to the majority of the observed mutations (Figure 4.1E). Interestingly, two or the three patients (Patients 6 and 7) with the highest observed mutational burden are young (age 25 and 14 respectively) suggesting intensity rather than duration of the mutational process is the contributing factor to the high number of observed mutations possibly related to a differing activation of the APOBEC system. Furthermore, there is also evidence from gene expression and/or immunohistochemistry studies that three members of the APOBEC family, APOBEC3C, APOBEC3D and APOBEC3G, are indeed expressed in parathyroid tissue (200,201), the latter supporting our data.

In summary, this study is also the first to outline the genetic landscape of PC and attempts to characterize the mutational processes shaping the PC genome and how these processes shape disease behavior. Whole-exome analysis revealed *PRUNE2* to be recurrently mutated on a germline and somatic manner with mutations clustered around a functionally unknown but evolutionary conserved region of the protein. *PRUNE2* mutation rate may be underestimated in PC as only exon 8 was sequenced and whole gene sequencing of *PRUNE2* will be helpful in determining its true mutation rate in PC. Further functional studies of PRUNE2 are warranted to

understand the role this protein in PC tumorigenesis. APOBEC mutational signature was found to be dominant in a subset of PC patients with high mutational burden and early age onset of disease. Further research will be needed to establish the role of the APOBEC family and its activation mechanism in the context of PC. While members of the kinase family related to cell migration and invasion were found to be mutated in PC, larger scale studies involving increased sample sizes and more comprehensive sequencing techniques such as whole genome sequencing, RNA sequencing and bisulfite sequencing will likely yield additional evidence of gene families and mutational processes occurring in PC.

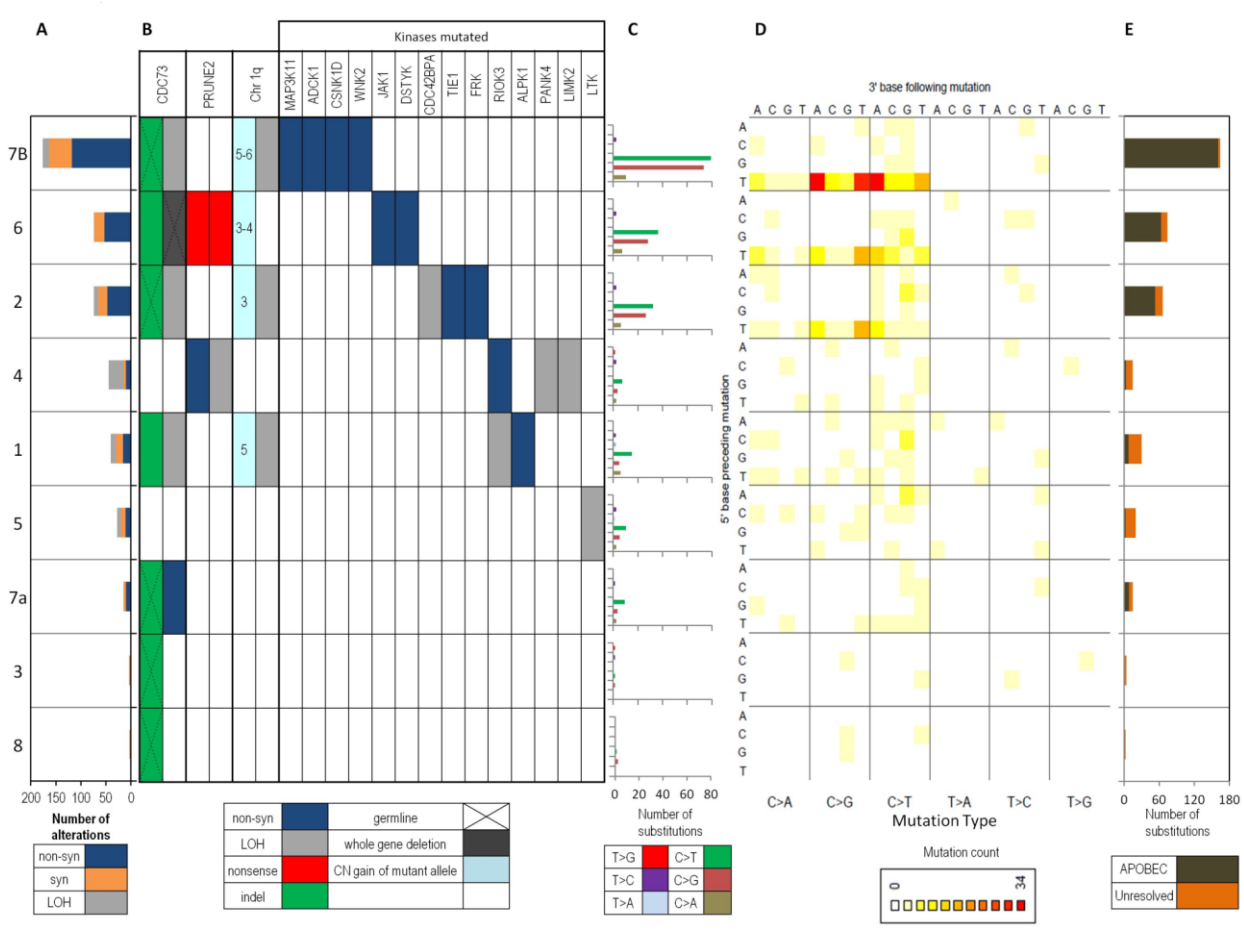
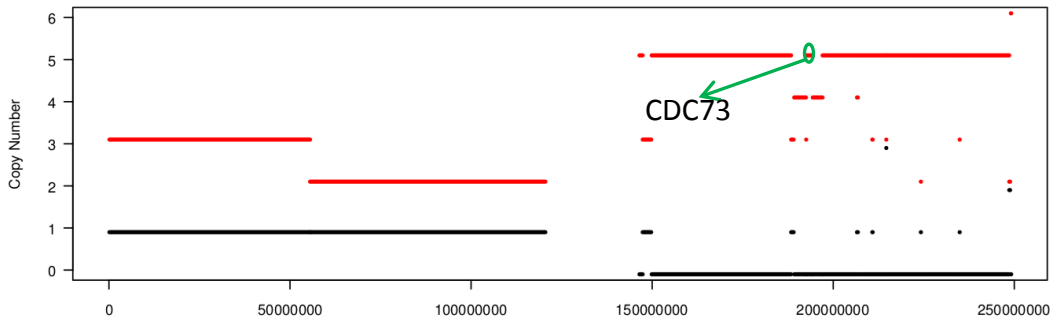


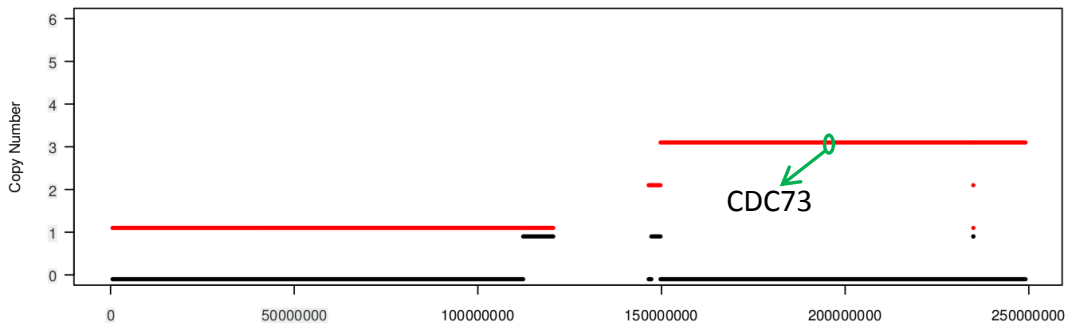
Figure 4.1: Mutational landscape of PC. A) Total number of somatic single nucleotide alterations B) Major mutational alterations C) Single nucleotide base substitutions spectra D) Heatmap of trinucleotide base substitution contexts E) Mutational contribution of APOBEC

Figure 4.2: Copy number estimation of chromosome 1 for each whole-exome sequenced PC sample using ASCAT 2.0

A. Sample 1:
Mutation status of *CDC73*: somatic indel
Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031
Predicted fragment location: chr1:192,552,160 - 194,325,878
Predicted fragment copy number gain: 5



B. Sample 2:
Mutation status of *CDC73*: germline indel
Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031
Predicted fragment location: chr1:149,732,207 - 234,853,921
Predicted fragment copy number gain: 3

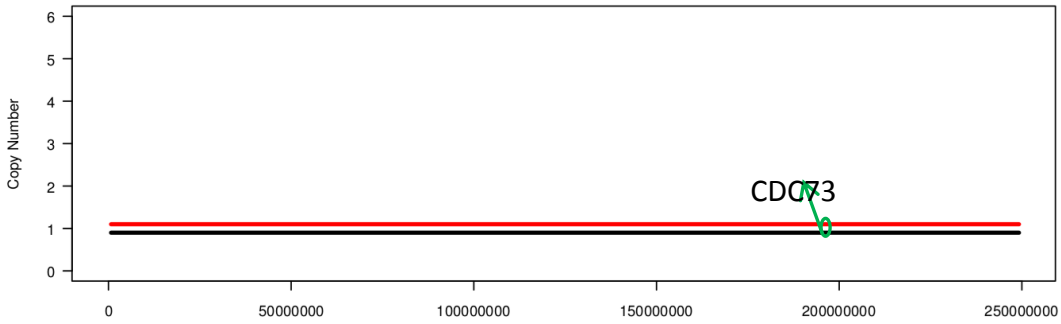


C. Sample 3:

Mutation status of *CDC73*: germline indel

Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031

No predicted copy number alterations for chromosome 1q

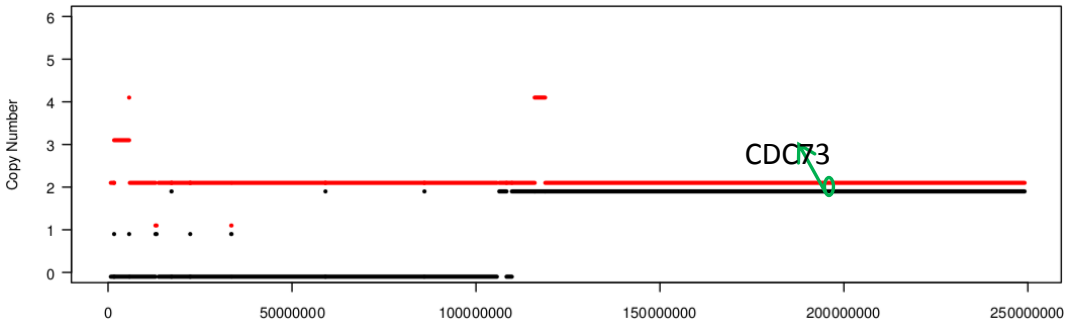


D. Sample 4:

Mutation status of *CDC73*: wildtype

Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031

No predicted copy number alterations where *CDC73* is located

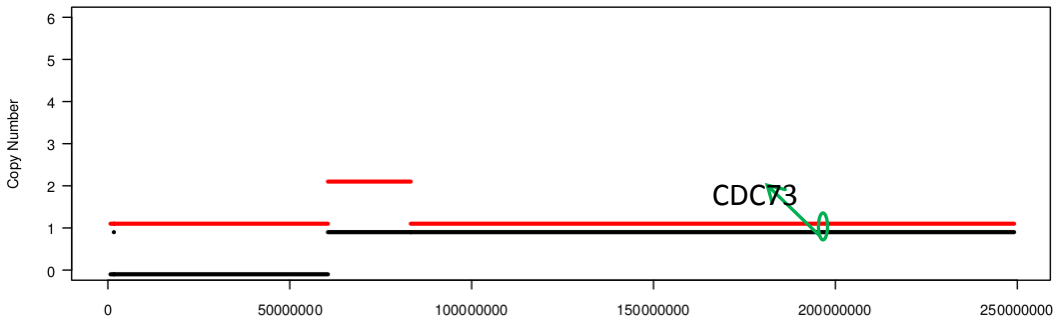


E. Sample 5:

Mutation status of *CDC73*: wildtype

Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031

No predicted copy number alterations where *CDC73* is located



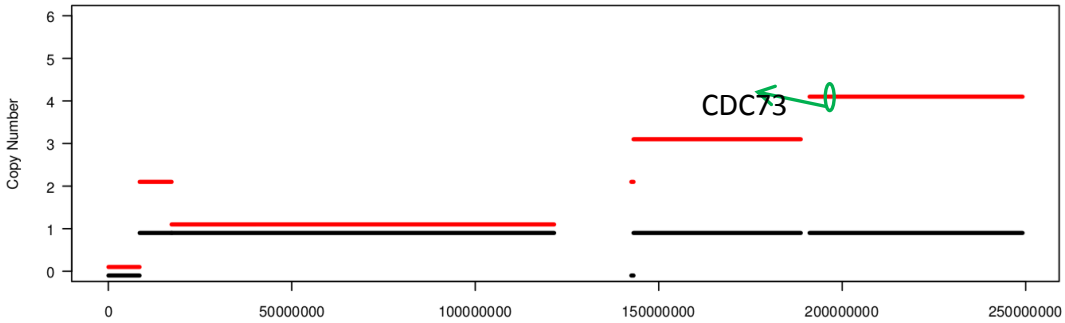
F. Sample 6:

Mutation status of *CDC73*: somatic indel

Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031

Predicted fragment location: chr1:191,115,965 - 249,150,330

Predicted fragment copy number gain: 4

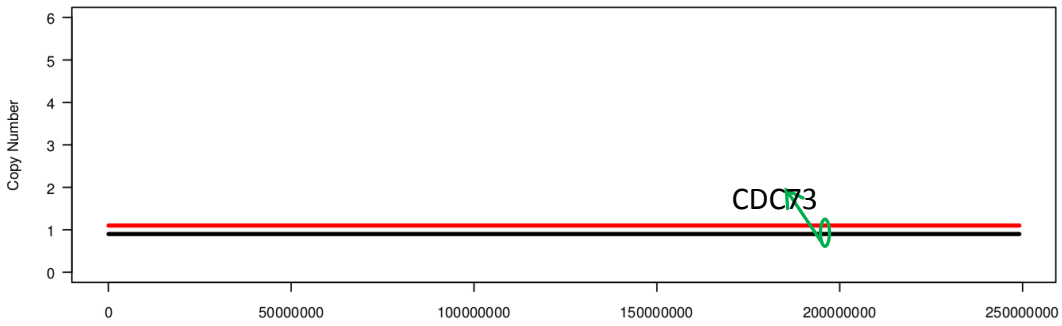


G. Sample 7a:

Mutation status of *CDC73*: germline indel / somatic SNV

Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031

No predicted copy number alterations for chromosome 1q



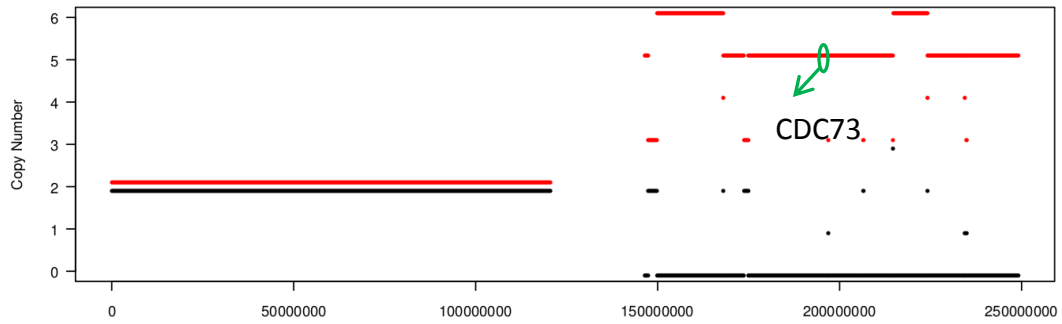
H. Sample 7b:

Mutation status of *CDC73*: germline indel

Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031

Predicted fragment location: chr1:174,927,388 - 196,876,458

Predicted fragment copy number gain: 5

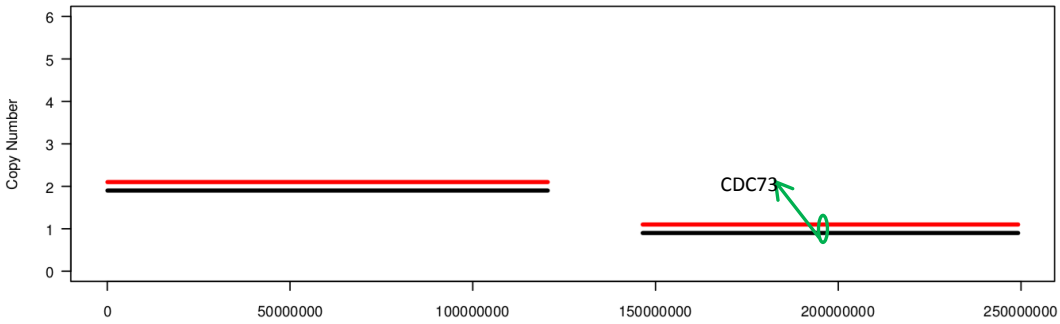


I. Sample 8:

Mutation status of *CDC73*: germline indel

Location of *CDC73* gene: Chr1: 193,091,147 - 193,223,031

No predicted copy number alterations where *CDC73* is located



Mutation status of *PRUNE2*: c.1354G>A
Location of *PRUNE2* gene: Chr9: 79,226,292 – 79,521,003

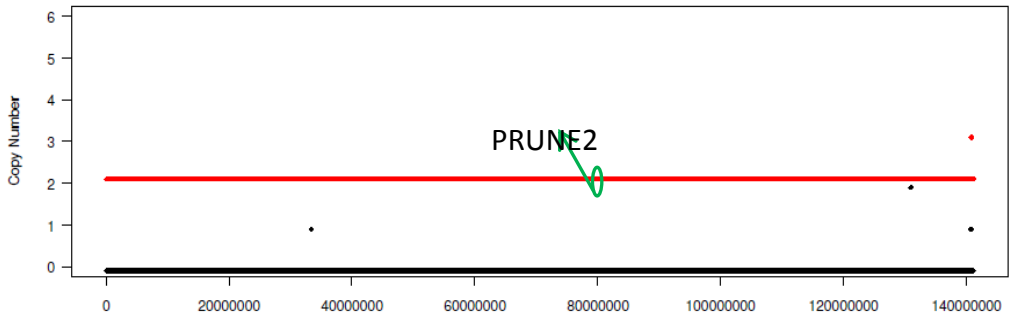


Figure 4.3: Predicted LOH of chromosome 9 for sample 4 using ASCAT 2.0.

Homo sapiens	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHTLLPGLDSY
Papio hamadryas	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHTLLPGLDSY
Callithrix jacchus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGLDSY
Gorilla gorilla	RSSRSSKESSVFLSDD**S**X**V**GE**G**AGPHHTLLPGLDSY
Macaca mulatta	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHTLLPGLDSY
Equus caballus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Pteropus vampyrus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Canis lupus familiaris	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Oryctolagus cuniculus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Felis catus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Bos taurus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Loxodonta africana	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Microcebus murinus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHTLLPGLDSY
Tupaia belangeri	RSSRSSKESSVFLSDD**S**P**V**GE**G**AVPHHSLLPGFDSY
Tursieps truncatus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Vicugna Pacos	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Cavia porcellus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Otolemur garnettii	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGLDSY
Rattus norvegicus	RSSRSSKESSVFLSDD**S**P**V**GE**G**GGPHHSLLPGFESY
Mus musculus	RSSRSSKESSVFLSDD**S**P**V**GD**G**GAPHHSLLPGFDSY
Myotis lucifugus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Procavia capensis	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Erinaceus europaeus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AVPHHNLLPGFDSY
Sorex araneus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHNLFPGFDSY
Spermophilus tridecemlineatus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGFDSY
Dipodomys ordii	RSSRSSKESSVFLSDD**S**P**V**GE**G**VGPHHSLLPGFDSY
Tarsius syrichta	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLPGLDSY
Dasypus novemcinctus	RSSRSSKESSVFLSDD**S**P**V**GE**G**AGPHHSLLP-FDSY

Figure 4.4: **Twenty eight mammalian species conservation analysis of PRUNE2 residue positions (Ser450, Val452, Gly455) corresponding to the three non-synonymous mutations (c.1349G>A, c.1354G>A, c.1364G>A) found in PC.** Bold and underlined letters are the residues predicted to be mutated in PRUNE2.

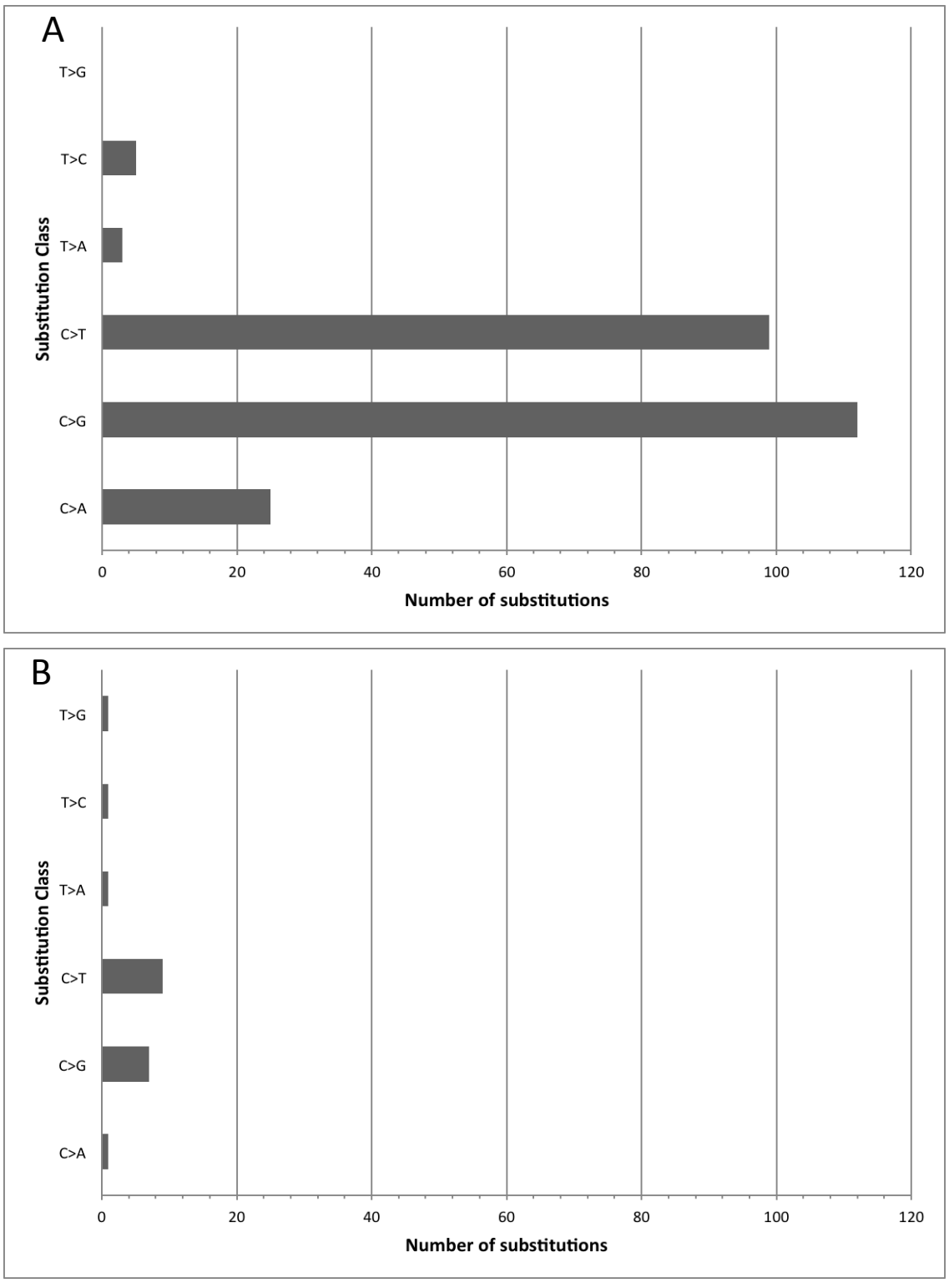
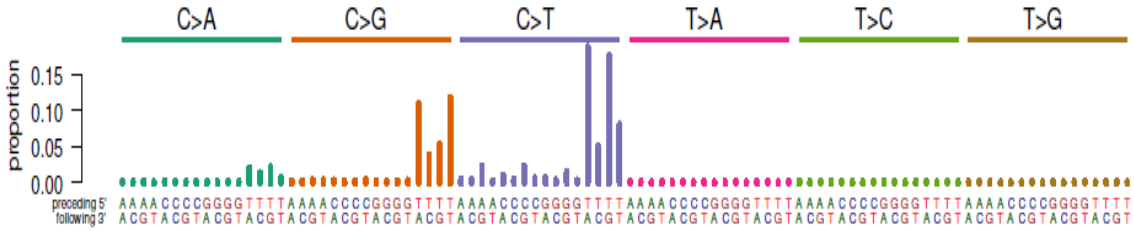


Figure 4.5: Distribution of base substitutions in PC. A) Base substitutions distribution of *CDC73* mutated samples. B) Base substitutions distribution of *CDC73* wildtype samples.

APOBEC



Unresolved

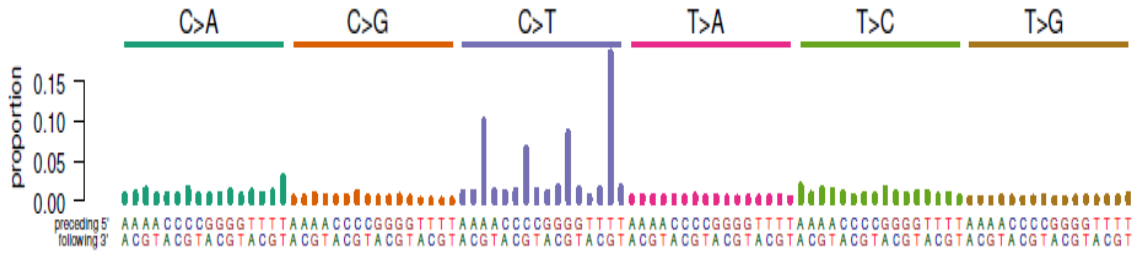


Figure 4.6: Mutational signatures found by EMu.

Sample	Previously reported	Previous Patient number	Age/Sex	Follow-up	Reference
1	YES	6	29/M	A/WD*	167
2	YES	III.10	59/M	D/other*	173
3	YES	K3	23/M	A/FOD*	174
4	NO	N/A	69/M	A/FOD*	N/A
5	NO	N/A	63/F	A/FOD*	N/A
6	YES	single patient study	25/M	A/FOD*	175
7a	YES	V.1	14/F	A/FOD*	173
7b					
8	NO	N/A	54/M	A/FOD*	N/A

Table 4.1: Patient information for PC discovery set. *: DOD = dead of disease; A/FOD = alive free of disease; A/WD = alive with disease, D/other = dead of other causes.

Sample	Previous patient number (ref. 167)	Age/ sex	Follow-up (ref.167)	Previous mutation screening of *CDC73* and *MEN1* (ref.176)	PRUNE2 mutations
A1	2	71/F	DOD*	Negative	Negative
A2	7	57/M	A/FOD*	Negative **	Negative
A3	8	48/M	A/FOD*	Negative **	Negative
A4	9	32/M	A/WD*	Negative **	Negative
A5	10	75/M	D/other*	Negative **	Negative
A6	12	66/F	DOD*	Negative	c.1364G>A, p.G455D
A7	13	34/M	DOD*	Negative	Negative
A8	14	50/F	DOD*	Somatic *MEN1*	Negative
A9	15	29/F	DOD*	Negative	c.1349G>A, p.S450N
A10	16	51/F	A/FOD*	Somatic *MEN1*	Negative
A11	21	41/M	DOD*	Germline+somatic *CDC73*	Negative
A12	22	62/M	A/FOD*	Negative	Negative
A13	23	36/F	D/other*	Germline *CDC73*	Negative

Table 4.2: Sample information for PC validation set. *: DOD = dead of disease; A/FOD = alive free of disease; A/WD = alive with disease, D/other = dead of other causes. **: targeted next generation sequencing on FFPE DNA confirmed the negative results for *CDC73*/*MEN1*.

Sample		Bases in Target Region	Reads mapped to target region	Ave. Depth Per Targeted Base	Targeted Bases with Depth at Least 10X	Targeted Bases with Depth at Least 20X
1	Normal	51,860,012	87,304,429	95	96.5	85
	Tumor		105,118,760	114	96.0	89
2	Normal		81,028,724	89	95.5	84
	Tumor		82,355,074	89	96.1	85
3	Normal		78,324,375	85	95.8	84
	Tumor		82,722,473	90	95.9	85
4	Normal		131,453,308	144	95.7	89
	Tumor		107,713,618	118	95.9	88
5	Normal		100,258,663	110	95.5	87
	Tumor		104,615,391	115	96.0	87
6	Normal		105,861,343	116	95.6	87
	Tumor		93,510,226	102	95.6	86
7A	Normal		92,468,242	101	95.5	87
	Tumor		102,365,127	112	95.8	88
7B	Tumor		113,665,328	124	95.8	88
8	Normal		80,191,918	87	95.6	85
	Tumor		82,835,644	90	95.6	85
Average			**95,987,803**	**105**	**95.8**	**86**

Table 4.3: PC whole-exome sequencing summary.

Sample pair	% dbSNPs in common
1N/1T	92.1
2N/2T	91.7
3N/3T	95.3
4N/4T	91.1
5N/5T	91.7
6N/6T	95.9
7N/7aT	95.7
7N/7bT	93.3
7aT/7bT	92.8
8N/8T	95.2
average	**93.48**

Table 4.4: Exome dbSNP concordance of whole-exome sequenced PC samples.

Table 4.5: Validated single nucleotide variants for whole-exome sequenced PC samples.

Sample	Gene Symbol	CCDS ID	Chr	Pos	Ref	Cons	AA Change	Change Type	mutation status
1	ADAMTS4	CCDS1223.1	1	161,163,439	T	C	T576A	Nonsyn	LOH
1	ALPK1	CCDS3697.1	4	113,345,145	A	G	N174S	Nonsyn	Somatic
1	ASTN2	CCDS48009.1	9	119188052	C	A	-	syn	somatic
1	C3orf32	CCDS2568.1	3	8,669,387	C	T	R202Q	Nonsyn	Somatic
1	CCDC17	CCDS44131.1	1	46088449	G	A	-	syn	somatic
1	CDX4	CCDS14424.1	x	72674244	C	A	-	syn	somatic
1	COPA	CCDS41424.1	1	160,261,683	C	A	V1071L	Nonsyn	LOH
1	DCAF6	CCDS1267.2	1	167,962,659	C	T	A295V	Nonsyn	LOH
1	DIAPH3	CCDS41898.1	13	60,582,786	A	T	-	3_prime_splice	Somatic
1	DUSP27	CCDS30932.1	1	167,096,337	A	G	I657V	Nonsyn	LOH
1	EVI5L	CCDS12188.1	19	7928321	G	A	-	syn	somatic
1	FAM187B	CCDS12448.1	19	35,719,283	G	A	R101C	Nonsyn	Somatic
1	FAM19A5	CCDS46728.1	22	49145791	C	T	-	syn	somatic
1	FANCA	CCDS32515.1	16	89805093	G	A	-	syn	somatic
1	FANCI	CCDS45346.1	15	89,850,745	A	T	K1192N	Nonsyn	Somatic
1	IGSF1	CCDS14629.1	x	130,412,488	C	A	W663L	Nonsyn	Somatic
1	IRF4	CCDS4469.1	6	397278	T	C	-	syn	somatic
1	LIPG	CCDS11938.1	18	47,110,081	G	A	R438H	Nonsyn	LOH
1	MAGEB17	ENST00000329538	x	16,189,318	C	A	F271L	Nonsyn	Somatic
1	MCF2L	CCDS45070.1	13	113,731,372	G	A	V560M	Nonsyn	LOH
1	MEG3	ENST00000398461	14	101,300,981	C	T	A143V	Nonsyn	Somatic
1	MICAL3	CCDS46659.1	22	18,371,839	G	A	Q618X	Nonsyn	Somatic
1	MOGAT3	ENST00000440203	7	100,839,452	G	A	P296L	Nonsyn	Somatic
1	MYO7B	CCDS46405.1	2	128,354,076	G	C	V762L	Nonsyn	Somatic
1	NCAN	CCDS12397.1	19	19335832	C	G	-	syn	somatic
1	NOTCH1	CCDS43905.1	9	139,412,330	G	A	Q439X	Nonsyn	Somatic

1	OFCC1	ENST00000460363	6	9,809,942	G	C	L117V	Nonsyn	Somatic
1	OR2G2	CCDS31092.1	1	247,751,881	C	G	R74G	Nonsyn	LOH
1	PLG	CCDS5279.1	6	161,152,814	G	C	K492N	Nonsyn	Somatic
1	PSMA7	CCDS13489.1	20	60718297	G	A	-	syn	somatic
1	RAB27B	CCDS11958.1	18	52556530	C	A	-	syn	somatic
1	RFWD2	CCDS30944.1	1	176,175,819	A	C	V99G	Nonsyn	LOH
1	RIOK3	CCDS11877.1	18	21,044,569	A	G	K174E	Nonsyn	LOH
1	SDCCAG1	CCDS9694.1	14	50,251,823	C	A	Q1020H	Nonsyn	Somatic
1	SEC16A	NM_014866	9	139,369,589	G	C	P827A	Nonsyn	Somatic
1	SEMA4A	CCDS1132.1	1	156,145,373	G	A	R540Q	Nonsyn	LOH
1	SLC15A3	CCDS7998.1	11	60718808	C	T	-	syn	somatic
1	SLC5A5	CCDS12368.1	19	17994700	G	A	-	syn	somatic
1	SLC8A2	CCDS33065.1	19	47960525	C	T	-	syn	somatic
1	UNC119	CCDS42232.1	17	26879390	C	T	-	syn	somatic
2	-	CCDS13527.1	20	62196677	G	A	-	syn	somatic
2	ABCB1	CCDS5608.1	7	87,183,219	C	G	R286T	Nonsyn	Somatic
2	ABLIM3	CCDS4294.1	5	148,579,994	G	A	-	5_prime_splice	Somatic
2	ACCN4	CCDS33384.1	2	220397639	C	T	-	syn	somatic
2	ANKRD35	CCDS919.1	1	145,560,250	C	T	R246W	Nonsyn	Somatic
2	CAMTA2	CCDS11063.1	17	4,872,096	C	G	Q1188H	Nonsyn	Somatic
2	CARD10	CCDS13948.1	22	37887785	A	G	-	syn	somatic
2	CCDC102B	CCDS11996.2	18	66,504,391	A	G	M131V	Nonsyn	Somatic
2	CDC42BPA	CCDS1558.1	1	227,216,756	C	T	R1310H	Nonsyn	LOH
2	CFH	CCDS1385.1	1	196,646,659	G	T	A161S	Nonsyn	LOH
2	CKAP4	CCDS9103.1	12	106641204	C	T	-	syn	somatic
2	CLRN3	CCDS7656.1	10	129,690,950	C	T	W33X	Nonsyn	Somatic
2	CLRN3	CCDS7656.1	10	129691034	C	T	-	syn	somatic
2	CRB1	CCDS1390.1	1	197,298,095	T	C	I205T	Nonsyn	LOH
2	DCAF13	CCDS34934.1	8	104,447,884	C	A	D424E	Nonsyn	Somatic
2	DCHS2	CCDS3785.1	4	155298567	G	C	-	syn	somatic

2	DDX31	CCDS6951.1	9	135,536,620	C	A	V248F	Nonsyn	Somatic
2	DDX31	CCDS6951.1	9	135,536,639	C	G	-	3_prime_splice	Somatic
2	DET1	CCDS45343.1	15	89,074,081	G	C	P297A	Nonsyn	Somatic
2	DOCK2	CCDS4371.1	5	169,494,529	C	G	L1495V	Nonsyn	Somatic
2	EFTUD1	CCDS42071.1	15	82,444,751	G	C	H682D	Nonsyn	Somatic
2	FAM89A	CCDS1590.1	1	231,155,682	C	T	R161Q	Nonsyn	LOH
2	FCRL3	CCDS1167.1	1	157,660,164	G	C	S524C	Nonsyn	Somatic
2	FRK	CCDS5103.1	6	116,265,579	C	A	G323V	Nonsyn	Somatic
2	HNRNPA1P4	ENST00000509706	8	83,204,580	C	G	G10A	Nonsyn	Somatic
2	HNRNPA3	CCDS2273.1	2	178,080,327	G	C	E45Q	Nonsyn	Somatic
2	HPD	CCDS9224.1	12	122,281,719	C	G	R284T	Nonsyn	Somatic
2	HSD17B2	CCDS10936.1	16	82132073	C	T	-	syn	somatic
2	HSPA6	CCDS1231.1	1	161,495,065	C	T	T206I	Nonsyn	LOH
2	KCNK1	CCDS1599.1	1	233,802,497	G	A	R171H	Nonsyn	LOH
2	KIAA2022	CCDS35337.1	x	73,961,334	C	G	D1020H	Nonsyn	Somatic
2	KLHDC9	CCDS30919.1	1	161,068,428	G	A	E35K	Nonsyn	LOH
2	LRP11	CCDS5220.1	6	150184722	C	T	-	syn	somatic
2	MACF1	CCDS436.1	1	39,896,470	G	C	E4183Q	Nonsyn	Somatic
2	MAFA	CCDS34955.1	8	144511513	G	A	-	syn	somatic
2	MBTPS1	CCDS10941.1	16	84,129,370	C	A	Q154H	Nonsyn	Somatic
2	NAT10	CCDS44568.1	11	34165001	C	T	-	syn	somatic
2	NTPCR	CCDS1597.1	1	233,105,700	G	A	E114K	Nonsyn	Somatic
2	NUDT17	CCDS30830.1	1	145,586,636	C	G	E314Q	Nonsyn	Somatic
2	OXGR1	CCDS9482.1	13	97,639,013	G	A	S334L	Nonsyn	Somatic
2	PALM	CCDS32857.1	19	746742	C	T	-	syn	somatic
2	PLXNB2	CCDS43035.1	22	50726207	G	A	-	syn	somatic
2	PRKD1	CCDS9637.1	14	30396704	C	T	-	syn	somatic
2	REEP4	CCDS6024.1	8	21,998,159	C	G	K28N	Nonsyn	Somatic
2	REXO4	CCDS6969.1	9	136276150	C	T	-	syn	somatic
2	RHOJ	CCDS9757.1	14	63,671,609	G	A	D8N	Nonsyn	Somatic

2	RPGRIP1	CCDS45080.1	14	21,796,613	G	A	E976K	Nonsyn	Somatic
2	RPGRIP1	CCDS45080.1	14	21,796,716	G	C	R1010T	Nonsyn	Somatic
2	RYR3	CCDS45210.1	15	33,941,432	G	A	E1380K	Nonsyn	Somatic
2	SCYL3	CCDS1287.1	1	169,831,884	C	T	R337Q	Nonsyn	LOH
2	SIGLEC1	CCDS13060.1	20	3,672,806	C	A	R1358S	Nonsyn	Somatic
2	SIGLEC15	CCDS32819.1	18	43418816	C	T	-	syn	somatic
2	SLC35F3	CCDS1600.1	1	234,041,480	C	T	Q87X	Nonsyn	Somatic
2	SLITRK5	CCDS9465.1	13	88,330,440	C	G	L933V	Nonsyn	Somatic
2	SNX32	CCDS8113.2	11	65,620,798	G	A	E402K	Nonsyn	Somatic
2	SRGAP1	ENST00000357825	12	64,485,680	C	T	R482X	Nonsyn	Somatic
2	SULT1C3	CCDS33267.1	2	108,875,210	G	C	D183H	Nonsyn	Somatic
2	SYNE1	CCDS5236.1	6	152,651,701	C	G	E4707Q	Nonsyn	Somatic
2	TACC2	CCDS7626.1	10	123,844,807	C	G	S931X	Nonsyn	Somatic
2	TIE1	CCDS482.1	1	43,774,729	G	A	C372Y	Nonsyn	Somatic
2	TMPRSS11F	CCDS3520.1	4	68,930,558	C	T	R287K	Nonsyn	Somatic
2	TOR1AIP2	CCDS1334.1	1	179,820,399	G	C	S45C	Nonsyn	Somatic
2	TOR1AIP2	CCDS1334.1	1	179,820,457	G	C	Q26E	Nonsyn	Somatic
2	TRMT12	CCDS6349.1	8	125,464,067	C	T	S300L	Nonsyn	Somatic
2	TUBA4B	NR_003063	2	220,135,939	C	G	R82G	Nonsyn	Somatic
2	UBN2	CCDS43655.1	7	138,957,071	G	T	D451Y	Nonsyn	Somatic
2	UBN2	CCDS43655.1	7	138,957,081	G	A	R454K	Nonsyn	Somatic
2	VTCN1	CCDS894.1	1	117690325	G	A	-	syn	somatic
2	WDR91	CCDS34758.1	7	134896257	C	T	-	syn	somatic
2	ZEB1	CCDS7169.1	10	31,815,684	G	C	G956A	Nonsyn	Somatic
2	ZNF202	CCDS8443.1	11	123600375	C	T	-	syn	somatic
2	ZNF259	CCDS8375.1	11	116658731	C	T	-	syn	somatic
2	ZNF334	CCDS33480.1	20	45,132,925	C	G	D57H	Nonsyn	Somatic
2	ZNF597	CCDS10505.1	16	3,487,020	G	C	L227V	Nonsyn	Somatic
3	BRWD3	CCDS14447.1	x	80064802	C	T	-	syn	somatic
3	C11orf35	CCDS7701.1	11	555835	A	C	-	syn	somatic

3	C14orf73	CCDS32163.1	14	103568900	C	G	-	syn	somatic
3	NUDT16L1	CCDS10519.1	16	4743711	A	G	-	syn	somatic
4	ABHD12B	CCDS9702.1	14	51,347,180	C	G	P39A	Nonsyn	LOH
4	ADAMTS13	CCDS6970.1	9	136,287,604	G	C	C14S	Nonsyn	LOH
4	AKAP9	CCDS5622.1	7	91,726,396	C	T	Q3375X	Nonsyn	Somatic
4	ANXA3	CCDS3584.1	4	79,494,338	G	C	G7A	Nonsyn	Somatic
4	API5	uc001mxg.2	11	43,357,545	G	A	G372D	Nonsyn	Somatic
4	ARX	CCDS14215.1	x	25031461	C	A	-	syn	somatic
4	ATG4C	CCDS623.1	1	63,282,474	G	A	G130D	Nonsyn	LOH
4	C14orf166B	CCDS9853.2	14	77,297,656	G	A	V110M	Nonsyn	LOH
4	C1orf59	CCDS787.1	1	109,191,376	G	C	P332A	Nonsyn	Somatic
4	C20orf151	CCDS13498.1	20	60989567	C	T	-	syn	somatic
4	C9orf174	CCDS35077.1	9	100,080,823	C	G	N529K	Nonsyn	LOH
4	CASZ1	CCDS41246.1	1	10,725,469	G	A	S59L	Nonsyn	LOH
4	CDC45	CCDS13762.1	22	19,470,327	G	A	V107I	Nonsyn	LOH
4	COL6A1	CCDS13727.1	21	47402598	C	T	-	syn	somatic
4	CRAT	CCDS6919.1	9	131,864,761	G	A	P183L	Nonsyn	LOH
4	CYB5R2	CCDS7780.1	11	7,690,921	C	G	V65L	Nonsyn	Somatic
4	EVL	CCDS9955.1	14	100,563,974	A	G	I113V	Nonsyn	LOH
4	EXD3	CCDS48066.1	9	140,243,678	G	A	R572C	Nonsyn	LOH
4	GBGT1	ENST00000372043	9	136,029,336	C	T	W218X	Nonsyn	LOH
4	HIATL1	CCDS6710.2	9	97,177,527	A	G	M66V	Nonsyn	LOH
4	HOXD10	CCDS2266.1	2	176,983,828	G	T	E298X	Nonsyn	Somatic
4	IGF2R	CCDS5273.1	6	160430070	A	G	-	syn	somatic
4	INPP5B	CCDS41306.1	1	38,397,626	G	C	S164W	Nonsyn	LOH
4	KIAA1539	CCDS6578.1	9	35,108,261	A	T	V4E	Nonsyn	LOH
4	LEPRE1	CCDS472.2	1	43,220,661	C	T	-	3_prime_splice	LOH
4	LIMK2	CCDS13891.1	22	31,674,324	C	G	S605C	Nonsyn	LOH
4	MAP7D1	ENST00000309824	1	36,645,502	C	T	P449S	Nonsyn	LOH
4	MOV10L1	ENST00000395843	22	50,596,655	G	A	R1024H	Nonsyn	LOH

4	PANK4	CCDS42.1	1	2,451,796	A	G	F222L	Nonsyn	LOH
4	PCSK5	NM_001190482	9	78,942,944	G	C	L1426F	Nonsyn	LOH
4	PLA2G6	ENST00000425347	22	38,539,175	G	A	P19L	Nonsyn	LOH
4	PLEKHG2	CCDS33022.2	19	39,913,768	G	A	D692N	Nonsyn	Somatic
4	PRUNE2	CCDS47982.1	9	79,325,836	C	T	V452M	Nonsyn	LOH
4	RHEB	CCDS5927.1	7	151,174,471	A	C	S75A	Nonsyn	Somatic
4	RIOK3	CCDS11877.1	18	21,053,547	A	G	I324V	Nonsyn	Somatic
4	SERPINA4	CCDS9927.1	14	95,035,841	G	A	R398Q	Nonsyn	LOH
4	SNCAIP	CCDS4131.1	5	121787062	G	A	-	syn	somatic
4	TAS1R1	CCDS81.1	1	6,639,297	G	C	G727R	Nonsyn	LOH
4	TCF20	CCDS14033.1	22	42,609,597	T	C	N572S	Nonsyn	LOH
4	TRIM14	CCDS6734.1	9	100,854,283	C	T	S234N	Nonsyn	LOH
4	TTC7B	CCDS32140.1	14	91,044,555	C	T	M735I	Nonsyn	LOH
4	TTN	NM_133378	2	179,459,139	G	A	A16793V	Nonsyn	Somatic
4	UBR7	CCDS9909.1	14	93,685,598	C	G	S284C	Nonsyn	LOH
4	WASF2	CCDS304.1	1	27,736,429	G	A	P366S	Nonsyn	LOH
4	WLS	CCDS30750.1	1	68,624,837	C	T	R156Q	Nonsyn	LOH
5	AKAP13	CCDS32320.1	15	86,283,483	G	A	E2534K	Nonsyn	LOH
5	ANO7	CCDS33423.1	2	242157740	C	T	-	syn	somatic
5	C10orf112	ENST00000377266	10	19,569,012	C	T	T335M	Nonsyn	Somatic
5	C3orf16	CCDS46933.1	3	149,508,696	G	A	P36S	Nonsyn	Somatic
5	CA6	CCDS30578.1	1	9027746	C	T	-	syn	somatic
5	CCNB2	CCDS10170.1	15	59,406,987	T	C	V170A	Nonsyn	LOH
5	CLEC1B	CCDS41752.1	12	10,145,809	C	T	C208Y	Nonsyn	Somatic
5	CSTF2T	CCDS44399.1	10	53457452	A	G	-	syn	somatic
5	CYP51A1	CCDS5623.1	7	91763625	C	A	-	syn	somatic
5	DMGDH	CCDS4044.1	5	78,294,107	G	A	T800M	Nonsyn	Somatic
5	ENTPD2	CCDS7025.1	9	139945711	C	T	-	syn	somatic
5	HLA-F	CCDS43437.1	6	29,693,221	C	G	-	3_prime_splice	Somatic
5	IFI35	CCDS11450.1	17	41,165,538	C	T	P143S	Nonsyn	Somatic

5	LL0XNC01-221F2.2	ENST00000440243	x	102,342,307	A	T	K62X	Nonsyn	Somatic
5	LTK	CCDS10077.1	15	41,797,670	G	A	R586C	Nonsyn	LOH
5	MAML1	CCDS34315.1	5	179,192,415	C	G	P135R	Nonsyn	Somatic
5	METTL14	CCDS34053.1	4	119,618,370	C	G	I179M	Nonsyn	Somatic
5	NPPA	CCDS139.1	1	11,907,171	C	T	R150Q	Nonsyn	LOH
5	PCDH19	CCDS43976.1	x	99657774	C	A	-	syn	somatic
5	PLG	CCDS5279.1	6	161,139,488	A	T	K317I	Nonsyn	LOH
5	POLA2	CCDS8098.1	11	65064680	G	A	-	syn	somatic
5	STT3B	CCDS2650.1	3	31,663,682	T	C	I474T	Nonsyn	Somatic
5	TNXB	CCDS47407.1	6	32064334	C	G	-	syn	somatic
5	TRDN	ENST00000265491	6	123,850,559	G	T	S167Y	Nonsyn	LOH
5	TTC16	CCDS6875.1	9	130,489,684	C	G	S568R	Nonsyn	Somatic
5	UTRN	CCDS34547.1	6	145,142,155	G	A	-	5_prime_splice	LOH
5	ZCCHC11	ENST00000466440	1	52,890,963	T	C	E172G	Nonsyn	LOH
5	ZNF467	CCDS5899.1	7	149463111	G	A	-	syn	somatic
6	ACSM5	CCDS10585.1	16	20,422,809	G	A	M1I	Nonsyn	Somatic
6	ADAMTS6	CCDS3983.2	5	64,537,947	C	T	W639X	Nonsyn	Somatic
6	AHSG	CCDS3278.1	3	186334255	C	T	-	syn	somatic
6	ARHGAP44	CCDS45616.1	17	12890454	G	A	-	syn	somatic
6	ARMC4	CCDS7157.1	10	28,149,754	G	A	H941Y	Nonsyn	Somatic
6	C15orf55	CCDS32190.1	15	34,645,940	C	A	F286L	Nonsyn	Somatic
6	C9orf114	CCDS6913.1	9	131,586,390	C	T	R292H	Nonsyn	Somatic
6	C9orf114	CCDS6913.1	9	131,588,369	C	G	E191Q	Nonsyn	Somatic
6	CAMKK2	CCDS44999.1	12	121691151	G	C	-	syn	somatic
6	CD83	CCDS4532.1	6	14,131,877	A	T	I94F	Nonsyn	Somatic
6	CHODL	CCDS13570.1	21	19,628,962	G	C	E72D	Nonsyn	Somatic
6	COL13A1	CCDS44419.1	10	71,697,439	G	C	E605Q	Nonsyn	Somatic
6	CRIM1	CCDS1783.1	2	36,764,511	G	C	K815N	Nonsyn	Somatic
6	CSRNP2	CCDS8807.1	12	51,470,312	C	G	R11S	Nonsyn	Somatic

6	DENND2D	CCDS831.1	1	111,730,865	G	C	F409L	Nonsyn	Somatic
6	DHX38	CCDS10907.1	16	72130158	C	T	-	syn	somatic
6	DSTYK	CCDS1451.1	1	205,131,298	G	C	L562V	Nonsyn	Somatic
6	EIF3A	CCDS7608.1	10	120,832,957	C	T	E125K	Nonsyn	Somatic
6	ELP3	CCDS6065.1	8	27,995,322	G	C	K338N	Nonsyn	Somatic
6	FAM92A1	CCDS47892.1	8	94,730,957	G	A	E200K	Nonsyn	Somatic
6	GPR27	CCDS2915.1	3	71803239	C	T	-	syn	somatic
6	GRIN1	CCDS7031.1	9	140061961	C	T	-	syn	somatic
6	HBG2	ENST00000380247	11	5,274,546	C	T	-	5_prime_splice	Somatic
6	HIST1H2AB	CCDS4574.1	6	26033599	C	T	-	syn	somatic
6	HNRNPH1	CCDS4446.1	5	179,044,114	G	A	-	3_prime_splice	Somatic
6	HNRNPK	CCDS6668.1	9	86,585,156	C	G	E428Q	Nonsyn	Somatic
6	JAK1	CCDS41346.1	1	65,304,210	G	T	P969T	Nonsyn	Somatic
6	KCNN4	CCDS12630.1	19	44271761	C	G	-	syn	somatic
6	KIAA1217	CCDS31165.1	10	24,822,087	C	T	P1112L	Nonsyn	Somatic
6	KISS1R	CCDS12049.1	19	920,612	C	G	S354W	Nonsyn	Somatic
6	LIPT2	CCDS44679.1	11	74204440	G	C	-	syn	somatic
6	MAMSTR	CCDS46137.1	19	49,216,598	G	T	P392T	Nonsyn	Somatic
6	MEGF6	CCDS41237.1	1	3,414,991	G	A	P1099L	Nonsyn	LOH
6	MFSD6L	CCDS11146.1	17	8701857	C	T	-	syn	somatic
6	MLXIPL	CCDS47605.1	7	73008169	G	A	-	syn	somatic
6	MN1	CCDS42998.1	22	28,193,988	G	T	F848L	Nonsyn	Somatic
6	MRFAP1	CCDS3389.1	4	6642610	C	T	-	syn	somatic
6	MS4A5	CCDS7987.1	11	60,201,278	G	C	G127A	Nonsyn	Somatic
6	MYL10	CCDS34713.1	7	101,256,856	G	C	-	3_prime_splice	Somatic
6	NCAPH	CCDS2021.1	2	97,001,577	C	T	P4S	Nonsyn	Somatic
6	OR2M1P	NR_002141	1	248,285,643	G	C	R69T	Nonsyn	Somatic
6	OR4P1P	ENST00000345013	11	55,451,658	C	G	F224L	Nonsyn	Somatic
6	OR4X2	CCDS31486.1	11	48,266,725	C	G	L24V	Nonsyn	Somatic
6	PAMR1	CCDS31460.1	11	35,456,055	C	G	-	5_prime_splice	Somatic

6	PCYT2	CCDS11791.1	17	79,863,543	C	T	-	5_prime_splice	Somatic
6	PHRF1	CCDS44507.1	11	608,155	C	T	S899F	Nonsyn	Somatic
6	POU6F2	CCDS34620.2	7	39,504,088	G	C	E627Q	Nonsyn	Somatic
6	PRUNE2	CCDS47982.1	9	79,325,581	C	A	E537X	Nonsyn	Somatic
6	PRUNE2	CCDS47982.1	9	79,325,770	C	A	E474X	Nonsyn	Somatic
6	PRUNE2	CCDS47982.1	9	79324907	C	T	-	syn	somatic
6	PTPRO	CCDS8675.1	12	15,673,198	G	A	D615N	Nonsyn	Somatic
6	SI	CCDS3196.1	3	164,758,725	C	T	-	5_prime_splice	Somatic
6	SIRPB1	CCDS46571.1	20	1592343	T	C	-	syn	somatic
6	SLC12A4	CCDS10855.1	16	67,995,569	C	T	G84E	Nonsyn	Somatic
6	SLC13A2	CCDS11231.1	17	26,817,855	G	A	E169K	Nonsyn	Somatic
6	SLFN11	CCDS11294.1	17	33,690,091	G	T	L246I	Nonsyn	Somatic
6	SLFNL1	CCDS460.1	1	41,486,295	G	C	S13X	Nonsyn	Somatic
6	SOX6	CCDS7821.1	11	16,362,637	G	C	S53C	Nonsyn	Somatic
6	SVEP1	CCDS48004.1	9	113,205,913	C	T	W1517X	Nonsyn	Somatic
6	TBC1D2B	CCDS32301.2	15	78369758	G	A	-	syn	somatic
6	TIMM22	CCDS32521.1	17	900427	A	G	-	syn	somatic
6	TM7SF3	CCDS8710.1	12	27148212	C	T	-	syn	somatic
6	TMEM22	CCDS3091.1	3	136,574,232	C	G	I310M	Nonsyn	Somatic
6	TMEM22	CCDS3091.1	3	136,574,392	C	G	L364V	Nonsyn	Somatic
6	TNFSF12-TNFSF13	CCDS11108.1	17	7452608	C	T	-	syn	somatic
6	TNS3	CCDS5506.2	7	47,407,959	C	T	-	5_prime_splice	Somatic
6	TRMT1L	CCDS1366.1	1	185,109,122	C	G	L364F	Nonsyn	Somatic
6	TSPYL6	NM_001003937	2	54,482,745	C	T	E182K	Nonsyn	Somatic
6	UBR7	CCDS9909.1	14	93673613	C	T	-	syn	somatic
6	USP44	CCDS9053.1	12	95927088	G	A	-	syn	somatic
6	YLPM1	CCDS45135.1	14	75,264,714	C	G	S905C	Nonsyn	Somatic
6	ZFP30	CCDS33005.1	19	38,126,286	C	G	E386Q	Nonsyn	Somatic
6	ZNF789	CCDS34693.1	7	99,084,563	C	T	L244F	Nonsyn	Somatic

6	ZNF860	CCDS46784.1	3	32,031,109	G	C	E180Q	Nonsyn	Somatic
7a	ALS2CR4	ENST00000426684	2	202,507,419	C	A	D9Y	Nonsyn	Somatic
7a	ATP6V0A2	CCDS9254.1	12	124,221,708	C	T	H310Y	Nonsyn	Somatic
7a	BNC2	CCDS6482.2	9	16583078	G	A	-	syn	somatic
7a	C20orf108	CCDS13450.1	20	54934067	C	T	-	syn	somatic
7a	CCNT2	CCDS2174.1	2	135,712,066	C	G	L681V	Nonsyn	Somatic
7a	CCRN4L	CCDS3743.1	4	139937271	C	T	-	syn	somatic
7a	CDC73	CCDS1382.1	1	193,099,350	T	C	L95P	Nonsyn	Somatic
7a	FAM83A	CCDS6340.1	8	124,195,265	G	A	E57K	Nonsyn	Somatic
7a	KIAA1409	CCDS9911.2	14	94,156,540	C	A	A2250E	Nonsyn	Somatic
7a	NEGR1	CCDS661.1	1	72,058,506	G	C	L312V	Nonsyn	Somatic
7a	OTOP1	CCDS3372.1	4	4,198,994	C	T	G523R	Nonsyn	Somatic
7a	SYNC	CCDS367.2	1	33,160,780	C	T	E307K	Nonsyn	Somatic
7a	TMEM200A	CCDS5140.1	6	130762743	G	A	-	syn	somatic
7a	VPS13B	CCDS6280.1	8	100,654,571	G	A	R1943Q	Nonsyn	Somatic
7a	ZNF546	CCDS12548.1	19	40513216	C	G	-	syn	somatic
7b	ABCA1	CCDS6762.1	9	107,547,923	G	A	-	3_prime_splice	Somatic
7b	ABCA2	CCDS43909.1	9	139916884	G	C	-	syn	somatic
7b	AC010872.2	ENST00000405799	2	21,364,830	G	C	M1497I	Nonsyn	Somatic
7b	ACADVL	CCDS11090.1	17	7,126,554	G	A	E394K	Nonsyn	Somatic
7b	ACCN3	CCDS5914.1	7	150747892	G	A	-	syn	somatic
7b	ACO1	CCDS6525.1	9	32,419,050	G	C	E225Q	Nonsyn	Somatic
7b	ADCK1	CCDS9869.1	14	78,399,608	C	G	I482M	Nonsyn	Somatic
7b	AFF4	CCDS4164.1	5	132,232,935	G	A	-	3_prime_splice	Somatic
7b	AKR1E2	CCDS31134.1	10	4888038	G	C	-	syn	somatic
7b	ALKBH1	CCDS32127.1	14	78,174,215	C	T	D45N	Nonsyn	Somatic
7b	ANAPC7	CCDS9145.2	12	110,811,950	C	G	VARIANT LONGER ORF	Nonsyn	Somatic
7b	ANKRD11	CCDS32513.1	16	89,346,755	G	C	F2065L	Nonsyn	Somatic

7b	ARID5B	CCDS31208.1	10	63,759,864	G	A	E173K	Nonsyn	Somatic
7b	ARSJ	CCDS43264.1	4	114,824,677	C	G	E185Q	Nonsyn	Somatic
7b	BCDIN3D	CCDS8790.1	12	50236666	G	A	-	syn	somatic
7b	BEND5	CCDS552.2	1	49,227,062	C	G	D103H	Nonsyn	Somatic
7b	BIRC2	CCDS8316.1	11	102,248,458	C	G	S533C	Nonsyn	Somatic
7b	C12orf26	CCDS9024.1	12	82,792,817	G	A	E259K	Nonsyn	Somatic
7b	C12orf51	CCDS44978.1	12	112,666,568	G	C	I1767M	Nonsyn	Somatic
7b	C16orf91	CCDS32360.1	16	1,470,575	C	A	R181L	Nonsyn	Somatic
7b	C1orf112	CCDS1285.1	1	169,806,228	G	A	R567H	Nonsyn	LOH
7b	C1orf201	CCDS253.1	1	24,710,419	C	G	K41N	Nonsyn	Somatic
7b	C7orf30	CCDS5381.1	7	23,340,484	G	A	D95N	Nonsyn	Somatic
7b	C8orf33	CCDS34974.1	8	146277850	C	T	-	syn	somatic
7b	CA5B	ENST00000474624	x	15,768,291	G	A	E49K	Nonsyn	Somatic
7b	CACNA1A	CCDS45998.1	19	13441104	G	C	-	syn	somatic
7b	CAPN8	ENST00000423927	1	223,718,170	C	T	D626N	Nonsyn	LOH
7b	CAT	CCDS7891.1	11	34,470,845	C	G	T58S	Nonsyn	Somatic
7b	CC2D2B	CCDS41555.1	10	97,779,514	C	G	S238X	Nonsyn	Somatic
7b	CCT7	CCDS42696.1	2	73461463	C	T	-	syn	somatic
7b	CDKN2AIP	CCDS34110.1	4	184,368,087	C	G	S417X	Nonsyn	Somatic
7b	CFH	CCDS1385.1	1	196,646,659	G	T	A161S	Nonsyn	LOH
7b	CHAF1A	CCDS32875.1	19	4,432,164	G	C	Q721H	Nonsyn	Somatic
7b	CHIT1	CCDS1436.1	1	203,186,979	C	A	K348N	Nonsyn	LOH
7b	CHTF18	CCDS45371.1	16	840615	C	G	-	syn	somatic
7b	CLEC14A	CCDS9667.1	14	38724514	G	A	-	syn	somatic
7b	CMAH	NR_002174	6	25,109,727	C	T	-	5_prime_splice	Somatic
7b	CMPK1	CCDS44135.1	1	47799689	G	C	-	syn	somatic
7b	COG2	CCDS1584.1	1	230827187	G	C	-	syn	somatic
7b	COL12A1	CCDS43482.1	6	75,904,651	G	C	S29X	Nonsyn	Somatic
7b	CRB1	CCDS1390.1	1	197,298,095	T	C	I205T	Nonsyn	LOH
7b	CSNK1D	CCDS11805.1	17	80,213,362	G	T	F93L	Nonsyn	Somatic

7b	CTNNA2	CCDS42703.2	2	79,971,603	G	C	E65Q	Nonsyn	Somatic
7b	CTSL2	CCDS6723.1	9	99,798,900	G	C	Q176E	Nonsyn	Somatic
7b	CTSL2	CCDS6723.1	9	99,798,917	G	A	S170L	Nonsyn	Somatic
7b	DEF6	CCDS4802.1	6	35,287,615	G	C	E468Q	Nonsyn	Somatic
7b	DIS3L2	ENST00000273009	2	233,208,145	C	T	R558C	Nonsyn	Somatic
7b	ECM1	CCDS953.1	1	150485832	G	A	-	syn	somatic
7b	EHD3	CCDS1774.1	2	31,489,100	G	C	E380Q	Nonsyn	Somatic
7b	EIF5	CCDS9980.1	14	103,803,037	G	A	E60K	Nonsyn	Somatic
7b	EMILIN2	CCDS11828.1	18	2,847,807	G	A	-	3_prime_splice	Somatic
7b	EMILIN2	CCDS11828.1	18	2847807	G	A	-	syn	somatic
7b	ERBB3	CCDS31833.1	12	56,495,772	C	G	S1321C	Nonsyn	Somatic
7b	ERBB3	CCDS31833.1	12	56495539	C	G	-	syn	somatic
7b	ERGIC1	CCDS34292.1	5	172324039	C	G	-	syn	somatic
7b	EWSR1	CCDS13851.1	22	29,692,363	G	C	-	5_prime_splice	Somatic
7b	FAM174A	CCDS4090.1	5	99,871,346	G	C	E38Q	Nonsyn	Somatic
7b	FAM75D5	NR_026851	9	84,530,026	G	C	-	5_prime_splice	Somatic
7b	FARSA	CCDS12287.1	19	13044529	C	A	-	syn	somatic
7b	FCGBP	CCDS12546.1	19	40354090	C	G	-	syn	somatic
7b	FCRL5	CCDS1165.1	1	157,516,861	C	T	R60Q	Nonsyn	LOH
7b	FER1L4	ENST00000454891	20	34,147,293	G	T	L1901M	Nonsyn	Somatic
7b	G6PD	CCDS14756.2	x	153761878	C	G	-	syn	somatic
7b	GAS2L2	CCDS11298.1	17	34074956	C	T	-	syn	somatic
7b	GATM	CCDS10122.1	15	45,668,858	C	A	E77X	Nonsyn	Somatic
7b	GCC2	CCDS33268.1	2	109,085,538	G	A	E107K	Nonsyn	Somatic
7b	GJA10	CCDS5025.1	6	90,604,959	C	G	P258A	Nonsyn	Somatic
7b	GPC5	CCDS9468.1	13	92,560,221	C	G	I437M	Nonsyn	Somatic
7b	GUCA1C	ENST00000393963	3	108,634,962	T	C	T152A	Nonsyn	LOH
7b	GUK1	CCDS1568.1	1	228334534	C	T	-	syn	somatic
7b	HEATR7B2	CCDS47202.1	5	41,047,825	G	A	Q576X	Nonsyn	Somatic
7b	HEPACAM2	CCDS43616.1	7	92848846	C	G	-	syn	somatic

7b	HIST1H1T	CCDS34349.1	6	26108124	G	C	-	syn	somatic
7b	HRAS	CCDS7698.1	11	534,209	C	A	-	5_prime_splice	Somatic
7b	HSPA5	CCDS6863.1	9	128003012	G	A	-	syn	somatic
7b	IFIH1	CCDS2217.1	2	163,123,875	C	G	M971I	Nonsyn	Somatic
7b	IGSF1	CCDS14629.1	x	130,409,661	C	T	R992Q	Nonsyn	Somatic
7b	IL18RAP	CCDS2061.1	2	103039733	G	A	-	syn	somatic
7b	INF2	CCDS9989.2	14	105,178,858	G	A	E860K	Nonsyn	Somatic
7b	INSC	CCDS41621.1	11	15260577	C	G	-	syn	somatic
7b	ITPR3	CCDS4783.1	6	33,662,698	C	G	-	3_prime_splice	Somatic
7b	KCNH1	CCDS1496.1	1	211,093,016	C	G	E476D	Nonsyn	Somatic
7b	KIAA0391	CCDS32063.1	14	35,592,714	G	A	G88E	Nonsyn	Somatic
7b	KIAA1377	CCDS31658.1	11	101,832,942	G	A	M392I	Nonsyn	Somatic
7b	KIAA1522	CCDS41298.1	1	33236208	C	A	-	syn	somatic
7b	KIF14	CCDS30963.1	1	200,534,786	T	C	K1225E	Nonsyn	LOH
7b	KRTAP3-3	CCDS32643.1	17	39150110	G	C	-	syn	somatic
7b	LARP4B	ENST00000263154	10	860,492	G	C	L707V	Nonsyn	Somatic
7b	LOH12CR1	CCDS8649.1	12	12,514,137	C	T	-	3_prime_splice	Somatic
7b	LOH12CR1	CCDS8649.1	12	12,514,145	C	T	P22S	Nonsyn	Somatic
7b	LRP2BP	CCDS3840.1	4	186,298,129	C	G	L50F	Nonsyn	Somatic
7b	LRRC16B	CCDS32054.1	14	24,523,651	C	T	S98L	Nonsyn	Somatic
7b	LUC7L2	CCDS43656.1	7	139,060,867	C	G	L41V	Nonsyn	Somatic
7b	MAP3K11	CCDS8107.1	11	65,375,255	C	G	D368H	Nonsyn	Somatic
7b	MCTP1	CCDS34203.1	5	94,248,628	C	G	W468C	Nonsyn	Somatic
7b	MFAP4	CCDS11208.1	17	19290360	C	T	-	syn	somatic
7b	MLEC	CCDS9206.1	12	121,125,332	G	A	R78Q	Nonsyn	Somatic
7b	MPI	CCDS10272.1	15	75,185,047	G	A	D131N	Nonsyn	Somatic
7b	MTSS1	CCDS6353.1	8	125565284	C	T	-	syn	somatic
7b	MYOC	CCDS1297.1	1	171,621,507	C	T	R82H	Nonsyn	LOH
7b	NADK	CCDS30565.1	1	1686019	C	A	-	syn	somatic
7b	NEB	CCDS46424.1	2	152,382,764	C	T	E5619K	Nonsyn	Somatic

7b	NFATC1	CCDS32850.1	18	77,170,463	C	G	S50C	Nonsyn	Somatic
7b	NFATC3	CCDS10860.1	16	68,225,541	C	T	S990L	Nonsyn	Somatic
7b	NFE2L1	CCDS11524.1	17	46136826	C	T	-	syn	somatic
7b	NKX2-6	ENST00000325017	8	23,563,868	C	G	E82Q	Nonsyn	Somatic
7b	OBSCN	ENST00000422127	1	228,467,119	G	A	R2457H	Nonsyn	LOH
7b	OBSCN	ENST00000366707	1	228,486,175	C	G	Q3894E	Nonsyn	Somatic
7b	OGT	CCDS14414.1	x	70,767,862	C	T	H213Y	Nonsyn	Somatic
7b	OR13C8	CCDS35090.1	9	107,332,253	G	A	D269N	Nonsyn	Somatic
7b	OR5AK3P	ENST00000326876	11	56,739,205	G	C	K227N	Nonsyn	Somatic
7b	OTOA	CCDS10600.2	16	21,698,917	G	C	D195H	Nonsyn	Somatic
7b	OXR1	CCDS47909.1	8	107,719,358	G	A	D537N	Nonsyn	Somatic
7b	OXR1	ENST00000497705	8	107,719,394	G	A	E482K	Nonsyn	Somatic
7b	PARP1	CCDS1554.1	1	226,561,965	C	G	D678H	Nonsyn	Somatic
7b	PITX1	CCDS4182.1	5	134,364,972	C	G	E148Q	Nonsyn	Somatic
7b	PKDREJ	CCDS14073.1	22	46,655,007	T	C	I1405V	Nonsyn	Somatic
7b	PLEKHA5	NM_001143821	12	19,522,730	G	A	-	5_prime_splice	Somatic
7b	PLEKHA5	CCDS44840.1	12	19522730	G	A	-	syn	somatic
7b	POLH	CCDS4902.1	6	43,550,805	G	A	D67N	Nonsyn	Somatic
7b	POLL	CCDS7513.1	10	103,344,496	G	C	H252D	Nonsyn	Somatic
7b	PPFIA1	CCDS31627.1	11	70,181,739	C	T	S456L	Nonsyn	Somatic
7b	PRDM5	CCDS3716.1	4	121,706,237	C	T	E400K	Nonsyn	Somatic
7b	PRM1	CCDS10547.1	16	11,375,025	C	T	R24K	Nonsyn	Somatic
7b	PSD	CCDS31272.1	10	104172320	C	T	-	syn	somatic
7b	PTPN14	CCDS1514.1	1	214,551,439	C	G	E851Q	Nonsyn	Somatic
7b	PTX4	CCDS32362.1	16	1,536,214	C	T	R383H	Nonsyn	Somatic
7b	RAB11FIP3	CCDS32351.1	16	476,726	G	A	-	5_prime_splice	Somatic
7b	RABL5	CCDS5719.1	7	100,959,650	G	A	S127F	Nonsyn	Somatic
7b	RBM26	CCDS9462.1	13	79,911,368	G	A	H841Y	Nonsyn	Somatic
7b	REPS1	CCDS47488.1	6	139251062	C	G	-	syn	somatic
7b	RGL3	CCDS12260.1	19	11493954	C	T	-	syn	somatic

7b	RP11-339B21.9	CCDS6902.1	9	131231430	G	A	-	syn	somatic
7b	RP11-464E15.4	ENST00000492461	3	148,804,124	G	A	M1I	Nonsyn	Somatic
7b	RPAP1	CCDS10079.1	15	41,809,998	C	T	E1344K	Nonsyn	Somatic
7b	RRM2B	CCDS34932.1	8	103,251,081	C	T	E8K	Nonsyn	Somatic
7b	SEMA4G	CCDS7501.1	10	102,739,648	G	A	E343K	Nonsyn	Somatic
7b	SF3A1	CCDS13875.1	22	30752746	C	T	-	syn	somatic
7b	SLC13A3	CCDS13400.1	20	45,194,865	C	T	-	5_prime_splice	Somatic
7b	SLC28A1	CCDS10334.1	15	85461799	C	T	-	syn	somatic
7b	SLC45A1	CCDS30577.1	1	8,398,033	C	A	F585L	Nonsyn	Somatic
7b	SNTG1	CCDS6147.1	8	51,449,319	C	G	L211V	Nonsyn	Somatic
7b	STAB1	CCDS33768.1	3	52548151	G	A	-	syn	somatic
7b	SYNCRIP	CCDS5005.1	6	86,351,076	G	C	Q28E	Nonsyn	Somatic
7b	TAF7L	CCDS35347.1	x	100,538,569	C	G	D136H	Nonsyn	Somatic
7b	TBL1XR1	CCDS46961.1	3	176755919	G	C	-	syn	somatic
7b	TDRD5	CCDS1332.1	1	179,609,549	C	T	P590L	Nonsyn	LOH
7b	TFCP2L1	CCDS2134.1	2	122,005,759	C	T	R162K	Nonsyn	Somatic
7b	THBS3	ENST00000469769	1	155,167,201	G	C	S44C	Nonsyn	Somatic
7b	TINF2	CCDS41936.1	14	24708961	G	C	-	syn	somatic
7b	TJP1	CCDS42007.1	15	30,010,985	C	T	E1121K	Nonsyn	Somatic
7b	TMBIM6	CCDS31797.1	12	50149423	C	T	-	syn	somatic
7b	TMEM87A	CCDS32205.1	15	42,503,939	C	T	M545I	Nonsyn	Somatic
7b	TMEM87B	CCDS33275.1	2	112,873,674	G	A	R541K	Nonsyn	Somatic
7b	TNKS1BP1	CCDS7951.1	11	57,070,020	C	G	W1532C	Nonsyn	Somatic
7b	TOR1A	CCDS6930.1	9	132586374	C	T	-	syn	somatic
7b	TP53BP1	CCDS45250.1	15	43,766,934	C	T	D373N	Nonsyn	Somatic
7b	TRPS1	CCDS6318.2	8	116,631,639	G	A	S229F	Nonsyn	Somatic
7b	TSGA10	CCDS2037.1	2	99,722,065	C	G	E102D	Nonsyn	Somatic
7b	TTC24	NM_001105669	1	156,551,248	G	A	R31Q	Nonsyn	LOH

7b	TULP3	CCDS8519.1	12	3,040,236	C	G	Q176E	Nonsyn	Somatic
7b	TULP4	CCDS34561.1	6	158,924,926	G	A	E1411K	Nonsyn	Somatic
7b	USP11	CCDS14277.1	x	47,101,007	C	G	S406X	Nonsyn	Somatic
7b	USP24	CCDS44154.1	1	55587145	G	A	-	syn	somatic
7b	VPS13D	CCDS30588.1	1	12,316,558	C	G	Q280E	Nonsyn	Somatic
7b	VPS37B	CCDS9239.1	12	123380553	G	A	-	syn	somatic
7b	WNK2	NM_006648	9	96,082,654	G	A	E2212K	Nonsyn	Somatic
7b	XRRA1	CCDS44680.1	11	74641382	C	T	-	syn	somatic
7b	ZCCHC6	NM_001185059	9	88,968,148	G	C	-	3_prime_splice	Somatic
7b	ZCCHC7	CCDS6608.2	9	37,305,714	G	C	-	5_prime_splice	Somatic
7b	ZIC1	CCDS3136.1	3	147,128,516	C	T	A206V	Nonsyn	Somatic
7b	ZNF271	NR_024565	18	32,887,775	G	C	R396T	Nonsyn	Somatic
7b	ZNF557	CCDS42485.1	19	7,083,276	G	T	E272X	Nonsyn	Somatic
7b	ZP2	CCDS10596.1	16	21,213,103	G	C	I476M	Nonsyn	Somatic
7b	ZSCAN5B	CCDS46203.1	19	56,704,274	A	G	F50L	Nonsyn	Somatic
7b	ZZEF1	CCDS11043.1	17	3,922,965	C	G	K2501N	Nonsyn	Somatic
8	GPSM1	CCDS48055.1	9	139251002	G	A	-	syn	somatic
8	PAQR4	CCDS10485.1	16	3019783	G	C	-	syn	somatic
8	SP8	CCDS43555.1	7	20823972	C	G	-	syn	somatic

	Validated SNVs		
Sample	**heterozygous**	**homozygous**	**Total**
1	30	0	30
2	65	1	66
3	3	1	4
4	14	1	15
5	20	0	20
6	71	2	73
7a	14	1	15
7b	164	0	164
8	3	0	3
TOTAL	**384**	**6**	**390**

Table 4.6: Zygosity summary of validated somatic mutations for whole-exome sequenced PC samples.

	CDC73					PRUNE2				
Sample	**exon**	**Mutation**	**Predicted effect**	**LOH of wildtype**	**mutant CN gain**	**exon**	**Mutation**	**Predicted effect**	**LOH of wildtype**	**mutant CN gain**
1	2	c.165delC (s)	p.Tyr55X	YES	5 copies			-		
2	4	c.356delA (g)	p.Gln119ArgfsX14	YES	3 copies			-		
3	1	c.30delG (g)	p.Gln10HisfsX11	NO	NO			-		
4			-			8	c.1354G>A(g)	p.Val452Met	YES	NO
5			-					-		
6	1	c.32delA (s)*	p.Tyr11SerfsX10	NO	4 copies	8	c.1609G>T (s)	p.Glu537X	NO	NO
						8	c.1420G>T (s)	p.Glu474X	NO	NO
7a	3	c.284T>C (s)	p.Leu95Pro	NO	NO			-		
	4	c.356delA (g)	p.Gln119ArgfsX14	NO	NO					
7b	4	c.356delA (g)	p.Gln119ArgfsX14	YES	5 copies			-		
8	7	c.539-544insA(g)	p.Ile182AsnfsX10	NO	NO			-		

Table 4.7: Recurrent mutations in *CDC73* and *PRUNE2* for whole-exome sequenced PC. (s) = somatic mutation, (g) = germline mutation, * = patient has a germline *CDC73* whole-gene deletion (see table3.1 for relevant reference)

Gene Name	AA change	Gene Description	Polyphen prediction	SIFT prediction
PARP1	D678H	poly (ADP-ribose) polymerase 1	probably damaging	Damaging
POLH	D67N	polymerase (DNA directed), eta	possibly damaging	Damaging
TP53BP1	D373N	tumor protein p53 binding protein 1	probably damaging	Tolerated
POLL	H252D	polymerase (DNA directed), lambda	benign	Tolerated
RRM2B	E8K	ribonucleotide reductase M2 B (TP53 inducible)	benign	Tolerated
CHAF1A	Q721H	chromatin assembly factor 1, subunit A (p150)	benign	Tolerated

Table 4.8: Mutated genes related to DNA damage repair in sample 7b.

Table 4.9: Gene classification analysis of validated somatic mutations in PC. Table below shows classification terms with P value < 0.01 as outputted by DAVID's gene classification analysis package for seven pairs and one triplet of PC samples. Classification terms related to kinases and the mutated genes corresponding to those terms are highlighted (dark grey highlight)

Classification Term	P value	RIOK3	CDC42BPA	FRK	TIE1	ALPK1	PANK4	LIMK2	LTK	JAK1	DSTYK	MAP3K11	ADCK1	CSNK1D	WNK2	KIF14	MOV10L1	AKAP9	HSPA6	SLFNL1	SLFN11	SCYL3	DDX31	EHD3	ACSM5
atp-binding	7.68E-23																								
GO:0005524~ATP binding	1.67E-22																								
GO:0032559~adenyl ribonucleotide binding	2.28E-22																								
GO:0030554~adenyl nucleotide binding	7.60E-22																								
GO:0001883~purine nucleoside binding	1.08E-21																								
GO:0001882~nucleoside binding	1.26E-21																								
nucleotide-binding	1.18E-20																								
GO:0032555~purine ribonucleotide binding	2.57E-20																								
GO:0032553~ribonucleotide binding	2.57E-20																								
GO:0017076~purine nucleotide binding	7.06E-20																								

Classification Term	P value	RIOK3	CDC42BPA	FRK	TIE1	ALPK1	PANK4	LIMK2	LTK	JAK1	DSTYK	MAP3K11	ADCK1	CSNK1D	WNK2	KIF14	MOV10L1	AKAP9	HSPA6	SLFNL1	SLFN11	SCYL3	DDX31	EHD3	ACSM5
nucleotide phosphate-binding region:ATP	9.64E-20																								
GO:0000166~nucleotide binding	2.69E-18																								
kinase	3.03E-15																								
GO:0006468~protein amino acid phosphorylation	4.11E-15																								
binding site:ATP	5.86E-15																								
GO:0016310~phosphorylation	5.05E-14																								
GO:0004672~protein kinase activity	1.11E-13																								
GO:0006796~phosphate metabolic process	7.50E-13																								
GO:0006793~phosphorus metabolic process	7.50E-13																								
domain:Protein kinase	1.79E-12																								
IPR017441:Protein kinase, ATP binding site	3.01E-12																								
IPR000719:Protein kinase, core	4.91E-12																								

Classification Term	P value	RIOK3	CDC42BPA	FRK	TIE1	ALPK1	PANK4	LIMK2	LTK	JAK1	DSTYK	MAP3K11	ADCK1	CSNK1D	WNK2	KIF14	MOV10L1	AKAP9	HSPA6	SLFNL1	SLFN11	SCYL3	DDX31	EHD3	ACSM5
serine/threonine-protein kinase	2.73E-10																								
transferase	8.29E-10																								
active site:Proton acceptor	1.67E-09																								
GO:0004674~protein serine/threonine kinase activity	2.39E-08																								
IPR008271:Serine/threonine protein kinase, active site	4.94E-06																								
IPR017442:Serine/threonine protein kinase-related	5.35E-06																								
SM00219:TyrKc	8.34E-06																								
tyrosine-protein kinase	8.55E-06																								
IPR001245:Tyrosine protein kinase	1.59E-05																								
ATP	1.63E-04																								
GO:0004713~protein tyrosine kinase activity	1.89E-04																								

Classification Term	P value	RIOK3	CDC42BPA	FRK	TIE1	ALPK1	PANK4	LIMK2	LTK	JAK1	DSTYK	MAP3K11	ADCK1	CSNK1D	WNK2	KIF14	MOV10L1	AKAP9	HSPA6	SLFNL1	SLFN11	SCYL3	DDX31	EHID3	ACSM5
IPR008266:Tyrosine protein kinase, active site	2.72E-04																								
tyrosine-specific protein kinase	1.49E-03																								
phosphotransferase	1.71E-03																								
phosphoprotein	6.95E-03																								
IPR007421:ATPase associated with various cellular activities	9.21E-03																								

Sample	Gene Symbol	Chr	Pos	Ref	Cons	Mutation status	AA Change	Gene Description
1	**RIOK3**	18	21,044,569	A	G	Germline*	K174E	RIO kinase 3 (yeast)
	ALPK1	4	113,345,145	A	G	Somatic	N174S	alpha-kinase 1
2	**CDC42BPA**	1	227,216,756	C	T	Germline*	R1310H	CDC42 binding protein kinase alpha (DMPK-like)
	TIE1	1	43,774,729	G	A	Somatic	C372Y	tyrosine kinase with IG-like and EGF-like domains 1
	FRK	6	116,265,579	C	A	Somatic	G323V	fyn-related kinase
4	**PANK4**	1	2,451,796	A	G	Germline*	F222L	pantothenate kinase 4
	LIMK2	22	31,674,324	C	G	Germline*	S605C	LIM domain kinase 2
	RIOK3	18	21,053,547	A	G	Somatic*	I324V	RIO kinase 3 (yeast)
5	**LTK**	15	41,797,670	G	A	Germline*	R586C	leukocyte receptor tyrosine kinase
6	**JAK1**	1	65,304,210	G	T	Somatic	P969T	Janus kinase 1
	DSTYK	1	205,131,298	G	C	Somatic	L562V	dual serine/threonine and tyrosine protein kinase
7b	**MAP3K11**	11	65,375,255	C	G	Somatic	D368H	mitogen-activated protein kinase kinase kinase 11
	ADCK1	14	78,399,608	C	G	Somatic	I482M	aarF domain containing kinase 1
	CSNK1D	17	80,213,362	G	T	Somatic	F93L	casein kinase 1, delta
	WNK2	9	96,082,654	G	A	Somatic	E2212K	WNK lysine deficient protein kinase 2

Table 4.10: Kinase mutations in PC. Dark grey highlight indicates mutation is predicted to be deleterious by both POLYPHEN and SIFT; light grey highlights indicates mutation is predicted to be deleterious by POLYPHEN only; no highlight indicates mutation is predicted to be benign by both POLYPHEN and SIFT. For mutation status column, Germline* status indicates mutation is validated to be heterozygous in the normal and homozygous in the tumor; Somatic status indicates mutation is validated to be wildtype in the normal with the mutation being validated to be heterozygous in the tumor; Somatic* status indicates mutation is validated to be wildtype in the normal with the mutation validated to be homozygous in the tumor.

Table 4.11: Gene classification analysis of validated somatic mutations in PC excluding sample 7b. Table below shows classification terms with P value < 0.01 as outputted by DAVID's gene classification analysis package for all whole-exome PC samples excluding sample 7b. Classification terms related to kinases and the mutated genes corresponding to those terms are highlighted (dark grey highlight)

Classification Term	P value	RIOK3	CDC42BPA	FRK	TIE1	ALPK1	PANK4	LIMK2	LTK	JAK1	DSTYK	AKAP9	HSPA6	SLFNL1	SLFN11	SCYL3
GO:0005524~ATP binding	5.76E-14															
GO:0032559~adenyl ribonucleotide binding	6.96E-14															
GO:0030554~adenyl nucleotide binding	1.45E-13															
GO:0001883~purine nucleoside binding	1.79E-13															
GO:0001882~nucleoside binding	1.97E-13															
atp-binding	8.78E-13															
GO:0032555~purine ribonucleotide binding	1.23E-12															
GO:0032553~ribonucleotide binding	1.23E-12															
GO:0017076~purine nucleotide binding	2.26E-12															
kinase	2.83E-12															
GO:0006468~protein amino acid phosphorylation	4.80E-12															
nucleotide-binding	1.53E-11															

Classification Term	P value	RIOK3	CDC42BPA	FRK	TIE1	ALPK1	PANK4	LIMK2	LTK	JAK1	DSTYK	AKAP9	HSPA6	SLFNL1	SLFN11	SCYL3
GO:0000166~nucleotide binding	2.07E-11															
GO:0016310~phosphorylation	2.94E-11															
GO:0004672~protein kinase activity	3.86E-11															
GO:0006796~phosphate metabolic process	2.05E-10															
GO:0006793~phosphorus metabolic process	2.05E-10															
IPR000719:Protein kinase, core	1.08E-09															
nucleotide phosphate-binding region:ATP	3.18E-09															
domain:Protein kinase	1.52E-08															
IPR017441:Protein kinase, ATP binding site	3.15E-08															
binding site:ATP	4.10E-08															
transferase	7.71E-08															
tyrosine-protein kinase	1.01E-06															
active site:Proton acceptor	3.86E-06															
serine/threonine-protein kinase	5.13E-06															

Classification Term	P value	RIOK3	CDC42BPA	FRK	TIE1	ALPK1	PANK4	LIMK2	LTK	JAK1	DSTYK	AKAP9	HSPA6	SLFNL1	SLFN11	SCYL3
GO:0004713~protein tyrosine kinase activity	2.33E-05															
GO:0004674~protein serine/threonine kinase activity	6.09E-05															
IPR008266:Tyrosine protein kinase, active site	6.65E-05															
SM00219:TyrKc	7.21E-05															
IPR001245:Tyrosine protein kinase	1.19E-04															
tyrosine-specific protein kinase	5.44E-04															
ATP	6.01E-04															
IPR008271:Serine/threonine protein kinase, active site	2.91E-03															
IPR017442:Serine/threonine protein kinase-related	3.03E-03															
IPR007421:ATPase associated with various cellular activities, AAA-4	5.87E-03															
phosphotransferase	9.10E-03															

Chapter Five: General Discussion and Future Work

5.1: General Discussion

There are several parallel themes that arose from the three papers I have presented as the body of my thesis; on the surface, there is a line of continuity in the basic methodology of these three studies. The common theme is the use of whole-exome capturing followed by massively parallel sequencing to obtain base level resolution data of the protein coding regions. The data is then aligned to the reference human genome to arrive at each sequenced base's chromosomal location. Based on each base's mapped location, a comparison can be made to determine concordance with the reference genome and to generate a list of variants. With the sequencing of matched tumor-normal pairs, the list of variants generated by the tumor can be compared with the list of variants generated by the corresponding normal tissue; The two lists of variants generated by this comparison: one containing the somatic elements, positions of variant nucleotides present in the tumor but not in the normal, and one containing the germline elements, positions of variant nucleotides present in both tumor and normal, constitute the fundamental starting point for all of the studies' analysis.

For the somatic elements, the analysis of nonsynonymous variants enables the identification of novel putative 'driver' mutations and prioritizes downstream molecular studies; the identification of *E2F1* somatic mutation in Chapter 2, of somatic mutations in *MLL3* and *RNF43* in Chapter 3, and of somatic mutations in *PRUNE2* in Chapter 4 demonstrates the fruitfulness of this analysis technique. Synonymous mutations has long been looked upon as “noise” in the search for driver mutations as these mutations, while somatic, do not result in a corresponding amino acid change of the gene product. Only recently have somatic synonymous mutations been demonstrated to play a role in the creation of gene enhancers and disruption of

gene silencers in cancer and light the way for a complete reassessment concerning the roles of this neglected mutation class in cancer development (202). The Cancer Genome Project (CGP) Group from the Wellcome Trust Sanger Institute took a step further back and conjectured that ALL somatic variants observed in any cancer are a result of N mutational processes leaving their marks over the lifetime of the tumor. In a series of publications, the CGP Group demonstrated the feasibility of computationally deconvoluting the mutational processes driving the observed somatic mutations pattern through the use of non-negative matrix factorization (54,57,197). The use of this computational technique culminated in the landmark study by Alexandrov et al. where it is shown that there are only 21 unique mutational signatures driving the observed ~5 million somatic mutations in over 7000 unique tumors comprising of 30 different cancer types (198). Of interest is the prevalence of APOBEC mediated mutational process across different types of cancer as shown by Roberts et al. and Burns et al. a few months prior to the seminal publication by Alexandrov et al. (55,56,197). Employing the idea of using the totality of the somatic variants data to arrive at the mutational processes driving the appearance of the observed somatic mutations, we have shown, in Chapter four, for the first time that the APOBEC mediated mutational process is dominant in a subset of parathyroid carcinomas with high mutational burden.

The three publications presented in this thesis each represents a time capsule, three snapshots of a four year period capturing the rapid development in somatic mutations analysis; I have shown the increasing depth and wealth of information that can be gleaned from analysis of whole-exome sequencing data from a single sample to multiple samples. I have also shown how the analysis evolved from the search for somatic driver mutations to the search for mutational processes that is the source of

these somatic mutations. Finally, I have shown the evolution from analyzing only a part of the somatic variants data, somatic nonsynonymous mutations, in the search for driver genes to the analysis almost all of the somatic variants data simultaneously in the search for driver mutational processes.

Development of techniques that leads to the fuller use of the data within whole-exome sequencing sets are ultimately sterile without considering the applications the results of these techniques can bring to scientists. The applications of the results found by analysis of somatic variants in cancer exomes represent the second theme of this thesis. Chapters two, three and four each represents a different application facet of a common somatic variant analysis theme. Perhaps the simplest application is presented in chapter four; the findings in the chapter offered a fresh breath to the research into parathyroid carcinoma. For the last decade since the seminal study implicating *CDC73* in the development of parathyroid carcinoma, other than functional studies of *CDC73* role as a tumor suppressor and also as a possible oncogene (168-170,203-206) there are no additional findings towards understanding this rare disease. Whole-exome sequencing study of this cancer not only revealed a novel recurrently mutated gene in the form of *PRUNE2* suggesting the possibility of a second driver gene in parathyroid carcinoma but also implicates the frequently mutated kinase family in the loss of control over cellular migration and invasion. Perhaps the most interesting and surprising finding is the dominance of the APOBEC mediated mutagenesis process in parathyroid carcinomas with high mutational burden. These results open up new avenues of inquiry not considered before and should provide fertile experimental grounds for further exploration where stagnation has been before. If there is one word to describe the application of Chapter four, it is "discovery".

Chapter two shows an amalgamation of computational and experimental techniques; the computational techniques presented there not only shows the power of bioinformatics in predicting the functional consequences of a somatic mutation but also to help prioritize the key experiments needed to be performed by the experimental biologist. The predicted somatic structural alteration to the key DNA binding site in E2F1 suggests impairment to the protein's DNA binding ability thus paved the way for the key chromatin immunoprecipation experiment showing the mutated E2F1 has severely impaired DNA binding ability. The results of the experimental work not only support the conjecture proposed by the bioinformatics analysis but also revealed additional findings; namely the increased stability of the mutated E2F1 leading to the additional conjecture that it behaves as a competitive inhibitor of Rb. The combined bioinformatics and experimental work as a whole is much greater than the sum of its parts and highlights the power of collaboration between computational and experimental research scientists. The one word to describe the application of Chapter two is “partnership”, the demonstration of the fruits of partnership between bioinformatics and experimental biology.

The publication of the Chapter three's results is initially an application of the “discovery” concept presented above. Analysis of multi-sample whole-exome sequencing of OV-related cholangiocarcinoma presents the landscape of recurrently altered genes with a mutational pattern similar to that pancreatic cancer on a mutated gene and nucleotide level. However, these initial results in combination with subsequent publications of additional targeted therapeutic-related findings related to the recurrently mutated genes found in OV-related CCA propels this study one step further. The PORCN inhibitor LGK974 and triptolide analog Minnelide has shown effectiveness in treating pancreatic cancer cell lines with *RNF43* inactivation and

KRAS/*TP53* mutations respectively and both treatments are currently at Phase I clinical trials in the United States (135,137,156,159). Due to the mutational similarities between OV-related CCA and pancreatic cancer as well as common mutations in *RNF43* inactivation and *KRAS*/*TP53* mutations, LGK974 and Minnelide can be re-purposed as targeted therapy for OV-related CCA with a background of either *RNF43* inactivation for the former or *KRAS*/*TP53* mutations for the latter. For cancer research, the application of Chapter three can be described as "translational" but for an economically disadvantaged population afflict with this rare and malignant cancer, Chapter 3 can only be described as "hope".

The thrill of scientific discovery that opens new avenues of research not considered before, the building of partnerships between bioinformatics and experimental biology that accelerates the discovery process, the translational application of basic scientific findings to a medical oncological setting and finally the hope that whatever I did in the last four years may bring about a quantum leap in the cure for cancer, these are the final fruits of labor that I will take away from my PhD study and these fruits will continue to provide the nourishment for the post-doctorate journeyman in his continuing journey of scientific discovery.

I have come to realize writing those last words above that there will never be an end to a candidate's thesis. For all thesis written before and will be written after mine, the end of a thesis can only mean an end of a beginning as the candidate looks to the future. With the publication of the first cancer kinomes followed by cancer exomes using Sanger sequencing technology, there has been, in the past six years, a seismic shift in the fundamental philosophical approach to cancer science (18,19). A paradigm shift from a "hypothesis first" driven approach to science to a "data first" driven approach to science; the arguments for hypothesis-driven or data-driven

approaches were eloquently and succinctly put forth by Prof. R.A. Weinberg and Prof. T.R. Golub respectively in a point-counterpoint opinion articles published in April 2010 (207,208).

In essence, a hypothesis driven approach starts with a conjecture followed by observations that support or refute said conjecture; while all conjectures were either proven false or have within it, the possibility of being proven false, the falsification of a conjecture inevitably grants the researcher some new conceptual insights into the problem he or she is studying. These new insights will allow the researcher to formulate a better conjecture than its predecessor thus allowing the cycle to repeat itself anew.

The data first driven approach seeks to take advantage of the rapid data generating capability of massively parallel sequencing to remove the step for hypothesis formulation by the researcher, a step that is necessarily biased by the researcher's *a priori* knowledge and preferences. The unbiased generation and analysis of sequencing data can act as a rapid survey and generator of hypothesis for previously unstudied cancer types or act to open unanticipated research directions in studied cancer types.

Having no knowledge of cancer biology as recently as six years ago, I have unquestionably benefited from the emergent data-first paradigm. Publications of somatic variant analysis of kinome then whole-exome data sets enabled me to quickly achieve an overall view of the important genes and pathways driving different cancer types and serves as an excellent starting point in the understanding of cancer biology.

The three publications presented in this thesis are also completely data-driven; while one cannot argue with the results and ideas generated by these three studies, one

cannot help but notice that, other than Chapter two, all that is offered in the remaining studies are a series of conjectures lacking in any serious attempts at confirmation or refutation through experimentation. There is a very real temptation, under the data-first paradigm, to fall under the spell of rapid large volume data production and to only consider the generation of conjectures to be of importance. After all, conjectures are easy to formulate while confirmation or refutation requires much time-consuming experimentation; to whit, Fermat's conjecture formulated in 1637 remained unresolved until 1995 by Prof. Andrew Wiles (209) and the Riemann hypothesis formulated in 1859 remains unresolved to this day. It is my belief that while the data-first approach during my PhD candidature has helped accelerate my understanding of cancer biology and has helped me immensely in identification of interesting ideas and hypothesis, what I have done thus far represents only one-half of the scientific method. Going forward, in a back-to-the-future manner, I believe a paradigm shift back to the venerable hypothesis-driven approach is needed for me to continue walking the path of scientific discovery. To demonstrate my understanding of this approach, a brief study proposal is included with this thesis based on a synthesis of the hypothesis-driven approach in combination with the partnership paradigm between bioinformatics and experimental biology.

5.2: Hypothetical research proposal

5.2.1: Title

Creation of LINE-1 retro-element inducible cell line as a surrogate for controllable APOBEC3 mediated mutagenesis.

5.2.2: Introduction

Long Interspersed Nuclear Elements-1 (LINE-1) is class of nuclear elements or retrotransposons that can amplify itself in a host genome using RNA intermediates and make up ~17% of the human genome (16,210,211). All LINE-1s are around 6000 base pairs and contains the following four discrete subunits:

1) 5' Untranslated Region (UTR) containing the RNA polymerase II promoter.

2) ORF1 gene encoding a 40kDa trimer forming RNA binding protein with nucleic acid chaparone activity.

3) ORF2 gene encoding a 149kDa protein with dual endonuclease and reverse transcriptase function that preferentially binds to LINE-1 RNA.

4) 3' UTR containing the polyadenylation signal (AATAAA) and poly-A tail.

There are currently only 80-120 currently active LINE-1s in the human genome with 99.9% of LINE-1s being inactive due to inversions, truncations and point mutations (212,213). A LINE-1 replicates by using the host's transcription machinery to produce a RNA copy of itself (Figure 5.1); this LINE-1 RNA will migrate to the cytoplasm where the host's translation machinery will produce the LINE-1 protein products, ORF1p and ORF2p, encoded by the ORF1 and ORF2 gene respectively. Multiple units of ORF1p and a single unit of ORF2p will bind to the LINE-1 RNA and this complex will migrate back to the nucleus. The endonuclease part of ORF2p will recognize the ATTTT DNA motif and make a cut between the

adenine and thymine base. The 3' TTTT overhang created by the endonuclease will allow the adenine rich tail of LINE-1 to attach, allowing the reverse transcriptase part of ORF2p to regenerate the original LINE-1 DNA sequence completing its replication cycle.

Due to the stochastic nature of LINE-1 re-integration into the host genome, uncontrolled LINE-1 replication should be deleterious to the host due to the possibility of random insertion into a critical protein coding gene. An example is the observations of LINE-1 insertions into teneurin transmembrane protein 3 (*ODZ3*), *ROBO2*, protein tyrosine phosphatase, receptor type, M (*PTPRM*), pericentriolar material 1 (*PCM1*), and cadherin 11, type 2, OB-cadherin (osteoblast) (*CDH11*) in colorectal cancer (214). Tubio et al. asked the general question whether LINE-1s can be somatically activated in cancer and if so, whether these observed activated LINE-1s contribute to cancer development; the answer to the former is yes with the answer to the latter being insufficient data for meaningful answer (215). LINE-1s were found to be activated in a somatic manner in at least 5 cancer types due aberrant hypomethylation of its promoter region. The activity of LINE-1 were found to wax and wane during the span of tumour evolution but appears to prefer reinsertion into intergenic or heterchromatic regions with low exon density and low expression genes. Thus while somatic LINE-1 retrotransposition is a new mutational process, the process appears to result in mainly passenger type alterations.

Of interest is the fact that promoter methylation is not the only manner in which LINE-1s are regulated; there are at least three more mechanisms in which the human cell have evolved to control LINE-1 activity. One mechanism is transcriptional silencing where Krueppel-associated box (KRAB) zinc-finger proteins recruits KRAB-associated protein-1 (KAP1) and its repressive complex to LINE-1 target sites

(216). The second mechanism is via miRNA processing complex, Drosha-DCGR8, where LINE-1 mRNAs are negatively repressed through cleavage of its structured regions by Drosha, an RNase enzyme (217). The third mechanism is the APOBEC3 family of deaminase enzymes has been shown to reduce LINE-1 retrotransposition frequency by up to 85% via an unknown mechanism independent from APOBEC deaminase activity (218,219,220).

There are three lines of evidence for consideration; one, somatic LINE-1 activation is a mutational process observed in many cancers (215). Two, one of roles of APOBEC3 is in the suppression of LINE-1 retrotransposition activity (218,219,220). Three, APOBEC3 mediated mutational signature is observed in a wide variety of cancers (55,56,198). These lines of evidence together logically generate the hypothesis below.

5.2.3: Hypothesis

Somatic activation of LINE-1 drives APOBEC3 mutagenesis in the cancer genome

5.2.4: Proposed mechanism

Somatic LINE-1 activation leads to APOBEC3 enzymes to be produced in response to suppress LINE-1 retrotransposition activity. While LINE-1 activity may increase or decrease during tumour evolution, it is persistent once activated. The persistent LINE-1 activity will necessitate persistent APOBEC3 presence; since several members of the APOBEC3 family are known to be mutagenic to genomic DNA, a side-effect of persistent APOBEC3 presence will be the increased mutations in the C>T or C>G at TpCpA or TpCpT context in the cancer genome.

5.2.5: Proposed milestones

1) To show LINE-1 activation induces corresponding APOBEC3 enzymes activation.

2) To show sustained LINE-1activation sustains APOBEC3 presence

3) Show increasing numbers of mutations corresponding to C>T or C>G at TpCpA or TpCpT context as a function of time with sustained LINE-1 activation.

5.2.6: Proposed experiments

The first step is to locate a cell line that is negative for LINE-1 activity; bisulphite sequencing of LINE-1 promoters is an option but the exhaustive search for and sequencing of all "live" LINE-1 promoters is a time consuming task. Rodic et al. recently showed protein expression of ORF1 gene is a marker of LINE-1 activity and is a common feature in many cancers but is absent in normal somatic tissues (221). Using the methodology proposed by Rodic et al., presence of ORF1p can be used as a marker of LINE-1 activity in cell lines to efficiently locate cell line(s) negative in LINE-1 activity, designated as L1-neg.

We will need a reliable method of controlling activation of LINE-1 through exogenous means. The Tetracycline-Controlled Transcription Activation (Tet) is a method to reversibly control gene transcription through the presence or absence of the antibiotic Tetracycline or its derivatives (222). The use of the Tet-On method or transcription activation in the presence of exogenous antibiotics will be selected to ensure complete control over transcription activation. There will be two components to this Tet-On system employed for this study: One component is the creation of a tetracycline inducible LINE-1 (Tet-On L1) through replacement of the original LINE-1 promoter region with a tetracycline responsive element (TRE). The second component will be the creation of a L1-neg cell line (rtTA_L1-neg) to contain the recombinant tetracycline controlled transcription factor (rtTA), a chimeric protein

requiring tetracycline for DNA binding to TRE to activate transcription. Insertion of Tet-On L1 into the genome of rtTA_L1-neg cell line through viral integration systems will create the final LINE-1 inducible cell line (L1-TetOn).

With the establishment of L1-TetOn cell line, testing for basal level of LINE-1 will be performed through protein expression of ORF1p, the surrogate marker for LINE-1 activity. Tet-On L1 will then be activated through tetracycline introduction and ORF1p expression observed to see if ORF1p expression increase with tetracycline activation which indicates Tet-On L1 has been activated. Tetracycline will then be withdrawn and ORF1p expression should fall back to basal level indicating deactivation of Tet-On L1. Once reliable control and detection of Tet-On L1 is established in L1-TetOn cell line, the expression of APOBEC3 family, both gene and protein levels, can be quantified first in basal L1-TetOn cell line, then in the presence of tetracycline and finally in the withdrawal of tetracycline. If the proposed conjecture is correct, tetracycline presence should bring about activation of APOBEC3 family of proteins to suppress the presence of activated Tet-On L1; while the withdrawal of tetracycline should suppress activation of Tet-On L1 thus APOBEC3 proteins should fall back to basal levels since they are no longer needed.

If APOBEC3 proteins do activate and deactivate in the presence or absence of LINE-1 activity then the next step will be sustained long term Tet-On L1 activation to mimic persistent LINE-1 presence during tumour evolution. Whole genome sequencing of the L1-TetOn cell line will be performed prior to tetracycline addition and then with tetracycline addition, whole genome sequencing will be performed every month for a one year period. If the conjecture is correct, there should be a steady increase in C>(T|G) at TpCpA or TpCpT context as a function of time

indicating that sustained activation of APOBEC3 family in suppressing LINE-1 activation will result in collateral DNA damage to the host genome.

5.2.7: Conclusion

Confirmation of this hypothesis will not only start to answer the fundamental question why APOBEC mutational signature is observed in a wide variety of cancers but also establish a definitive genome-wide mutational signature for APOBEC3 mutational process for comparison with computational derivation of the signature proposed before (58,198). In addition, studies of carcinogens or gene mutations can be performed against the background of inducible APOBEC3 activation allowing for mutational processes to be studied in combination. Combination studies will enable a more realistic study of cancer development and generate much needed experimentally derived mutational data for bioinformatics to refine the computational process of mutational signature separation and identification.

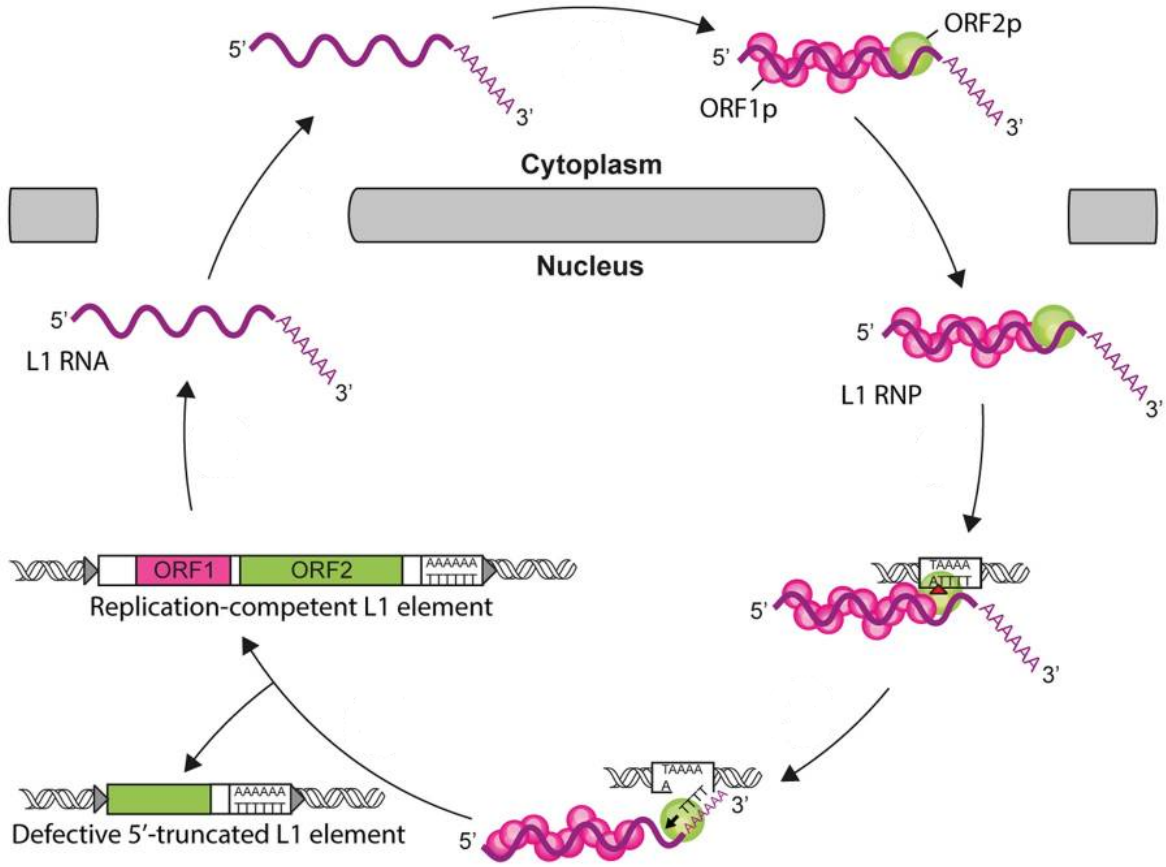


Figure 5.1: Life cycle of LINE-1 retrotransposon. Figure extracted and modified from Figure 1 of Viollet S, Monot C, Cristofari G: L1 retrotransposition: The snap-velcro model and its consequences. *Mob Genet Elem* 2014, **4**:e28907. Article is originally published by Landes Bioscience.

References

1 von Hansemann D. **Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung**. *Virchows Arch Path Anat* 1890, **119**:299.

2 Boveri T. **Zur Frage der Entstehung Maligner Tumoren**. Gustav Fischer; 1914. pp. 1–64.

3 Goodman MF, Fygenson KD. **DNA polymerase fidelity: from genetics toward a biochemical understanding**. *Genetics* 1998, **148**:1475–1482.

4 Ananthaswamy HN, Pierceall WE. **Molecular mechanisms of ultraviolet radiation carcinogenesis**. *Photochem Photobiol* 1990, **53**:1119-1136.

5 Loeb LA, Harris CC. **Advances in chemical carcinogenesis: a historical review and prospective**. *Cancer Res* 2008, **68**:6863–6872.

6 Reddy EP, Reynolds RK, Santos E, Barbacid M. **A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene**. *Nature* 1982, **300**:149–152.

7 Rowley J. **A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining**. *Nature* 1973, **243**:290–293.

8 Kurzrock R, Kantarjian HM, Druker BJ, Talpaz M. **Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics**. *Ann Intern Med* 2003, **138**:819-830.

9 Hanahan D, Weinberg RA. **The hallmarks of cancer**. *Cell* 2000, **100**:57-70.

10 Hanahan D, Weinberg RA. **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**:646-674.

11 Maxam AM, Gilbert W. **A new method for sequencing DNA**. *Proc Natl Acad Sci USA* 1977, **74**:560-564.

12 Sanger F, Coulson AR. **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. *J Mol Biol* 1975, **94**:441-448.

13 Sanger F, Nicklen S, Coulson AR. **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci USA* 1977, **74**:5463-5467.

14 Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, Arnheim N. **Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia**. *Science* 1985, **230**:1350-1354.

15 Saiki R, Gelfand D, Stoffel S, Scharf S, Higuchi R, Horn G, Mullis K, Erlich H. **Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase**. *Science* 1988, **239**:487-491.

16 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP,

Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium. **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

17 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M,

Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. **The sequence of the human genome**. *Science* 2001, **291**:1304-1351.

18 Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR. **Patterns of somatic mutations in human cancer genomes**. *Nature* 2007, **446**:153-158.

19 Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. **Core signalling pathways in human pancreatic cancers revealed by global genomic analyses**. *Science* 2008, **321**:1801-1806.

20 Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander

S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. **Somatic mutations affect key pathways in lung adenocarcinoma**. *Nature* 2008, **455**:1069-1075.

21 Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA Jr, Hartigan J, Smith DR, Strausberg RL, Marie SK, Shinjo SM, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. **An integrated genomic analysis of human glioblastoma multiforme**. *Science* 2008, **321**:1807-1812.

22 Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JW, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA. **Mutations of the BRAF gene in human cancer**. *Nature* 2002, **417**:949-954.

23 Tsai J, Lee JT, Wang W, Zhang J, Cho H, Mamo S, Bremer R, Gillette S, Kong J, Haass NK, Sproesser K, Li L, Smalley KS, Fong D, Zhu YL, Marimuthu A, Nguyen H, Lam B, Liu J, Cheung I, Rice J, Suzuki Y, Luu C, Settachatgul C, Shellooe R, Cantwell J, Kim SH, Schlessinger J, Zhang KY, West BL, Powell B, Habets G, Zhang C, Ibrahim PN, Hirth P, Artis DR, Herlyn M, Bollag G. **Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity**. *Proc Natl Acad Sci USA* 2008, **105**:3041-3046.

24 Bollag G, Hirth P, Tsai J, Zhang J, Ibrahim PN, Cho H, Spevak W, Zhang C, Zhang Y, Habets G, Burton EA, Wong B, Tsang G, West BL, Powell B, Shellooe R, Marimuthu A, Nguyen H, Zhang KY, Artis DR, Schlessinger J, Su F, Higgins B, Iyer R, D'Andrea K, Koehler A, Stumm M, Lin PS, Lee RJ, Grippo J, Puzanov I, Kim KB, Ribas A, McArthur GA, Sosman JA, Chapman PB, Flaherty KT, Xu X, Nathanson KL, Nolop K. **Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma**. *Nature* 2010, **467**:596-599.

25 Bollag G, Tsai J, Zhang J, Zhang C, Ibrahim P, Nolop K, Hirth P. **Vemurafenib: the first drug approved for BRAF-mutant cancer**. *Nat Rev Drug Discov* 2012, **11**:873-886.

26 Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD, Fulton LA, Locke DP, Magrini VJ, Abbott RM, Vickery TL, Reed JS, Robinson JS, Wylie T, Smith SM, Carmichael L, Eldred JM, Harris CC, Walker J, Peck JB, Du F, Dukes AF, Sanderson GE, Brummett AM, Clark E, McMichael JF, Meyer RJ, Schindler JK, Pohl CS, Wallis JW, Shi X, Lin L, Schmidt H, Tang Y, Haipek C, Wiechert ME, Ivy JV, Kalicki J, Elliott G, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson MA, Baty J, Heath S, Shannon WD,

Nagarajan R, Link DC, Walter MJ, Graubert TA, DiPersio JF, Wilson RK, Ley TJ. **Recurring mutations found by sequencing an acute myeloid leukemia genome**. *N Engl J Med* 2009, **361**:1058-1066.

27 Borger DR, Tanabe KK, Fan KC, Lopez HU, Fantin VR, Straley KS, Schenkein DP, Hezel AF, Ancukiewicz M, Liebman HM, Kwak EL, Clark JW, Ryan DP, Deshpande V, Dias-Santagata D, Ellisen LW, Zhu AX, Iafrate AJ. **Frequent mutation of isocitrate dehydrogenase (IDH)1 and IDH2 in cholangiocarcinoma identified through broad-based tumor genotyping**. *Oncologist* 2012, **17**:72-79.

28 Rohle D, Popovici-Muller J, Palaskas N, Turcan S, Grommes C, Campos C, Tsoi J, Clark O, Oldrini B, Komisopoulou E, Kunii K, Pedraza A, Schalm S, Silverman L, Miller A, Wang F, Yang H, Chen Y, Kernytsky A, Rosenblum MK, Liu W, Biller SA, Su SM, Brennan CW, Chan TA, Graeber TG, Yen KE, Mellinghoff IK. **An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells**. *Science* 2013, **340**:626-630.

29 Prahl M; Agresta S. **Study of Orally Administered AG-120 in Subjects With Advanced Solid Tumors, Including Glioma, With an IDH1 Mutation**. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2014- [cited 2014 Dec 20]. Available from: https://clinicaltrials.gov/ct2/show/NCT02073994 NLM Identifier: NCT02073994.

30 Prahl M; Agresta S. **Study of Orally Administered AG-120 in Subjects with Advanced Hematologic Malignancies With an IDH1 Mutation**. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2014- [cited 2014 Dec 20]. Available from: https://clinicaltrials.gov/ct2/show/NCT02074839 NLM Identifier: NCT02074839.

31 Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. **Whole-genome random sequencing and assembly of Haemophilus influenza Rd**. *Science* 1995, **269**:496-512.

32 Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J,

Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**:2185-2195.

33 Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**:872-876.

34 Harbour JW, Onken MD, Roberson ED, Duan S, Cao L, Worley LA, Council ML, Matatall KA, Helms C, Bowcock AM. **Frequent mutation of BAP1 in metastasizing uveal melanomas**. *Science* 2010, **330**:1410-1413.

35 Stratton MR, Campbell PJ, Futreal PA. **The cancer genome**. *Nature* 2009, **458**:719-724.

36 Langmead B, Salzburg S. **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357-359.

37 Li H, Durbin R. **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**:589-595.

38 Liu CM, Wong T, Wu E, Luo R, Yiu SM, Li Y, Wang B, Yu C, Chu X, Zhao K, Li R, Lam TW. **SOAP3: ultra-fast GPU-based parallel alignment tool for short reads**. *Bioinformatics* 2012, **28**:878-879.

39 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup. **The Sequence alignment/map (SAM) format and SAMtools**. *Bioinformatics* 2009, **25**:2078-2079.

40 https://broadinstitute.github.io/picard/

41 DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M. **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat Genet* 2011, **43**:491-498.

42 https://www.ncbi.nlm.nih.gov/SNP/

43 The 1000 Genomes Project Consortium. **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061-1073.

44 Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ. **COSMIC: exploring the world's knowledge of somatic mutations in human cancer**. *Nucleic Acids Res* 2015, 43(Database issue):D805-11.

45 http://sourceforge.net/projects/primer3/

46 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**:248-249.

47 H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. **The Protein Databank**. *Nucleic Acids Res* 2000, **28**:235-242.

48 Arnold K, Bordoli L, Kopp J, and Schwede T. **The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling**. *Bioinformatics* 2006, **22**:195-201.

49 Guex N, Peitsch MC. **SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling**. *Electrophoresis* 1997, **18**:2714-2723.

50 Anders S. **Visualization of genomic data with the Hilbert curve.** *Bioinformatics* 2009, **25**:1231-1235.

51 Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Børresen-Dale AL, Kristensen VN. **Allele-specific copy number analysis of tumors**. *Proc Natl Acad Sci USA* 2010, **107**:16910-16915.

52 Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, Weng WH, Siew EY, Liu Y, Heng HL, Chong SC, Gan A, Tay ST, Lim WK, Cutcutache I, Huang D, Ler LD, Nairismägi ML, Lee MH, Chang YH, Yu KJ, Chan-On W, Li BK, Yuan YF, Qian CN, Ng KF, Wu CF, Hsu CL, Bunte RM, Stratton MR, Futreal PA, Sung WK, Chuang CK, Ong CK, Rozen SG, Tan P, Teh BT. **Genome-wide mutational signatures of aristolochic acid and its application as a screening tool**. *Sci Transl Med* 2013, **5**:197ra101.

53 Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH, Moriya M, Niknafs N, Douville C, Karchin R, Turesky RJ, Pu YS, Vogelstein B, Papadopoulos N, Grollman AP, Kinzler KW, Rosenquist TA. **Mutational signature of aristolochic acid exposure as revealed by exome sequencing**. *Sci Transl Med* 2013, **5**:197ra102.

54 Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jönsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerød A, Tutt A, Martens JW, Aparicio SA, Borg Å, Salomon AV, Thomas G, Børresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR; Breast

Cancer Working Group of the International Cancer Genome Consortium. **Mutational processes molding the genomes of 21 breast cancers**. *Cell* 2012, **149**:979-993.

55 Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, Harris S, Shah RR, Resnick MA, Getz G, Gordenin DA. **An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers**. *Nat Genet* 2013, **45**:970-976.

56 Burns MB, Temiz NA, Harris RS. **Evidence for APOBEC3B mutagenesis in multiple human cancers**. *Nat Genet* 2013, **45**:977-983.

57 Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. **Deciphering signatures of mutational processes in human cancer**. *Cell Rep* 2013, **3**:246-259.

58 Fischer A, Illingworth CJ, Campbell PJ, Mustonen V. **EMu: probabilistic inference of mutational processes and their localization in the cancer genome**. *Genome Biol* 2013, **14**:R39.

59 Kim H, Park H. **Sparse non-negative matrix factorization via alternating non-negativity-constrained least squares for microarray data analysis**. *Bioinformatics* 2007, **23**:1495-1502.

60 Lee DD, Seung HS. **Learning the parts of objects by non-negative matrix factorization**. *Nature* 1999, **401**:788-791.

61 Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. **Algorithms and applications for approximate nonnegative matrix factorization**. *Comput Stat Data Anal* 2007, **52**:155-173.

62 Hoekstra A, Riben M, Frumovitz M, Liu J, Ramirez PT. **Well differentiated papillary mesothelioma of the peritoneum: a pathological analysis and review of the literature**. *Gynecol Oncol* 2005, **98**:161–167.

63 Wagner JC, Sleggs CA, Marchant P. **Diffuse pleural mesothelioma and asbestos exposure in the north west Cape province**. *Br J Ind Med* 1960, **17**:260-271.

64 Selikoff J, Hammond EC, Seidman H. **Mortality experiences of insulation workers in the United States**. *Cancer* 1980, **46**:2736-2740.

65 Ashcroft T. **Epidemiological and quantitative relationships between mesothelioma and asbestos on Tyneside**. *J Clin Pathol* 1973, **26**:832-840.

66 Maher B. **Fear in the Dust**. *Nature* 2010, **468**:884-885.

67 Bani-Hani K, Gharaibeh K. **Malignant peritoneal mesothelioma**. *J Surg Oncol.* 2005, **91**:17–25.

68 Hanrahan JB. **A combined papillary mesothelioma and adenomatoid tumor of the omentum: report of a case**. *Cancer* 1963, **11**:1497–1500.

69 Clarke JM, Helft P. **Long-term survival of a woman with well differentiated papillary mesothelioma fo the peritoneum: a case report and review of the literature**. *J Med Case Reports*. 2010, **4**:346.

70 Daya D, McCaughey WTE. **Well-differentiated papillary mesothelioma of the peritoneum**. *Cancer* 1990, **65**:292-296.

71 Malpica A, Sant'Ambrogio S, Deavers MT, Silva EG. **Well-differentiated papillary mesothelioma of the female peritoneum: a clinicopathogic study of 26 cases**. *Am J Surg Pathol* 2012, **36**:117-127.

72 Hoekman K, Tognon G, Risse EK, Bloemsma CA, Vermorken JB. **Well-differentiated papillary mesothelioma of the peritoneum: a separate entity**. *Eur J Cancer* 1996, **32A**:255-258.

73 Burrig KF, Pfitzer P, Hort W. **Well differentiated papillary mesothelioma of the peritoneum: a borderline mesothelioma**. *Virchows Archiv A Pathol Anat* 1990, **417**:443–447.

74 Jaurand MC. **Mechanisms of fiber-induced genotoxicity**. *Environmental Health Perspectives* 1997, **105(Suppl 5)**:1073-1084.

75 Pisick E, Salgia R. **Molecular Biology of Malignant Mesothelioma: A Review**. *Hematol Oncol Clin N Am* 2005, **19**:997-1023.

76 Sugarbaker DJ, Richards WG, Gordon GJ, Dong L, De Rienzo A, Maulik G, Glickman JN, Chirieac LR, Hartman ML, Taillon BE, Du L, Bouffard P, Kingsmore SF, Miller NA, Farmer AD, Jensen RV, Gullans SR, Bueno R. **Transcriptome sequencing of malignant pleural mesothelioma tumors**. *Proc Natl Acad Sci USA* 2008, **105**:3521-3526.

77 Yang H, Rivera Z, Jube S, Nasu M, Bertino P, Goparaju C, Franzoso G, Lotze MT, Krausz T, Pass HI, Bianchi ME, Carbone M. **Programmed necrosis induced by asbestos in human mesothelial cells causes high-mobility group box 1 protein release and resultant inflammation**. *Proc Natl Acad Sci USA* 2010, **107**:12611-12616.

78 Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. **Targeted capture and massively parallel sequencing of 12 human exomes**. *Nature* 2009, **461**:272-276.

79 Musti M, Kettunen E, Dragonieri S, Lindholm P, Cavone D, Serio G, Knuutila S. **Cytogenetic and molecular genetic changes in malignant mesothelioma**. *Cancer Genetics and Cytogenetics* 2006, **170**:9-15.

80 Thompson JD, Higgins DG, Gibson TJ. **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.

81 Yue P, Melamud E, Moult J. **SNPs3D: Candidate gene and SNP selection for association studies**. *BMC Bioinformatics* 2006, **7**:166.

82 Kumar P, Henikoff S, Ng PC. **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nat Protoc. 2009*,**4**:1073-1081.

83 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, 46(3):310-315.

84 Zheng N, Fraenkel E, Pabo CO, Pavletich NP. **Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP**. *Genes Dev.* 1999, **13**:666-674.

85 Wang C, Chen L, Hou X, Li Z, Kabra N, Ma Y, Nemoto S, Finkel T, Gu W, Cress WD, Chen J. **Interactions between E2F1 and SirT1 regulate apoptotic response to DNA damage**. *Nat Cell Biol* 2006, 8(9):1025-31.

86 Moroni MC, Hickman ES, Lazzerini Denchi E, Caprara G, Colli E, Cecconi F, Müller H, Helin K. **Apaf-1 is a transcriptional target of E2F and p53**. *Nat Cell Biol* 2001, 3(6):552-558.

87 Bracken AP, Ciro M, Cocito A, Helin K. **E2F target genes: unraveling the biology**. *Trends Mol Med* 2004 August, **29**:409-417.

88 Chen WY, Wang DH, Yen RC, Luo J, Gu W, Baylin SB. **Tumor suppressor HIC1 directly regulates SIRT1 to modulate p53-dependent DNA-damage responses**. *Cell* 2005, 123(3):437-48.

89 Yuan J, Minter-Dykhouse K, Lou Z. **A c-Myc-SIRT1 feedback loop regulates cell growth and transformation**. *J Cell Biol* 2009, 185(2):203-11.

90 Robles AI, Bemmels NA, Foraker AB, Harris CC. **APAF-1 is a transcriptional target of p53 in DNA damage-induced apoptosis**. *Cancer Res* 2001, 61(18):6660-4.

91 Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, Hopf C, Huhse B, Mangano R, Michon AM, Schirle M, Schlegl J, Schwab M, Stein MA, Bauer A, Casari G, Drewes G, Gavin AC, Jackson DB, Joberty G, Neubauer G, Rick J, Kuster B, Superti-Furga G. **A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway**. *Nat Cell Biol* 2004, **6**:97-105.

92 Xu LG, Li LY, Shu HB. **TRAF7 Potentiates MEKK3-induced AP1 and CHOP Activation and Induces Apoptosis**. *J Biol Chem* 2004, **274**:17278-17282.

93 Shaulian E, Karin M. **AP-1 as a regulator of cell life and death**. *Nat Cell Biol* 2002, **4**:E131-E136.

94 Clark VE, Erson-Omay EZ, Serin A, Yin J, Cotney J, Ozduman K, Avşar T, Li J, Murray PB, Henegariu O, Yilmaz S, Günel JM, Carrión-Grant G, Yilmaz B, Grady C, Tanrikulu B, Bakircioğlu M, Kaymakçalan H, Caglayan AO, Sencar L, Ceyhun E, Atik AF, Bayri Y, Bai H, Kolb LE, Hebert RM, Omay SB, Mishra-Gorur K, Choi M, Overton JD, Holland EC, Mane S, State MW, Bilgüvar K, Baehring JM, Gutin PH, Piepmeier JM, Vortmeyer A, Brennan CW, Pamir MN, Kiliç T, Lifton RP, Noonan JP, Yasuno K, Günel M. **Genomic analysis of non-*NF2* meningiomas reveals mutations in *TRAF7, KLF4, AKT1,* and *SMO***. *Science* 2013, **339**:1077-1080.

95 Reuss DE, Piro RM, Jones DT, Simon M, Ketter R, Kool M, Becker A, Sahm F, Pusch S, Meyer J, Hagenlocher C, Schweizer L, Capper D, Kickingereder P, Mucha J, Koelsche C, Jäger N, Santarius T, Tarpey PS, Stephens PJ, Andrew Futreal P, Wellenreuther R, Kraus J, Lenartz D, Herold-Mende C, Hartmann C, Mawrin C, Giese N, Eils R, Collins VP, König R, Wiestler OD, Pfister SM, von Deimling A. **Secretory meningiomas are defined by combined KLF K409Q and TRAF7 mutations**. *Acta Neuropathol* 2013, **125**: 351-358.

96 Serra-Pagès C, Medley QG, Tang M, Hart A, Streuli M. **Liprins, a family of LAR transmembrane protein-tyrosine phosphatase-interacting proteins**. *J. Biol. Chem.* 1998, **273**:15611-15620.

97 Colas E, Perez C, Cabrera S, Pedrola N, Monge M, Castellvi J, Eyzaguirre F, Gregorio J, Ruiz A, Llaurado M, Rigau M, Garcia M, Ertekin T, Montes M, Lopez-Lopez R, Carreras R, Xercavins J, Ortega A, Maes T, Rosell E, Doll A, Abal M, Reventos J, Gil-Moreno A. **Molecular markers of endometrial carcinoma detected in uterine aspirates**. *Int. J. Cancer* **201**1, **129**:2435-2444.

98 Johnson DG, Ohtani K, Nevins JR. **Autoregulatory control of E2F1 expression in response to positive and negative regulators of cell cycle progression**. *Genes & Dev.* 1994, **8**:1514-1525.

99 Stanelle J, Putzer BM. **E2F1-induced apoptosis: turning killers into therapeutics**. *Trends Mol Med* 2006 April, **12**:177-185.

100 Field SJ, Tsai FY, Kuo F, Zubiaga AM, Kaelin WG Jr, Livingston DM, Orkin SH, Greenberg ME. **E2F-1 Functions in Mice to Promote Apoptosis and Suppress Proliferation**. *Cell* 1996, **85**:549-561.

101 Yamasaki L, Jacks T, Bronson R, Goillot E, Harlow E, Dyson NJ. **Tumor Induction and Tissue Atrophy in Mice Lacking E2F-1**. *Cell* 1996, **85**:537-548.

102 Cress WD, Johnson DG, Nevins JR. **A Genetic Analysis of the E2F1 Gene Distinguishes Regulation by Rb, p107 and Adenovirus E4**. *Mol. Cell Biol.* 1993, **13**: 6314-6325.

103 Halaban R, Cheng E, Zhang Y, Mandigo CE, Miglarese MR. **Release of cell cycle constraints in mouse melanocytes by overexpressed mutant $E2F1_{E132}$ but not by deletion of $p16^{INK4A}$ or $p21^{WAF1/CIP1}$**. *Oncogene* 1998, **16**:2489-2501.

104 Nevins JR. **The Rb/E2F pathway and cancer**. *Hum Mol Genet* 2001, **10**:669-703.

105 Illei PB, Rusch VW, Zakowski MF, Ladanyi M. **Homozygous deletion of *CDKN2A* and codeletion of the methyltheioadenosine phosphorylase gene in the majority of pleural mesotheliomas**. *Clin Cancer Res* 2003, **9**:2108-2113.

106 Bott M, Brevet M, Taylor BS, Shimizu S, Ito T, Wang L, Creaney J, Lake RA, Zakowski MF, Reva B, Sander C, Delsite R, Powell S, Zhou Q, Shen R, Olshen A, Rusch V, Ladanyi M. **The nuclear deubiquitinase BAP1 is commonly inactivated by somatic mutations and 3p21.1 losses in malignant pleural mesothelioma.** *Nat Genet* 2011, **43**:668-672.

107 Ribeiro C, Campelos S, Moura CS, Machado JC, Justino A, Parente B. **Well differentiated peritoneal mesothelioma: clustering in a Portuguese family with a germline BAP1 mutation**. *Ann Oncol* 2013, **24**:2147-2150.

108 Test JR, Cheung M, Pei J, Below JE, Tan Y, Sememtino E, Cox NJ, Dogan AU, Pass HI, Trusa S, Hesdorffer M, Nasu M, Powers A, Rivera Z, Comertpay S, Tanji M, Gaudino G, Yang H, Carbone M. **Germline BAP1 mutations predispose to malignant mesothelioma**. *Nat Genet* 2011, **43**:1022-1025.

109 Nemoto H, Tate G, Kishomoto K, Saito M, Shirahata A, Umemoto T, Matsubara T, Goto T, Mizukami H, Kigawa G, Mitsuya T, Hibi K. **Heterozygous loss of NF2 is an early molecular alteration in a well differentiated papillary mesothelioma of the peritoneum**. *Cancer Genet* 2012, **205**:594-598.

110 Patel T. **Cholangiocarcinoma**. *Nat Clin Pract Gastroenterol Hepatol* 2006, **3**:33-42.

111 Bragazzi MC, Cardinale V, Carpino G, Venere R, Semeraro R, Gentile R, Gaudio E, Alvaro D. **Cholangiocarcinoma: Epidemiology and risk factors**. *Transl Gastrointest Cancer* 2012,1:21-32.

112 Sithithaworn P, Yongvanit P, Duenngai K, Kiatsopit N, Pairojkul C. **Roles of liver fluke infection as risk factor for cholangiocarcinoma**. *J Hepatobiliary Pancreat Sci* 2014, **21**:301-308.

113 Kurathong S, Lerdverasirikul P, Wongpaitoon V, Pramoolsinsap C, Kanjanapitak A, Varavithya W, Phuapradit P, Bunyaratvej S, Upatham ES, Brockelman WY. ***Opisthorchis viverrini* infection and cholangiocarcinoma**. *Gastroenterology* 1985, **89**:151-6.

114 Haswell-Elkins MR, Mairiang E, Mairiang P, Chaiyakum J, Charmadol N, Loapaiboon V, Sithithaworn P, Elkins DB. **Cross-sectional study of *Opisthorchis viverrini* infection and cholangiocarcinoma in communities within a high risk area in northeast Thailand**. *Int J Cancer* 1994, **59**:505-509.

115 Grundy-Warr C, Andrews RH, Sithithaworn P, Petney TN, Sripa B, Laithavewat L, Ziegler AD. **Raw attitudes, wetland cultures, life-cycles: socio-cultural dynamics relating to *Opisthorchis viverrini* in the Mekong Basin**. *Parasitol Int* 2012, **61**:65-70.

116 Bhamarapravati N, Thammavit W, Vajrasthira S. **Liver changes in hamsters infected with a liver fluke of man, *Opisthorchis vivverini***. *Am J Trop Med Hyg* 1978, **27**:787-794.

117 Mairiang E, Elkins DB, Mairiang P, Chaiyakum J, Chamadol N, Loapaiboon V, Posri S, Sithithaworn P, Haswell-Elkins M. **Relationship between intensity of *Opisthorchis vivverini* infection and hepatobiliary disease detected by ultrasonography**. *J Gastoenterol Hepatol* 1992, **7**:17-21.

118 Mairiang E, Haskwell-Elkins MR, Mairiang P, Sithithaworn P, Elkins DB. **Reversal of biliary tract abnormalities associated with *Opisthorchis vivverini* infection following praziquantel treatment**. *Trans R Soc Trop Med Hyg* 1993, **87**:194-197.

119 Chernrungroj G. **Risk factors for cholangiocarcinoma: a case-control study**. New Haven, CT: Yale University; 2000.

120 Mark Feldman, Lawrence S. Friedman, Lawrence J. Brandt, ed. (21 July 2006). **Sleisenger and Fordtran's Gastrointestinal and Liver Disease (8th ed.)**. Saunders. pp. 1493–6. ISBN 978-1-4160-0245-1.

121 Farley DR, Weaver AL, Nagorney DM. **"Natural history" of unresected cholangiocarcinoma: patient outcome after noncurative intervention**. *Mayo Clin Proc* 1995, **70**:425-429.

122 Yamamoto M, Takasaki K, Yoshikawa T. **Lymph Node Metastasis in Intrahepatic Cholangiocarcinoma**. *Jpn J Clin Oncol* 1999, **29**:147-150.

123 Valle J, Wasan H, Palmer DH, Cunningham D, Anthoney A, Maraveyas A, Madhusudan S, Iveson T, Hughes S, Pereira SP, Roughton M, Bridgewater J; ABC-02 Trial Investigators. **Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer**. *N Engl J Med* 2010, **362**:1273:1281.

124 Tannapfel A, Benicke M, Katalinic A, Ulhmann D, Kockerling F, Hauss J, Wittekind C. **Frequencing of p16^{ink4A} alterations and k-*ras* mutations in intrahepatic cholangiocarcinoma of the liver**. *Gut* 2000, **47**:721-727.

125 Ahrendt SA, Eisenberger CF, Yip L, Rashid A, Chow JT, Pitt HA, Sidransky D. **Chromosome 9p21 loss and p16 inactivation in primary sclerosing cholangitis-associated cholangiocarcinoma**. *J Surg Res* 1999, **84**:88-93.

126 Li M, Zhao H, Zhang X, Wood LD, Anders RA, Choti MA, Pawlik TM, Daniel HD, Kannangai R, Offerhaus GJ, Velculescu VE, Wang L, Zhou S, Vogelstein B, Hruban RH, Papadopoulos N, Cai J, Torbenson MS, Kinzler KW. **Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma**. *Nat Genet* 2011, **43**:828-829.

127 Venkatachalam S, Shi YP, Jones SN, Vogel H, Bradley A, Pinkel D, Donehower LA. **Retention of wild-type p53 in tumors from p53 heterozygous mice: reduction of p53 dosage can promote cancer formation**. *EMBO J* 1998, **17**:4657-4667.

128 Cook WD, McCaw BJ. **Accomodating haploinsufficient tumour suppressor genes in Knudson's model**. *Oncogene* 2000, **19**:3434-3438.

129 Venkatachalam S, Tyner SD, Pickering CR, Boley S, Recio L, French JE, Donehower LA. **Is p53 haploinsufficient for tumor suppression? Implications for the p53+/- mouse model in carcinogenicity testing**. *Toxicol Pathol* 2001, **29**(Suppl):147-154.

130 Teoh PJ, Chung TH, Sebastien S, Choo SN, Yan J, Ng SB, Fonseca R, Chng WJ. **p53 haploinsufficiency and functional abnormalities in multiple myeloma**. *Leukemia* 2014, **28**:2066-2074.

131 van Oijen MG, Slootweg PJ. **Gain-of-function mutations in the tumor suppressor gene p53**. *Clin Cancer Res* 2000, **6**:2138-45.

132 Kim NH, Kim HS, Kim NG, Lee I, Choi HS, Li XY, Kang SE, Cha SY, Ryu JK, Na JM, Park C, Kim K, Lee S, Gumbiner BM, Yook JI, Weiss SJ. **p53 and**

microRNA-34 are suppressors of canonical Wnt signaling. *Sci Signal* 2011, **4**:ra71.

133 Cha YH, Kim NH, Park C, Lee I, Kim HS, Yook JI. **miRNA-34 intrinsically links p53 tumor suppressor and Wnt signaling**. *Cell Cycle* 2012, **11**:1273-1281.

134 Shinada K, Tsukiyama T, Sho T, Okimura F, Asaka M, Hatakeyama S. **RNF43 interacts with NEDL1 and regulates p53-mediated transcription**. *Biochem Biophys Res Commun* 2011, **404**:143-7.

135 Jiang X, Hao HX, Growney JD, Woolfenden S, Bottiglio C, Ng N, Lu B, Hsieh MH, Bagdasarian L, Meyer R, Smith TR, Avello M, Charlat O, Xie Y, Porter JA, Pan S, Liu J, McLaughlin ME, Cong F. **Inactivating mutations of RNF43 confer Wnt dependency in pancreatic ductal adenocarcinoma**. *Proc Natl Acad Sci USA* 2013, 110:12649-12654.

136 Giannakis M, Hodis E, Jasmine Mu X, Yamauchi M, Rosenbluh J, Cibulskis K, Saksena G, Lawrence MS, Qian ZR, Nishihara R, Van Allen EM, Hahn WC, Gabriel SB, Lander ES, Getz G, Ogino S, Fuchs CS, Garraway LA. **RNF43 is frequently mutated in colorectal and endometrial cancers**. *Nat Genet* 2014, doi: 10.1038/ng.3127. [Epub ahead of print]

137 Novartis Pharmaceuticals. **A Study of Oral LGK974 in Patients With Malignancies Dependent on Wnt Ligands**. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2011- [cited 2014 Dec 20]. Available from: https://clinicaltrials.gov/ct2/show/NCT01351103 NLM Identifier: NCT01351103.

138 Deng Y, Wu X. **Peg3/Pw1 promotes p53-mediated apoptosis by inducing Bax translocation from cytosol to mitochondria**. *Proc Natl Acad Sci USA* 2000, **97**:12050-12055.

139 Jiang X, Yu Y, Yang HW, Agar NY, Frado L, Johnson MD. **The imprinted gene PEG3 inhibits Wnt signaling and regulates glioma growth**. *J Biol Chem* 2010, **285**:8472-8480.

140 Aoki K, Aoki M, Sugai M, Harada N, Miyoshi H, Tsukamoto T, Mizoshita T, Tatematsu M, Seno H, Chiba T, Oshima M, Hsieh CL, Taketo MM. **Chromosomal instability by beta-catenin/TCF transcription in APC or beta-catenin mutant cells**. *Oncogene* 2007, **26**:3511-3520.

141 Rashid A, Ueki T, Gao YT, Houlihan PS, Wallace C, Wang BS, Shen MC, Deng J, Hsing AW. **K-ras mutation, p53 overexpression, and microsatellite instability in biliary tract cancers: a population-based study in China**. *Clin Cancer Res* 2002, **8**:3156-3163.

142 Idziaszczyk S, Wilson CH, Smith CG, Adams DJ, Cheadle JP. **Analysis of the frequency of GNAS codon 201 mutations in advanced colorectal cancer**. *Cancer Genet Cytogenet* 2010, **202**:67-69.

143 Weinstein LS, Shenker A, Gejman PV, Merino MJ, Friedman E, Spiegel AM. **Activating mutations of the stimulatory G protein in the McCune-Albright syndrome**. *New Eng J Med* 1991, **325**:1688-1695.

144 Wu J, Matthaei H, Maitra A, Dal Molin M, Wood LD, Eshleman JR, Goggins M, Canto MI, Schulick RD, Edil BH, Wolfgang CL, Klein AP, Diaz LA Jr, Allen PJ, Schmidt CM, Kinzler KW, Papadopoulos N, Hruban RH, Vogelstein B. **Recurrent GNAS Mutations Define an Unexpected Pathway for Pancreatic Cyst Development**. *Sci Transl Med* 2011, **3**:92ra66.

145 Hahn SA, Schutte M, Hoque AT, Moskaluk CA, da Costa LT, Rozenblum E, Weinstein CL, Fischer A, Yeo CJ, Hruban RH, Kern SE. **DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1**. *Science* 1996, 271:351-353.

146 Lagna G, Hata A, Hemmati-Brivanlou A, Massagué J. **Partnership between DPC4 and SMAD proteins in TGF-beta signaling pathways**. *Nature* 1996, **383**:832-836.

147 de Caestecker MP, Hemmati P, Larisch-Bloch S, Ajmera R, Roberts AB, Lechleider RJ. **Characterization of functional domains within Smad4/DPC4**. *J Biol Chem* 1997, **272**:13690-13696.

148 Hahn SA, Bartsch D, Schroers A, Galehdari H, Becker M, Ramaswamy A, Schwarte-Waldhoff I, Maschek H, Schmiegel W. **Mutations of the DPC4/Smad4 gene in biliary tract carcinoma**. *Cancer Res* 1998. **58**:1124-1126.

149 Xu X, Kobayashi S, Qiao W, Li C, Xiao C, Radaeva S, Stiles B, Wang RH, Ohara N, Yoshino T, LeRoith D, Torbenson MS, Gores GJ, Wu H, Gao B, Deng CX. **Induction of intrahepatic cholangiocellular carcinoma by liver-specific disruption of Smad4 and PTEN in mice**. *J Clin Invest* 2006, **116**:1843-1852.

150 Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC, Boca SM, Carter H, Samayoa J, Bettegowda C, Gallia GL, Jallo GI, Binder ZA, Nikolsky Y, Hartigan J, Smith DR, Gerhard DS, Fults DW, VandenBerg S, Berger MS, Marie SK, Shinjo SM, Clara C, Phillips PC, Minturn JE, Biegel JA, Judkins AR, Resnick AC, Storm PB, Curran T, He Y, Rasheed BA, Friedman HS, Keir ST, McLendon R, Northcott PA, Taylor MD, Burger PC, Riggins GJ, Karchin R, Parmigiani G, Bigner DD, Yan H, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. **The genetic landscape of the childhood cancer medulloblastoma**. *Science* 2011, **331**:435-439.

151 Roskams T. **Liver stem cells and their implications in hepatocellular and cholangiocarcinoma**. *Oncogene* 2006, **25**:3818-3822.

152 Cardinale V, Wang Y, Carpino G, Alvaro D, Reid L, Gaudio E. **Multipotent stem cells in the biliary tree**. *Int J Anat Embryol* 2010, **115**:85-90.

153 Nakanuma Y. **A novel approach to biliary tract pathology based on similarities to pancreatic counterparts: is the biliary tract an incomplete pancreas?**.*Pathol Int* 2010, **60**:419-429.

154 Pfeifer GP. **Mutagenesis at methylated CpG sequences**. *Curr Top Microbiol Immunol* 2006, **301**:259-281.

155 Karran P, Lindahl T. **Hypoxanthine in deoxyribonucleic acid: generation by heat-induced hydolysis of adenine residues and release in free form by deoxyribonucleic acid glycosylase from calf thymus**. *Biochemistry* 1980, **19**:6005-6011.

156 Velagapudi M; Vocila L. **Study of Minnelide™ in Patients With Advanced GI Tumors**. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2013- [cited 2014 Dec 20]. Available from: https://clinicaltrials.gov/ct2/show/NCT01927965 NLM Identifier: NCT01927965.

157 Shamon LA, Pezzuto JM, Graves JM, Mehta RR, Wangcharoentrakul S, Sangsuwan R, Chaichana S, Tuchinda P, Cleason P, Reutrakul V. **Evaluation of the mutagenic, cytotoxic and antitumor potential of triptolide, a highly oxygenated diterpene isolated from Tripterygium wilfordii**. *Cancer Lett* 1997, **112**:113-117.

158 Tengchaisri T, Chawengkirttikul R, Rachaphaew N, Reutrakul V, Sangsuwan R, Sirisinha S. **Antitumor activity of triptolide against cholangiocarcinoma growth in vitro and in hamsters**. *Cancer Lett* 1998, **133**: 169-175.

159 Chugh R, Sangwan V, Patil SP, Dudeja V, Dawra RK, Banerjee S, Schumacher RJ, Blazar BR, Georg GI, Vickers SM, Saluja AK. **A preclincal evaluation of Minnelide as a therapeutic agent against pancreatic cancer**. *Sci Transl Med* 2012, **4**:156ra139.

160 Sharrets JM, Kebebew E, Simonds WF. **Parathyroid Cancer**. *Seminoncol* 2010, **37**:580-590.

161 Rawat N, Khetan N, Williams DW, Baxter JN. **Parathyroid carcinoma**. *Br J Surg* 2005, **92**:1345-1353.

162 Carpten JD, Robbins CM, Villablanca A, Forsberg L, Prescuittini S, Bailey-Wilson J, Simonds WF, Gillanders EM, Kennedy AM, Chen JD, Agarwal SK, Sood R, Jones MP, Moses TY, Haven C, Petillo D, Leotlela PD, Harding B, Cameron D, Pannett AA, Hoog A, Heath III H, James-Newton LA, Robinson B, Zarbo RJ, Cavaco BM, Wassif W, Perrier ND, Rosen IB, Kristoffersson U, Turnpenny PD, Farnebo LO, Besser GM, Jackson CE, Morreau H, Trent JM, Thakker RV, Marx SJ, Teh BT, Larsson C, Hobbs MR. ***HRPT2,* encoding parafibromin, is mutated in hyperparathyroidism-jaw tumor syndrome**. *Nature Genet* 2002, **32**:676-680.

163 Howell VM, Haven CJ, Kahnoski K, Khoo SK, Petillo D, Chen J, Fleuren GJ, Robinson BG, Delbridge LW, Philips J, Nelson AE, Krause U, Hammje K, Dralle H, Hoang-Vu C, Gimm O, Marsh DJ, Morreau H, Teh BT. ***HRPT2* mutations are associated with malignancy in sporadic parathyroid tumours**. *J Med Genet* 2003, **40**:657-663.

164 Shattuck TM, Valimaki S, Obara T, Gaz RD, Clark OH, Shoback D, Wierman ME, Tojo K, Robbins CM, Carpten JD, Farnebo LO, Larsson C, Arnold A. **Somatic and Germ-Line Mutations of the *HRPT2* Gene in Sporadic Parathyroid Carcinoma**. *N Engl J Med* 2003, **349**:1722-1729.

165 Sato M, Miyauchi A, Namihira H, Bhuiyan MR, Imachi H, Murao K, Takahara J. **A newly recognized germline mutation of *MEN1* gene identified in a patient with parathyroid adenoma and carcinoma**. *Endocrine* 2000, **12**:223-226.

166 Jenkins PJ, Satta MA, Simmgen M, Drake WM, Williamson C, Lowe DG, Britton K, Chew SL, Thakker RV, Besser GM. **Metastatic parathyroid carcinoma in the MEN2A syndrome**. *Clin Endocrinol* 1997, **47**:747-751.

167 Witteveen JE, Hamdy NAT, Dekkers OM, Kievit J, van Wezel T, Teh BT, Romijn JA, Morreau H. **Downregulation of CASR expression and global loss of parafibromin staining are strong negative determinants of prognosis in parathyroid carcinoma**. *Modern Pathol* 2011, **24**:688-697.

168 Woodard GE, Lin L, Zhang J, Agarwali SK, Marx SJ, Simonds WF. **Parafibromin, product of the hyperparathyroidism-jaw tumor syndrome gene *HRPT2*, regulates cyclin D1/PRAD1 expression**. *Oncogene* 2005, **24**:1272-1276.

169 Lin L, Zhang J, Panicker LM, Simonds WF. **The parafibromin tumor suppressor protein inhibits cell proliferation by repression of the *c-myc* proto-oncogene**. *Proc. Natl. Acad. Sci. USA.* 2008, **105**:17420-17425.

170 Takahashi A, Tsutsumi R, Kikuchi I, Obuse C, Saito Y, Seidi A, Karisch R, Fernandez M, Cho T, Ohnishi N, Rozenblatt-Rosen O, Meyerson M, Neel BG, Hatakeyama M. **SHP2 Tyrosine Phosphatase Converts Parafibromin/Cdc73 from a Tumor Suppressor to an Oncogenic Driver**. *Mol Cell* 2011, **43**:45-56.

171 Tan MH, Morrison C, Wang P, Yang X, Haven CJ, Zhang C, Zhao P, Tretiakova MS, Korpi-Hyovalti E, Burgess JR, Soo KC, Cheah W, Cao B, Resau J, Morreau H, Teh BT. **Loss of Parafibromin Immunoreactivity is a Distinguishing Feature of Parathyroid Carcinoma**. *Clin Cancer Res* 2004, **10**:6629-6637.

172 Kasaian K, Wiseman SM, Thiessen N, Mungall KL, Corbett RD, Qian JQ, Nip KM, He A, Tse K, Chuah E, Varhol RJ, Pandoh P, McDonald H, Zeng T, Tam A, Schein J, Birol I, Mungall AJ, Moore RA, Zhao Y, Hirst M, Marra MA, Walker BA, Jones SJ. **Complete genomic landscape of a recurring sporadic parathyroid carcinoma**. *J Pathol* 2013, **230**:249-260.

173 Williamson C, Cavaco BM, Jauch A, Dixon PH, Forbes S, Harding B, Holtgreve-Grez H, Schoell B, Pereira MC, Font AP, Loureiro MM, Sobrinho LG, Santos MA, Thakker RV. **Mapping the gene causing hereditary primary hyperthyroidism in a portuguese kindred to Chromosome 1q22-q31**. *J Bone Miner Res* 1999, **14**:230-239.

174 Bradley KJ, Hobbs MR, Buley ID, Carpten JD, Cavaco BM, Fares JE, Laidler P, Manek S, Robbins CM, Salti IS, Thompson NW, Jackson CE, Thakker RV. **Uterine tumours are a phenotypic manifestation of the hyperparathyroidism-jaw tumour syndrome**. *J Int Medicine* 2005, **257**:18-26.

175 Domingues R, Tomaz RA, Martins C, Nunes C, Bugalho MJ, Cavaco BM. **Identification of first germline *HRPT2* whole-gene deletion in a patient with primary hyperthyroidism**. *Clin Endocrinol (Oxf)* 2012, **76**:33-38.

176 Haven CJ, van Puijenbroek M, Tan MH, Teh BT, Fleuren GJ, van Wezel T, Morreau H. **Identification of *MEN1* and *HRPT2* somatic mutations in paraffin-**

embedded (sporadic) parathyroid carcinomas. *Clin Endocrinol(Oxf)* 2007, **67**:370-376.

177 Newey PJ, Nesbit MA, Rimmer AJ, Attar M, Head RT, Christie PT, Gorvin CM, Stechman M, Gregory L, Mihai R, Sadler G, McVean G, Buck D, Thakker RV. **Whole-exome sequencing studies of nonhereditary (sporadic) parathyroid adenomas**. *J Clin Endocrinol Metab* 2012, **97**:E1995-2005.

178 Huang DW, Sherman BT, Lempicki RA. **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature Protocols* 2009, **4**:44-57.

179 The Cancer Genome Atlas Research Network. **Comprehensive molecular characterization of urothelial bladder carcinoma**. *Nature* 2014, **507**:315-322.

180 Guo G, Sun X, Chen C, Wu S, Huang P, Li Z, Dean M, Huang Y, Jia W, Zhou Q, Tang A, Yang Z, Li X, Song P, Zhao X, Ye R, Zhang S, Lin Z, Qi M, Wan S, Xie L, Fan F, Nickerson ML, Zou X, Hu X, Xing L, Lv Z, Mei H, Gao S, Liang C, Gao Z, Lu J, Yu Y, Liu C, Li L, Fang X, Jiang Z, Yang J, Li C, Zhao X, Chen J, Zhang F, Lai Y, Lin Z, Zhou F, Chen H, Chan HC, Tsang S, Theodorescu D, Li Y, Zhang X, Wang J, Yang H, Gui Y, Wang J, Cai Z. **Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation**. *Nat Genet* 2013, **45**:1459-1463.

181 Agarwal SK, Schrock E, Kester MB, Burns AL, Heffess CS, Ried T, Marx SJ. **Comparative genomic hybridication analysis of human parathyroid tumors**. *Cancer Genet Cytogenet* 1998, **106**:30-36.

182 Kytola S, Farnebo F, Obara T, Isola J, Grimelious L, Farnebo LO, Sandelin K, Larsson C. **Patterns of chromosomal imbalances in parathyroid carcinomas**. *Am J Pathol* 2000, **157**:579-586.

183 Woodard GE, Lin L, Zhang J, Agarwali SK, Marx SJ, Simonds WF. **Parafibromin, product of the hyperthyroidism-jaw tumor syndrome gene *HRPT2*, regulates cyclin D1/PRAD1 expression**. *Oncogene* 2005, **24**:1272-1276.

184 Lin L, Czarpiga M, Nini L, Zhang J, Simonds WF. **Nuclear localization of the Parafibromin Tumor Suppressor Protein Implicated in the Hyperthyroidism-Jaw Tumor Syndrome Enhances Its Proapoptotic Function**. *Mol Cancer Res* 2007, **5**:193-193.

185 Lin L, Zhang J, Panicker LM, Simonds WF. **The parafibromin tumor suppressor protein inhibits cell proliferation by repression of the *c-myc* proto-oncogene**. *Proc. Natl. Acad. Sci. USA.* 2008, **105**:17420-17425

186 Machida T, Fujita T, Ooo ML, Ohira M, Isogai E, Mihara M, Hirato J, Tomotsune D, Hirata T, Fujimori M, Adachi W, Nakagawara A. **Increased expression of proapoptotic BMCC1, a novel gene with the BNIP2 and Cdc42GAP homology (BCH) domain, is associated with favorable prognosis in human neuroblastomas**. *Oncogene* 2006, **25**:1931-1942.

187 Zhao LR, Tian W, Wang GW, Chen KX, Yang JL. **The prognostic role of PRUNE2 in leiomyosarcoma**. *Chin J Cancer* 2013, DOI: 10.5732/cjc.013.10069.

188 Soh UJ, Low BC. **BNIP2 extra long inhibits RhoA and cellular transformation by Lbc RhoGEF via its BCH domain**. *J Cell Sci* 2008, **121**:1739-1749.

189 Heng YW, Lim HH, Mina T, Utomo P, Zhong S, Lim CT, Koh CG. **TPPP acts downstream of RhoA-ROCK-LIMK2 to regulate astral microtubul organization and spindle orientation**. *J Cell Sci* 2012, **125**:1579-1590.

190 Sanz-Moreno V, Gaggioli C, Yeo M, Albrengues J, Wallberg F, Viros A, Hooper S, Mitter R, Féral CC, Cook M, Larkin J, Marais R, Meneguzzi G, Sahai E, Marshall CJ. **ROCK and JAK1 signaling cooperate to control actomyosin contractility in tumor cells and stroma**. *Cancer Cell* 2011, **20**:229-245.

191 Kimmelman AC, Hezel AF, Aguirre AJ, Zheng H, Paik JH, Ying H, Chu GC, Zhang JX, Sahin E, Yeo G, Ponugoti A, Nabioullin R, Deroo S, Yang S, Wang X, McGrath JP, Protopopova M, Ivanova E, Zhang J, Feng B, Tsao MS, Redston M, Protopopov A, Xiao Y, Futreal PA, Hahn WC, Klimstra DS, Chin L, DePinho RA. **Genomic alterations link Rho family of GTPases to the highly invasive phenotype of pancreas cancer**. *Proc Natl Acad Sci USA* 2008, **105**:19372-19327.

192 Wilkinson S, Paterson HF, Marshall CJ. **Cdc42-MRCK and Rho-ROCK signaling cooperate in myosin phosphorylation and cell invasion**. *Nat Cell Biol* 2005, **7**:255-261.

193 Zhou X, Hua L, Zhang W, Zhu M, Shi Q, Li F, Zhang L, Song C, Yu R. **FRK controls migration and invasion of human glioma cells by regulating JNK/c-JUN signaling**. *J Neurooncol* 2012, **110**:9-19.

194 Garcia J, Sandi MJ, Cordelier P, Binétruy B, Pouysségur J, Iovanna JL, Tournaire R. **Tie1 deficiency induces endothelial-mesenchymal transition**. *EMBO Rep* 2012, **13**:431-439.

195 Roll JD, Reuther GW. **ALK-activating homologous mutations in LTK induce cellular transformation**. *PLoS One* 2012, **7**:e31733.

196 Yu K, Ganesan K, Tan LK, Laban M, Wu J, Zhao XD, Li H, Leung CH, Zhu Y, Wei CL, Hooi SC, Miller L, Tan P. **A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers**. *PLoS Genet* 2008, **4**:e1000129.

197 Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jönsson G, Mariani O, Boyault S, Miron P, Fatima A, Langerød A, Aparicio SA, Tutt A, Sieuwerts AM, Borg Å, Thomas G, Salomon AV, Richardson AL, Børresen-Dale AL, Futreal PA, Stratton MR, Campbell PJ; Breast Cancer Working Group of the International Cancer

Genome Consortium. **The life history of 21 breast cancers**. *Cell* 2012, **149**:994-1007.

198 Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. **Signatures of mutational processes in human cancer**. *Nature* 2013, **500**:415-421.

199 Waters TR, Swann PF. **Thymine-DNA glycosylase and G to A transition mutations at CpG sites**. *Mut Res* 2000, **462**:137-147.

200 Au AY, McDonald K, Gill A, Sywak M, Diamond T, Conigrave AD, Clifton-Bligh RJ. **PTH mutation with primary hyperparathyroidism and undetectable intact PTH**. *N Engl J Med* 2008, **359**:1184-1186.

201 Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, Ponten F. **Towards a knowledge-based human protein atlas**. *Nat Biotechnol* 2010, **28**: 1248-1250.

202 Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. **Synonymous mutations often acts at driver mutations in human cancers**. *Cell* 2014, **156**:1324-1335.

203 Yang YJ, Han JW, Youn HD, Cho EJ. **The tumor suppressor, parafibromin, mediates histone H3 K9 methylation for cyclin D1 repression**. *Nucleic Acids Res* 2010, **38**:382-390.

204 Rozenblatt-Rosen O, Hughes CM, Nannepaga SJ, Shanmugam KS, Copeland TD, Guszczynski T, Resau JH, Meyerson M. **The parafibromin tumor suppressor protein is part of a human Paf1 complex**. *Mol Cell Biol* 2005, **25**:612-620.

205 Yart A, Gstaiger M, Wirbelauer C, Pecnik M, Anastasiou D, Hess D, Krek W. **The HRTP2 tumor suppressor gene product parafibromin associates with human PAF1 and RNA polymerase II**. *Mol Cell Bio* 2005, **25**:5052-5060.

206 Zhu B, Mandal SS, Pham AD, Zheng Y, Erdjument-Bromage H, Batra SK, Tempst P, Reinberg D. **The human PAF complex coordinates transcription with events downstream of RNA synthesis**. *Genes Dev* 2005, **19**:1668-1673.

207 Weinberg R. **Point: Hypotheses first**. *Nature* 2010, **464**:678.

208 Golub T. **Counterpoint: Data first**. *Nature* 2010, **464**:679.

209 Wiles A. **Modular elliptic curves and Fermat's last theorem**. *Annals of mathematics* 1995, **141**:443-551.

210 Doucet AJ, Hulme AE, Sahinovic E, Kulpa DA, Moldovan JB, Kopera HC, Athanikar JN, Hasnaoui M, Bucheton A, Moran JV, Gilbert N. **Characterization of LINE-1 ribonuclearprotein particles**. *PLoS Genet* 2010, **6**: pii:e1001150.

211 Craig NL, Craigie R, Gellert M, Lambowitz AM. **Mobile DNA II**. American Society for Microbiology Press; Washington: 2002.

212 Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH Jr. **Many human L1 elements are capable of retrotransposition**. *Nat Genet* 1997, **16**:37-43.

213 Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr. **Hot L1s account for the bulk of retrotransposition in the human population**. *Proc Natl Acad Sci USA* 2003, **100**:5280-5285.

214 Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, Wheelan S, Upton KR, Shukla R, Faulkner GJ, Largaespada DA, Kazazian HH Jr. **Extensive L1 retrotransposition in colorectal tumors**. *Genome Res* 2012, **22**:2328-2338.

215 Tubio JM, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K, Menzies A, Roman-Garcia P, Fullam A, Gerstung M, Shlien A, Tarpey PS, Papaemmanuil E, Knappskog S, Van Loo P, Ramakrishna M, Davies HR, Marshall J, Wedge DC, Teague JW, Butler AP, Nik-Zainal S, Alexandrov L, Behjati S, Yates LR, Bolli N, Mudie L, Hardy C, Martin S, McLaren S, O'Meara S, Anderson E, Maddison M, Gamble S; ICGC Breast Cancer Group; ICGC Bone Cancer Group; ICGC Prostate Cancer Group, Foster C, Warren AY, Whitaker H, Brewer D, Eeles R, Cooper C, Neal D, Lynch AG, Visakorpi T, Isaacs WB, van't Veer L, Caldas C, Desmedt C, Sotiriou C, Aparicio S, Foekens JA, Eyfjörd JE, Lakhani SR, Thomas G, Myklebost O, Span PN, Børresen-Dale AL, Richardson AL, Van de Vijver M, Vincent-Salomon A, Van den Eynden GG, Flanagan AM, Futreal PA, Janes SM, Bova GS, Stratton MR, McDermott U, Campbell PJ. **Mobile DNA in cancer. Extensive transduction of of nonrepetitive DNA mediated by L1 retrotranposition in cancer genomes**. *Science* 2014, **345**:1251343.

216 Jacobs FM, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. **An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons**. *Nature* 2014, **516**:242-245.

217 Heras SR, Macias S, Plass M, Fernandez N, Cano D, Eyras E, Garcia-Perez JL, Cáceres JF. **The Microprocessor controls the activity of of mammalian retrotransposons**. *Nat Struct Mol Biol* 2013, **20**:1173-1181.

218 Muckenfuss H, Hamdorf M, Held U, Perkovic M, Löwer J, Cichutek K, Flory E, Schumann GG, Münk C. **APOBEC3 proteins inhibit human LINE-1 retrotransposition**. *J Biol Chem* 2006, **281**:22161-22172.

219 Lovsin N, Peterlin BM. **APOBEC3 proteins inhibit LINE-1 retrotransposition in the absence of ORF1p binding**. *Ann N Y Acad Sci* 2009, **1178**:268-275.

220 Horn AV, Klawitter S, Held U, Berger A, Vasudevan AA, Bock A, Hofmann H, Hanschmann KM, Trösemeier JH, Flory E, Jabulowsky RA, Han JS, Löwer J, Löwer R, Münk C, Schumann GG. **Human LINE-1 restriction by APOBEC3C is deaminase independent and mediated by an ORF1p interaction that affects LINE reverse transcriptase activity.** *Nucleic Acids Res* 2014, **42**:396-416.

221 Rodić N, Sharma R, Sharma R, Zampella J, Dai L, Taylor MS, Hruban RH, Iacobuzio-Donahue CA, Maitra A, Torbenson MS, Goggins M, Shih IeM, Duffield AS, Montgomery EA, Gabrielson E, Netto GJ, Lotan TL, De Marzo AM, Westra W, Binder ZA, Orr BA, Gallia GL, Eberhart CG, Boeke JD, Harris CR, Burns KH. **Long intersperced element-1 protein expression is a hallmark of many cancers**. *Am J Path* 2014, **184**:1280-1286.

222 Gossen M, Bujard H. **Tight control of gene expression in mammalian cells by tetracycline responsive promoters**. *Proc Natl Acad Sci USA* 1992, **89**:5547-5551.