The Era of Next-Generation Sequencing in Clinical Oncology

Job van Riet

The Era of Next-Generation Sequencing in Clinical Oncology

Job van Riet

The Era of Next-Generation Sequencing in Clinical Oncology

Het tijdperk van Next-Generation Sequencing in de klinische oncologie

Colofon

ISBN: 978-94-6458-584-1

Cover: Ridderprint | www.ridderprint.nl, Sandra Tukker

Print: Ridderprint | www.ridderprint.nl

The studies described in this thesis were performed within the framework of the Erasmus MC Molecular Medicine (MolMed) Graduate School at the department of Urology, department of Medical Oncology and the Cancer Computational Biology Center of the Erasmus Medical Center, Rotterdam, The Netherlands.

Printing of this thesis was financially supported by the Erasmus University Rotterdam (EUR) and the Department of Medical Oncology of the Erasmus MC Cancer Institute, Erasmus University Rotterdam.

An electronic version of this dissertation is available at
https://repub.eur.nl/.

The Era of Next-Generation Sequencing in Clinical Oncology

Het tijdperk van Next-Generation Sequencing in de klinische oncologie

Proefschrift

ter verkrijging van de graad van doctor aan de

Erasmus Universiteit Rotterdam

op gezag van

de rector magnificus

Prof. dr. A.L. Bredenoord

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

woensdag 16 november 2022 om 10:30 uur

door

Job van Riet

geboren te Soest, Nederland.

Erasmus University Rotterdam

Promotiecommissie

Promotoren:	Prof. dr. R. de Wit
	Prof. dr. ir. G.W. Jenster
Overige leden:	Prof. dr. H.R. Delwel
	Prof. dr. J.B.J. van Meurs
	Prof. dr. V. van Noort
Copromotoren:	Dr. M.P.J.K. Lolkema
	Dr. ir. H.J.G. van de Werken

Table of contents

List of Abbreviations

ADT Androgen deprivation therapy 163

AhR Aryl hydrocarbon receptor 175

AI Allelic imbalances 21, 23, 26, 29, 33, 36, 38, 45, 46

aNEN Locally advanced or metastatic (advanced) neuroendocrine neoplasm 15, 124–128, 130, 132, 137–142, 145, 152, 193, 197, 198

ashr Adaptive shrinkage estimator 181

AU Approximately Unbiased 116

BAF B-Allele Frequency 22, 23, 25–39, 44–57, 96, 97, 99

BAM Binary Alignment Map 96, 179, 181, 193

BER Base excision repair 131, 139

BH Benjamini-Hochberg 85, 115

bp Base pair 96, 102

CDS Coding sequence 69, 71, 72, 89

cfDNA Circulating free DNA 11

CGAs Cancer germline antigens 165, 171, 173, 174

CGI Cancer Genome Interpreter 97

ChIP-seq Chromatin immunoprecipitation sequencing 78, 80, 86, 87, 93, 94, 102–104, 113

CIVIC Clinical Interpretation of Variants in Cancer 97

CN Copy-number 97, 100

CNA Copy-number alteration(s) 11, 80

COSMIC Catalogue Of Somatic Mutations In Cancer 91, 97, 99, 110, 111, 115, 118, 119

CPTAC Clinical Proteomic Tumor Analysis Consortium 62–64

CRPC Castration-resistant prostate cancer 84

ctDNA Circulating tumor DNA 11, 14, 194–196

CTL CD8 T lymphocytes 165, 166, 171, 174, 175, 181

CUP Cancers of unknown primary 14, 195

DC Dendritic cells 164, 168, 175

DEGs Differentially expressed genes 167, 169, 174, 176, 182

DNA Deoxyribonucleic acid 1–3, 5, 7–11, 21, 23, 24, 29, 40, 79, 93, 95–97, 102, 130, 138, 139, 195

DRUP Drug Rediscovery Protocol 12, 192

EGA European Genome-phenome Archive 176, 182

EMT Epithelial-mesenchymal transition 168, 175

ENCODE The Encyclopedia of DNA Elements 7, 12

ETS E26 transformation-specific 82–84, 89–92, 101, 109, 112, 119

FAIR Findability, Accessibility, Interoperability, and Reusability 192

FBS Fetal bovine serum 102

FFPE Formalin-Fixed Paraffin-Embedded 165, 177

FISH Fluorescence *in situ* hybridization 25, 31, 32, 41

GATK Genome Analysis Toolkit 179

GDC Genomic Data Commons 177

GEO Gene Expression Omnibus 104

GEP Gastroenteropancreatic 140

GS Gleason score 103

GSEA Gene-set enrichment analysis 165, 167, 182

GTF Gene transfer format v2 26, 28, 38, 70

HE Hematoxylin-eosin 96

HGP Human Genome Project 5, 12

HMF Hartwig Medical Foundation 16, 192

HRD Homologous recombination deficiency 11, 78, 79, 88, 128, 129, 133, 139, 147

ICI Immune Checkpoint Inhibitors 163, 168, 171, 176

IHC Immunohistochemistry 25, 33, 41, 56

IHW Independent hypothesis weighting 181

InDels Small insertions and deletions 80, 82–87, 89–91, 96–98, 100, 101, 103, 104, 107, 108, 112, 114, 119, 120

IP Immunoprecipitation 102

IQR Interquartile range 81, 82, 85, 107, 108, 110, 115, 126, 153

kbp Kilo-base pair 86, 88, 104, 114, 119, 120

KWF Koningin Wilhelmina Fonds voor de Nederlandse Kankerbestrijding 192

LAF Lower allele-frequency 117, 156

log$_2$FC Log$_2$ fold-change 167, 169, 181, 182, 188

LOH Loss of heterozygosity 21, 29, 33, 34, 36–38, 47, 56, 57, 197, 199

lsei Least Squares with Equality and Inequality 180

Mbp Mega-base pair 82–85, 87, 97, 98, 100, 104, 110, 113, 114, 117, 120

mCRPC Metastatic castration-resistant prostate cancer 15, 16, 78, 80, 82–87, 90–95, 100, 103, 104, 107–110, 113, 115, 117–120, 163, 174, 193, 197

MDSC Myeloid-derived suppressor cells 164, 175

METC Medical ethical committee 95

MMR Mismatch repair 94

MNV Multi-nucleotide variation(s) 80, 82, 83, 86, 87, 90, 91, 96–98, 100, 101, 107, 108, 112, 114, 119

mRNA Messenger RNA 7, 92, 119

MSI Microsatellite instability 11, 78, 83, 86–88, 91, 92, 94, 98, 102, 112, 119, 164

NAP Normal-adjacent prostate 16, 164–180, 186, 188, 189

ncRNA Non-coding RNA 12

NEC Neuroendocrine carcinomas 123–125, 131, 138, 140, 198

NEN Neuroendocrine neoplasms 123, 133, 135, 137, 140, 142, 145

NES Normalized Enrichment Scores 169

NET Neuroendocrine tumors 123–125, 131, 132, 137, 140, 198

NGS Next-Generation Sequencing 11–14, 16, 21–23, 26, 29, 33, 38, 40, 41, 61, 191, 194, 195, 197, 199

NK Natural killer 164, 168, 175

NMF Non-negative matrix factorization 111

OLO Optimal leaf ordering 90, 100, 120

OR Objective response 163

OS Overall survival 171, 190

PC Principal component 114, 120

PCA Principal component analysis 114, 120

PCAWG Pan-Cancer Analysis of Whole Genomes 12, 123, 192

PCR Polymerase chain reaction 3

PFS Progression-free survival 171, 190

pNET Pancreatic neuroendocrine tumors 123, 124, 140

PON Panel of normals 178, 179

PPAR Peroxisome proliferator-activated receptor 167, 174

PRAD (localized) prostate adenocarcinoma 16, 163–179, 186, 188, 189

RNA Ribonucleic acid 3, 5, 7, 8, 11

SBS Single-base substitution 129, 130, 146

SCC Squamous cell carcinoma 35, 37

SI-NET Small intestine NET 123, 124

SJ Splice-junctions 74

SNP Single-nucleotide polymorphism 21, 103

SNV Single-nucleotide variation(s) 80, 82–87, 89–91, 96–101, 103, 104, 107, 108, 110–112, 114, 119, 120

SV Structural variant(s) 80, 82, 83, 86, 89, 100, 101, 112, 114, 120

t-NEPC Treatment-induced neuroendocrine prostate cancer 195

TAMs Tumor-associated macrophages 163, 164, 175

TC Tumor cell percentage 117

TCGA The Cancer Genome Atlas 12, 88, 92, 101, 119, 192

TCR T cell receptor 165, 166, 168, 171, 172, 174, 175

Ti Transition 108, 110

TILs Tumor infiltrating lymphocytes 163, 164, 168, 170

TMB Tumor mutational burden 9, 82–84, 86, 87, 89–93, 103, 112, 119, 120, 127, 134, 138, 139, 141, 163, 197, 198

TME Tumor microenvironment 14–16, 175, 176, 194, 195, 198, 200

TPM Transcripts per million 178, 180

Tv Transversion 108, 110

UCSC University of California, Santa Cruz 63, 96, 97

UTR Untranslated region 7, 72

UV Ultraviolet light 9

VCF Variant call format 22, 23, 25, 26, 28, 38, 40, 63, 69, 70

VEP Variant Effect Predictor 179

VST Variance-stabilizing transformation 173, 179, 180, 187

WES Whole-exome sequencing 12, 26, 29, 36, 38–40, 123, 124, 193, 195

WGD Whole-genome duplication 128, 133

WGS Whole-genome sequencing 12, 16, 26, 38, 78–81, 84, 86, 87, 92–95, 100, 101, 104, 115, 116, 123, 124, 141, 148, 152, 192, 193, 195, 197–200

WIDE WGS Implementation in the standard Diagnostics for Every cancer patient 12, 192

wss Within-cluster sum of square 116

Chapter 1

Introduction

The Whole is Greater than the Sum of its Parts

Attributed to Aristotle

Brief history of the gene

The hereditary instructions for the development, direction and maintenance of a cellular organism are encoded within the deoxyribonucleic acid (DNA) of the species. This genetic blueprint is encased within its double helical molecular structure, as revealed and made famous by Watson & Crick (as inspired by the work of Rosalind E. Franklin) in 1953.[1] The DNA resides within the nucleus of the cell as compact structures termed *chromosomes*.

The chemical makeup of DNA, and thus the hereditary and biochemical properties of terrestrial life, is derived from only four basic constituents known as *nucleotide bases*. These nucleotides can be recognized as adenine, thymine, cytosine and guanine, shortened and canonically denoted by their characteristic acronym as A, T, C, G respectively (Figure 1.1).

The order of these nucleotides (A, T, C and G) determines the genetic messages which are to be followed and carried out by the complex molecular machinery of the cell. The DNA consists of two intertwined strands, each strand recognizable by the orientation of the nucleotides in regard to the phosphate backbone of the DNA; going up-/ or downstream the rigid backbone (5′ -> 3′). Nucleotide bases on opposite strands are paired in complementary fashion, an adenosine (A) is always paired with a thymine (T) and each cytosine (C) is similarly paired with a guanine (G) through hydrogen bond interactions.

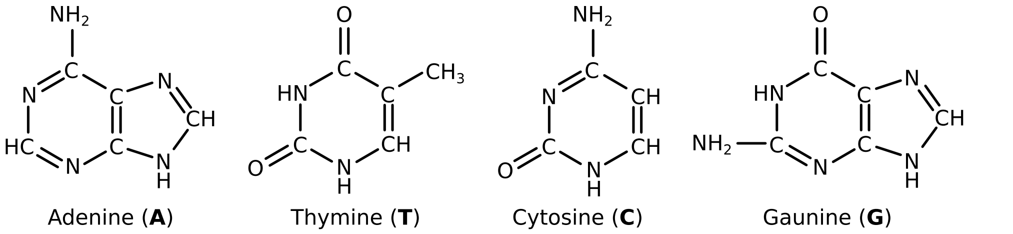


Figure 1.1: The major building blocks of life and members of the nucleotide family: adenine (A), thymine (T), cytosine (C) and guanine (G).

Prior to the characterization of the physical structure of DNA, the scientific community already possessed extensive hypotheses and models of the hereditary nature of phenotypical traits observed in life, within both animal and plant kingdoms. Unrecognized for many years after its initial publication in 1865 and rediscovered only ~40 years later, the experiments into the proposed patterns of inheritance within the common garden pea (*Pisum sativum*) by Johann Gregor Mendel captured much of the abstract foundations of modern genetics, including the description of exchangeable "*Zellelemente*" as minute and abstract units of inheritance; later these elements would be redefined into our current understanding of *genes*.[2] However, the first usage of the term "*gene*" would only be seen much later in the work of Hugo de Vries in 1901, proposing that mutations in *pangenes* were the drivers of genetic diversity and the possible origin of species.[3] In 1909, Wilhelm Ludvig Johannsen extended upon this reasoning and proposed the term *gene* to describe a more exact definition of these units of inheritance in regards to phenotypical changes relating to underlying genotypical changes within species.[4] The whole of these genetic messages within the species is termed the *genome*, containing all DNA with its underlying genes and genetic information.

Following the notion of Mendel's observation on his Second Law (inherited traits are able to segregate independently), Walter S. Sutton (in 1903) discovered that the inheritance of genes was in close relation to the outcome of chromosomal segmentation during cell division, leading to the first observations that certain genes are harbored on specific chromosomes.[5] Several years later in 1910, utilizing the genetic model of the fruit fly (*Drosophila melanogaster*), Thomas Hunt Morgan and colleagues discovered that genes indeed lie upon fixed positions within chromosomes (in this case sex-linked) and following this discovery, further employed the fruit fly to publish the first-ever genetic map detailing the chromosomal location of several genes within its genome.[6,7]

With these observations, it also became evident that the number of genes is vastly greater than the number of corresponding chromosomes. Early cytogenetics revealed the distinctive karyogram of the diploid human chromosomal landscape, 22 autosomal chromosomes, denoted based upon decreasing chromosomal length, and two allosomes (XX for females and XY for males). Deviations from this canonical chromosomal pattern within the parental germ-cells (giving rise to the zygote) or aberrations during embryogenesis are linked to a wide range of genetic and phenotypic abnormalities within individuals. A small overview of such common genetic disorders due to chromosomal aberrations is given in Table 1.1.

Since then, advancements in molecular techniques and technological innovations have elucidated much of the complex molecular mechanics and interplay of cellular machinery driving genetic inheritance and messaging.[20,21] Several major advances which aided in revealing the genetic code of life include several Nobel-prize winning works within the field of chemistry, physics and physiology or medicine. In 1957, using *Escherichia coli* models, Arthur Kornberg and colleagues discovered the family of enzymes (DNA polymerase) involved in DNA replication[22] and utilized these DNA polymerases to invent various supporting molecular techniques to ultimately decipher the ribonucleic acid (RNA) codon table. Har Gobind Khorana and colleagues synthesized the first oligonucleotides, and in 1976, the first synthetic gene.[23] Discoveries of more accurate and thermally-stable DNA polymerases within other species, such as *Thermus aquaticus* (*Taq*), and modifications to improve the replicative potential of these enzymes allowed for the invention of several key sequencing principles. In 1977, Frederick Sanger and colleagues revealed their work on a DNA sequencing technique revolving around the selective incorporation of chain-terminating dideoxynucleotides to sequentially determine the nucleotide sequence of a given DNA molecule, known as Sanger sequencing.[24] Subsequently, they used this technique to fully characterize the first DNA-based genome, that of the bacteriophage *ϕX174* (*PhiX*).[25] However, these techniques required the input of large quantities of DNA molecules for accurate detection which warranted extensive time and effort to quantify and isolate. This issue was alleviated in 1983 by Kary Mullis and colleagues with the invention of polymerase chain reaction (PCR); a rapid and accurate DNA template replication process which is still the *de facto* method for producing the large concentrations of pure DNA necessary for sequencing. During these major discoveries which led to more accessible and automated sequencing approaches, the first complete protein-coding gene sequence to be revealed (through nuclease digestion of the respective RNA molecule and subsequent

Table 1.1: Overview of common genetic diseases associated with large-scale germline chromosomal aberrations. Mean prevalence per 10.000 newborns and with 95% confidence interval given (if available).

Disease	**Prevalence**	**Symptom(s)**	**Chromosomal aberration(s)**
Klinefelter syndrome	17 in male births (14-20; metastudy)[8]	Among others; Genital abnormalities, hypogonadism and infertility.	XXY aneuploidy[9,10]
Down's Syndrome	13.83 (13.63, 14.03; US)[11]; 14.57 (14.43, 14.73; NL)[12]	Among others; intellectual disability, developmental delays, hypotonia, heart and gastrointestinal disorders and craniofacial abnormalities	Trisomy 21[13]
Turner's Syndrome	4 in female births (DE)[14]	Among others; Development disorders incl. ovarian failure, infertility, osteoporosis, hypothyroidism, and renal and gastrointestinal disease	Structurally abnormal X chromosome, monosomy X or 45,X/46,XY mosaicism.[14,15]
Edwards Syndrome	2.34 (2.26, 2.42; US)[11]	Among others; intellectual disability, heart and gastrointestinal disorders, increased risk of certain types of cancers and craniofacial abnormalities	(Mosiac) Trisomy 18[16,17]
Patau syndrome	1.08 (1.02, 1.13; US)[11];	Among others; intellectual disability, developmental delays, hypotonia, heart and urogenital disorders and musculoskeletal / craniofacial abnormalities	Trisomy 13[16]
Cri du chat Syndrome	.6 (JP)[18]	Among others; intellectual disability, craniofacial abnormalities and characteristic cat-like crying	Partial or complete deletion of 5p.[18,19]

separation by electrophoresis) was that of the small bacteriophage *MS2* in 1972, which was shortly expanded upon with the first fully sequenced genome (constituting 3569 bases and present as single-stranded RNA molecule) in 1976, both by the laboratory of Walter Fiers within the university of Gent.[26,27] These prior efforts were pivotal to the future of modern genetics in revealing the essence of genetic mapping but showed that much manual labor was required to genotype only few genes and/or small genomes.

Advancements within the molecular and technical instruments required to simultaneously sequence large batches of DNA, allowed for the promise of fully sequencing and investigating larger genomes; including the full genetic sequence of man. The Human Genome Project (HGP), the largest international scientific research project to date, sought to fully determine every single nucleotide of the human genome. The HGP, initiated in 1990, revealed the first draft version of the human genome on June 26, 2000 and provided a more finalized human genome on April 14, 2003. The total cost of this enormous project is estimated to be around 2.7 billion U.S. dollars. This huge collaborative effort has sparked much technological and biological innovation and is to-this-day paramount to many current landmark studies and routine diagnostics.[28,29] Continued research by the HGP and many laboratories around the world has yielded a complete human genome (currently version GRCh38.p13) which is used as a healthy reference genome to detect genomic abnormalities within patients suffering from a wide scale of genetic diseases. This genome consensus has been assembled from the DNA derived from the white blood cells of four randomized healthy individuals (two male and two females). The current draft of the human reference genome is ~3.1 billion nucleotides in length and contains 19,982 protein-coding genes (with experimental evidence), according to the GENCODE consortium (v33)[30]. Only several challenging repeat-like genomic regions and correct placement of several contigs are left in revealing the complete genetic code of man. However, the current genomic sequence is more than sufficient in serving as a critical and high-quality reference in distinguishing functions and clinically-relevant mutations within genetic disease and malignancies. A small overview of common germline single-gene disorders, in which genes only slightly deviate from this reference genome due to small single base-substitution or insertion/deletion mutations, is given in Table 1.2.

To further underscore the scientific and societal importance of unobstructed access to this resource, the United States Supreme Court (2013; Association for

Table 1.2: Overview of common genetic diseases associated with base-substitution or insertion/deletion mutations within single genes. Mean prevalence per 10.000 newborns was retrieved from Orphanet[31] on May 31st 2020.

Disease	Prevalence	Symptom(s)	Affected gene	Inheritance
Cystic Fibrosis	0.1 - 0.9	Chloride impermeability leading to (hyper)production of viscid mucus leading to progressive respiratory and digestive damage	Cystic fibrosis conductance transmembrane regulator (*CFTR; 7q31*)	Autosomal recessive
Sickle-cell anemia	1-5	Anemia, bacterial infections and vaso-occlusive crisis	Beta hemoglobin (*HBB*; 11p15)	Autosomal recessive
Huntington's disease	0.1 - 0.9	Neurodegenerative disorder of the central nervous system characterized by unwanted choreatic movements, behavioral and psychiatric disturbances and dementia.	Huntingtin (*HTT*; 4p16)	Autosomal dominant
Autosomal dominant polycystic kidney disease	1-5	Development of multiple cysts within the kidney leading to a range of renal complications	Polycystic kidney disease 1 (*PKD1*; 16p13) and polycystic kidney disease 2 (*PKD2*; 4q22)	Autosomal dominant
Phenylketonuria	1-5	Intellectual disability, developmental delays and motor-related disorders	Phenylalanine hydroxylase (*PAH*; 12q22)	Autosomal recessive
Fabry disease	1-5	Multisystemic lysosomal storage disease leading to accumulation of sphingolipids	Alpha-galactosidase A (*GLA*; Xq21)	X-linked recessive
Tay-Sachs disease	<1	Accumulation of G2 gangliosides due to hexosaminidase A defiency, leading to progressive neurodegradation.	Hexosaminidase A (*HEXA*; 15q23)	Autosomal recessive
Duchenne muscular dystrophy	0.1 - 0.9	Rapidly progressive muscular weakness due to degeneration of skeletal, smooth and cardiac muscle.	Dystrophin (*DMD*; Xq21)	X-linked recessive

Molecular Pathology v. Myriad Genetics, Inc.) ruled that naturally occurring human genes are not an invention and therefore cannot be patented; ensuring that no single individual, company, nation or country can make claim to this resource.

In similar fashion, large collaborative genome-related efforts such as The Encyclopedia of DNA Elements (ENCODE) project have mapped many genetic regulators, such as proximal and distal regulatory elements which bind to the DNA based on sequence-contexts (e.g., POL2RA, EZH2 and SETDB1), promoter activities (e.g., *H3K27ac*), and chromosomal interactions.[32] All this research and data have been made available to other researchers to further our understanding of the fundamentals of the human genome and their relationship to genetic diseases.

The central dogma of molecular biology

The eukaryotic DNA is comprised out of a myriad of genetic elements including *cis/-trans*-acting elements, genes, introns, exons, enhancers, motif-sites, centromeres, telomeres and many others; each with their own characteristic function and essential purpose. Genes are transcribed into RNA molecules through the elaborate process of transcription, and subsequently, messenger RNA (mRNA) molecules are followed by translation into amino-acid structures termed *proteins*. All cells within the species harbor near-identical DNA, yet based on their localization, environment and function, differently regulate and transcribe distinct genes through molecular mechanisms affecting their transcriptome. The general and basic structure of a gene, as exists in the human genome, consists out of several of these genetic elements. Such genes consist out of one or multiple exons containing the protein-coding sequence(s) and are interspersed by non-coding sequences (introns). In addition, the starting and terminal exon(s) contain non-coding sequences known as untranslated region (UTR). These non-coding sequences will not be incorporated into the final protein sequence and serve other purposes, such as regulatory roles and to allow alternate conjugation of exonic sequences (rather than only the linear follow-through; *alternative splicing*), which greatly expands the number of protein configurations (isoforms) derived from a single gene.

Within eukaryotes, the transcription of DNA into mature messenger RNA capable of subsequent translation is facilitated through an intricate and efficient multi-step process.[33] Briefly, the canonical mRNA transcription process is facilitated by RNA polymerase II, as promoted by one or more transcription factors, which generates a complementary RNA molecule based upon the DNA template. Post-transcriptional modifications further ready these pre-messenger RNA molecules for export out of the nucleus and subsequent translation, most commonly through capping, polyadenylation and splicing. At the 5′-side of the pre-mRNA molecule, a 7-methylguanylate cap (m^7G) is attached serving multiple functions: 1) nuclear exportation and further processing through interactions with the nuclear cap-binding complex, 2) recruitment of the 43S pre-initiation complex through interactions with the 40S ribosomal subunit, 3) prevents endonucleolytic cleavage, and 4) assisting in the excision of the 5′ proximal intron through splicing.[33–36] At the 3′-side of the pre-mRNA molecule, additional adenine nucleotides are attached to generate the polyadenylate (poly(A)) tail. This poly(A)-tail serves to stabilize and protect the RNA molecule from degradation.[33] After these post-transcriptional modifications, the mature messenger RNA is capable of being translated into proteins by the ribosomal machinery and facilitating factors.

These processes allow the human genome to produce a great arsenal of RNA molecules and proteins which maintain and propagate cellular life; an arsenal even greater than the significant number of genes present on the genome.

Cancer: malignancy of the tissue

Repair and maintenance of the proper state of genes and cellular function(s) is essential to all cellular life to enable the correct transfer of genetic instructions throughout life and evolution. Spontaneous (or driven) mutational processes within the somatic cells of an individual may give rise to an malignancy of the tissue; known as cancer. The disease manifests itself as an uncontrolled spread and malignant transformation of cells, both within and beyond its primary site of origin; as made migratory through the blood and lymphoid systems. These uncontrolled clusters of malignant cells hijack the vital resources necessary for proper organ functionality, leading to disruptions within the careful equilibrium of the healthy cellular systems and ultimately progressing to organ failure or otherwise fatal conditions.

Cancer is the second leading cause of death (world-wide), responsible for an estimated 9.6 million deaths in 2018 and surpassed only in incidence by heart diseases.[37–39] With an estimated 18.1 million new cases each year (and rising), both clinical and fundamental research into the underlying molecular biology, diagnostics and treatment of this malignancy is worthwhile.

As summarized in tables 1.1 and 1.2, genetic disorders rarely deviate from the canonical genomic status but, instead, stretch the extent of healthy genetic makeup due to strict cellular regulation upon the embryonic and fetal gestation process. Malignant cells however, have acquired several key principles which evade and manipulate these protective cellular processes. These hallmarks include sustained proliferation, evasion of growth suppressors, replicative immortality, induced angiogenesis, resistance to apoptotic processes, promotion of supportive micro-environments, metabolic rewiring, immune modulation and acquirement of invasive and metastatic potential.[40,41] In addition, certain tumors (e.g., prostate cancer) exhibit extensive genomic aberrations which are only rarely seen in germline diseases, such as catastrophic chromosomal re-arrangements leading to chromoanagenesis (chromothipsis, chromoplexis and chromoanasynthesis).[42–44]

Many of these hallmarks benefiting the evolutionary progression towards malignancy have been acquired by somatic alterations accrued within the human genome; twisting and (re-)activating the genetic harbingers of cellular instruction. Genomic alterations can arise from various internal and external origins and can accrue over time if left uncorrected and without penalty. These alterations can arise from spontaneous events due to cellular aging and common errors during routine processes such as DNA replication or mis-repair, by enzymatic induction (e.g., APOBEC activity), or by environmental/chemical induction through stimuli such as carcinogens and radiation (e.g., from ultraviolet light (UV)).[45,46] The minimal number of genomic mutations within coding regions required for the malignant formation of tumors within primary lesions is observed to be dependent on the tissue and site of origin.[47] The median tumor mutational burden (TMB) of bone marrow myelodysplastic syndrome is observed to be as little as 0.8 (0.8 somatic mutation(s) per coding megabase) whilst the median TMB of skin melanoma is observed to be as high as 14.4.[47]

The genetic (mis-)instructions contributing to the malignant progression of cells ranges per primary site and tissue of origin, with specific alterations seen mostly only within certain tissues; e.g., the *TMPRSS2-ERG* gene-fusion event within malignant prostate tissue and observed within 50% of prostate adenocarcinoma.[48,49] The identification and experimental validation of key recurrent somatic mutations within genes benefiting the evolutionary trajectory of malignant cells have yielded large sets of cancer-associated genes.[50–53] These driver genes can be categorized into two categories; (proto-)oncogenes and tumor suppressors.[54] In this distinction, (proto-)oncogenes are those genes that stimulate cell-growth, division and survival and which accrue somatic mutations which alter their proper operation(s). Adversely, tumor suppressor genes serving the prevention of malignant progression are often inactivated entirely. In addition to DNA-mediated (mis-)instructions, epigenetic changes affecting the chromatin state can also disrupt the careful equilibrium which modulates the underlying transcription of genes and can thereby promote malignant progression and/or differentiation.[55,56]

Depending on the primary site of origin, time of clinical diagnosis, existing treatment options and overall health of a patient, the 5-year survival outcomes differs greatly between malignancies. Improvements in the diagnosis, treatment and prevention of localized disease is steadily increasing the 5-year survival rate of cancer patients.[37–39] As the diseases progresses, the leading cause of cancer-related death is attributed to the undisturbed spread of malignant cells beyond their primary site to distant nodes; known as *metastasis*.[57] Death following metastatic progression accounts for roughly 66 percent (and possibly upwards to 90%) of all cases.[58] Exceptions of fatal primary disease are often restricted to malignancies which are particularly difficult to detect early and are often only noticeable at later and advanced stages, such as pancreatic or central nervous system malignancies.[59]

The constitution of malignant cells within a tumor is heterogeneous as these have not all propagated from the same parental lineage, leading to several distinct clonal populations within the tumor; each with their own diverging and malignant path of tumor evolution.[60] This tumor heterogeneity can be evidenced by distinct somatic aberrations observed only within clonal fractions of the malignant population. Conversely, clinical treatment of tumors can give rise to certain subclones which have evolved (by random mutagenesis) and have the evolutionary advantage of becoming (more) resistant to the treatment given. This field of research is slowly being advanced by the introduction of single-cell sequencing techniques which captures more fully this heterogeneity, yet the interpretation and possible effects on clinical decision-making are still undergoing.[61]

The underlying biology of these malignancies is slowly being unraveled and taken advantage of in novel therapies, yet much remains to be explored. Recent discoveries and diagnosis are closely tied with the advancement of new or improved molecular methodologies and sequencing techniques, together with the experience of interpreting these results, and it stands to reason that even more biological processes will be elucidated in coming years. This advancement will be, in part, made possible by the availability of large sequencing data-sets which increase our statistical power to detect rare aberrations and biological mechanisms.

Next generation sequencing in oncology

Detection of somatic mutations within a tumor genome through Next-Generation Sequencing (NGS) of DNA reveals the evolutionary history detailing the malignant progression and cellular origin.[62] Conversely, whole-transcriptome and epigenetic analyses allow for supplemental examination of the cellular origin and present state of the cell. These NGS techniques allow for personalized diagnosis and putative treatment options. The shift from the 'one size fits all' treatment paradigm to more personalized approaches, utilizing prognostic and predictive biomarkers, can prevent unnecessary costs due to inappropriate therapy and help reduce treatment-related toxicity. In addition, this could extend the range of putative therapies for late-stage metastatic disease on a per-patient basis.[63,64]

Applications of NGS can reveal the tumor heterogeneity, characterize microsatellite instability (MSI), homologous recombination deficiency (HRD), regional hypermutation (kataegis) and key cancer-related somatic aberrations, including structural rearrangements and copy-number alteration(s) (CNA), nucleotide substitutions and small insertions and deletions within specific genes or regulatory elements. Analysis of RNA sequencing furthermore reveals abnormal expression or modifications within the transcriptome, including biomarkers distinguishing healthy from malignant tissues.

With increasing reports of genetic components associated with genetic disease, it has become routine to perform targeted genome profiling on sets of *a priori* clinically-relevant genes within patients, such as common drivers in cancer (e.g., *TP53*, *ERBB2*, *MET*, *BRCA2*, *KRAS*, *SF3B1*, *PTEN*, *MSH2*) or those associated with epilepsy (e.g., *SOX6*, *PICK1* and *SLC1A3*).[52,65,66] Recent advances in isolating circulating tumor DNA (ctDNA), circulating free DNA (cfDNA) and exosomes secreted from cancer cells within peripheral blood even allow for the non-invasive detection, classification and monitoring of (early-stage) malignancies and prenatal genetic diseases.[67–70]

With the arrival of more cost-friendly, parallel, and sensitive second- and even third-generation sequencing techniques and facilitating platforms, the cost for sequencing an individual's entire exome or genome is steadily decreasing.[20,21] However, additional costs such as the storage, computational processing and trained personnel makes whole-exome sequencing (WES) and whole-genome sequencing (WGS) still primarily worthwhile for cancer research purposes and remains out of reach for routine diagnostics.[71] Recent and current studies, such as Drug Rediscovery Protocol (DRUP), WGS Implementation in the standard Diagnostics for Every cancer patient (WIDE) and CPCT-02 studies, are testing the feasibility of performing and interpreting WGS to broaden therapeutic options and clinical outcome for cancer patients within the Netherlands.[53,64]

Landmark initiatives such as the HGP and ENCODE have inspired similar collaborative efforts within the field of oncology to pool together their resources and available tumor (and matched normal) tissues to generate large and uniform NGS data-sets which span both localized and more recently, metastatic malignancies. An overview of several of these major publicly-available cancer cohorts is given in table 1.3.

Table 1.3: Overview of major publicly-available next-generation sequenced cancer cohorts, with an estimated number of unique samples as of Mar. 2020.

Cohort	**Disease Stage**	**Sequencing Focus**	**# Samples**
TCGA	Localized disease	Exome, Transcriptome and Methylome	10.511[51]
PCAWG	Localized disease	Genome	2778[72]
CPCT-02	Metastatic disease	Genome, Transcriptome	3953[53]

These large data-sets already harbor the key to uncovering novel genes, aberrations, and biological mechanisms relating to cancer biology, including several regulated by events within the non-coding regions of the genome.[51,53,72] Likewise, the molecular complexity of the disease is ever increasing with observations that also non-coding RNA (ncRNA) play critical roles through recurrent somatic mutation, relocation and deregulation.[73,74]

The availability of large uniform NGS data-sets enabled the discovery of distinct genome-wide mutational signatures which could be associated with genomic stress, somatic variation, enzymatic activities, given treatments and cellular aging.[75,76] Computational procedures to deconvolute and annotate these mutational signatures quickly became available and allowed for detection of these biologically and clinically-relevant characteristics within single tumor genomes.

These large cohorts and the broad applications of existing and upcoming NGS are yet to be fully explored and will undoubtedly grant new insights into the biological intricacies of these malignancies and will allow for novel approaches of battling this dreadful malady which afflicts an ever-growing number of people.

The rise of computational biology

The umbrella term of ”computational biology” or ”bioinformatics” signifies a rather broad field of closely-related scientific focuses and interdisciplinary skills. These terms could be applicable within the evolutionary sciences, -omics sciences or any such scientific discipline focusing on the computational analysis of large-scale datasets or when a systematic approach (i.e., automated or scripted) is warranted. This role has historically been attempted by researchers taken up the additional mantle of data-scientist next to their other multitude of responsibilities. However, the sheer data-deluge of current-day enormous sequencing efforts and likewise ambitions[53,72,77–79] require sophisticated, structured and documented computational workflows coupled with sufficient *fingerspitzengefühl* for accurate interpretations and reproducible results. As the complexity and need for such workflows and technical requirements are increasing, the need for dedicated staff and centers to facilitate the storage, computational power and analysis of large-scale and in-depth research is expanding. In turn, this led to distinct and full-time roles for bioinformaticians to bridge the fields of (molecular) biology and computer science.

The importance of bioinformatics within current-day science is noteworthy, with Wren et al. (2016)[80] highlighting that over one third (34%) of the most-cited scientific papers relate to bioinformatics. These fundamental works delve in such topics as sequencing alignment[81,82], germline and somatic variant callers[83,84], prediction of 2D/3D molecular structures[85], local sequence similarities[86–88], phylogenetic reconstruction and the conceptualization of accompanying statistical methods such as bootstrapping techniques[89] and large-scale collaborations and subsequent databasing[30,32,90]. A recent major interest (and hype) has been placed onto the implementation of machine-learning based methods to aid the automated image-based classification of tissue-slides or medical scans and for the recognition of complex patterns underlying gene-expressions by utilizing artificial neural networks and feature extraction methods.[85,91,92]

As such, modern medicine is intertwined with the use of NGS and the accompanying bioinformatics for the daily operations of molecular diagnostics and clinical decision-making.[93] The current-day costs of NGS coupled with optimized workflows allows for the discovery of the genetic layout and drivers underlying patient-specific disease(s) and can thereby provide additional options such as personalized medicine[64,94,95], monitoring of the mechanisms of treatment-resistance[96], detection of viral integration and components[97] and can provide extensive molecular classification which is even capable of revealing the likely tissue-of-origin for cancers of unknown primary (CUP)[98].

Recent technical innovations such as single-cell sequencing and the increased utility of non-invasive collection and sequencing of ctDNA have opened new promising avenues for the (longitudinal) monitoring and interrogation of the complex and dynamic clonal interactions of malignancies, coupled with extensive interrogation of the tumor microenvironment (TME). Due to the even-greater data-deluge and intricacies of such biological investigation, these new avenues are again paved with the intrinsic dependency on computational biology and will likely spark the next era of sequencing in the field of oncology.

Scope of this thesis

Cancer is a malignant state of the tissue which has become errant and unresponsive to the internal and external checkpoints maintaining the otherwise intrinsic and tightly-regulated processes of DNA replication, repair, and division. This malignant state is greatly orchestrated and maintained through deregulated harbingers of genetic information known as oncogenes whilst silencing the tumor suppressors and its guardian roles.[40] Therefore, a potential remedy or inhibition of this malignant state lies in uncovering the complex interplay between the genetic template (genotype) and its malignant representation (phenotype) whilst also taking mind of the dynamic interaction with the surrounding TME and (treatment-driven) clonal evolution. Using a variety of molecular techniques, we can already exploit these malignant hallmarks of cancer by utilizing a wide range of genetic elements or features which are unique, absent or over-represented within malignant tissue. Discovering these distinct features allows us to perform molecular diagnosis and classification of current and retrospective disease-burdens and to deduce potential patient-specific treatment options in order to improve overall survival and quality of life for patients.

Within this thesis, we set out to design open-source software and algorithms to unburden the processing and interpretation of the large quantities of biological data derived from molecular diagnosis and experimental setups. This biological data can range from limited targeted panels of *a priori*-selected known oncogenes and tumor suppressor genes to the massive data-deluge of modern-day whole genome and transcriptome sequencing approaches. With the improved accuracy and volume of clinically-relevant somatic markers, we set out to increase the ease of interpreting such genomic markers for daily molecular diagnostics purposes. In addition, due to the increased volume of detectable somatic aberrations, we set out to provide an accurate and robust approach to translate genomic aberrations into it's respective protein sequence variant(s) to improve the detection and quantification of (potentially immunogenic) protein-variants unique to certain malignancies and genotypes.

To better understand the scores of genomic aberrations underlying the continuation or progression of malignancies and the divergent paths to treatment-resistance(s), we sought to interrogate the somatic inventories of two large-scale cohorts of whole-genome sequenced metastatic castration-resistant prostate cancer (mCRPC) and locally advanced or metastatic (advanced) neuroendocrine neoplasm (aNEN) for new potential avenues of patient-specific treatment; as make possible through the combined and massive effort of the CPCT-02 study and Hartwig Medical Foundation (HMF). As WGS allows for the interrogation of the non-coding genome, we also sought to investigate the presence of recurrent non-coding aberrations driving castration-resistance in mCRPC.

To further investigate potential treatment strategies for (localized) prostate adenocarcinoma (PRAD), we sought to investigate the as-of-yet unknown roles of the transcription factor *ERG* regarding immune-related mechanisms such as immune evasion or suppression or altered dynamics of the TME; in comparison to normal-adjacent prostate (NAP) tissues. Whilst the genomic fusion between *TMPRSS2* and *ERG* is a prevalent somatic event in PRAD (~50% of cases), any major significance regarding overall survival or treatment-strategies remains lacking. As the transcriptomic and epigenetic landscape of *TMPRSS2-ERG*$^+$ PRAD differs significantly from it's *TMPRSS2-ERG*$^-$ PRAD counterpart[99], differences in regards to immune-regulatory systems could provide evidence for clinical impact and immune-based therapies in PRAD.

As evidenced by the introduction and scope of this thesis, the sheer utility of NGS allows us to delve into many scientific inquiries still left unanswered in our battle against this dreadful malady known under the common moniker of cancer.

References

[1] J. D. Watson and F. H. C. Crick, *Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid,* Nature **171**, 737–738 (1953).

[2] G. Mendel, *Versuche über pflanzen-hybriden,* Verhandlungen des naturforschenden Vereines in Brünn. **Bd.4 (1865-1866)**, 3 ((1865-1866)).

[3] H. De Vries, *Die mutationstheorie. Versuche und beobachtungen über die entstehung von arten im pflanzenreich,,* Vol. Bd.1 (1901) (Leipzig,Veit & comp.„ 1901) p. 684, https://www.biodiversitylibrary.org/bibliography/11336 — "Literatur": v. 2, p. [715]-717. — 1. bd. Die entstehung der arten durch mutation.–2. bd. Elementare bastardlehre.

[4] W. Johansen, *Elemente der exakten erblichkeitslehre,* Zeitschrift fur Induktive Abstammungs- und Vererbungslehre **2**, 136 (1909).

[5] W. S. Sutton, *The chromosomes in heredity,* The Biological Bulletin **4**, 231 (1903).

[6] T. H. Morgan, *Chromosomes and heredity,* The American Naturalist **44**, 449 (1910).

[7] A. H. Sturtevant, *The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association,* Journal of experimental zoology **14**, 43 (1913).

[8] R. H. Scofield, G. R. Bruner, B. Namjou, R. P. Kimberly, R. Ramsey-Goldman, *et al.*, *Klinefelter's syndrome (47, xxy) in male systemic lupus erythematosus patients: support for the notion of a gene-dose effect from the x chromosome,* Arthritis & Rheumatism: Official Journal of the American College of Rheumatology **58**, 2511 (2008).

[9] P. A. Jacobs and J. A. Strong, *A case of human intersexuality having a possible xxy sex-determining mechanism,* Nature **183**, 302 (1959).

[10] C. M. Smyth and W. J. Bremner, *Klinefelter syndrome,* Archives of Internal Medicine **158**, 1309 (1998).

[11] C. T. Mai, J. L. Isenburg, M. A. Canfield, R. E. Meyer, A. Correa, *et al.*, *National population-based estimates for major birth defects, 2010–2014,* Birth Defects Research **111**, 1420 (2019).

[12] H. B. van Gameren-Oosterom, S. Buitendijk, C. Bilardo, K. M. van der Pal-de Bruin, J. Van Wouwe, *et al.*, *Unchanged prevalence of down syndrome in the netherlands: results from an 11-year nationwide birth cohort,* Prenatal diagnosis **32**, 1035 (2012).

[13] S. E. Antonarakis, R. Lyle, E. T. Dermitzakis, A. Reymond, and S. Deutsch, *Chromosome 21 and down syndrome: from genomics to pathophysiology,* Nature reviews genetics **5**, 725 (2004).

[14] J. Nielsen and M. Wohlert, *Chromosome abnormalities found among 34910 newborn children: results from a 13-year incidence study in århus, denmark,* Human genetics **87**, 81 (1991).

[15] M. Elsheikh, D. Dunger, G. Conway, and J. Wass, *Turner's syndrome in adulthood,* Endocrine reviews **23**, 120 (2002).

[16] A. I. Taylor, *Autosomal trisomy syndromes: a detailed study of 27 cases of edwards' syndrome and 27 cases of patau's syndrome.* Journal of Medical Genetics **5**, 227 (1968).

[17] A. Cereda and J. C. Carey, *The trisomy 18 syndrome,* Orphanet journal of rare diseases **7**, 81 (2012).

[18] M. Higurashi, M. Oda, K. Iijima, S. Iijima, T. Takeshita, *et al.*, *Livebirth prevalence and follow-up of malformation syndromes in 27,472 newborns,* Brain and Development **12**, 770 (1990).

[19] E. Niebuhr, *The cri du chat syndrome,* Human genetics **44**, 227 (1978).

[20] C. S. Pareek, R. Smoczynski, and A. Tretyn, *Sequencing technologies and genome sequencing,* Journal of applied genetics **52**, 413 (2011).

[21] J. M. Heather and B. Chain, *The sequence of sequencers: The history of sequencing dna,* Genomics **107**, 1 (2016).

[22] I. R. Lehman, M. J. Bessman, E. S. Simms, and A. Kornberg, *Enzymatic synthesis of deoxyribonucleic acid i. preparation of substrates and partial purification of an enzyme from escherichia coli,* Journal of Biological Chemistry **233**, 163 (1958).

[23] H. G. Khorana, K. Agarwal, P. Besmer, H. Büchi, M. Caruthers, *et al.*, *Total synthesis of the structural gene for the precursor of a tyrosine suppressor transfer rna from escherichia coli. 1. general introduction.* Journal of Biological Chemistry **251**, 565 (1976).

[24] F. Sanger, S. Nicklen, and A. R. Coulson, *Dna sequencing with chain-terminating inhibitors,* Proceedings of the national academy of sciences **74**, 5463 (1977).

[25] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, *et al.*, *Nucleotide sequence of bacteriophage $□$x174 dna,* nature **265**, 687 (1977).

[26] W. M. Jou, G. Haegeman, M. Ysebaert, and W. Fiers, *Nucleotide sequence of the gene coding for the bacteriophage ms2 coat protein,* Nature **237**, 82 (1972).

[27] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, *et al.*, *Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene,* NATURE **260**, 500 (1976).

[28] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, *et al.*, *The sequence of the human genome,* science **291**, 1304 (2001).

[29] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, *et al.*, *Initial sequencing and analysis of the human genome,* (2001).

[30] J. Harrow, A. Frankish, J. Gonzalez, E. Tapanari, M. Diekhans, *et al.*, *Gencode: The reference human genome annotation for the encode project,* Genome Research **22**, 1760 (2012).

[31] A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, *et al.*, *Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users,* Human mutation **33**, 803 (2012).

[32] E. P. Consortium *et al.*, *An integrated encyclopedia of dna elements in the human genome,* Nature **489**, 57 (2012).

[33] N. J. Proudfoot, A. Furger, and M. J. Dye, *Integrating mrna processing with transcription,* Cell **108**, 501 (2002).

[34] V. H. Cowling, *Regulation of mrna cap methylation,* Biochemical Journal **425**, 295 (2010).

[35] A. Ramanathan, G. B. Robb, and S.-H. Chan, *mrna capping: biological functions and applications,* Nucleic acids research **44**, 7511 (2016).

[36] A. Ghosh and C. D. Lima, *Enzymology of rna cap synthesis,* Wiley Interdisciplinary Reviews: RNA **1**, 152 (2010).

[37] N. G. Zaorsky, T. Churilla, B. Egleston, S. Fisher, J. Ridge, *et al.*, *Causes of death among cancer patients,* Annals of Oncology **28**, 400 (2017).

[38] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, *et al.*, *Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,* CA: a cancer journal for clinicians **68**, 394 (2018).

[39] R. L. Siegel, K. D. Miller, and A. Jemal, *Cancer statistics, 2020,* CA: A Cancer Journal for Clinicians **70**, 7 (2020).

[40] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: the next generation,* cell **144**, 646 (2011).

[41] Y. A. Fouad and C. Aanei, *Revisiting the hallmarks of cancer,* American journal of cancer research **7**, 1016 (2017).

[42] F. Pellestor, *Chromoanagenesis: cataclysms behind complex chromosomal rearrangements,* Molecular cytogenetics **12**, 6 (2019).

[43] I. Cortés-Ciriano, J. J.-K. Lee, R. Xi, D. Jain, Y. L. Jung, *et al.*, *Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing,* Nature Genetics **52**, 331 (2020).

[44] W. P. Kloosterman, V. Guryev, M. van Roosmalen, K. J. Duran, E. de Bruijn, *et al.*, *Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline,* Human molecular genetics **20**, 1916 (2011).

[45] T. Helleday, S. Eshtad, and S. Nik-Zainal, *Mechanisms underlying mutational signatures in human cancers,* Nature Reviews Genetics **15**, 585 (2014).

[46] J. E. Kucab, X. Zou, S. Morganella, M. Joel, A. S. Nanda, *et al.*, *A compendium of mutational signatures of environmental agents,* Cell **177**, 821 (2019).

[47] Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, *et al.*, *Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden,* Genome medicine **9**, 34 (2017).

[48] S. A. Tomlins, B. Laxman, S. Varambally, X. Cao, J. Yu, *et al.*, *Role of the tmprss2-erg gene fusion in prostate cancer,* Neoplasia (New York, NY) **10**, 177 (2008).

[49] M. Fraser, V. Y. Sabelnykova, T. N. Yamaguchi, L. E. Heisler, J. Livingstone, *et al.*, *Genomic hallmarks of localized, non-indolent prostate cancer,* Nature **541**, 359 (2017).

[50] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, *Emerging patterns of somatic mutations in cancer,* Nature reviews Genetics **14**, 703 (2013).

[51] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, *et al.*, *The cancer genome atlas pan-cancer analysis project,* Nature genetics **45**, 1113 (2013).

[52] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, *et al.*, *Comprehensive characterization of cancer driver genes and mutations,* Cell **173**, 371 (2018).

[53] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, *et al.*, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[54] E. Y. Lee and W. J. Muller, *Oncogenes and tumor suppressor genes,* Cold Spring Harbor perspectives in biology **2**, a003236 (2010).

[55] A. P. Feinberg, R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer,* Nature reviews genetics **7**, 21 (2006).

[56] W. L. Tam and R. A. Weinberg, *The epigenetics of epithelial-mesenchymal plasticity in cancer,* Nature medicine **19**, 1438 (2013).

[57] A. W. Lambert, D. R. Pattabiraman, and R. A. Weinberg, *Emerging biological principles of metastasis,* Cell **168**, 670 (2017).

[58] H. Dillekås, M. S. Rogers, and O. Straume, *Are 90% of deaths from cancer caused by metastases?* Cancer Medicine **8**, 5574 (2019).

[59] S. P. Pereira, L. Oldfield, A. Ney, P. A. Hart, M. G. Keane, *et al.*, *Early detection of pancreatic cancer,* The lancet Gastroenterology & hepatology **5**, 698 (2020).

[60] C. E. Meacham and S. J. Morrison, *Tumour heterogeneity and cancer cell plasticity,* Nature **501**, 328 (2013).

[61] T. Baslan and J. Hicks, *Unravelling biology and shifting paradigms in cancer with single-cell sequencing,* Nature Reviews Cancer **17**, 557 (2017).

[62] P. Polak, R. Karlić, A. Koren, R. Thurman, R. Sandstrom, *et al.*, *Cell-of-origin chromatin organization shapes the mutational landscape of cancer,* Nature **518**, 360 (2015).

[63] M. J. Duffy and J. Crown, *A personalized approach to cancer treatment: how biomarkers can help,* Clinical chemistry **54**, 1770 (2008).

[64] D. Van der Velden, L. Hoes, H. van der Wijngaart, J. van Berge Henegouwen, E. van Werkhoven, *et al.*, *The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs,* Nature **574**, 127 (2019).

[65] D. H. Spencer, J. K. Sehn, H. J. Abel, M. A. Watson, J. D. Pfeifer, *et al.*, *Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens,* The Journal of molecular diagnostics **15**, 623 (2013).

[66] J. Noebels, *Pathway-driven discovery of epilepsy genes,* Nature neuroscience **18**, 344 (2015).

[67] T. Forshew, M. Murtaza, C. Parkinson, D. Gale, D. W. Tsui, *et al.*, *Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma dna,* Science translational medicine **4**, 136ra68 (2012).

[68] S. Chetty, M. J. Garabedian, and M. E. Norton, *Uptake of noninvasive prenatal testing (nipt) in women following positive aneuploidy screening,* Prenatal Diagnosis **33**, 542 (2013).

[69] S. T. Kim, W. S. Lee, R. B. Lanman, S. Mortimer, O. A. Zill, *et al.*, *Prospective blinded study of somatic mutation detection in cell-free dna utilizing a targeted 54-gene next generation sequencing panel in metastatic solid tumor patients,* Oncotarget **6**, 40360 (2015).

[70] H. Osumi, E. Shinozaki, Y. Takeda, T. Wakatsuki, T. Ichimura, *et al.*, *Clinical relevance of circulating tumor dna assessed through deep sequencing in patients with metastatic colorectal cancer,* Cancer medicine **8**, 408 (2019).

[71] K. Schwarze, J. Buchanan, J. C. Taylor, and S. Wordsworth, *Are whole-exome and whole-genome sequencing approaches cost-effective? a systematic review of the literature,* Genetics in Medicine **20**, 1122 (2018).

[72] I. The, T. P.-C. A. of Whole, G. Consortium, *et al.*, *Pan-cancer analysis of whole genomes,* Nature **578**, 82 (2020).

[73] T. Gutschner and S. Diederichs, *The hallmarks of cancer: a long non-coding rna point of view,* RNA biology **9**, 703 (2012).

[74] X. Chen, S. Fan, and E. Song, *Noncoding rnas: new players in cancers,* in *The Long and Short Non-coding RNAs in Cancer Biology* (Springer, 2016) pp. 1–47.

[75] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, *et al.*, *Landscape of somatic mutations in 560 breast cancer whole-genome sequences,* Nature **534**, 47 (2016).

[76] L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. T. Ng, *et al.*, *The repertoire of mutational signatures in human cancer,* Nature **578**, 94 (2020).

[77] L. F. van Dessel, J. van Riet, M. Smits, Y. Zhu, P. Hamberg, *et al.*, *The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact,* Nature communications **10**, 1 (2019).

[78] L. Angus, M. Smid, S. M. Wilting, J. van Riet, A. V. Hoeck, *et al.*, *The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies,* Nature genetics , 1 (2019).

[79] M. Fraser, V. Sabelnykova, T. Yamaguchi, L. Heisler, J. Livingstone, *et al.*, *Genomic hallmarks of localized, non-indolent prostate cancer,* Nature **541**, 359 (2017).

[80] J. D. Wren, *Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades,* Bioinformatics **32**, 2686 (2016).

[81] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, *Star: Ultrafast universal rna-seq aligner,* Bioinformatics **29**, 15 (2013).

[82] H. Li and R. Durbin, *Fast and accurate short read alignment with burrows-wheeler transform,* Bioinformatics **25**, 1754 (2009).

[83] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, *The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data,* Genome Research **20**, 1297 (2010).

[84] S. Kim, *Strelka2: Fast and accurate variant calling for clinical sequencing applications,* bioRxiv (2017).

[85] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, *et al.*, *Highly accurate protein structure prediction with alphafold,* Nature **596**, 583 (2021).

[86] S. B. Needleman and C. D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins,* J Mol Biol **48**, 443 (1970).

[87] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool,* J Mol Biol **215**, 403 (1990).

[88] J. D. Thompson, D. G. Higgins, and T. J. Gibson, *Clustal w improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,* Nucleic Acids Res **22**, 4673 (1994).

[89] J. Felsenstein, *Confidence limits on phylogenies: An approach using the bootstrap,* Evolution **39**, 783 (1985).

[90] S. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, *et al.*, *Cosmic: Somatic cancer genetics at high-resolution,* Nucleic Acids Research **45**, D777 (2017).

[91] L. Nguyen, J. W. M. Martens, A. V. Hoeck, and E. Cuppen, *Pan-cancer landscape of homologous recombination deficiency,* Nature Communications **11**, 5584 (2020).

[92] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, *Machine learning applications in cancer prognosis and prediction,* Computational and structural biotechnology journal **13**, 8 (2015).

[93] P. Roepman, E. de Bruijn, S. van Lieshout, L. Schoenmaker, M. C. Boelens, *et al.*, *Clinical validation of whole genome sequencing for cancer diagnostics,* The Journal of Molecular Diagnostics (2021).

[94] S. Mullane and E. Van Allen, *Precision medicine for advanced prostate cancer,* Current Opinion in Urology **26**, 231 (2016).

[95] S. Jones, V. Anagnostou, K. Lytle, S. Parpart-Li, M. Nesselbush, *et al.*, *Personalized genomic analyses for cancer mutation discovery and interpretation,* Science Translational Medicine **7** (2015), 10.1126/scitranslmed.aaa7161.

[96] M. Murtaza, S.-J. Dawson, D. W. Tsui, D. Gale, T. Forshew, *et al.*, *Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma dna,* Nature **497**, 108 (2013).

[97] M. Zapatka, I. Borozan, D. S. Brewer, M. Iskar, A. Grundhoff, *et al.*, *The landscape of viral associations in human cancers,* Nature genetics **52**, 320 (2020).

[98] W. Jiao, G. Atwal, P. Polak, R. Karlic, E. Cuppen, *et al.*, *A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns,* Nature communications **11**, 1 (2020).

[99] A. Abeshouse, J. Ahn, R. Akbani, A. Ally, S. Amin, *et al.*, *The molecular taxonomy of primary prostate cancer,* Cell **163**, 1011 (2015).

Chapter 2

SNPitty: An Intuitive Web Application for Interactive B-Allele Frequency and Copy Number Visualization of Next-Generation Sequencing Data

J. van Riet[a,b], N.M.G. Krol[a,c], P.N. Atmodimedjo[c], E. Brosens[d], W.F.J. van IJcken[f], M.P.H.M. Jansen[e], J.W.M. Martens[e], L.H. Looijenga[c], G. W. Jenster[b], H.J. Dubbink[c], W.N.M.Dinjens[c], H.J.G. van de Werken[a,b]

a Cancer Computational Biology Center, Erasmus MC, University Medical Center, Rotterdam, Netherlands.
b Department of Urology, Erasmus MC, University Medical Center, Rotterdam, Netherlands.
c Department of Pathology, Erasmus MC, University Medical Center, Rotterdam, Netherlands.
d Department of Clinical Genetics, Erasmus MC, University Medical Center, Rotterdam, the Netherlands.
e Department of Medical Oncology, Erasmus MC, University Medical Center, Rotterdam, the Netherlands.
f Center for Biomics, Erasmus MC, University Medical Center, Rotterdam, the Netherlands.

This chapter has been published in **The Journal of Molecular Diagnostics** Volume 20, Issue 2, March 2018, Pages 166-176

Abstract

Exploration and visualization of next-generation sequencing data are crucial for clinical diagnostics. Software allowing simultaneous visualization of multiple regions of interest coupled with dynamic heuristic filtering of genetic aberrations is, however, lacking. Therefore, the authors developed the web application SNPitty that allows interactive visualization and interrogation of variant call format files by using B-allele frequencies of single-nucleotide polymorphisms and single-nucleotide variants, coverage metrics, and copy numbers analysis results.

SNPitty displays variant alleles and allelic imbalances with a focus on loss of heterozygosity and copy number variation using genome-wide heterozygous markers and somatic mutations. In addition, SNPitty is capable of generating predefined reports that summarize and highlight disease-specific targets of interest.

SNPitty was validated for diagnostic interpretation of somatic events by showcasing a serial dilution series of glioma tissue. Additionally, SNPitty is demonstrated in four cancer-related scenarios encountered in daily clinical practice and on whole-exome sequencing data of peripheral blood from a Down syndrome patient. SNPitty allows detection of loss of heterozygosity, chromosomal and gene amplifications, homozygous or heterozygous deletions, somatic mutations, or any combination thereof in regions or genes of interest. Furthermore, SNPitty can be used to distinguish molecular relationships between multiple tumors from a single patient.

On the basis of these data, the authors demonstrate that SNPitty is robust and user friendly in a wide range of diagnostic scenarios.

Introduction

Genetic instabilities such as somatic copy number alterations, loss of heterozygosity (LOH), copy-neutral LOH/uniparental disomy, or mutational changes in proto-oncogenes, tumor suppressor genes, and genetic regulatory elements are of putative relevance in tumor development and progression.[1,2] Somatic events can give rise to allelic imbalances (AI) by the gain or loss of alleles due to errors in mitotic segregation, through single-nucleotide mutations, or through insertions and deletions of chromosomal segments, possibly as causal factors in cancers.[3–5]

Striking examples are tumor suppressors such as *TP53* and *RB1*, which are inactivated in many cancers by a deletion of one allele coupled with a mutational change in the other allele.[6–8] Copy-neutral LOH/uniparental disomy, which occurs due to loss of one parental allele and gain of the other allele, cannot be detected by calculating copy number state alone. It is therefore of paramount importance to extend the investigation of AI to establish correct molecular diagnosis and prognosis.

The authors have previously investigated and validated the diagnostic potential of Next-Generation Sequencing (NGS) to detect allelic losses and imbalances using heterozygous markers.[9] Using heterozygous single-nucleotide polymorphism (SNP)s, LOH and AI could be reliably detected with higher sensitivity and with a lower amount of input deoxyribonucleic acid (DNA) (1 to 10 ng) than other molecular techniques such as microsatellite marker analysis and multiplex ligation-dependent probe amplification. In addition, the combination of SNPs analysis and gene analysis by NGS was found to be a very powerful strategy for detection of large chromosomal aberrations and mutations relevant for molecular classification of tumors, clinical diagnosis, treatment, and prognosis.[9]

By utilizing informative heterozygous markers, NGS provides cost-effective and reliable diagnostic insights into somatic AI on a per-sample basis. Reliable results from targeted sequencing can even be achieved without the absolute necessity of matched normal samples, albeit slightly less accurately.[10] Owing to the admixture of both normal and malignant cells in a tissue-slice section, a single tissue biopsy sample can be used for diagnostic investigation by utilizing heterozygous markers while taking the tumor cell percentage into account. This added benefit can be especially helpful in scenarios where the acquisition of matched normal tissue is challenging, for example in revisiting historical samples from biobanks or brain tumor preparations.

Polymorphisms detected by NGS are routinely stored in generalized and standardized variant call format (VCF) files.[11] Predefined standard fields for storing the number of sequenced reads or number of observations per reference and alternative allele(s) are available in this format. VCF files are part of the output of most industry-standard variant calling and annotation suites. B-Allele Frequency (BAF) for each variant are computed based on standard VCF fields. The BAF formula is a simple division of the observations per alternative allele over the sum of observations for both reference and alternative alleles. BAF thus represents the ratio of each alternative allele per variant present in a sequenced sample and has been applied for copy number analysis of SNP arrays.[12]

Besides using heterozygous variants, copy number analysis can be performed based on the covered genome of the sequenced sample. Results of copy number analysis, derived from segmentation-based algorithms such as the ONCOCNV software package version 6.6 (ONCOCNV, Paris, France; http://oncocnv.curie.fr),[13] are generally stored in segment files. These copy number alteration segment files contain the absolute and/or $\log_2$ ratio of copy numbers per loci or region as estimated by platform- or genome-wide analysis.

In the context of NGS-based targeted multigene panels using heterozygous markers, BAF and copy number analysis can be used to estimate tumor cell percentages, somatic aberrations and imbalances, quality of amplicons, and heterogeneity of tumors as described below.

Currently, there are several publicly available tools to display variants from VCF files alongside additional genetic information and annotation. For example, the Integrated Genome Viewer, JBrowse, and the UCSC Genome Browser are commonly used genome browsers.[14–16] However, these tools visualize variants on their respective reference genome based on their exact genomic positions. For targeted sequencing of distant sites throughout the human genome, these tools are not always suited for apparent genome-wide diagnostic interpretation using their default settings. For instance, viewing distant or interchromosomal regions of interest spread throughout the genome requires separate examinations. Clinical interpretation often requires a holistic view of the relationships between observed aberrations, e.g., determining whether a glioma sample with an observed IDH1 mutation has additional aberrations such as a somatic mutation in the *TP53* gene or a 1p/19q co-deletion and/or loss of the *CDKN2A* gene.[17]

To simplify and accelerate the genome-wide diagnostic interpretation of AI, a visualization approach based on the relative positioning of variants on chromosomes has been proposed. Using this approach, variants will be displayed based on relative positioning to neighboring variants. For example, variants on chromosome 10 will be displayed next to each other without any fixed distance between them, ordered on ascending genomic position and chromosome. To this end, an easy-to-install and user-friendly web application that uses relative positioning to display variants and their respective BAF from user-submitted VCF files, called SNPitty, was developed.

Material and Methods

Sample Preparation and Processing

Tissues were microdissected manually, and all samples contain at least 70% to 80% tumor cells as indicated by our local pathologists. Dependent on the tissue, between 1 and 10 ng of DNA was isolated and subsequently sequenced on the Ion Torrent PGM platform with supplier's materials and protocols (Life Technologies, Carlsbad, CA) as described previously.9 Generally, library and template preparations were performed consecutively with the AmpliSeq Library Kit 2.0 to 384 LV and the Ion PGM Template OT2 200 kit. Templates were sequenced using the Ion PGM Sequencing 200 Kit v2 on an Ion 318v2 chip. Custom in-house primer designs utilizing heterozygous markers on the autosomal chromosomes from NCBI dbSNP database build 138 (https://www.ncbi.nlm.nih.gov/projects/SNP) with at least 45% global minor allele frequency were used to create panel-specific assays targeting known genetic aberrations associated with tumor formation, progression, and classification.[18]

NGS reads were subsequently aligned against the human reference genome (hg19; UCSC Genome Browser, last accessed February 2009) using the Torrent Mapping Alignment Program (TMAP) software version 5.2 (Life Technologies) with default settings. Torrent Variant Caller software version 5.2 (Life Technologies) was used to determine and measure both novel and predefined heterozygous (hotspot) variants using the Generic - PGM - Somatic - Low Stringency settings. Additional heuristic filtering discarded variants with a total read depth <100. AI was assigned using the criterion of at least two consecutive informative SNPs.[9]

Copy numbers were estimated using ONCOCNV software version 6.6 with default settings using the amplicon coordinates of the respective panels.[13] The malignant tissues were compared against panel-specific copy number baselines of seven normal tissues. Briefly, read coverages per targeted regions were generated from BAM files and normalized against the respective baseline as well as GC content of the reference genome (hg19). Segments were aggregated per region and a single region-level copy number estimate was generated.

A glioma tissue sample with near-100% neoplastic cells hosting 1p/19q codeletions was diluted as a proof of concept. A serial dilution with adjacent normal tissue was performed to establish glioma samples with varying tumor cell percentage mixtures (near 100%, 60%, 40%, 20%, and 10%), accompanied by a single matched normal sample. DNA was extracted from peripheral blood of a girl with Down syndrome (SE14-0562) using standard protocols (Qiagen, Venlo, the Netherlands). The target (exome) was captured with the HaloPlex exome target enrichment kit (Agilent Technologies, Santa Clara, CA) and sequenced on a HiSeq2000 system (Illumina, San Diego, CA) using the TruSeq software version 3 paired-end 100 bp sequencing protocol. The reads were trimmed for the Illumina adapter, and 245M reads were subsequently aligned against the human reference genome build 19 (hg19) using BWA[19] software version 0.6.2 (SourceForge; http://bio-bwa.sourceforge.net/bwa.shtml) and the NARWHAL pipeline software version 1.0 (Netherlands Bioinformatics Center, Nijmegen, the Netherlands; https://trac.nbic.nl/narwhal)[20] resulting in an average target base coverage of 280× (and 94% of the target bases were covered at least 20×). Variants were called using GATK software version 2.4 (Broad Institute, Cambridge, MA; https://software.broadinstitute.org/gatk/).[21] Only informative heterozygous markers present in the dbSNP database with at least 175× read coverage were kept for analysis. Chromosomal AI was assigned using the criterion of at least 500 informative markers representative for the entire chromosome.

The Complete Genomics whole-genome sequence of the prostate cancer cell line VCaP was processed and visualized as previously described.[22]

All samples were assessed anonymously according to the code for adequate secondary use of tissue code of conduct established by the Dutch Federation of Medical Scientific Societies (https://www.federa.org/codes-conduct, last accessed February 10, 2017).

VCF files have been submitted to the European Variation Archive (EVA; https://www.ebi.ac.uk/eva) under accession number PRJEB21914.

Immunohistochemistry and FISH for p53 Overexpression and *EGFR* Amplification

Fluorescence *in situ* hybridization (FISH) was performed to validate *EGFR* amplification using the Poseidon Repeat Free EGFR, Her-1 (7p11) & SE 7 Control probe (Kreatech Diagnostics, Amsterdam, the Netherlands). Slides were examined under a Zeiss Axio-Images M2 microscope with the Piezo scanning stage, slide images were captured with a Zeiss AxioCam MRm rev.3 camera (Zeiss, Jena, Germany).

According to standard protocols, overexpression of p53 was assessed by immunohistochemistry (IHC) using mouse monoclonal antibody (clone BP53-11; Ventana Medical Systems, Mountain View, CA) on the automated Ventana BenchMark ULTRA platform.

Technical Design of SNPitty

SNPitty was implemented in the statistical platform R software version 3.4.2 (R Project for Statistical Computing; http://www.r-project.org)[23] using the Shiny framework and several BioConductor packages: VariantAnnotation version 1.22.0 and Biobase version 2.36.0 (BioConductor; https://bioconductor.org).[24,25] A Docker image was generated to facilitate the entire installation of SNPitty and required dependencies.[26] Multiple single (or multi-) sample VCF or VCF.gz files were merged on the union of sets using BCFtools version 1.4 (https://samtools.github.io/bcftools).[11] Nonintersecting variants after merging were set to NA in the respective samples lacking these variants.

BAF per variant was calculated based on a combination of available genotype fields in the uploaded VCF files. The BAF for a specific alternative allele per sample is calculated as follows:

$$BAF = \frac{\text{Observations for alternative allele}}{\text{Observations for all alleles}} \tag{2.1}$$

The BAF value displayed in SNPitty is a summation of the BAF for all alternative alleles to represent the total difference to their respective reference allele. Each independent variant (row) in the VCF file is plotted as an individual data point.

Additional regional information was added by uploading a gene transfer format v2 (GTF) file containing the genomic ranges of each region.

Results

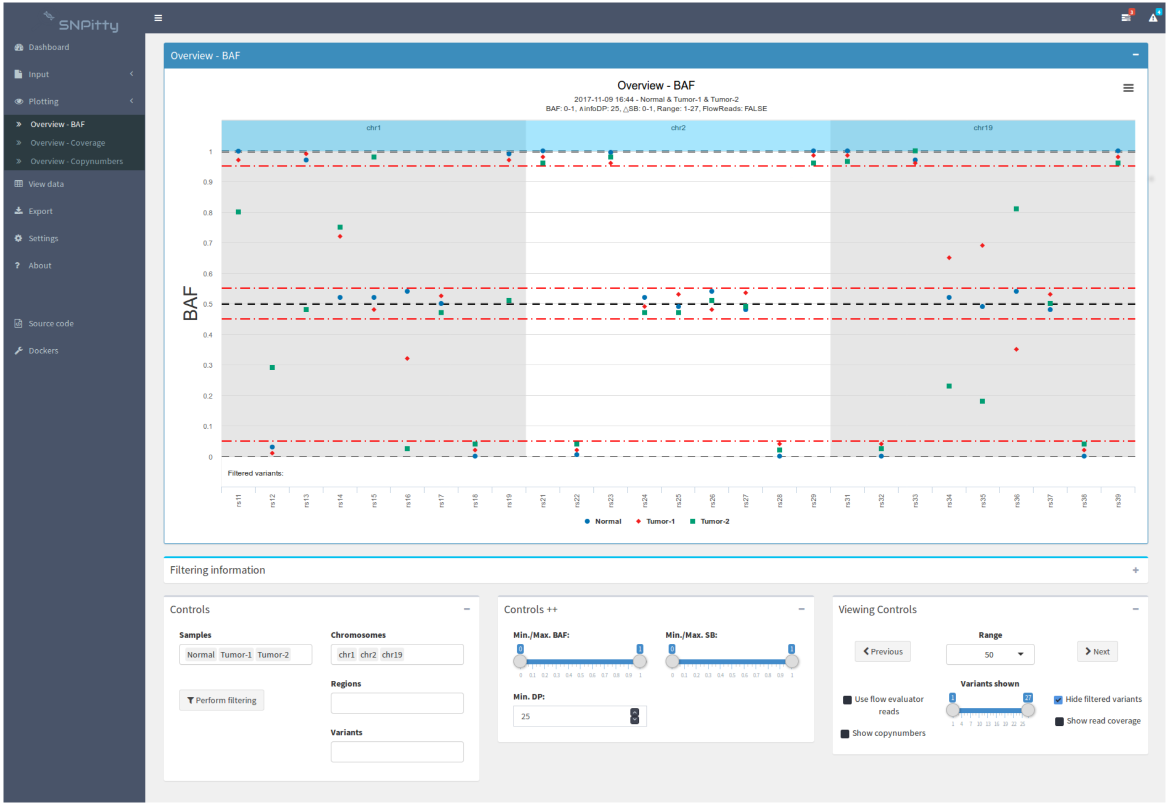
SNPitty Web Application

The user-friendly web application SNPitty (CCBC, Rotterdam, the Netherlands; https://bitbucket.org/ccbc/snpitty, last accessed January 10, 2017) was developed to support and improve diagnostic interpretation of AI. SNPitty visualizes BAF of somatic variants and heterozygous markers detected by NGS-based targeted multigene panels, whole-exome sequencing (WES), or whole-genome sequencing (WGS) efforts. The web application is accessible by all modern web browsers; figure 2.1 shows an impression of the web interface. Note that all session information and data are automatically deleted after closing a session (e.g., closing the web interface) to ensure privacy.

A relative-positioning approach was used to provide insight into multiple distant or interchromosomal regions of interest spread throughout the human genome in an interactive and comprehensive manner.

SNPitty processes single VCF files and is also capable of merging multiple VCF files, which are obtained from most of the industry-standard variant calling suites for NGS. Submitted VCF files should contain the variants of scientific or diagnostic interest with optional information, such as the amount of forward and reverse reads to calculate strand bias. Genotype fields that are used for BAF calculations can be selected manually or be inferred automatically based on available fields in the submitted VCF files. The option of using Ion Torrent–specific flow evaluator reads is also implemented in SNPitty.

SNPitty also allows visualization of ratios, means, or absolute counts of copy number segments derived from segmentation-based copy number variation detection algorithms. These copy number results can be visualized simultaneously with BAF to increase insight into germline or somatic aberrations.



Overview - BAF

2017-11-09 16:44 - Normal & Tumor-1 & Tumor-2
BAF: 0-1, ∧infoDP: 25, △SB: 0-1, Range: 1-27, FlowReads: FALSE

Filtered variants:

Filtering information

Controls

Samples: Normal Tumor-1 Tumor-2

Chromosomes: chr1 chr2 chr19

Regions

Variants

Controls ++

Min./Max. BAF: 0 – 1

Min./Max. SB: 0 – 1

Min. DP: 25

Viewing Controls

Range: 50

Variants shown: 1 – 27

■ Use flow evaluator reads

■ Show copynumbers

☑ Hide filtered variants

■ Show read coverage

Figure 2.1: **Overview of the web interface of SNPitty.**
BAF visualization is shown using artificial data for three samples with a selection for human chromosomes 1, 2, and 19. Chromosomes or regions are indicated with alternating gray and white backgrounds; regional names are displayed on top. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.

Optionally, predefined regional information for visualizing regions of diagnostic interest can be supplied in a file using GTF format. Variants and segments are assigned to these regions based on their genomic overlap, e.g., a predefined *PTEN* region on the human chromosome 10 starting at base position 87863113 and ending at base position 87971930 will cover all the variants and segments that overlap this region. Therefore, disease-specific variants of diagnostic interest can be easily categorized and quickly selected for visualization. Moreover, multiple GTF files can be uploaded that are disease- or sample-context specific.

All submitted variants can be simultaneously displayed with BAF, read coverage, and copy number information using interactive charts. The charts are scalable and can be saved to various high-resolution output formats, such as PDF, SVG, PNG, and JPEG. Variants and regions can be dynamically filtered based on various user criteria to quickly answer diverse diagnostic questions. Heuristic filtering is easily performed using, among others, minimal and/or maximal BAF, read depth, strand bias, or annotation status, or via manual selection of variants, regions, and chromosomes.

Specific variants can be highlighted using regular expressions, e.g., all variants containing a specific tag such as "rs" can be highlighted to indicate all dbSNP database variants having an rsID. Read coverage information per sample is shown per variant, both via overlay and/or mouse hovering. A distinct plotting window to display the amount of forward and reverse reads (if applicable) and total read depth is also implemented. With this window, the user can easily focus on amplifications and deletions of interest, or display strand bias and quality of amplicons. All variants present in the (combined) VCF file can be viewed simultaneously or via an intuitive system of sliding windows using intervals, e.g., consecutively show 50, 100, or 1000 variants. Users can also choose to only display specific variants of interest based on their respective identifiers.

PDF reports can be generated for user-specified samples based on LaTeX/knitr templates. These reports highlight the targets of interest of the respective diagnostic panel with full support of the statistical platform R and the BioConductor suite.[25,27] The reports are fully customizable but require some experience with R and LaTeX for advanced functionality. Several global templates are provided with SNPitty that host a number of useful features such as sample summaries, chromosomal or regional overviews of variants showing BAF and coverage, and ideograms of the human chromosomes (hg19) showing the location of variants on the human genome.

SNPitty is open source software under the GNU GPLv3 license and freely available (https://bitbucket.org/ccbc/snpitty, last accessed February 10, 2017), including documentation on usage and installation.

To ease local software deployment, a Docker image of SNPitty has been generated (https://hub.docker.com/r/ccbc/snpitty, last accessed February 10, 2017), which can be used as a lightweight virtual machine deployable on both Unix-based and Windows machines to provide reproducible environments.

SNPitty and BAF-dependent interpretation was applied on NGS-based targeted multigene panels utilizing heterozygous markers.[9] Four cancer-related diagnostic scenarios encountered in daily clinical practice and a glioblastoma dilution series have been showcased to validate SNPitty as a robust and all-round BAF viewer for routine diagnostic purposes. Furthermore, a germline chromosomal amplification of chromosome 21 is shown using WES and extended in silico validation by reproducing BAF results of a previously published study on the prostate cancer cell line VCaP.[22]

Proof of Principle of the Features Present in SNPitty

To validate the robustness of BAF in diagnostic scenarios, a serial dilution of DNA was generated from malignant glioma tissue. This glioma tumor tissue was serially diluted with an increasing amount of adjacent normal tissue to generate mixtures with decreasing tumor cell percentages (n = 5) and sequenced using in-house targeted NGS glioma panel (Figure 2.2). The heterozygous markers that are unaltered by AI retain a heterozygous genotype (BAF = 0.5) in both normal and (diluted) malignant tissue. Markers with a germline homozygous genotype retain BAF of 0 or 1 in respect to the exclusive presence of the reference allele or the alternative allele(s).

A 1p/19q LOH co-deletion is present in the malignant tissue and absent in the normal tissue, as evident by the increasing deviation from a heterozygous genotype (BAF = 0.5) reaching a homozygous genotype (BAF = 0 or BAF = 1) in the 1p/19q regions in respect to tumor cell percentage. The malignant tissue also harbors a heterozygous somatic *IDH1* c.395G>A (p.R132H) mutation. A homozygous genotype is present in the nonmalignant tissue and reaches a heterozygous genotype in the malignant tissues with respect to the tumor cell percentages. This validation

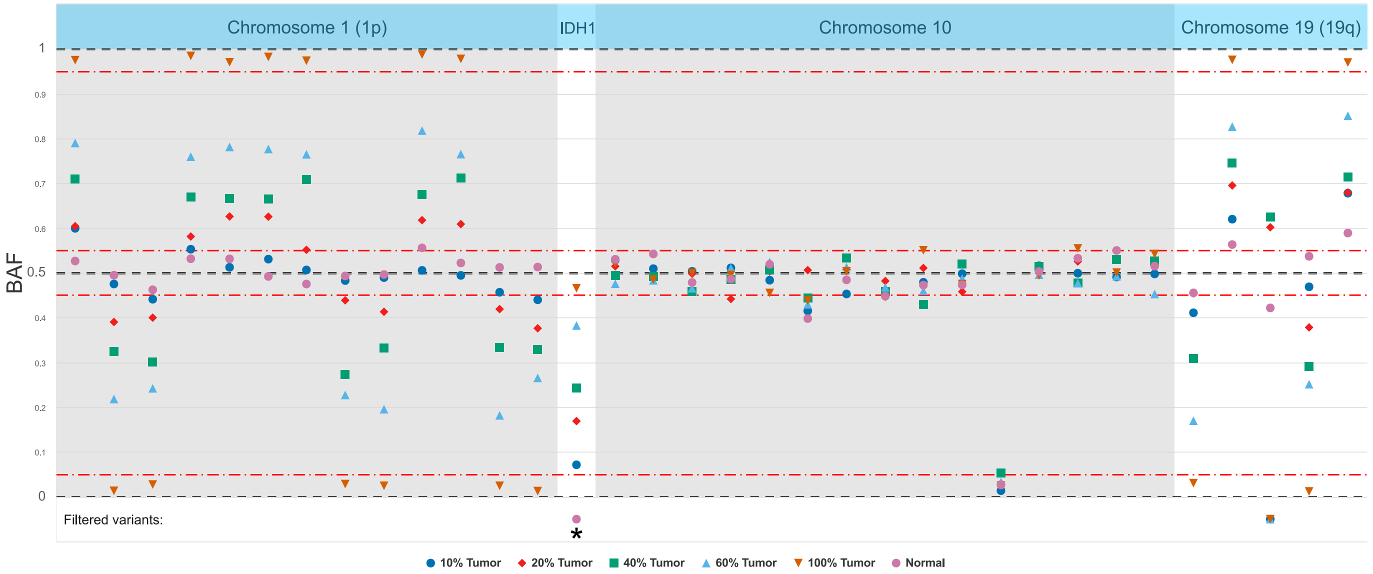


Figure 2.2: **Validation of SNPitty by serial dilution of glioma tumor tissue with matched normal.** Glioma tumor tissue was serially diluted with an increasing amount of near-adjacent normal tissue, 0% (purple circles), 10% (blue circles), 20% (red rhombi), 40% (green squares), 60% (blue triangles), to 100% (orange triangles), respectively. The dilution series show an increasing BAF deviation for heterozygous markers on 1p and 19q, with respect to a higher tumor cell percentage. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. Markers that retained homozygous state in all six samples are filtered. **IDH1* c.395G>A (p.R132H) somatic mutation.

shows that BAF is a robust metric to visualize somatic aberrations in the subtype of oligodendrogliomas, which are typically characterized by high occurrences of 1p/19q co-deletions and somatic mutations in the *IDH1*/*IDH2* gene and in the promoter region of *TERT*.[28]

Detection of Genomic Amplification and Heterozygous Deletion Using SNPitty

To illustrate the visualization of read coverages and copy number segments to assess somatic copy number alterations in SNPitty, the *EGFR* gene copy number status was analyzed in a glioblastoma sample. *EGFR* amplification is a common genetic aberration in glioblastomas.[29] By comparing the number of reads in the *EGFR* locus to the surrounding regions, *EGFR* amplification can be appreciated and further corroborated by platform-wide copy number analysis (Figure 2.3A and Supplemental Figure S2.1). We further hypothesize that this is an *EGFR* amplification of a single allele because several *EGFR* markers are not reaching a homozygous state but retain a semiheterozygous BAF of 0.05 or 0.95. These markers might be germline heterozygous; as a consequence, the nonamplified allele is only present in a single copy and therefore sequenced in a lesser amount. *EGFR* amplification was confirmed by showing increased copy numbers of *EGFR* in the vast majority of the malignant cells using FISH (Figure 2.3B). The two *EGFR* markers not showing concordant read coverage originate from amplicons with lower performance on this specific panel design.

Next to the *EGFR* amplification, this sample contains a heterozygous deletion of *PTEN*, which is detected by deviations of germline heterozygous markers toward a more homozygous state in both flanking and coding regions of *PTEN*. This heterozygous deletion of *PTEN* is further corroborated by genome-wide copy number analysis. Furthermore, two somatic missense mutations in the coding region of *TP53*, namely a c.817C>T (p.R273C) and a c.215C>G (p.P72R) mutation were detected (data not shown).

An additional pleura adenocarcinoma with *EGFR* amplification, coupled with a p.E746_A750delELREA and p.T790M (c.2369C>T) mutation in the *EGFR* region, can be seen in Supplemental Figure S2.2.

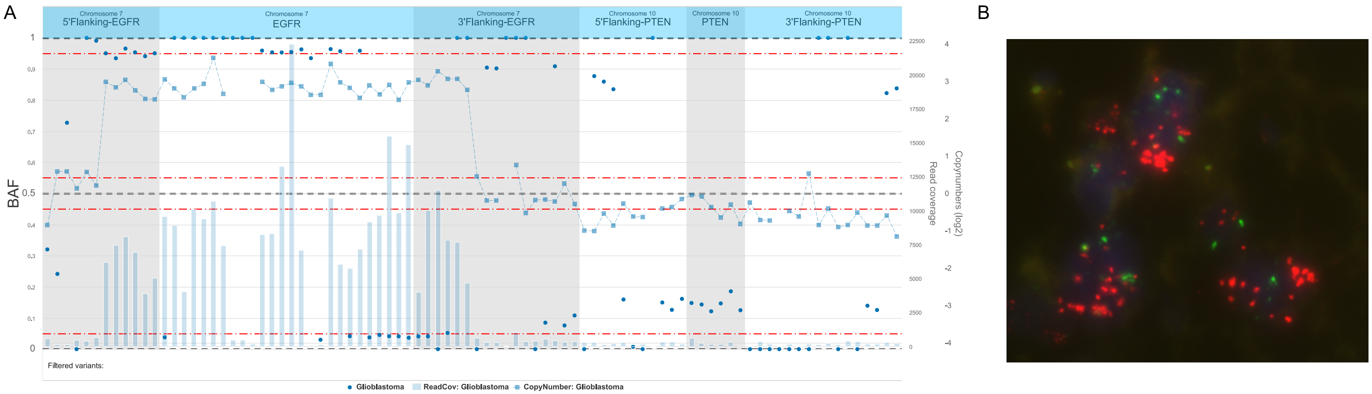


Figure 2.3: **(a)**: BAF visualization of heterozygous markers (blue circles), accompanied with respective read coverage (blue bars) and $\log_2$ copy number ratio (dashed blue line with blue squares), on the *EGFR* and *PTEN* regions for a single glioblastoma sample. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. **(b)**: *EGFR* gene amplification by FISH in respective sample. FISH analysis was performed using EGFR, Her-1 (7p11) probe (red), and SE 7 control probe (green); original magnification (×63).

Detection of Homozygous Deletions Using SNPitty

Homozygous deletions of chromosomal regions are of clinical interest and can also be assessed and visualized with SNPitty. For example, homozygous deletion of *CDKN2A*, which encodes the tumor suppressor p16 and p14ARF, is known to be a driver of glioblastoma development.[17] A homozygous deletion of *CDKN2A* was detected in glioblastoma tissue using SNPitty (Figure 2.4). This homozygous deletion can be appreciated by the marked decrease in read coverage of the *CDKN2A*-covering amplicons. Moreover, heterozygous deletion/LOH of *CDKN2A* flanking region can be appreciated by seven informative markers showing BAF deviations. LOH in *CDKN2A* flanking regions, combined with decreased read coverages and two heterozygous markers (BAF = approximately 0.5) in the *CDKN2A* locus, suggest a homozygous loss of *CDKN2A*. A heterozygous loss of *CDKN2A* can be discarded because the two heterozygous markers do not show BAF deviations. Therefore, we hypothesize that these NGS results originate from admixed normal tissue, which explains a marked decrease in read coverage while retaining heterozygosity of the two markers.

Hence, SNPitty is capable of visualizing LOH and homozygous deletion simultaneously from a single admixed tumor sample.

An additional glioblastoma with a homozygous *CDKN2A* deletion, coupled with regions of heterozygosity on 1p and 19q, and AI on chromosome 7 can be seen in Supplemental Figure S2.3.

Discovering Two-Hit Models in *TP53* Using SNPitty

SNPitty is able to clearly display evidence for genes having undergone a two-hit model of inactivation, e.g., in which a loss-of-function mutation is found in combination with LOH, as is common in tumor suppressors.[8] This scenario is shown for the frequently mutated *TP53* gene in squamous cell carcinoma tissue composed of at least 70% neoplastic cells (Figure 2.5A). In this scenario, the 5′ and 3′ flanking regions of *TP53* show LOH as indicated by four informative markers differing between matched normal and tumor samples, accompanied by a somatic missense mutation [c.536A>G (p.H179R)] in the coding region of *TP53*. Reduced read coverage of the markers on the flanking regions indicate a possible heterozygous deletion. IHC of p53 confirmed the presence of a putative stabilizing *TP53* mutation in the malignant cell tissue as evident by the abnormally high presence of p53 (Figure 2.5B).

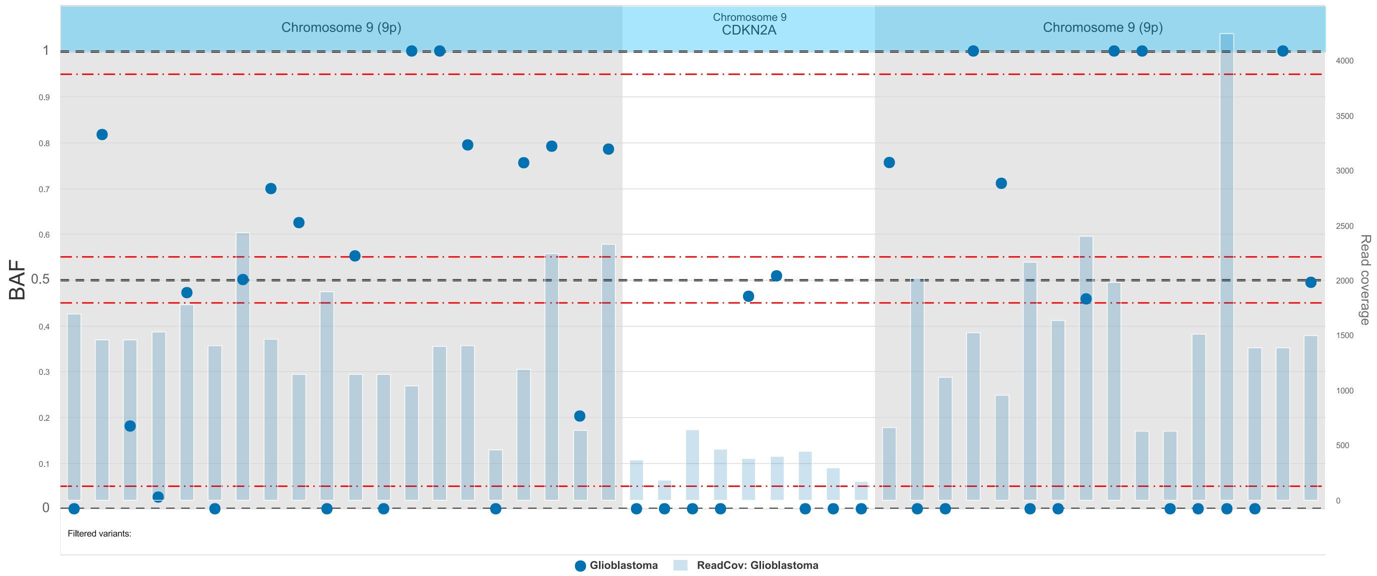


Figure 2.4: **Detection of homozygous deletions using SNPitty.**
BAF visualization of heterozygous markers (blue circles), accompanied with respective read coverage (blue bars), on the chromosome arm 9p and *CDKN2A* region for a single glioblastoma sample. LOH on 9p is accompanied by a homozygous deletion of *CDKN2A*. The remaining heterozygous state of *CDKN2A* is a reflection of the nonmalignant tissue present in the sample. Read coverage is shown in transparent blue bars. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.

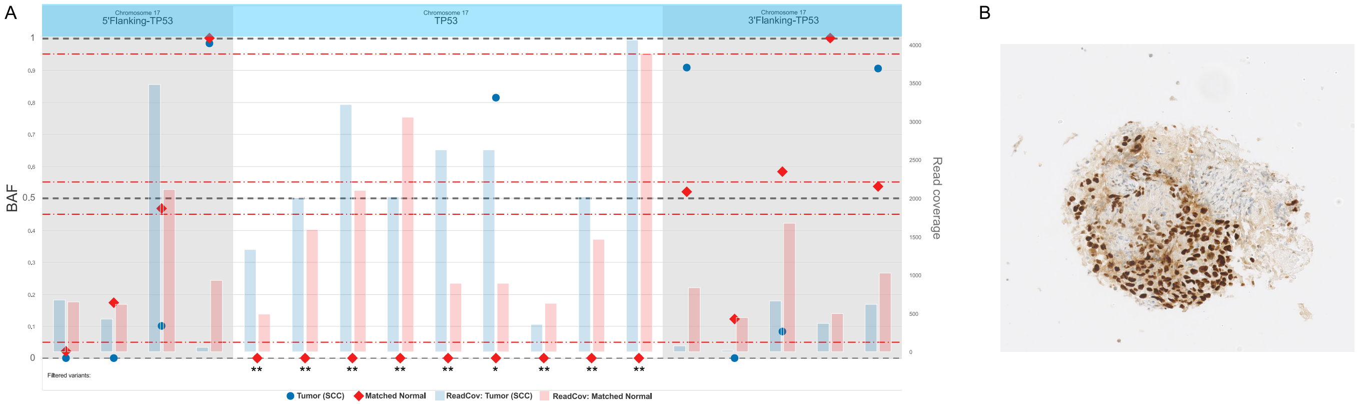


Figure 2.5: **Discovering a two-hit model of *TP53* in mediastinal squamous cell carcinoma using SNPitty.**
(a): BAF visualization of heterozygous markers, accompanied with respective read coverage, on the flanking and coding regions of *TP53* for mediastinal squamous cell carcinoma (SCC) (blue circles and blue bars) and near-adjacent prostate tissue samples (matched normal, red rhombi and red bars). * Additional homozygous nucleotide positions to display genomic read coverage in TP53 region. ** c.536A>G (p.H179R) somatic missense mutation. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. **(b)**: Immunohistochemistry of p53 shows an overexpression of p53 in malignant tissue of the respective mediastinal squamous cell carcinoma sample; original magnification (×63).

An additional lung adenocarcinoma with a detected somatic p.K132R (c.395A>G) mutation in the *TP53* region, coupled with two markers showing LOH, can be seen in Supplemental Figure S2.4.

SNPitty Allows Detection of Distinct LOH in Multiple Tumor Samples

Due to recent major therapy improvements, cancer is increasingly becoming a chronic disease.[30] This means patients successfully treated for cancer have a greater lifespan and as a consequence, have a higher rate of recurrence. For the treatment of patients with multiple synchronous or metachronous tumors, it is of utmost importance to obtain an accurate diagnosis; does a patient showing multiple tumors have metastatic disease (single primary tumor accompanied by metastasis), or does the patient have multiple independent primary malignancies developed by separate carcinogenic events.[31]

Comparing the evolutionary history of patient-derived tumors can distinguish between synchronous or metachronous tumors. A clinical case is shown, with SNPitty, where a patient presented multiple tumors in the colon, lung, bladder, and vertebrae in a 12-year period.

From this patient, two biopsies from the lung and vertebrae were obtained and subsequently sequenced. Here, these two biopsies were visualized alongside a single matched normal tissue sample for multiple chromosomes showing AI (Figure 2.6). By visualizing the BAF of informative heterozygous markers on these chromosomes, a divergent evolutionary history can be seen due to nonoverlapping AI events. Most strikingly, the *ATM* region on chromosome 11 is affected by LOH in both tumors on opposing alleles. This indicates that these tumors are a reflection of distinct tumor entities.

Additional clonality assessments (n = 6), using SNPitty, from diverse patients that presented multiple tumors in distinct locations can be seen in Supplemental Figures S2.5, S2.6, S2.7, S2.8, S2.9, S2.10 and S2.11.

SNPitty Visualizes Germline Trisomy 21 Using WES Data

Large chromosomal abnormality can be identified by large regions of BAF imbalance of heterozygous markers. This principle is shown using a Down syndrome patient with an expected germline abnormality of chromosome 21. WES of periph-

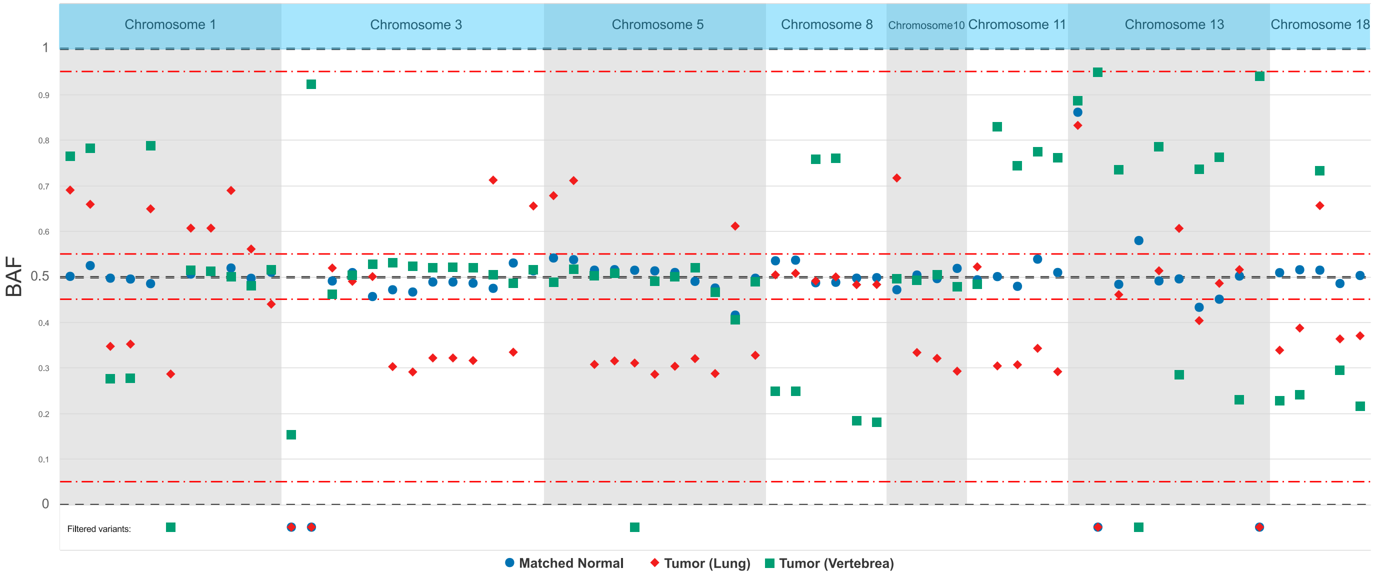


Figure 2.6: **Detection of distinct LOH in multiple tumor samples using SNPitty.**
BAF visualization of heterozygous markers on chromosomes 1, 3, 5, 8, 10, 11, 13, and 18 for SCC, originating from lung (red rhombi) and vertebrae tumors (green squares), and single matched normal prostate sample (blue circles). Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.

eral blood from this individual shows this germline abnormality as evident by AAB (BAF = approximately 0.33) and ABB (BAF = approximately 0.66) genotypes on chromosome 21 (Figure 2.7). Chromosome 21 displayed an additional copy of the entire chromosome, whereas all other autosomal chromosomes were confirmed to diploid copy number status (only chromosome 5 is shown). Using this scenario of a germline chromosomal abnormality, the capability of SNPitty to detect specific germline aberrations is shown.

Additional *in silico* Validation of SNPitty

The robustness of SNPitty was further validated by reproducing the BAF results of a previously published study on the instability of chromosome 5 in the prostate cancer cell line VCaP.[22] SNPitty is capable of reproducing the BAF visualization of large-scale genomic aberrations on chromosome 5 in the prostate cancer cell line VCaP as shown by Teles Alves *et al.*[22] in WGS data (Supplemental Figure S2.12). The same near-triploid state of the VCaP genome on chromosome 5 is highlighted by large-scale chromosomal AI coupled with clustered rearrangements on 5q.

Furthermore, a colon adenocarcinoma and cecum adenocarcinoma with somatic abnormalities in mismatch-repair–related regions in diagnostic scenarios as encountered in daily practice are shown in Supplemental Figures S2.13 and S2.14.

Discussion

Due to complex and diverse molecular mechanisms driving tumor development and progression, correctly interpreting data generated from NGS-based genome-wide or targeted multigene panels is crucial for daily diagnostics. By applying SNPitty in scenarios encountered in daily practice, the added value of SNPitty for detecting LOH, chromosomal and gene amplifications, homozygous or heterozygous deletions, single-nucleotide mutations, and clonality assessment is demonstrated. Additionally, a scenario in which the capability to interrogate WES, WGS, and germline datasets, for example, to detect chromosome-wide LOH or amplification, is demonstrated.

Using the industry-standard VCF format, coupled copy number segments, and the GTF format to define regions of interest, SNPitty facilitates a flexible and simple method for users to explore variants and copy numbers, and to visualize aberrations without in-depth knowledge of bioinformatics. It realizes this through a simple

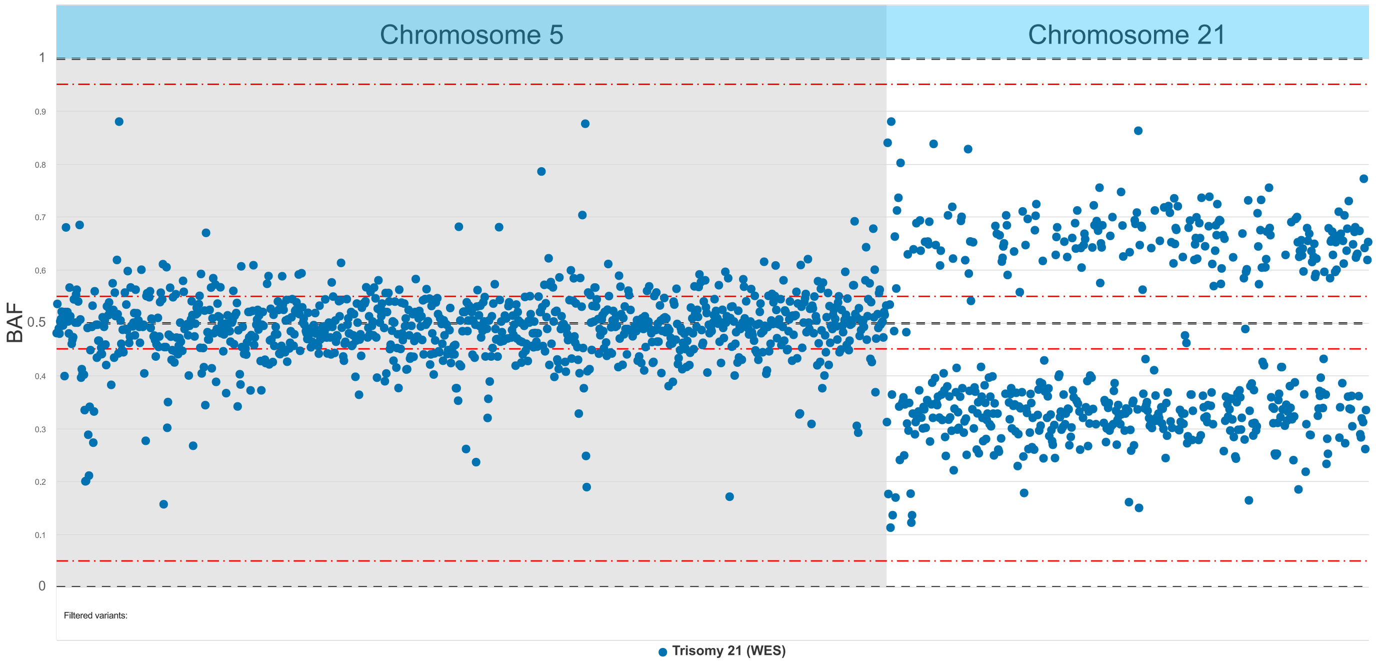


Figure 2.7: **Detection of chromosomal abnormality in whole-exome sequenced peripheral blood from a Down syndrome patient using SNPitty.** BAF visualization of informative heterozygous markers (covered with >175 reads) from WES peripheral blood (blue circles) on chromosome 5 and chromosome 21. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. *n* = 954 (chromosome 5); *n* = 554 (chromosome 21).

and user-friendly, dynamic, graphical web interface to visualize and filter variants from one or multiple VCF files. High-resolution images of any plot and viewpoint can be exported as multiple industry-standard output formats such as SVG and PDF. Custom-made reports based on LaTeX templates can highlight results of clinical interest in a standardized manner for a wide range of diagnostic scenarios.

Here, human malignant and germline tissue was focused on using targeted NGS-based multigene panels and WES; however, VCF files containing variants from other species can also be visualized and interpreted using SNPitty due to uniformity of file standards. Currently, only LaTeX report templates for human reference genomes (hg19 and hg38) have been added to automatically generate reports. Support of alternative genomes can be added by customizing or adding additional LaTeX templates.

User-specific and dynamic heuristic filtering of variants can be applied using SNPitty to interactively filter low-quality or erroneous sites. These low-quality sites most often arise due to poor quality of the DNA libraries, nonspecific binding of amplicon/primer (if applicable), technical noise generated during the sequencing procedure, usage of formalin-fixed, paraffin-embedded material, and low coverage of the genetic aberrations of interest.[32]

Currently, the methodology used by SNPitty not been as rigorously validated in-house for use in germline malignancies as it has been for somatic events. Further optimization, extension, and practical use will likely be needed to also provide a more flexible and robust toolkit for a wide range of use cases involving germline aberrations.

Applying dynamic heuristic filtering coupled with the ability to display various regions of interest using relative positioning, rather than absolute positioning, can more quickly give insight into potentially causal genetic aberrations. Expert interpretation of these aberrations present in a clinical sample plays a crucial role in clinical decision making. SNPitty therefore allows viewing and interpreting the various genomic aberrations simultaneously to formulate a more holistic hypothesis of sample-specific causal factors.

SNPitty is not aimed to replace the role of conventional genome browsers such as Integrated Genome Viewer, JBrowse, and the UCSC Genome Browser for research purposes, because these powerful tools are aimed to handle and display a greater variety of data from a myriad of molecular techniques. However, due to the extensive configuration required to simultaneously view multiple (distant and/or interchromosomal) regions of interest in the aforementioned genome browsers, SNPitty can be used as a robust alternative in these scenarios.

Overall, SNPitty is a user-friendly, open source, and Docker deployable web application that can aid and accelerate research and daily diagnostic interpretation by visualizing the results of NGS-based experiments utilizing heterozygous markers, single-nucleotide variants, and copy number results.

Acknowledgments

We thank the Department of Pathology (Erasmus Medical Center) for providing the samples and generating the IHC and FISH images, in particular Hein F. Sleddens for performing FISH, and Sharmiela Ramlal, Erina de Leeuw, Hans Stoop, Janine Shukla, Jeanine Roos, and Boukje van Gils for performing IHC; Saskia Hiltemann, Youri Hoogstrate, and Andrew Stubbs (Erasmus Medical Center) for sharing and helping to interpret the VCaP data; and Rutger W. Brouwer (Erasmus Medical Center) for processing and analyzing the exome data.

References

[1] D. Hanahan and R. Weinberg, *Hallmarks of cancer: The next generation,* Cell **144**, 646 (2011).

[2] T. Zack, S. Schumacher, S. Carter, A. Cherniack, G. Saksena, *et al.*, *Pan-cancer patterns of somatic copy number alteration,* Nature Genetics **45**, 1134 (2013).

[3] C. Lu, M. Xie, M. Wendl, J. Wang, M. McLellan, *et al.*, *Patterns and functional implications of rare germline variants across 12 cancer types,* Nature Communications **6** (2015), 10.1038/ncomms10086.

[4] G. Tate, T. Tajiri, T. Suzuki, and T. Mitsuya, *Mutations of the kit gene and loss of heterozygosity of the pten region in a primary malignant melanoma arising from a mature cystic teratoma of the ovary,* Cancer Genetics and Cytogenetics **190**, 15 (2009).

[5] S. Thiagalingam, S. Laken, J. Willson, S. Markowitz, K. Kinzler, *et al.*, *Mechanisms underlying losses of heterozygosity in human colorectal cancers,* Proceedings of the National Academy of Sciences of the United States of America **98**, 2698 (2001).

[6] N. Sato, H. Tsunoda, M. Nishida, Y. Morishita, Y. Takimoto, *et al.*, *Loss of heterozygosity on 10q23.3 and mutation of the tumor suppressor gene pten in benign endometrial cyst of the ovary: Possible sequence progression from benign endometrial cyst to endometrioid carcinoma and clear cell carcinoma of the ovary,* Cancer Research **60**, 7052 (2000).

[7] S. Baker, A. Preisinger, J. Jessup, C. Paraskeva, S. Markowitz, *et al.*, *p53 gene mutations occur in combination with 17p allelic deletions as late events in colorectal tumorigenesis,* Cancer Research **50**, 7717 (1990).

[8] A. Berger, A. Knudson, and P. Pandolfi, *A continuum model for tumour suppression,* Nature **476**, 163 (2011).

[9] H. Dubbink, P. Atmodimedjo, R. van Marion, N. Krol, P. Riegman, *et al.*, *Diagnostic detection of allelic losses and imbalances by next-generation sequencing: 1p/19q co-deletion analysis of gliomas,* Journal of Molecular Diagnostics **18**, 775 (2016).

[10] S. Jones, V. Anagnostou, K. Lytle, S. Parpart-Li, M. Nesselbush, *et al.*, *Personalized genomic analyses for cancer mutation discovery and interpretation,* Science Translational Medicine **7** (2015), 10.1126/scitranslmed.aaa7161.

[11] P. Danecek, A. Auton, G. Abecasis, C. Albers, E. Banks, *et al.*, *The variant call format and vcftools,* Bioinformatics **27**, 2156 (2011).

[12] P. Van Loo, S. Nordgard, O. Lingjærde, H. Russnes, I. Rye, *et al.*, *Allele-specific copy number analysis of tumors,* Proceedings of the National Academy of Sciences of the United States of America **107**, 16910 (2010).

[13] V. Boeva, T. Popova, M. Lienard, S. Toffoli, M. Kamal, *et al.*, *Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data,* Bioinformatics **30**, 3443 (2014).

[14] H. Thorvaldsdóttir, J. Robinson, and J. Mesirov, *Integrative genomics viewer (igv): High-performance genomics data visualization and exploration,* Briefings in Bioinformatics **14**, 178 (2013).

[15] O. Westesson, M. Skinner, and I. Holmes, *Visualizing next-generation sequencing data with jbrowse,* Briefings in Bioinformatics **14**, 172 (2013).

[16] K. Rosenbloom, J. Armstrong, G. Barber, J. Casper, H. Clawson, *et al.*, *The ucsc genome browser database: 2015 update,* Nucleic Acids Research **43**, D670 (2015).

[17] D. Parsons, S. Jones, X. Zhang, J.-H. Lin, R. Leary, *et al.*, *An integrated genomic analysis of human glioblastoma multiforme,* Science **321**, 1807 (2008).

[18] S. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, *et al.*, *Dbsnp: The ncbi database of genetic variation,* Nucleic Acids Research **29**, 308 (2001).

[19] H. Li and R. Durbin, *Fast and accurate short read alignment with burrows-wheeler transform,* Bioinformatics **25**, 1754 (2009).

[20] R. Brouwer, M. van den hout, F. Grosveld, and W. van ijcken, *Narwhal, a primary analysis pipeline for ngs data,* Bioinformatics **28**, 284 (2012).

[21] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, *The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data,* Genome Research **20**, 1297 (2010).

[22] I. Teles Alves, S. Hiltemann, T. Hartjes, P. Van Der Spek, A. Stubbs, *et al.*, *Gene fusions by chromothripsis of chromosome 5q in the vcap prostate cancer cell line,* Human Genetics **132**, 709 (2013).

[23] R. Core Team, *R: A language and environment for statistical computing,* R: A Language and Environment for Statistical Computing (2013).

[24] V. Obenchain, M. Lawrence, V. Carey, S. Gogarten, P. Shannon, *et al.*, *Variantannotation: A bioconductor package for exploration and annotation of genetic variants,* Bioinformatics **30**, 2076 (2014).

[25] W. Huber, V. Carey, R. Gentleman, S. Anders, M. Carlson, *et al.*, *Orchestrating high-throughput genomic analysis with bioconductor,* Nature Methods **12**, 115 (2015).

[26] C. Boettiger, *An introduction to docker for reproducible research,* (2015) pp. 71–79.

[27] R. Ihaka and R. Gentleman, *R: A language for data analysis and graphics,* Journal of Computational and Graphical Statistics **5**, 299 (1996).

[28] P. Killela, Z. Reitman, Y. Jiao, C. Bettegowda, N. Agrawal, *et al.*, *Tert promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal,* Proceedings of the National Academy of Sciences of the United States of America **110**, 6021 (2013).

[29] K. Hatanpaa, S. Burma, D. Zhao, and A. Habib, *Epidermal growth factor receptor in glioma: Signal transduction, neuropathology, imaging, and radioresistance1,* Neoplasia **12**, 675 (2010).

[30] R. Siegel, K. Miller, and A. Jemal, *Cancer statistics, 2016,* CA Cancer Journal for Clinicians **66**, 7 (2016).

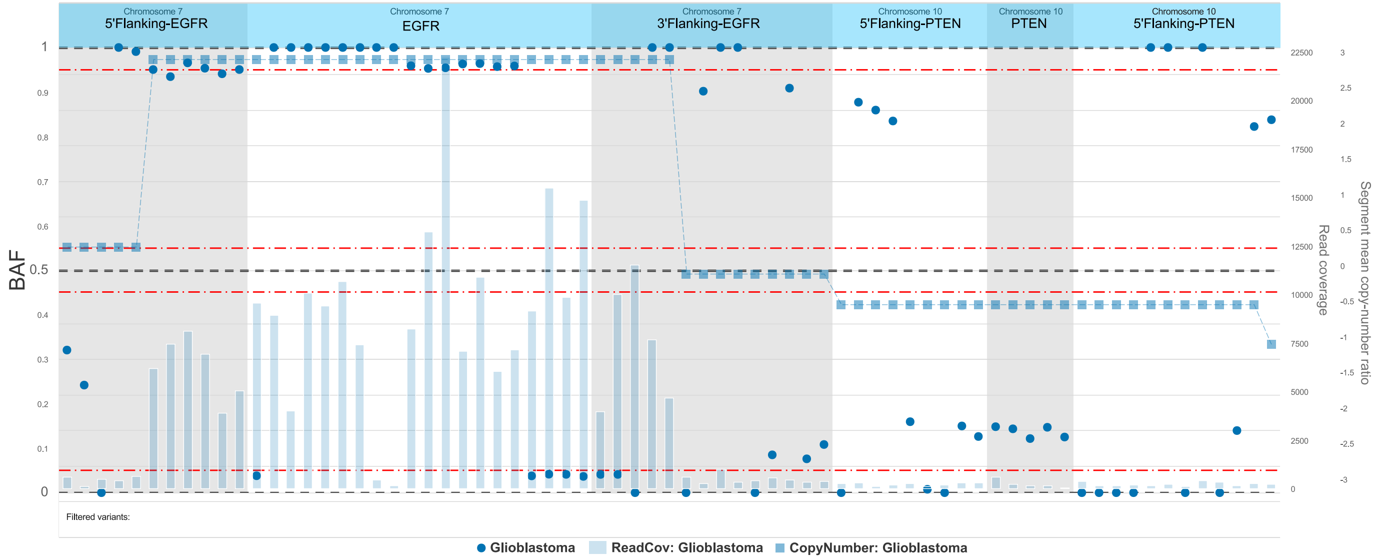
[31] F. Li, W.-Z. Zhong, F.-Y. Niu, N. Zhao, J.-J. Yang, *et al.*, *Multiple primary malignancies involving lung cancer,* BMC Cancer **15** (2015), 10.1186/s12885-015-1733-8.

[32] M. Srinivasan, D. Sedmak, and S. Jewell, *Effect of fixatives and tissue processing on the content and integrity of nucleic acids,* American Journal of Pathology **161**, 1961 (2002).

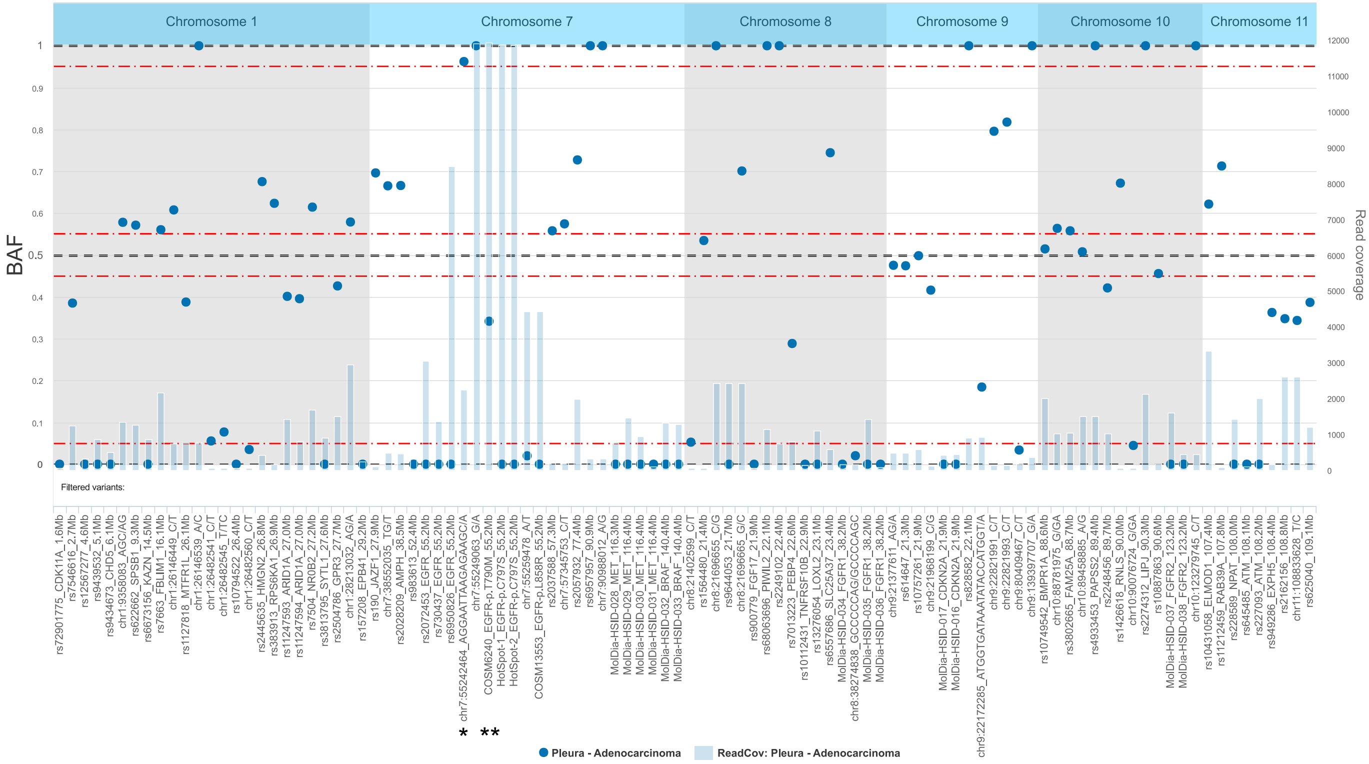
Supplemental Data

Supplementary data and figures accompanying the chapter:

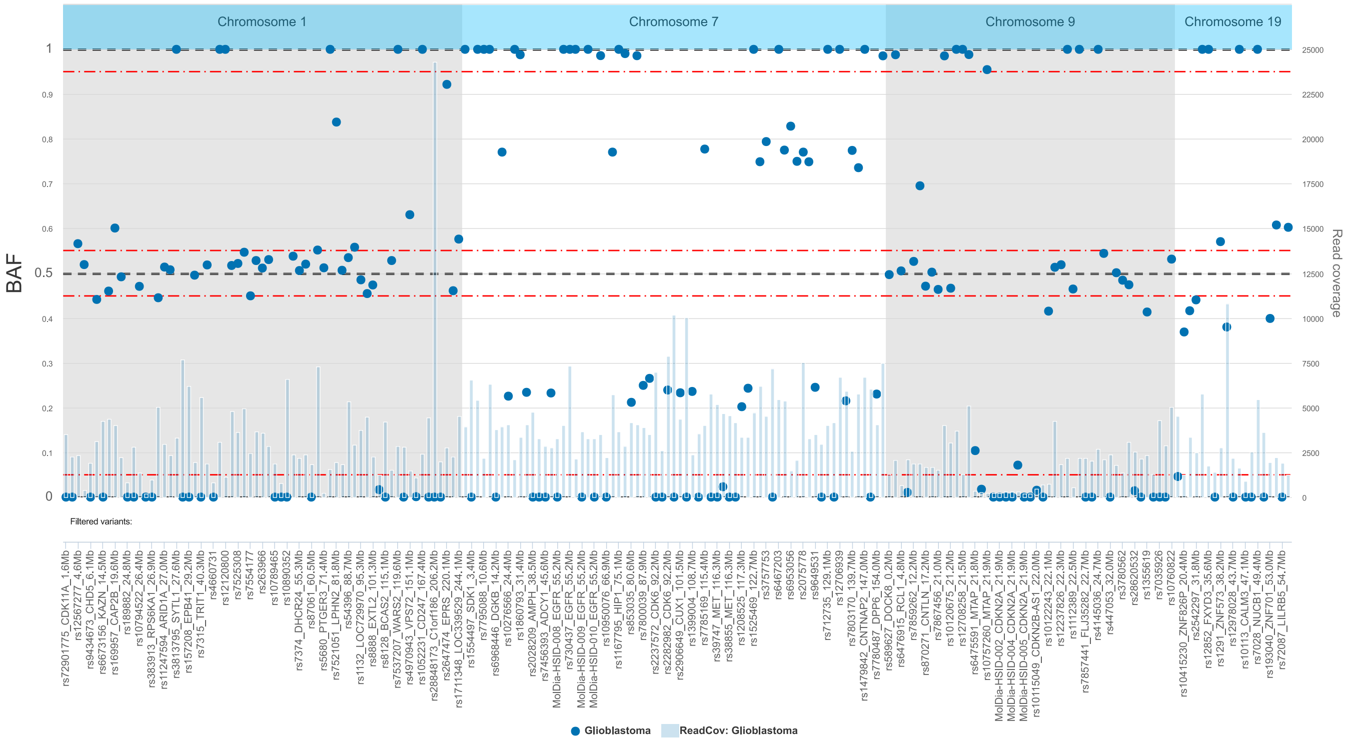
”SNPitty: An Intuitive Web Application for Interactive B-Allele Frequency and Copy Number Visualization of Next-Generation Sequencing Data”



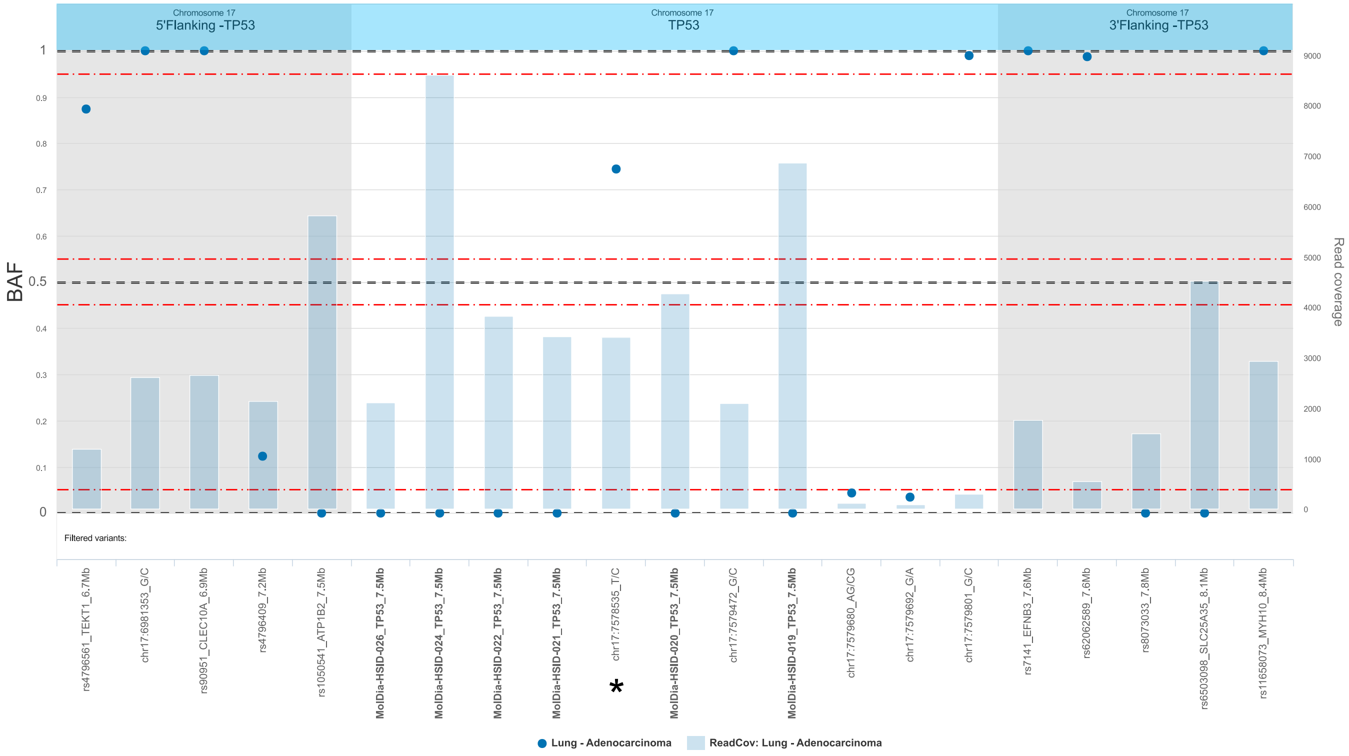
Supplementary figure S2.1: **Detection of *EGFR* amplification and heterozygous *PTEN* deletion in glioblastoma using SNPitty.** BAF visualization of heterozygous markers (blue circles), accompanied with respective read coverage (blue bars) and $\log_2$ segment, mean copy number ratio (dashed blue line with blue squares), on the *EGFR* and *PTEN* regions for a single glioblastoma sample. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



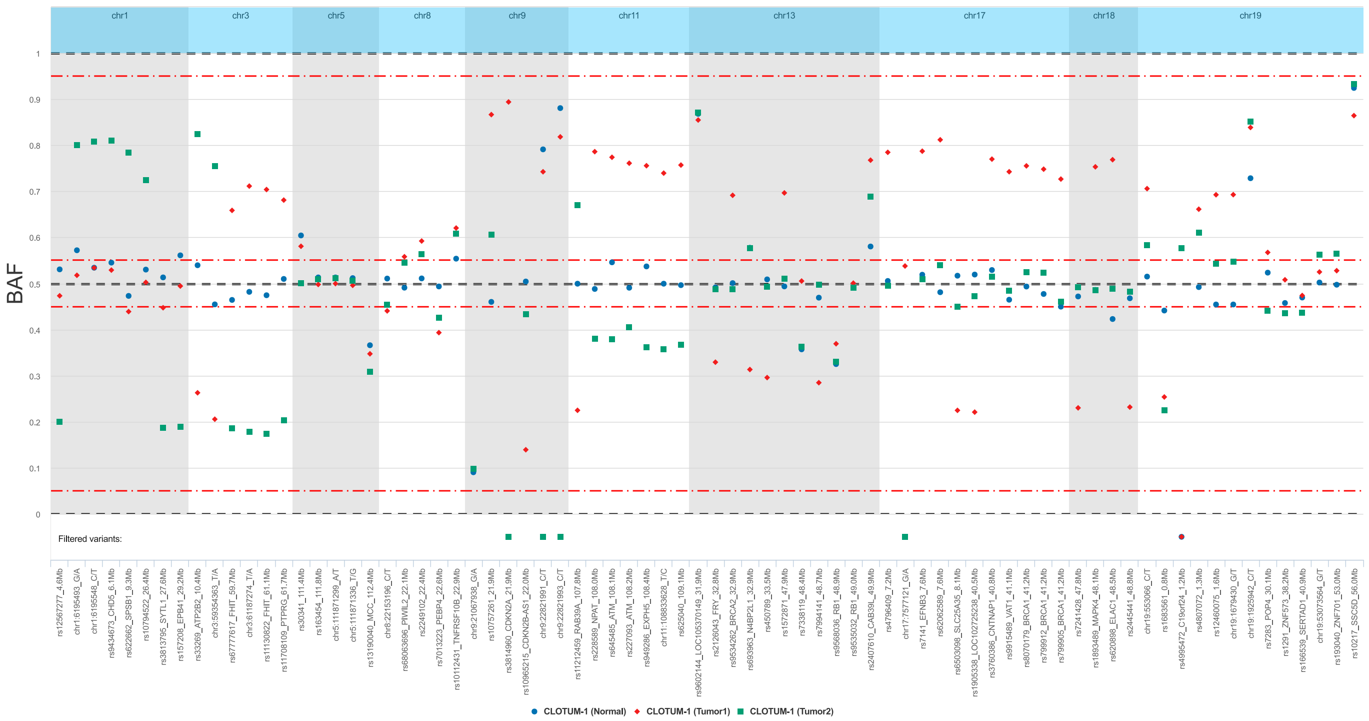
Supplementary figure S2.2: **Detection of *EGFR* amplification coupled with additional mutations in the *EGFR* region and AI on various chromosomes.**
BAF visualization of heterozygous markers (blue circles), accompanied with respective read coverage (blue bars) on chromosome 1, 7, 8, 9, 10, and 11 for a single pleura adenocarcinoma sample. Black dashed lines show the homozygous BAF ranges (0,1) whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. **EGFR* p.E746_A750delELREA somatic mutation. ***EGFR* p.T790M (c.2369C>T) somatic mutation.



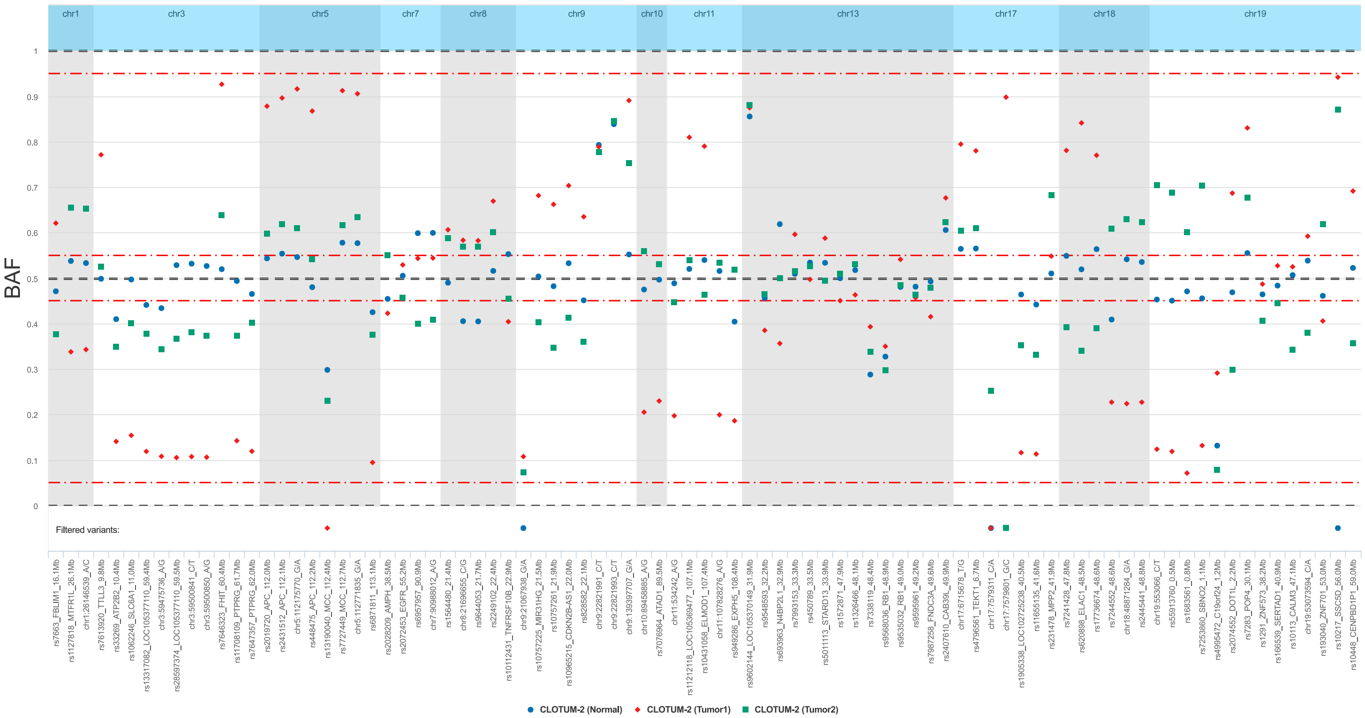
Supplementary figure S2.3: **Detection of homozygous deletion of *CDKN2A* region and regional allelic imbalances.**
BAF visualization of markers (covered with >50 reads) on chromosomes 1, 7, 9, and 19 for a single glioblastoma (blue circles) accompanied with respective read coverage (blue bars). Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. Briefly, a region of heterozygosity on 1p and 19q can be seen coupled with AI on chromosome 7 and a homozygous *CDKN2A* deletion.



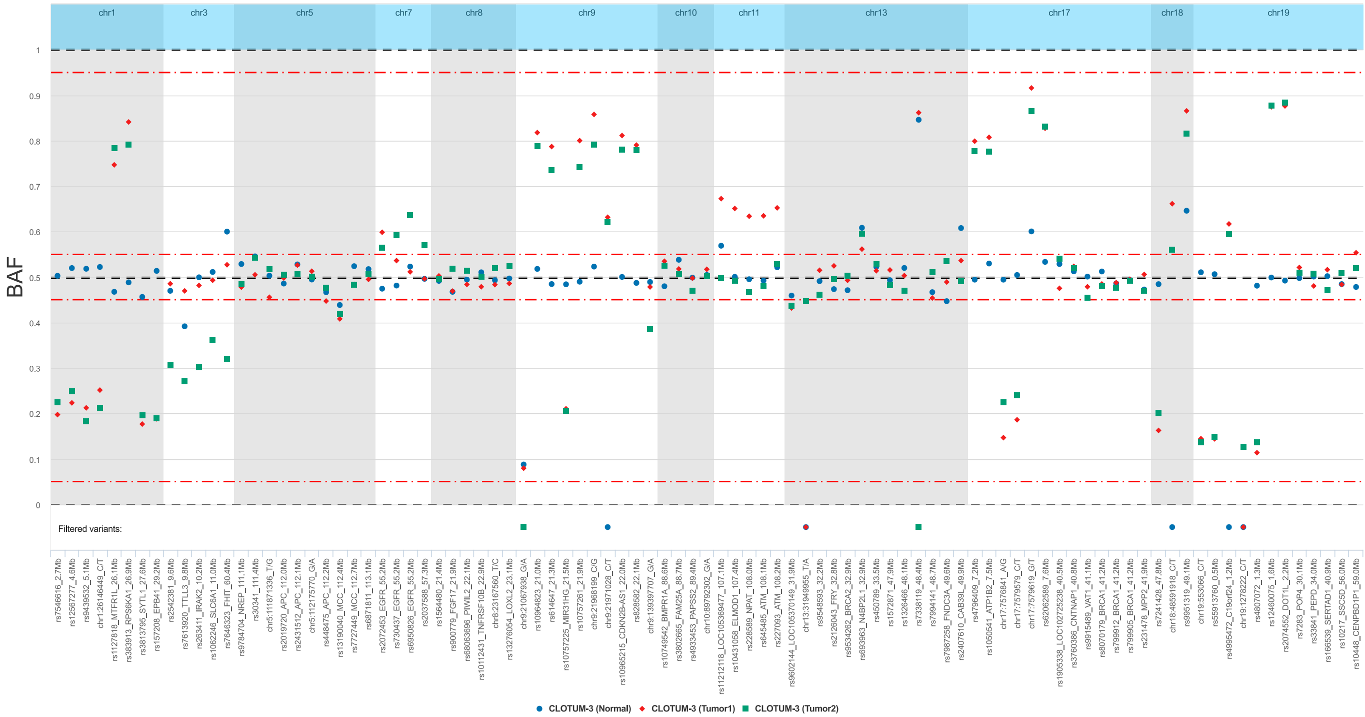
Supplementary figure S2.4: **Detection of *TP53* mutation in lung adenocarcinoma.**
BAF visualization of markers (covered with >50 reads) on the *TP53* region for a single lung adenocarcinoma (blue circles) accompanied with respective read coverage (blue bars). Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. Additional homozygous genomic positions for genomic coverage purposes are shown in bold. Briefly, a p.K132R (c.395A>G) mutation in the *TP53* region was found coupled with two markers showing LOH on the 5′ flanking region of *TP53*. IHC confirmed overexpression of p53 (data not shown). *TP53 p.K132R (c.395A>G) somatic mutation.



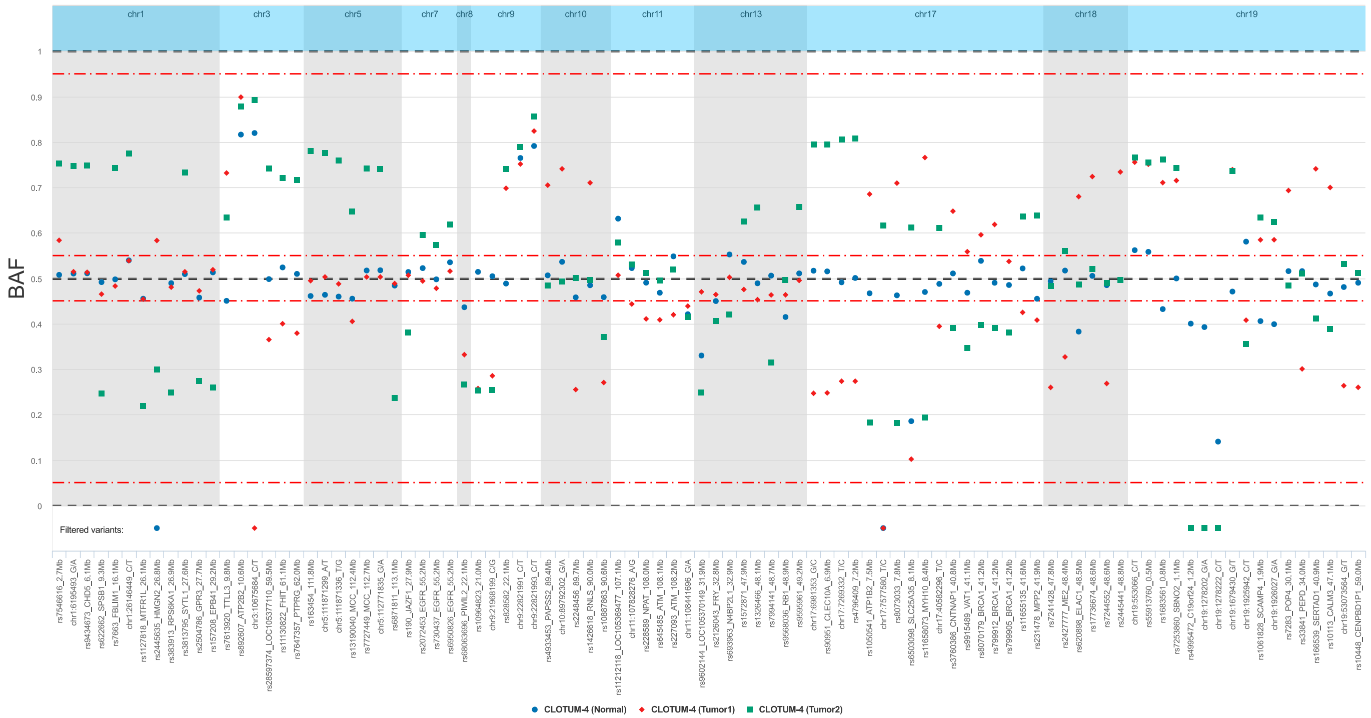
Supplementary figure S2.5: **Clonality assessment of diagnostic sample CLOTUM-1.**
BAF visualization of heterozygous markers (covered with >50 reads and BAF between 0.05 and 0.95) on chromosomes 1, 3, 5, 8, 9, 11, 13, 17, 18, and 19 for two tumors (red rhombi and green squares) and a single matched normal sample (blue circles) derived from a single patient. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



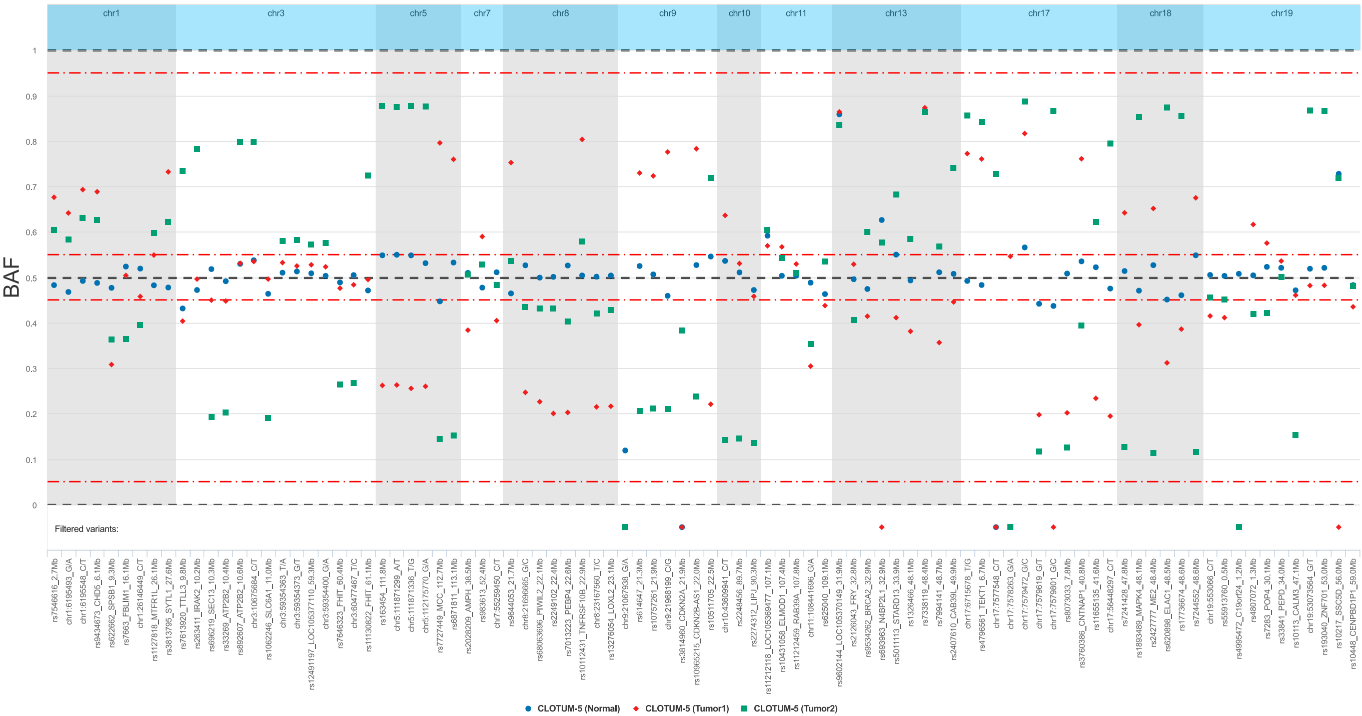
Supplementary figure S2.6: **Clonality assessment of diagnostic sample CLOTUM-2.**
BAF visualization of heterozygous markers (covered with >50 reads and BAF between 0.05 and 0.95) on chromosomes 1, 3, 5, 7, 8, 9, 10, 11, 13, 17, 18, and 19 for two tumors (red rhombi and green squares) and a single matched normal sample (blue circles) derived from a single patient. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



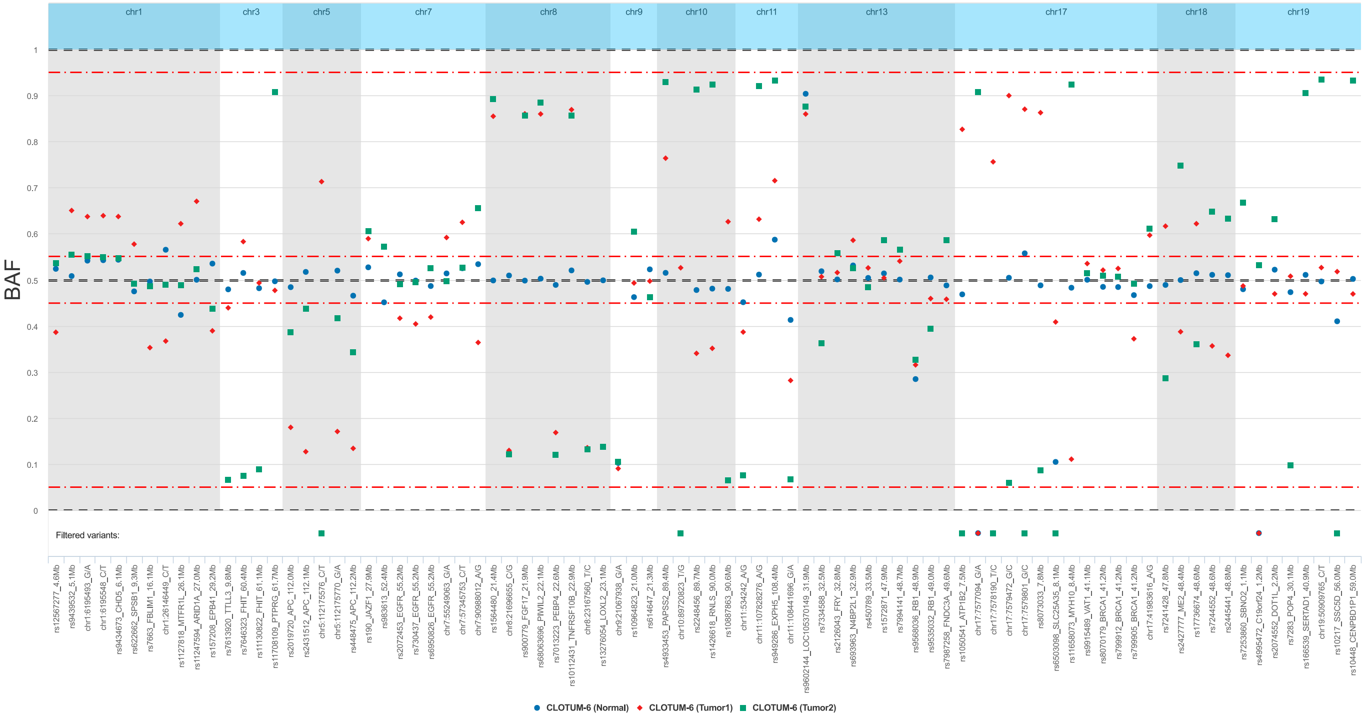
Supplementary figure S2.7: **Clonality assessment of diagnostic sample CLOTUM-3.**
BAF visualization of heterozygous markers (covered with >50 reads and BAF between 0.05 and 0.95) on chromosomes 1, 3, 5, 7, 8, 9, 10, 11, 13, 17, 18, and 19 for two tumors (red rhombi and green squares) and a single matched normal sample (blue circles) derived from a single patient. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



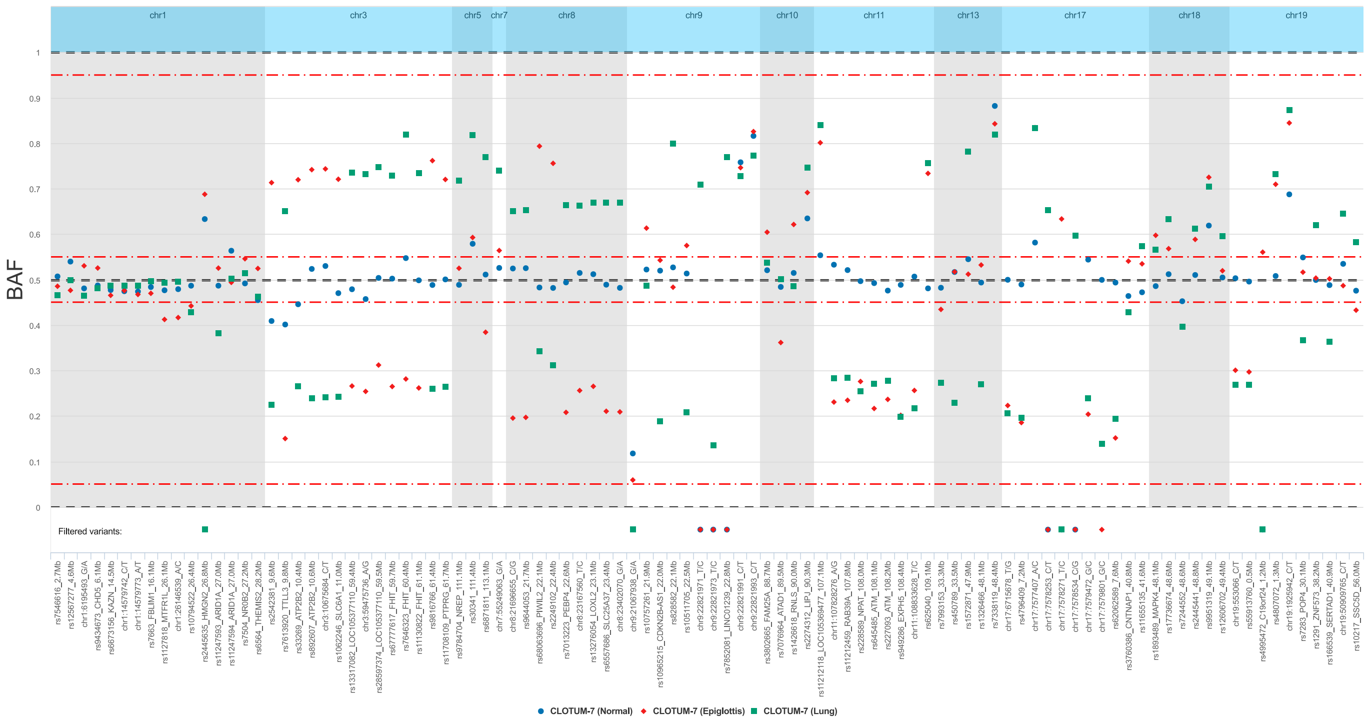
Supplementary figure S2.8: **Clonality assessment of diagnostic sample CLOTUM-4.**
BAF visualization of heterozygous markers (covered with >50 reads and BAF between 0.05 and 0.95) on chromosomes 1, 3, 5, 7, 8, 9, 10, 11, 13, 17, 18, and 19 for two tumors (red rhombi and green squares) and a single matched normal sample (blue circles) derived from a single patient. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



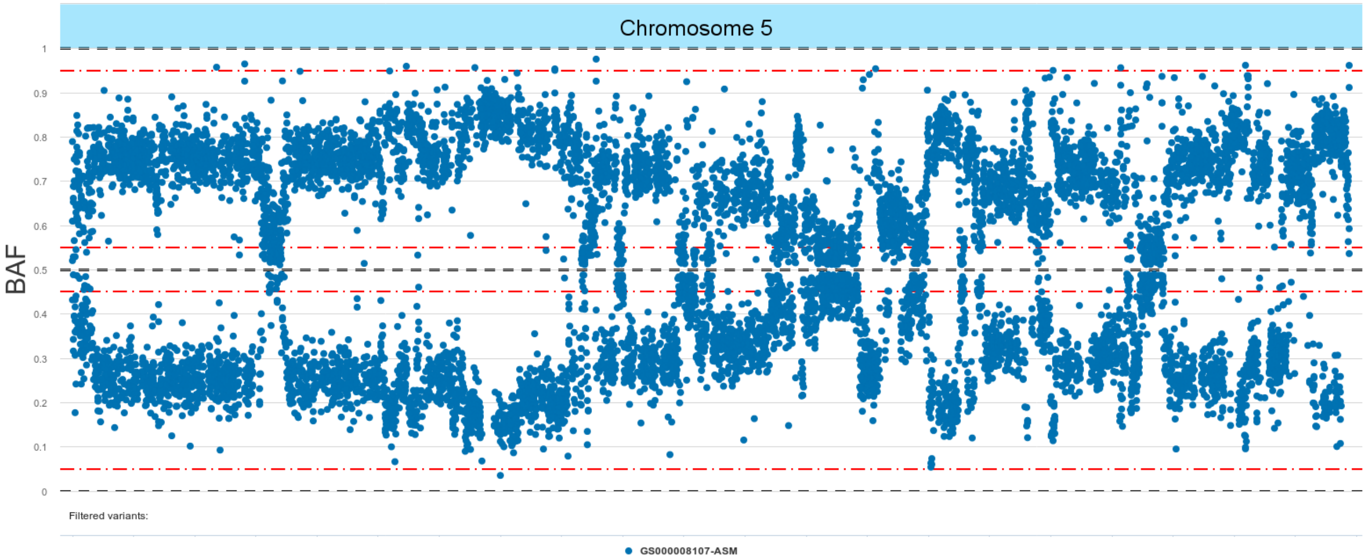
Supplementary figure S2.9: **Clonality assessment of diagnostic sample CLOTUM-5.**
BAF visualization of heterozygous markers (covered with >50 reads and BAF between 0.05 and 0.95) on chromosomes 1, 3, 5, 7, 8, 9, 10, 11, 13, 17, 18, and 19 for two tumors (red rhombi and green squares) and a single matched normal sample (blue circles) derived from a single patient. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



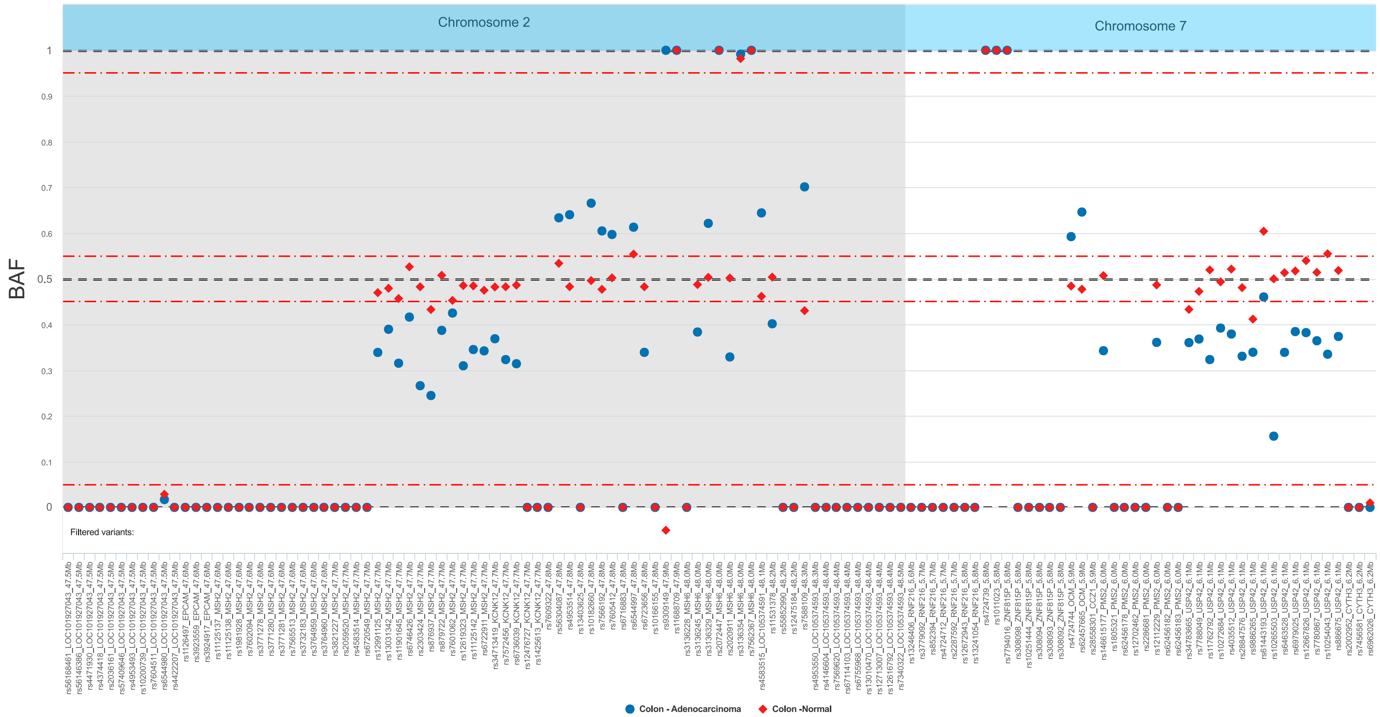
Supplementary figure S2.10: **Clonality assessment of diagnostic sample CLOTUM-6.**
BAF visualization of heterozygous markers (covered with >50 reads and BAF between 0.05 and 0.95) on chromosomes 1, 3, 5, 7, 8, 9, 10, 11, 13, 17, 18, and 19 for two tumors (red rhombi and green squares) and a single matched normal sample (blue circles) derived from a single patient. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



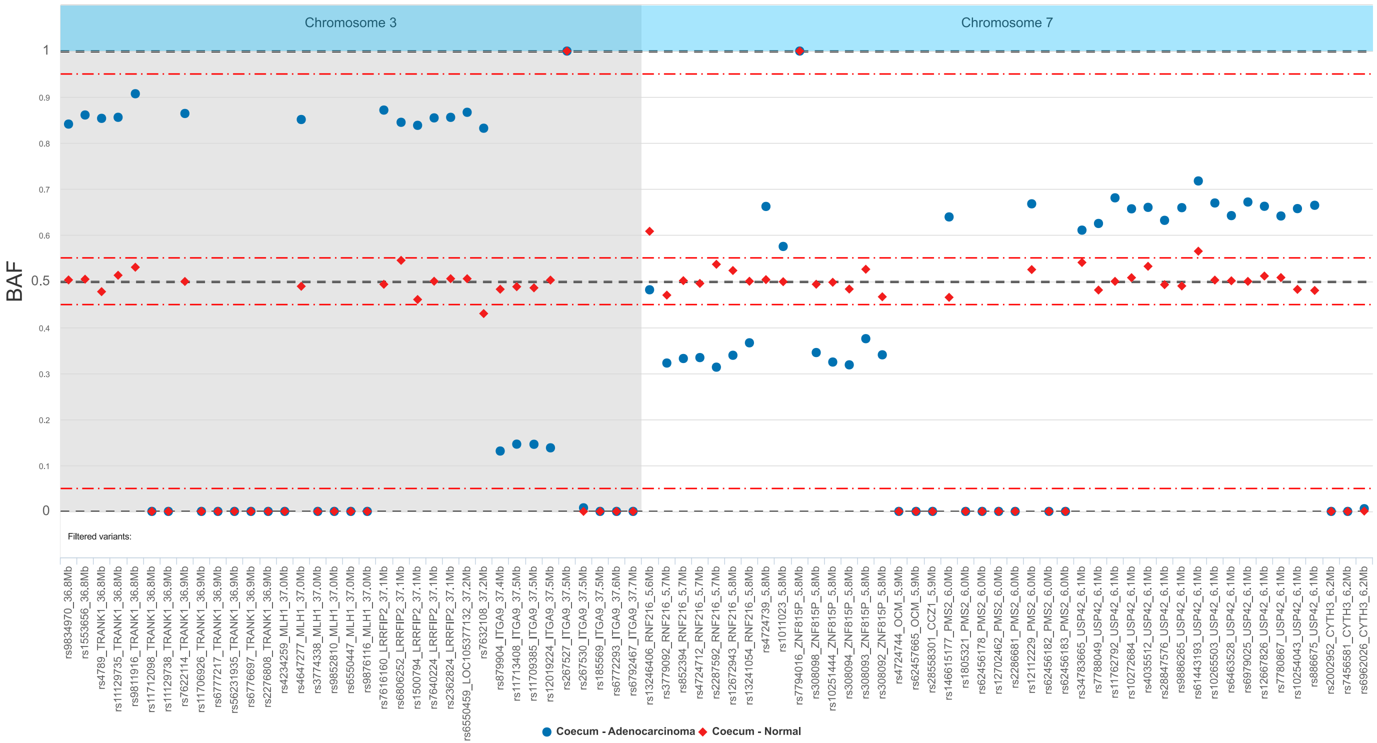
Supplementary figure S2.11: **Clonality assessment of diagnostic sample CLOTUM-7.**
BAF visualization of heterozygous markers (covered with >50 reads and BAF between 0.05 and 0.95) on chromosomes 1, 3, 5, 7, 8, 9, 10, 11, 13, 17, 18, and 19 for two tumors (red rhombi and green squares) and a single matched normal sample (blue circles) derived from a single patient. Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



Supplementary figure S2.12: **Reproduction of BAF overview of chromosome 5 in the prostate cancer cell line VCaP.**
BAF visualization of informative heterozygous markers (covered with >150 reads and BAF between 0.05 and 0.95) on chromosome 5 of the VCaP genome (blue circles). Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively.



Supplementary figure S2.13: **Assessment of diagnostic colon adenocarcinoma sample.**
BAF visualization of markers (covered with >50 reads) on chromosomes 2 and 7 for a colon adenocarcinoma (blue circles) and single matched (red rhombi). Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. Briefly, LOH of the mismatch-repair genes *MSH2*, *MSH6*, and *PMS2*, and the *USP42* region can be appreciated. Subsequent IHC confirmed lack of *MSH2* and *MSH6* expression in the nucleus of malignant tissue but did not detect *PMS2* abnormalities (data not shown).



Supplementary figure S2.14: **Assessment of diagnostic cecum adenocarcinoma sample.**
BAF visualization of markers (covered with >50 reads) on chromosomes 3 and 7 for a cecum adenocarcinoma (blue circles) and single matched normal sample (red rhombus). Black dashed lines show the homozygous BAF ranges (0,1), whereas the red dash-dot lines reflect the putative borders of balanced heterozygosity or homozygosity BAF values 0.05, 0.45, 0.55, 0.95, respectively. Briefly, LOH of chromosome 3 can be seen coupled with allelic imbalances on chromosome 7.

Chapter 3

ProteoDisco: A flexible R approach to generate customized protein databases for extended search space of novel and variant proteins in proteogenomic studies

W.S. van de Geer[a,b*], **J. van Riet**[a-c*], H.J.G. van de Werken[a,c,d]

[a] Cancer Computational Biology Center, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.

[b] Department of Medical Oncology, Erasmus MC Cancer Institute, University Medical Center, Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands.

[c] Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.

[d] Department of Immunology, Erasmus MC Cancer Institute, University Medical Center, Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands.

[*] These authors contributed equally.

This chapter has been published in **Bioinformatics** Volume 38, Issue 5, 1 March 2022, Pages 1437–1439, https://doi.org/10.1093/bioinformatics/btab809

Abstract

Summary: We present an R-based open-source software termed ProteoDisco that allows for flexible incorporation of genomic variants, fusion-genes and (aberrant) transcriptomic variants from standardized formats into protein variant sequences. ProteoDisco allows for a flexible step-by-step workflow allowing for in-depth customization to suit a myriad of research approaches in the field of proteogenomics, on all organisms for which a reference genome and transcript annotations are available.

Availability: ProteoDisco (R package version ≥ 0.99) is available from https://github.com/ErasmusMC-CCBC/ProteoDisco/ and https://doi.org/doi:10.18129/B9.bioc.ProteoDisco.

Introduction

The rise and ease of current Next-Generation Sequencing (NGS) techniques, coupled with reduced costs in both NGS and high-resolution mass-spectrometry, offers opportunity to incorporate sample-specific protein variants during proteomics experiments for increased accuracy and detection rates of, for instance, distinctive proteotypic peptides in bottom-up proteomics experiments. Expanding the repertoire of proteins and these proteotypic peptides can provide novel insights into disease-specific protein variants, their underlying molecular profiles and regulation, neoantigen prediction and expand our knowledge on the genetic variations encoded in proteomes.[1–5] This is further fueled by the standardization and publication of proteomics resources which allows for the interrogation and combination of existing datasets.[6,7] Rising global efforts in capturing the genetic sequences of diverse organisms, disease-related genotypes and their transcriptomes with subsequent proteome-resources warrants the implementation of a flexible yet intuitive toolset. This toolset should provide a bridge between genomic and transcriptomic variants and their incorporation within respective protein variants (proteogenomics) using industry-standard infrastructure, such as Bioconductor[8], and allow for flexibility in facilitating the myriad experimental settings applied in research. Therefore, we designed and developed ProteoDisco, an open-source R software-package using existing Bioconductor class-infrastructures to allow for the accurate and flexible generation of variant protein sequences and their derived proteotypic peptides from the incorporation of sample-specific genomic and transcriptomic information. In addition, we present the results of ProteoDisco and two similar open-source tools which are frequently utilized within proteogenomics (customProDB[9] and QUILTS[3]) with their performance in generating correct protein variants and respective proteotypic peptides from supplied genomic variants.

Approach

ProteoDisco incorporates genomic variants, splice-junctions (derived from transcriptomics) and fusion genes within provided reference genome sequences and transcript-annotations to generate their respective protein variant sequence(s). These sequences can be curated, altered and subsequently exported into a database in FASTA-format for use in downstream analysis. To limit the number of generated protein variants, ProteoDisco provides filtering options based on a minimal number of distinct proteotypic (identifiable) peptides. The global workflow of ProteoDisco is summarized in six steps as depicted within Figure 3.1. In addition, an extended overview of how (novel) splice-junctions and gene-fusion events are incorporated is shown in Supplementary Figure S3.1.

To compare the accuracy of ProteoDisco against two common alternatives for proteogenomics studies (customProDB[9] and QUILTS[3]), we utilized a manually-curated dataset and two large independent proteomics studies. The manually-curated dataset contained 28 genomic variants reported in COSMIC[10] comprising multiple variant classes; synonymous and nonsynonymous single-nucleotide variants (SNVs), multi-nucleotide variants (MNVs) and in- and out-of-frame insertions/deletions (InDels). In addition, we utilized recently-published results from large-scale colon and breast cancer cohorts within the Clinical Proteomic Tumor Analysis Consortium (CPTAC) to illustrate the accuracy of ProteoDisco in generating identical proteotypic peptides as detected within these studies.[2,5] This comparison revealed that ProteoDisco correctly generated proteotypic peptides from their respective genomic variants after thorough checking and yielded the highest number of expected and reconstructed proteotypic peptides within all three datasets (Supplementary Figure S3.2). This difference can be attributed to ProteoDisco's native flexibility in reference genome selection, multiple incorporation strategies, sanity-checks such as reference base verification and the correct incorporation of stop-loss variants. In total, only four enigmatic genomic variants (of three fragments) from Mertins *et al.* could not be reconstructed to reproduce their proteotypic peptide(s).

Conclusion

In this article, we present ProteoDisco, a suitable, open-source and flexible suite for the generation of protein variant databases usable in downstream proteogenomic studies and capable of correctly incorporating a diverse range of genomic variants and transcriptomic splice-junctions. We report that ProteoDisco accurately produces protein variant sequences harboring previously-identified proteotypic fragments from their respective genomic variants. Further examples and use-cases can be found in the vignette of the ProteoDisco package.

Methods

Technical design of ProteoDisco

ProteoDisco was programmed within the R statistical language (v4.1.1) and built upon existing classes within the Bioconductor infrastructure (v3.13) to allow flexible inheritance and future extensions. Additional information on the usage and design of ProteoDisco can found in the extended methodology (**Supplementary M&M**).

Assessment of the correct integration of genomic variants into protein variants

We generated a custom validation-dataset containing established somatic variants (SNVs, MNVs and InDels; *n* = 28) and their respective protein variants as listed within COSMIC[10] (v92; GRCh37; **Online Suppl. Table 1**). In addition, we utilized recent proteogenomics studies from the CPTAC cancer cohorts containing genomic variants and their respective *in silico* generated proteotypic peptides which had been measured and identified using high-throughput proteomics approaches.[2,5] In the Wen *et al.* dataset[5] (CPTAC - Colon Cancer), genomic variants (and their respective proteotypic peptides) were split into sample-specific variant call format (VCF)-files based on the data present within their published Suppl. Data S15 (see reference, sheet 1: 'prospective_colon_label_free_in'). The Mertins *et al.* dataset[2] (CPTAC - Breast Cancer) was aggregated into a single VCF-file based on the data present within their published Suppl. Table S2 (see reference, sheet 2: 'Variants').

Using these three datasets, we ran ProteoDisco (v0.99), customProDB (v1.30.1) and the web-interface of QUILTS (v3.0; accessed 13-04-2021) to generate custom protein-variant databases using uniform University of California, Santa Cruz (UCSC)RefSeq[11] (GRCh37) transcript-annotations and settings. The custom protein-variant databases were generated based on two approaches within ProteoDisco. The first approach incorporated each genomic variant independently and the second allowed for the simultaneous incorporation of all genomic variants per overlapping transcript-annotation, e.g., two variants on different coding exons would both be incorporated within the resulting variant protein-sequence. Incorporation of all possible combinations of mutant exons yields too many combinations and is therefore not included amongst the options. The generated variant protein sequences and respective proteotypic peptides from each customized protein-variant database were compared against the proteotypic peptides as expected from COSMIC or as detected within the respective CPTAC-studies using all three tools (Supplementary Figure S3.1). E.g., if ProteoDisco generated three distinct proteotypic peptides for a given genomic variant and one of those was identified within CPTAC (or COSMIC), it was counted as a concordant result.

Code availability

All source-code has been made available within Bioconductor (`https://doi.org/doi:10.18129/B9.bioc.ProteoDisco`) and deposited within GitHub (`https://github.com/ErasmusMC-CCBC/ProteoDisco`) under the GPL-3 license.

Data availability

The custom validation dataset (GRCh37) which has been used in the analysis as presented within this manuscript has been stored within ProteoDisco and is accessible at `https://github.com/ErasmusMC-CCBC/ProteoDisco/main/inst/extdata`. COSMIC (v92; accessed on 14-04-2021) was used to derive the validation dataset (GRCh37), the external validation datasets based on CPTAC (colon and breast cancer) were generated based on the supplementary data published by Wen *et al.*[5] and Mertins *et al.*[2].

Author contributions

All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: van Riet, van de Geer, van de Werken. Acquisition of data: van Riet, van de Geer. Analysis and interpretation of data: All authors. Drafting of the manuscript: All authors. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: All authors. Obtaining funding: None. Administrative, technical, or material support: All authors. Supervision: van de Werken. Other: None.

Acknowledgements

We would like to thank S. Eldeb and W. M. Hoska for their preliminary work on the integration of mutations into transcript sequences. Figures 3.1 and S3.1 were

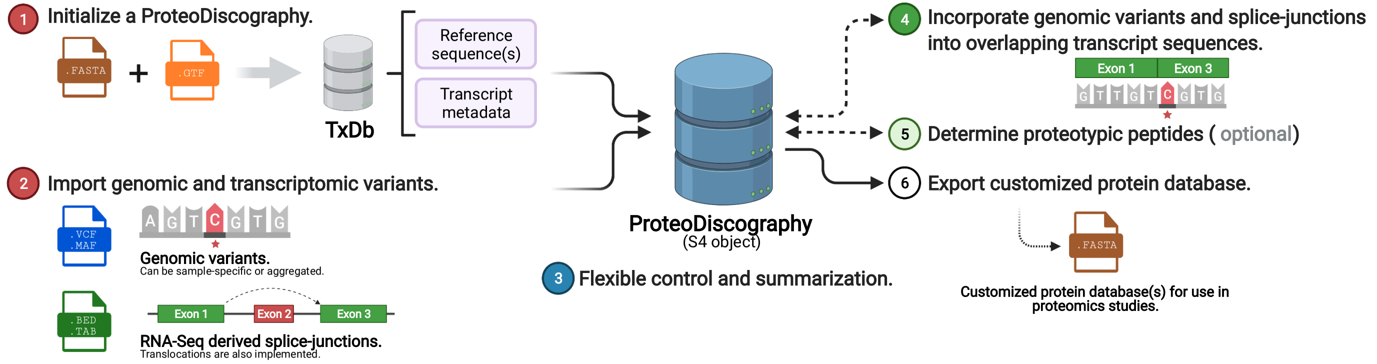

1 Initialize a ProteoDiscography.
.FASTA
+
.GTF
TxDb
Reference sequence(s)
Transcript metadata
2 Import genomic and transcriptomic variants.
.VCF
.MAF
A G T C G T G
Genomic variants.
Can be sample-specific or aggregated.
.BED
.TAB
Exon 1
Exon 2
Exon 3
RNA-Seq derived splice-junctions.
Translocations are also implemented.
ProteoDiscography
(S4 object)
3 Flexible control and summarization.
4 Incorporate genomic variants and splice-junctions into overlapping transcript sequences.
Exon 1
Exon 3
G T T G T C G T G
5 Determine proteotypic peptides (optional)
6 Export customized protein database.
.FASTA
Customized protein database(s) for use in proteomics studies.

Figure 3.1: **Schematic overview of the ProteoDisco workflow.**
The global workflow of ProteoDisco can be categorized as six major steps. **1)** Initialize a ProteoDiscography by utilizing custom references sequence(s) and gene-annotation(s) or using pre-existing TxDb objects. **2)** Import (sample-specific) genomic variants, splice-junctions or manual sequences. Several sanity-checks are performed during importation, including the validation of matching reference nucleotide(s). **3)** Dynamically view, extend, alter and customize imported records and derived sequences. **4)** Incorporation of genomic variants and splice-junctions into overlapping transcript-annotations, translocations between chromosomes can also be processed. The incorporation can be performed in a sample-specific manner, exon or transcript-specific manner or in an aggregated manner. **5)** Cleave derived protein variant sequences and determine proteotypic peptides, per protein, which are not present within the reference protein sequences (TxDb) or additional protein databases (e.g., UniProKB). **6)** Export the derived protein variant sequences into an external protein-sequence database(s) using FASTA format.

created using BioRender.com.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] A. I. Nesvizhskii, *Proteogenomics: Concepts, applications and computational strategies,* Nature Methods **11**, 1114 (2014).

[2] P. Mertins, D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, *et al.*, *Proteogenomics connects somatic mutations to signalling in breast cancer,* Nature **534**, 55 (2016).

[3] K. V. Ruggles, K. Krug, X. Wang, K. R. Clauser, J. Wang, *et al.*, *Methods, tools and current perspectives in proteogenomics,* Molecular and Cellular Proteomics **16**, 959 (2017).

[4] S. Vasaikar, C. Huang, X. Wang, V. A. Petyuk, S. R. Savage, *et al.*, *Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities,* Cell **177**, 1035 (2019).

[5] B. Wen, K. Li, Y. Zhang, and B. Zhang, *Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis,* Nature Communications **11** (2020), 10.1038/s41467-020-15456-w.

[6] E. W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, *et al.*, *The proteomexchange consortium in 2017: Supporting the cultural change in proteomics public data deposition,* Nucleic Acids Research **45**, D1100 (2017).

[7] M. Zahn-Zabal, P. A. Michel, A. Gateau, F. Nikitin, M. Schaeffer, *et al.*, *The nextprot knowledgebase in 2020: Data, tools and usability improvements,* Nucleic Acids Research **48**, D328 (2020).

[8] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, *et al.*, *Bioconductor: open software development for computational biology and bioinformatics.* Genome biology **5**, R80 (2004).

[9] X. Wang, B. Zhang, and J. Wren, *Customprodb: An r package to generate customized protein databases from rna-seq data for proteomics search,* Bioinformatics **29**, 3235 (2013).

[10] S. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, *et al.*, *Cosmic: Somatic cancer genetics at high-resolution,* Nucleic Acids Research **45**, D777 (2017).

[11] A. Frankish, M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, *et al.*, *Gencode reference annotation for the human and mouse genomes,* Nucleic Acids Research **47**, D766 (2019).

[12] V. Obenchain, M. Lawrence, V. Carey, S. Gogarten, P. Shannon, *et al.*, *Variantannotation: A bioconductor package for exploration and annotation of genetic variants,* Bioinformatics **30**, 2076 (2014).

[13] C. Trapnell, L. Pachter, and S. L. Salzberg, *Tophat: Discovering splice junctions with rna-seq,* Bioinformatics **25**, 1105 (2009).

[14] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, *Star: Ultrafast universal rna-seq aligner,* Bioinformatics **29**, 15 (2013).

Supplemental Data

Supplementary data and figures accompanying the chapter:

”ProteoDisco: A flexible R approach to generate customized protein databases for extended search space of novel and variant proteins in proteogenomic studies”

Extended Materials and Methodology on the design of ProteoDisco.

The major workflow of ProteoDisco can be divided into six steps;

1. Generation of the ProteoDiscography containing the reference genome sequences and transcript annotations of choice, as detailed below. This ProteoDiscography will also house the imported genomic and transcriptomic input and subsequent *in silico* generated protein variants and related information.

2. Import of genomic variants (either VCF and MAF files or VRanges objects[12]) or splice-junctions from transcriptomics such as .BED output from TopHat[13], SJ.out.tab output from STAR[14] or manual entries following a simple format to for instance denote translocations and/or fusion-gene events (e.g., *TMPRSS2-ERG*). ProteoDisco is capable of handling SNVs, InDels, MNV variants of both non-synonymous as synonymous variants.

3. Integration of genomic variants and splice-junctions into their respective transcripts and coding sequence (CDS). Translation of *in silico* generated transcript variants into their respective protein variants, the genetic code used for translation can be altered to allow for divergent translation tables for non-standard organisms.

4. Determine the number of proteotypic peptides per transcript variant, this can be determined based against the given reference database (as given to the ProteoDiscography) or be extended with additional protein-sequence databases. In addition, ProteoDisco can also check for proteotypic peptides compared to the other generated protein variants.

5. Export of the generated protein variants into a distinct FASTA database for use in downstream proteomics analysis to extend the (sample-specific or cohort-wide) search-space.

1. Generation and design of the ProteoDiscography; the internal datastructure.

All reference genome sequences (BSGenome objects), transcript annotations (TxDb objects) and generated results (BioStrings, tibbles and DataFrames) throughout ProteoDisco are housed within a custom (S4-class) termed ProteoDiscography. The reference database and transcript annotations for the ProteoDiscography can be generated in two ways; using pre-generated BSGenome (reference sequences) and TxDb (transcript annotations) objects, for instance available from BioConductor, or by supplying the reference genome sequences and transcript annotations (FASTA and gene transfer format v2 (GTF)/GFF file, respectively) which in turn generates these objects. In addition, the genetic code can be specified (as detailed by Biostrings) to also allow for non-standard translation tables.

2. Import genomic variants and splice-junctions within the ProteoDiscography.

After initialization of a ProteoDiscography, genomic variants and splice-junctions can be imported. Genomic variants (or somatic mutations) can be imported from .VCF or .MAF files or VRanges objects containing the genomic positions, strand and reference/variant alleles. By default, all given reference anchors (genomic position(s) and reference allele) of the genomic variants are checked against the provided reference genome and nucleotide at the respective position(s) to prevent inconsistencies. If non-matching reference anchors are detected, ProteoDisco will either halt the import-process and whilst displaying the erroneous records or, by setting ignoreNonMatch = TRUE, it will report and remove these non-matching records and continue with the remainder.

Splice-junctions can be imported from standard .BED (e.g., TopHat) and.SJ.out.tab (e.g., STAR) files or manually supplied using a simple format. Each of these formats should detail the genomic position (and optionally, strand information) of the donor and acceptor junction-sites for each splice-junction (junctionA and junctionB, respectively). Manual input can be supplied using the following format:

1. junctionA: Genomic coordinates of the 5'-junction (i.e., the position of the first intronic base). Format: chr:start:strand, i.e.: chr1:100:+
2. junctionB: Genomic coordinates of the 3'-junction (i.e., the position of the last intronic base). Format: chr:start:strand, i.e.: chr1:150:+
3. sample: Sample-identifier. (*optional*)
4. identifier: Identifier for the splice-junction, this identifier will be used to denote the splice-junction in downstream analysis. (*optional*)

This manual-input can also be used to supply splice-junctions from translocation events such as *BCR-ABL* which result in a protein variant containing exonic sequences from two chromosomes.

In addition, users can also supply pre-determined full-length transcript sequences into the ProteoDiscography. These manually-supplied transcript sequences can then also be used to determine proteotypic peptides compared to the reference database and/or protein variants. ProteoDisco houses functions to detect duplicate samples and overwrite these (if required) or append new genomic variants and/or splice-junctions to existing samples (based on sample names). In addition, it can also be toggled to remove all pre-existing samples within the ProteoDiscography prior to importation of new input.

3. Incorporation of genomic variants and splice-junctions within the coding sequence of overlapping transcripts.

ProteoDisco facilitates options to incorporate all supplied genomic variants (incl. synonymous variants) for all samples simultaneously or to perform this on a per-sample basis (aggregateSamples). Similarly flexible, users can choose between incorporating all mutations (per-sample or all samples aggregated) within the same transcript (e.g., a single RNA transcript containing 5 mutations; aggregateWithinTranscript = TRUE) or to generate separate transcripts, each harboring only a single mutation (e.g., 5 transcripts for 5 mutations; aggregateWithinTranscript = FALSE). Finally, users have similar functionality at exon-level (aggregateWithinExon).

Based on the parameters set by the user, genomic variants are overlapped with the coding sequences (CDS) of each transcript within the supplied TxDb. Per variant, all overlapping CDS (from one or multiple transcripts) will be altered by incorporating the overlapping genomic variant(s) at the correct coding position. The reference anchor (reference allele) will be checked if this conforms to the nucleotide at the coding position, taking in mind the orientation of the CDS. Genomic variants (e.g., InDels) overlapping the intron-exon or exon-intron boundary of a CDS will be split and only the CDS-overlapping portion will be incorporated.

After all genomic variants have been incorporated within their overlapping CDS in the transcript(s), the transcript sequence is generated by stitching all CDS of the transcript from 5’ to 3’ together. Based on the parameters set by the user, this will either results in a single transcript variant containing all mutant CDS or multiple transcripts with distinct mutant CDS. In addition, the 3’ untranslated region

(3′ untranslated region (UTR)) is also added to the mutant transcript sequence to capture additional coding nucleotides after a possible loss of the canonical stop-codon.

Splice-junctions are handled by determining the nearest adjacent (5′ and 3′) or overlapping CDS sequences. If a splice-junction overlaps with an existing CDS, that CDS will be used as assigned CDS and be altered to start (if 5′; junctionA) or end (3′; junctionB) at the genomic position of the respective junction, resulting in a shortened CDS. If the junctions do not overlap with an existing CDS, it will be assigned to the nearest adjacent CDS, taking in mind the orientation and strand of the splice-junction and CDS. This will either retrieve a canonical CDS directly flanking the splice-junctions or assign a CDS further away. The nucleotides spanning the splice-junction to the assigned CDS will then be added to assigned CDS and thereby effectively extending the CDS. If the splice-junction is further away than a max. distance (as set by the user; default 250 nt), a cryptic exon (of a size set by the user; default = 99 nt) will be generated and incorporated within all overlapping transcripts. As we cannot discern frame-status for cryptic exons, a three or six-frame translation (if splice-junction has no strand information) will be performed.

The generated splice-junction-derived transcript can also span two distinct genes; e.g., if one junction is most adjacent to gene X and the second junction is most adjacent to gene Y. These ‘fusion’-genes are then generated by stitching (taking the strand into account) of all upstream CDS of gene X (ending at the assigned 5′ CDS) with all downstream CDS of gene Y (starting at the assigned 3′ CDS); extensions and/or shortenings are also incorporated in these situations.

Post-incorporation, all generated transcripts (genomic variants, splice-junctions and/or manual sequences) can be curated and altered using the setMutantTranscripts function. After the generation of transcript variants (or manual alteration thereof), all mutant transcript sequences are translated into their respective protein sequences and cleaved at the earliest stop-codon. If the canonical stop-codon is lost, it will continue translating into the 3′ UTR until the next-earliest stop-codon or stop at the end of the 3′ UTR. Generated splice-junctions transcripts without known translation frame(s) will generate a three-frame (if orientations are known and concordant) or a six-frame (if strand orientations is unknown or disconcordant) translation of the transcript sequence(s).

4. Filtering for protein variants based on proteotypic peptides.

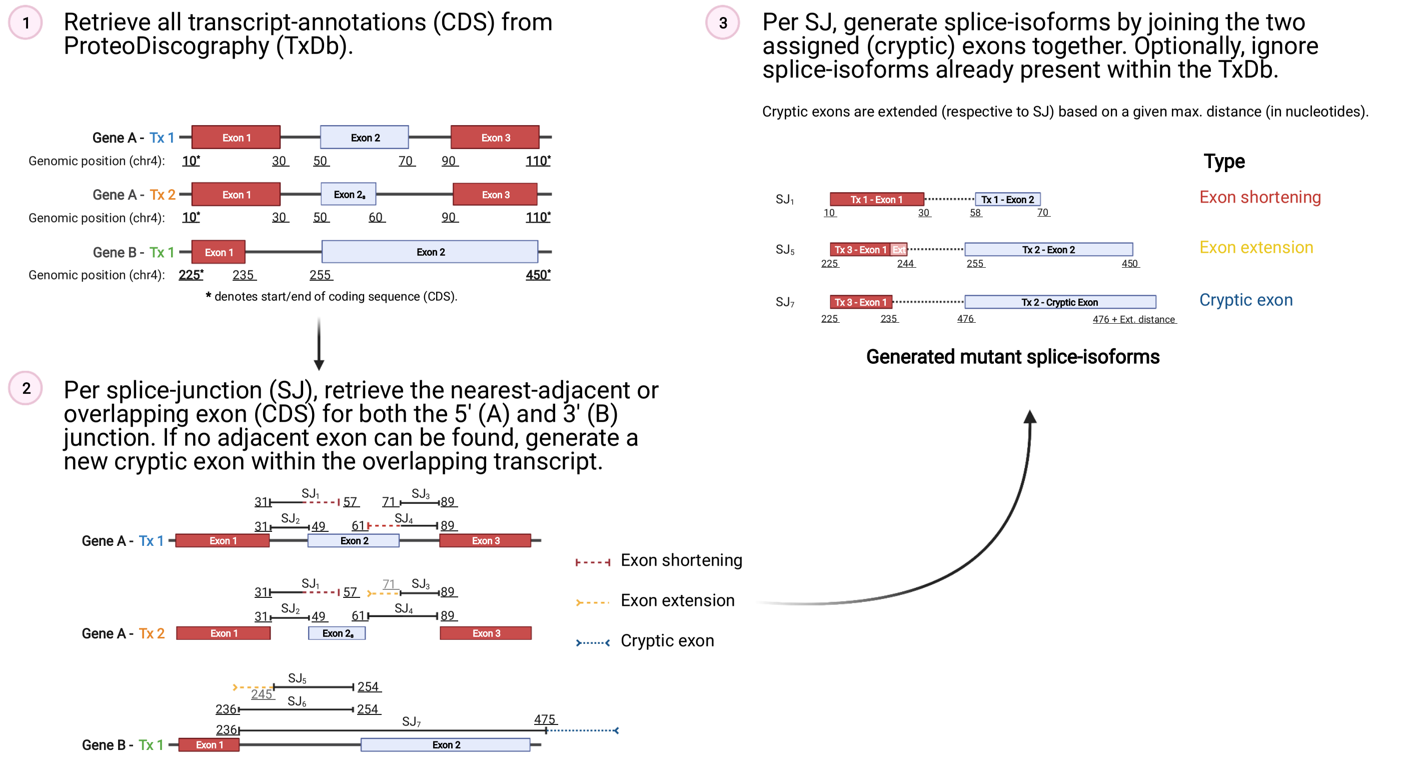
To reduce number of potential protein variants, ProteoDisco provides an optional filtering procedure to retain protein variants containing a min. number of proteotypic peptides not seen in the supplied TxDb (and additional) protein databases and thereby identifiable in subsequent MS/MS analysis.

Conceptually, we cleave the protein variants with the same protease as would be used in the respective MS/MS experiment (e.g., Trypsin) and compare the resulted cleaved peptides against the input TxDb (and additional databases) cleaved in the same manner (allowing user-set missed cleavages) and, subsequently, determine the number of distinct cleaved fragments not detected in the reference protein database(s). In addition, it can also be toggled to check for uniqueness against all other generated protein variants within the ProteoDiscography. This extends the ProteoDiscography with the number of proteotypic peptides per protein variants which can be used to filter protein variants prior to exporting the protein sequences to a FASTA file (step 5).

5. Export protein variants into a customized protein database (FASTA).

Generated protein variants can be exported into an external (FASTA) database. As mentioned, users can subset exported proteins based on the minimum number of proteotypic peptides. This optional filtering step removes identical peptide variants of homologous proteins and sequences that are indistinguishable due to mutations. Users can output all generated protein variants into the same aggregated file or generate distinct files containing sample-specific protein variants.

The FASTA headers for each protein sequence contain identifiers and information on the incorporated variant(s) or splice-junctions which can be easily related back to the ProteoDiscography.



1 Retrieve all transcript-annotations (CDS) from ProteoDiscography (TxDb).

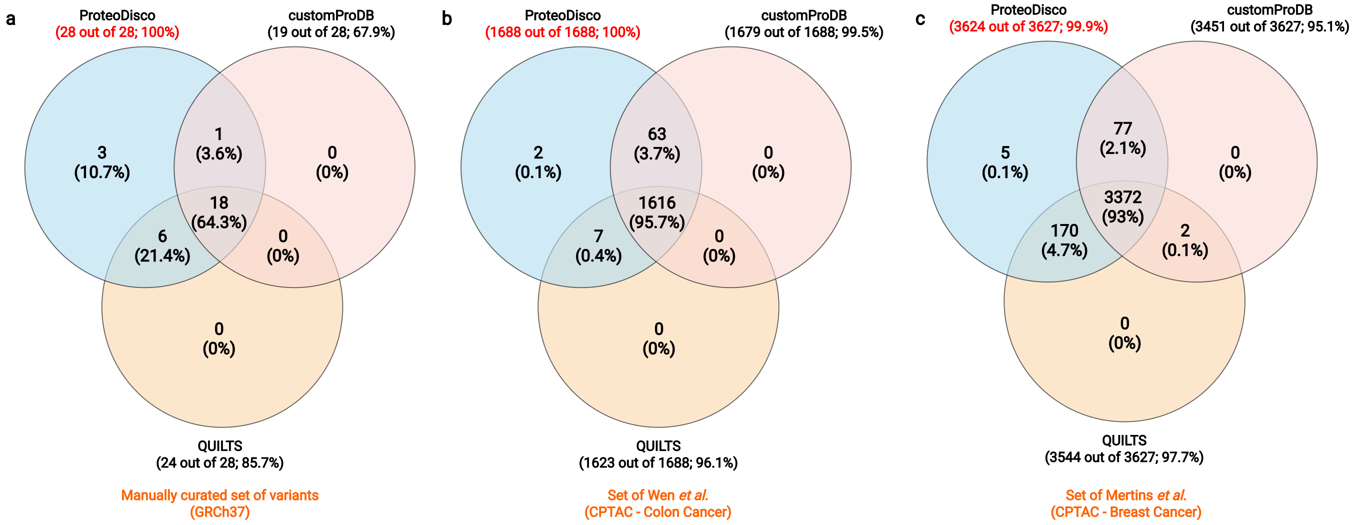
2 Per splice-junction (SJ), retrieve the nearest-adjacent or overlapping exon (CDS) for both the 5' (A) and 3' (B) junction. If no adjacent exon can be found, generate a new cryptic exon within the overlapping transcript.

3 Per SJ, generate splice-isoforms by joining the two assigned (cryptic) exons together. Optionally, ignore splice-isoforms already present within the TxDb.

Cryptic exons are extended (respective to SJ) based on a given max. distance (in nucleotides).

Gene A - Tx 1
Exon 1
Exon 2
Exon 3
Genomic position (chr4): 10* 30 50 70 90 110*
Gene A - Tx 2
Exon 1
Exon 2$_a$
Exon 3
Genomic position (chr4): 10* 30 50 60 90 110*
Gene B - Tx 1
Exon 1
Exon 2
Genomic position (chr4): 225* 235 255 450*
* denotes start/end of coding sequence (CDS).
SJ$_1$ 31 57
SJ$_3$ 71 89
SJ$_2$ 31 49
SJ$_4$ 61 89
Gene A - Tx 1
Exon 1
Exon 2
Exon 3
SJ$_1$ 31 57
SJ$_3$ 71 89
SJ$_2$ 31 49
SJ$_4$ 61 89
Gene A - Tx 2
Exon 1
Exon 2$_a$
Exon 3
SJ$_5$ 245 254
SJ$_6$ 236 254
SJ$_7$ 236 475
Gene B - Tx 1
Exon 1
Exon 2
Exon shortening
Exon extension
Cryptic exon
Type
SJ$_1$ Tx 1 - Exon 1 Tx 1 - Exon 2 10 30 58 70 Exon shortening
SJ$_5$ Tx 3 - Exon 1 Ext Tx 2 - Exon 2 225 244 255 450 Exon extension
SJ$_7$ Tx 3 - Exon 1 Tx 2 - Cryptic Exon 225 235 476 476 + Ext. distance Cryptic exon
Generated mutant splice-isoforms

Supplementary figure S3.1: **Overview of the procedure of generation mutant splice-isoforms based on inter- and intrachromosomal splice-junctions.**
Schematic overview on the handling of splice-junctions (SJ) to generate splice-isoforms. Optionally, users can opt to only generate non-canonical splice-isoforms, thereby ignoring canonical forms already present within the ProteoDiscography TxDb.



Supplementary figure S3.2: **The number of concordant proteotypical fragments for ProteoDisco, customProDb and QUILTS for our manually-curated test-set and two CPTAC-datasets (colon and breast cancer).** Venn-diagrams displaying the number and relative percentage of correctly-incorporated (and total number of) genomic variants and accompanying proteotypical fragments per dataset and tool. We tested ProteoDisco (v1.0), customProDb (v1.30.1) and QUILTS (v3.0) using uniform annotations and settings (as much as possible).
a) Overlap of concordant results based on our validation dataset (COSMIC; GRCh37). **b)** Overlap of concordant results based on the CPTAC colon cancer dataset (Wen *et al.*). **c)** Overlap of concordant results based on the CPTAC breast cancer dataset (Mertins *et al.*).

Chapter 4

The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact

L.F. van Dessel[a*], **J. van Riet**[a-c*], M. Smits[d], Y. Zhu[e-f], P. Hamberg[g], M.S. van der Heijden[h-j], A.M. Bergman[e,j], I.M. van Oort[k], R. de Wit[a], E.E. Voest[h,j], N. Steeghs[h,j], T.N. Yamaguchi[l], J. Livingstone[l], P.C. Boutros[l-q], J.W.M. Martens[a,h], S. Sleijfer[a,h], E.P.J.G. Cuppen[r,s], W. Zwart[e,f,t], H.J.G. van de Werken[b,c], N. Mehra[d¶] & M.P.J.K. Lolkema[a,h¶]

a Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands.
b Cancer Computational Biology Center, Erasmus MC, University Medical Center, Rotterdam, Netherlands.
c Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands.
d Department of Medical Oncology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands.
e Division on Oncogenomics, The Netherlands Cancer Institute, Amsterdam, The Netherlands.
f Oncode Institute, Utrecht, The Netherlands.
g Department of Internal Medicine, Franciscus Gasthuis & Vlietland, Rotterdam, The Netherlands.
h Center for Personalized Cancer Treatment, Rotterdam, The Netherlands.
i Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands.
j Department of Medical Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands.
k Department of Urology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands.
l Informatics and Biocomputing Program, Ontario Institute for Cancer Research, Toronto, Canada.
m Department of Medical Biophysics, University of Toronto, Toronto, Canada.
n Department of Pharmacology and Toxicology, University of Toronto, Toronto, Canada.
o Department of Human Genetics, University of California Los Angeles, Los Angeles, USA.
p Department of Urology, University of California Los Angeles, Los Angeles, USA.
q Jonsson Comprehensive Cancer Centre, University of California Los Angeles, Los Angeles, USA.
r Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Utrecht, The Netherlands.
s Hartwig Medical Foundation, Amsterdam, The Netherlands.
t Laboratory of Chemical Biology and Institute for Complex Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands.

* These authors contributed equally.
¶ These authors jointly supervised this work.

This chapter has been published in **Nature Communications** Volume 10, November 2019, Article number: #5251

Abstract

Metastatic castration-resistant prostate cancer (mCRPC) has a highly complex genomic landscape. With the recent development of novel treatments, accurate stratification strategies are needed. Here we present the whole-genome sequencing (WGS) analysis of fresh-frozen metastatic biopsies from 197 mCRPC patients. Using unsupervised clustering based on genomic features, we define eight distinct genomic clusters. We observe potentially clinically relevant genotypes, including microsatellite instability (MSI), homologous recombination deficiency (HRD) enriched with genomic deletions and *BRCA2* aberrations, a tandem duplication genotype associated with *CDK12*$^{-/-}$ and a chromothripsis-enriched subgroup. Our data suggests that stratification on WGS characteristics may improve identification of MSI, *CDK12*$^{-/-}$ and HRD patients. From WGS and chromatin immunoprecipitation sequencing (ChIP-seq) data, we show the potential relevance of recurrent alterations in non-coding regions identified with WGS and highlight the central role of AR signaling in tumor progression. These data underline the potential value of using WGS to accurately stratify mCRPC patients into clinically actionable subgroups.

Introduction

Prostate cancer is known to be a notoriously heterogeneous disease and the genetic basis for this interpatient heterogeneity is poorly understood.[1,2] The ongoing development of new therapies for metastatic prostate cancer that target molecularly defined subgroups further increases the need for accurate patient classification and stratification.[3–5] Analysis of whole-exome sequencing data of metastatic prostate cancer tumors revealed that 65% of patients had actionable targets in non-androgen receptor related pathways, including PI3K, Wnt, and deoxyribonucleic acid (DNA) repair.[6] Several targeted agents involved in these pathways, including mTOR/AKT pathway inhibitors[7] and PARP inhibitors[8], are currently in various phases of development and the first clinical trials show promising results. Therefore, patients with metastatic prostate cancer could benefit from better stratification to select the most appropriate therapeutic option. More extensive analysis using WGS-based classification of tumors may be useful to improve selection of patients for different targeted therapies. The comprehensive nature of WGS has many advantages, including the detection of mutational patterns, as proven by the successful treatment of patients with high-tumor mutational burden with immune check-point blockade therapy.[9–12] Moreover, WGS unlike exome sequencing, can detect structural variants and aberrations in non-coding regions, both important features of prostate cancer.

The stratification of prostate cancer patients, based on differences in the mutational landscape of their tumors, has mainly focused on mutually exclusive mutations, copy-number alterations, or distinct patterns in RNA-sequencing caused by the abundant *TMPRSS2-ERG* fusion, which is recurrent in 50% of primary prostate tumors[6,13–18]. More recently, WGS of metastatic prostate cancer tumors demonstrated that structural variants arise from specific alterations such as *CDK12*$^{-/-}$ and *BRCA2*$^{-/-}$ genotypes, and are strongly associated with genome-wide events such as large tandem duplications or small genomic deletions, respectively.[19–23] Advances in WGS analysis and interpretation have revealed rearrangement signatures in breast cancer relating to disease stage, HRD, and *BRCA1*/*BRCA2* defects based on size and type of structural variant[22,24]. Thus, WGS enables the identification of patterns of DNA aberrations (i.e., genomic scars) that may profoundly improve classification of tumors that share a common etiology, if performed in a sufficiently powered dataset.

In this study, we analyzed the WGS data obtained from 197 mCRPC patients. We describe the complete genomic landscape of mCRPC, including tumor specific single- and multi-nucleotide variants (single-nucleotide variation(s) (SNV)s and multi-nucleotide variation(s) (MNV)s), small insertions and deletions (InDels), copy-number alteration(s) (CNA), mutational signatures, kataegis, chromothripsis, and structural variant(s) (SV). Next, we compared the mutational frequency of the detected driver genes and genomic subgroups with an unmatched WGS cohort of primary prostate cancer ($n = 210$), consisting of exclusively of Gleason score 6–7 tumors.[15,25] We investigated the presence of possible driver genes by analyzing genes with enriched (non-synonymous) mutational burdens and recurrent or high-level copy-number alterations.[26,27] By utilizing various basic genomic features reflecting genomic instability and employing unsupervised clustering, we were able to define eight distinct genomic subgroups of mCRPC patients. We combined our genomic findings with AR, FOXA1, and H3K27me ChIP-seq data, and confirmed that important regulators of AR-mediated signaling are located in non-coding regions with open chromatin and highlight the central role of AR signaling in tumor progression.

Results

Characteristics of the mCRPC cohort and sequencing approach

We analyzed fresh-frozen metastatic tumor samples and matched blood samples from 197 castration-resistant prostate cancer patients using WGS generating to date the largest WGS dataset for mCRPC (Figure 4.1a). Clinical details on biopsy site, age, and previous treatments of the included patients are described in Figure 4.1b, 4.1c and Supplementary Data 1 (available online). WGS data was sequenced to a mean coverage of 104X in tumor tissues and 38X in peripheral blood (Supplementary Figure S4.1a). The median estimated tumor cell purity using *in silico* analysis of our WGS data was 62% (range: 16–96%; Supplementary Figure S4.1b). Tumor cell purity correlated weakly with the frequency of called SNVs (Spearman correlation; rho = 0.2; $p = 0.005$), InDels (Spearman correlation; rho = 0.35; $p < 0.001$), MNVs (Spearman correlation; rho = 0.25; $p < 0.001$) and structural variants (Spearman correlation; rho = 0.22; $p = 0.002$; Supplementary Figure S4.1c).

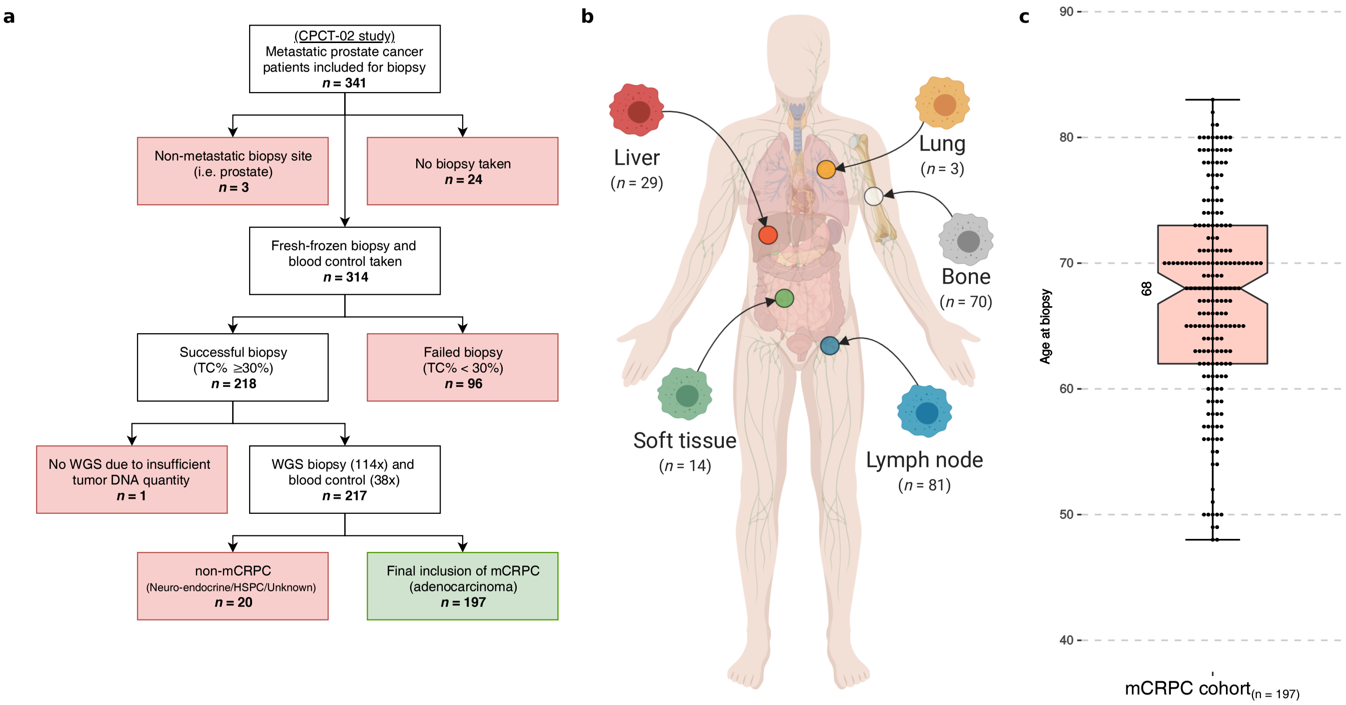

a
(CPCT-02 study)
Metastatic prostate cancer patients included for biopsy
n = 341
Non-metastatic biopsy site (i.e. prostate)
n = 3
No biopsy taken
n = 24
Fresh-frozen biopsy and blood control taken
n = 314
Successful biopsy (TC% ≥30%)
n = 218
Failed biopsy (TC% < 30%)
n = 96
No WGS due to insufficient tumor DNA quantity
n = 1
WGS biopsy (114x) and blood control (38x)
n = 217
non-mCRPC (Neuro-endocrine/HSPC/Unknown)
n = 20
Final inclusion of mCRPC (adenocarcinoma)
n = 197
b
Liver
(n = 29)
Lung
(n = 3)
Bone
(n = 70)
Soft tissue
(n = 14)
Lymph node
(n = 81)
c
Age at biopsy
90
80
70
60
50
40
68
mCRPC cohort(n = 197)

Figure 4.1: **Overview of study design and patient cohort ($n = 197$).**
(a): Flowchart of patient inclusion. From the CPCT-02 cohort, patients with metastatic prostate cancer were selected. Patients were excluded if data from metastatic samples were not available and if clinical data indicated that patients had hormone-sensitive or neuro-endocrine prostate cancer or unknown disease status at the time of analysis. **(b):** Overview of the biopsy sites. Number of biopsies per metastatic site analyzed with WGS. **(c):** Age of patients at biopsy. Bee-swarm boxplot with notch of the patient age distribution. Boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the interquartile range (IQR). Data points outside the IQR are shown.

Landscape of mutational and structural variants in mCRPC

The median tumor mutational burden (TMB) at the genomic level (SNVs and InDels per mega-base pair (Mbp)) was 2.7 in our mCRPC cohort, including 14 patients with high TMB (>10). We found a median of 6621 SNVs (IQR: 5048–9109), 1008 small InDels (IQR: 739–1364), 55 MNVs (IQR: 34–86) and 224 SVs (IQR: 149–370) per patient (Supplementary Figure S4.2a–c). We observed a highly complex genomic landscape consisting of multiple driver mutations and structural variants in our cohort.

We confirmed that known driver genes of prostate cancer were enriched for non-synonymous mutations (Figure 4.2 and Supplementary Figure S4.2e)[13,15,28]. In total, we detected 11 genes enriched with non-synonymous mutations: *TP53*, *AR*, *FOXA1*, *SPOP*, *PTEN*, *ZMYM3*, *CDK12*, *ZFP36L2*, *PIK3CA*, and *APC*. *ATM* was mutated in 11 samples, but after multiple-testing correction appeared not to be enriched.

Our copy-number analysis revealed distinct amplified genomic regions, including 8q and Xq and deleted regions including 8p, 10q, 13q, and 17p (Supplementary Figure S4.2d). Well-known prostate cancer driver genes[8,16], such as *AR*, *PTEN*, *TP53*, and *RB1*, are located in these regions. In addition to large-scale chromosomal copy-number alterations, we could identify narrow genomic regions with recurrent copy-number alterations across samples, which could reveal important prostate cancer driver genes (Supplementary Data 1 (available online)).

TMPRSS2-ERG gene fusions were the most common fusions in our cohort ($n = 84$ out of 197; 42.6%) and were the majority of ETS family fusions ($n = 84$ out of 95; 88.4%; Figure 4.2 and Supplementary Figure S4.3). This is comparable to primary prostate cancer, where ETS fusions are found in approximately 50% of tumors.[13,15] The predominant break point was located upstream of the second exon of *ERG*, which preserves its ETS-domain in the resulting fusion gene.

In 42 patients (21.3%), we observed regional hypermutation (kataegis; Figure 4.2 and Supplementary Figure S4.4). In addition, we did not observe novel mutational signatures specific for metastatic disease or possible pre-treatment histories (Supplementary Figure S4.5).[29]

To further investigate whether our description of the genome-wide mutational

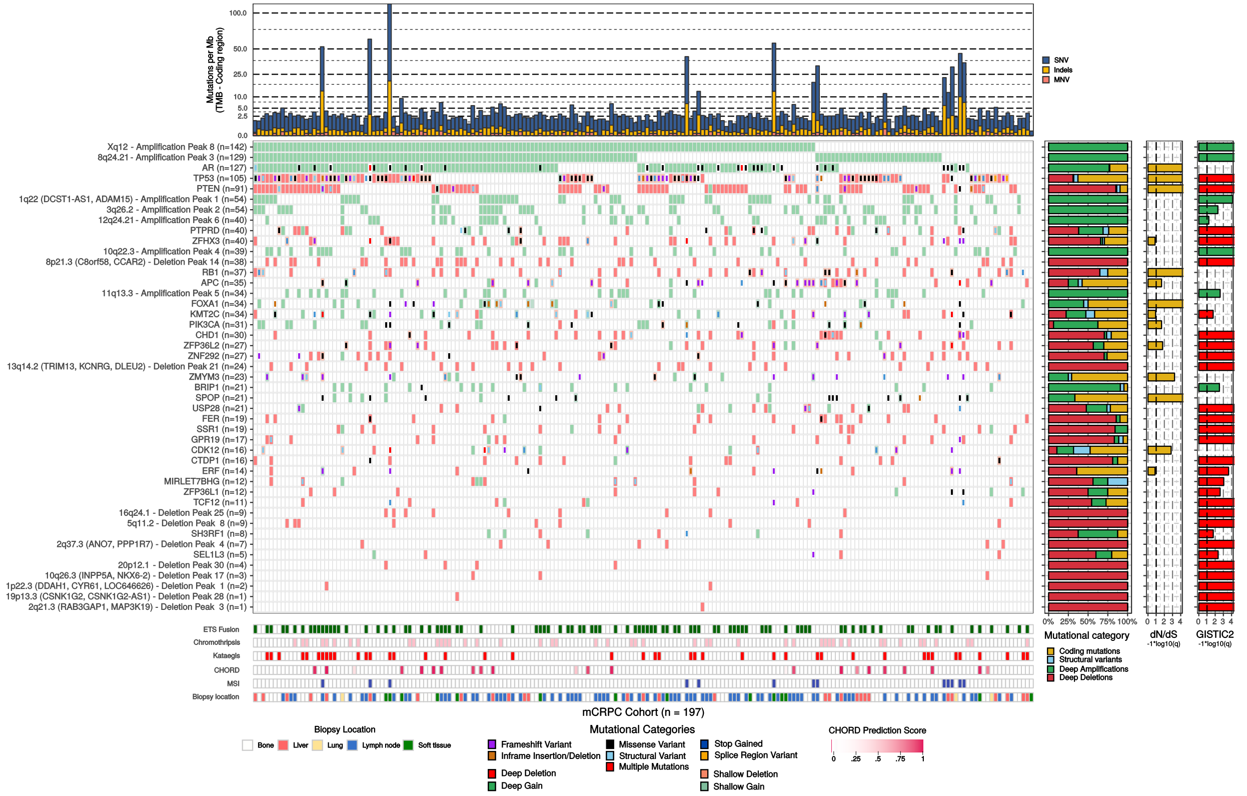


Figure 4.2: **mCRPC shows multiple recurrent somatic alterations affecting several oncogenic pathways.** Based on dN/dS ($q \leq 0.1$) and GISTIC2 focal peak ($q \leq 0.1$) criteria, we show the genes and focal genomic foci that are recurrently mutated, amplified, or deleted in our mCRPC cohort of 197 patients. The upper track (top bar plot) displays the number of genomic mutations per Mbp (TMB) per SNV (blue), InDels (yellow), and MNV (orange) category in coding regions (square-root scale). Samples are sorted based on mutual-exclusivity of the depicted genes and foci. The heatmap displays the type of mutation(s) per sample; (light-)green or (light-)red backgrounds depict copy-number aberrations while the inner square depicts the type of (coding) mutation(s). Relative proportions of mutational categories (coding mutations [SNV, InDels, and MNV] (yellow), SV (blue), deep amplifications [high-level amplifications resulting in many additional copies] (green), and deep deletions [high-level losses resulting in (near) homozygous losses] (red)) per gene and foci are shown in the bar plot next to the heatmap. Narrow GISTIC2 peaks covering ≤ 3 genes were reduced to gene-level rows if one of these genes is present in the dN/dS ($q \leq 0.1$) analysis or is a known oncogene or tumor-suppressor. For GISTIC2 peaks covering multiple genes, only deep amplifications and deep deletions are shown. Recurrent aberrant focal genomic foci in gene deserts are annotated with their nearest gene. Significance scores ($-1*\log_{10}(q)$) of the dN/dS and GISTIC2 analysis are shown on the outer-right bar plots; bars in the GISTIC2 significance plot are colored red if these foci were detected as a recurrent focal deletion and green if detected as a recurrent focal gain. Per sample, the presence of (predicted) E26 transformation-specific (ETS) fusions (green), chromothripsis (light pink), kataegis (red), CHORD prediction score (HR-deficiency) (pink gradient), MSI status (dark blue), and biopsy location are shown as bottom tracks.

burden and observed alterations in drivers and/or subtype-specific genes in mCRPC were metastatic specific, we compared our data against an unmatched WGS cohort of primary prostate cancer ($n = 210$)[15,25], consisting of Gleason score 6–7 disease. Comparison of the median genome-wide TMB (SNVs and InDels per Mbp) revealed that the TMB was roughly 3.8 times higher in mCRPC (Figure 4.3a) and the frequency of structural variants was also higher between disease stages (Figure 4.3b), increasing as disease progresses. Analysis on selected driver and subtype-specific genes showed that the mutational frequency of several genes (*AR*, *TP53*, *MYC*, *ZMYM3*, *PTEN*, *PTPRD*, *ZFP36L2*, *ADAM15*, *MARCOD2*, *BRIP1*, *APC*, *KMT2C*, *CCAR2*, *NKX3-1*, *C8orf58*, and *RYBP*) was significantly altered ($q \leq 0.05$) between the primary and metastatic cohorts (Figure 4.3c–e). All genes for which we observed significant differences in mutational frequency, based on coding mutations, were enriched in mCRPC (Figure 4.3d). We did not identify genomic features that were specific for the metastatic setting, beyond androgen deprivation therapy-specific aberrations revolving *AR* (no aberrations in hormone-sensitive setting versus 137 aberrations in castration-resistant setting). We cannot exclude from these data that matched sample analysis or larger scale analysis could reveal such aberrations.

We next determined whether previous treatments affect the mutational landscape. Using treatment history information, we grouped prior secondary anti-hormonal therapy, taxane-based chemotherapy and systemic radionucleotide therapy into different groups (Supplementary Figure S4.6). This analysis did not reveal systematic biases due to pre-treatment in aberrations, such as TMB, kataegis, chromothripsis, ETS fusions, or somatically altered genes (Supplementary Data 1 (available online)).

The role of the AR-pathway in mCRPC

Focusing on the AR-pathway revealed that aberrant AR signaling occurred in 80% of our patients. In 57.3% of patients both *AR* and the *AR*-enhancer (˜66.13 Mb on chromosome X; located about 629 kbp upstream of the AR gene[20]) were affected (Figure 4.4a). In an additional 6.6% and 14.7% of tumors only *AR* gene alterations or *AR*-enhancer amplification occurred, respectively. The percentage of mCRPC patients with the exclusive *AR*-enhancer amplification (29 out of 197; 14.7%) versus exclusively *AR*-locus amplification (13 out of 197; 6.6%) is similar to previous observations, which showed 21 out of 94 castration-resistant prostate cancer (CRPC) patients (10.3%) with exclusively *AR*-enhancer amplification versus

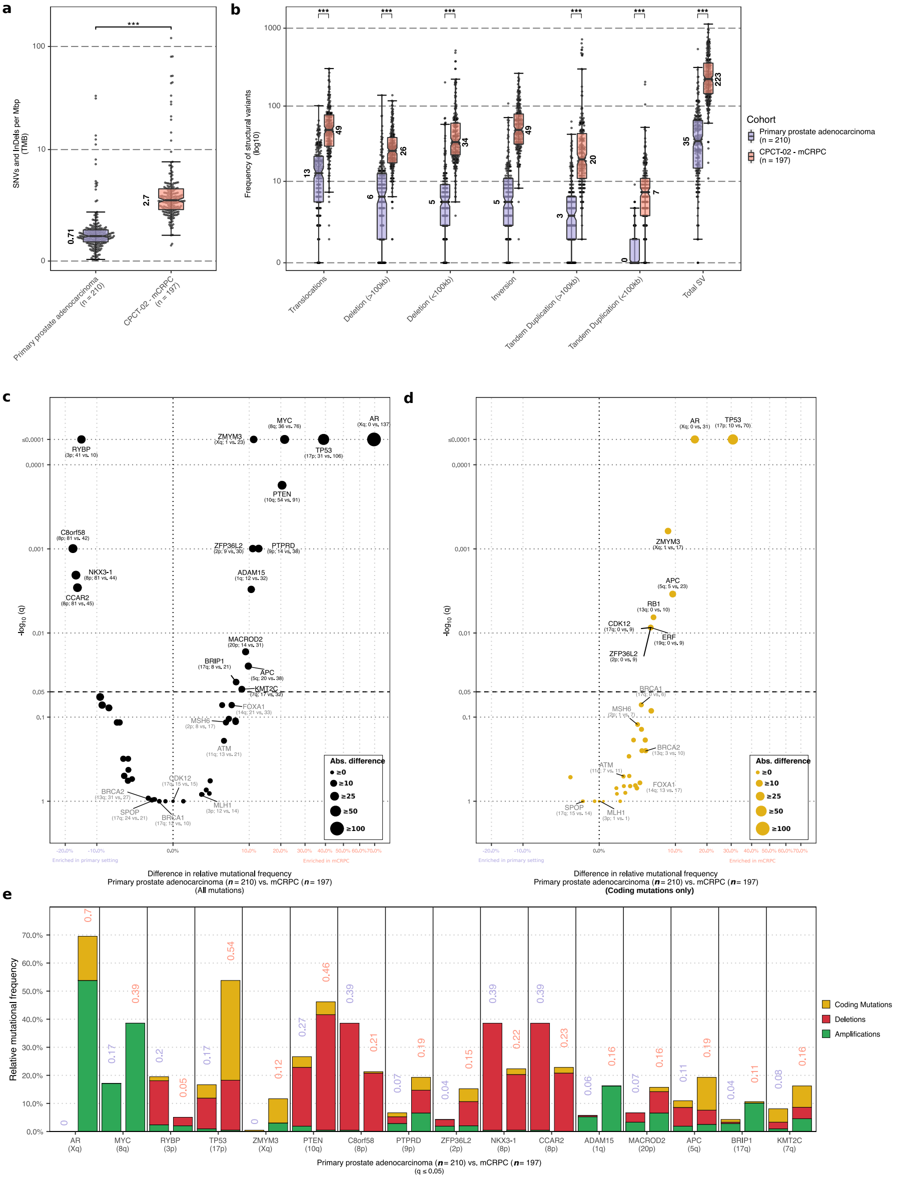


Figure 4.3: **Comparison of the mutational landscape between primary prostate cancer and mCRPC.**
(a): Tumor mutational burden (SNVs and InDels per Mbp) from a primary prostate cancer ($n = 210$) and the CPTC-02 mCRPC cohort ($n = 197$). Bee-swarm boxplot with notch of the tumor mutational burden. Boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. Statistical significance was tested with Wilcoxon rank-sum test and $p \leq 0.001$ is indicated as ***. **(b)**: Frequency of structural variant events from an unmatched cohort of primary prostate cancer ($n = 210$) and the CPTC-02 mCRPC cohort ($n = 197$). **(c)**: Comparison of the mutational frequencies for driver genes detected by dN/dS and/or GISTIC2, or subtype-specific genes, enriched in mCRPC relative to primary prostate cancer or vice-versa. The difference in relative mutational frequency is shown on the x-axis and the adjusted p-value (two-sided Fisher's Exact Test with Benjamini-Hochberg (BH) correction) is shown on the y-axis. **(d)**: Same as in **c** but using only coding mutations. **(e)**: Overview of the mutational categories of the driver genes detected by dN/dS and/or GISTIC2, or subtype-specific genes, enriched in mCRPC relative to primary prostate cancer ($q \leq 0.05$). For each gene the frequency in primary prostate cancer is displayed followed by the frequency in mCRPC.

4 out of 94 CRPC patients (4.3%) with exclusively *AR*-locus amplification.[20] Concurrent amplification of the *AR* gene and the *AR*-enhancer was not necessarily of equal magnitude, which resulted in differences in copy number enrichment of these loci (Figure 4.4b).

To date, no AR ChIP-seq data has been reported in human mCRPC samples and evidence of increased functional activity of the amplified enhancer thus far is based on cell line models.[30] To resolve this, we performed AR ChIP-seq on two selected mCRPC patient samples with *AR*-enhancer amplification based on WGS data. As controls we used two prostate cancer cell-lines (LNCaP and VCaP) and three independent primary prostate cancer samples that did not harbor copy-number alterations at this locus (Supplementary Figure S4.7).[31] We observed active enhancer regions (H3K27ac) in the castration-resistant setting, co-occupied by AR and FOXA1, at the amplified *AR*-enhancer. This is substantially stronger when compared to the hormone-sensitive primary prostate cancer samples without somatic amplifications (Figure 4.4c and Supplementary Figure S4.7). Furthermore, a recurrent focal amplification in a non-coding region was observed at 8q24.21 near *PCAT1*. This locus bears similar epigenetic characteristics to the *AR*-enhancer with regard to H3K27ac and, to a lesser extent, binding of AR and/or FOXA1 in the mCRPC setting (Figure 4.4d and Supplementary Figure S4.7).

WGS-based stratification defines genomic subgroups in mCRPC

Our comprehensive WGS data and large sample size enabled us to perform unsupervised clustering on several WGS characteristics to identify genomic scars that can define subgroups of mCRPC patients. We clustered our genomic data using the total number of SVs, relative frequency of SV category (translocations, inversions, insertions, tandem duplications, and deletions), genome-wide TMB encompassing SNV, InDels and MNV, and tumor ploidy. Prior to clustering, we subdivided tandem duplications and deletions into two major categories based on the respective genomic size of the aberration (smaller and larger than 100 kilo-base pair (kbp)) since previous studies revealed distinctions based on similar thresholds for these structural variants in relation to specific-mutated genes.[19–21,32] Similarly, we observed a difference in genomic size and number in our subgroups of mCRPC patients (Supplementary Figure S4.8).

This analysis defined eight distinct subgroups (Figures 4.5, 4.6 and Supplemen-

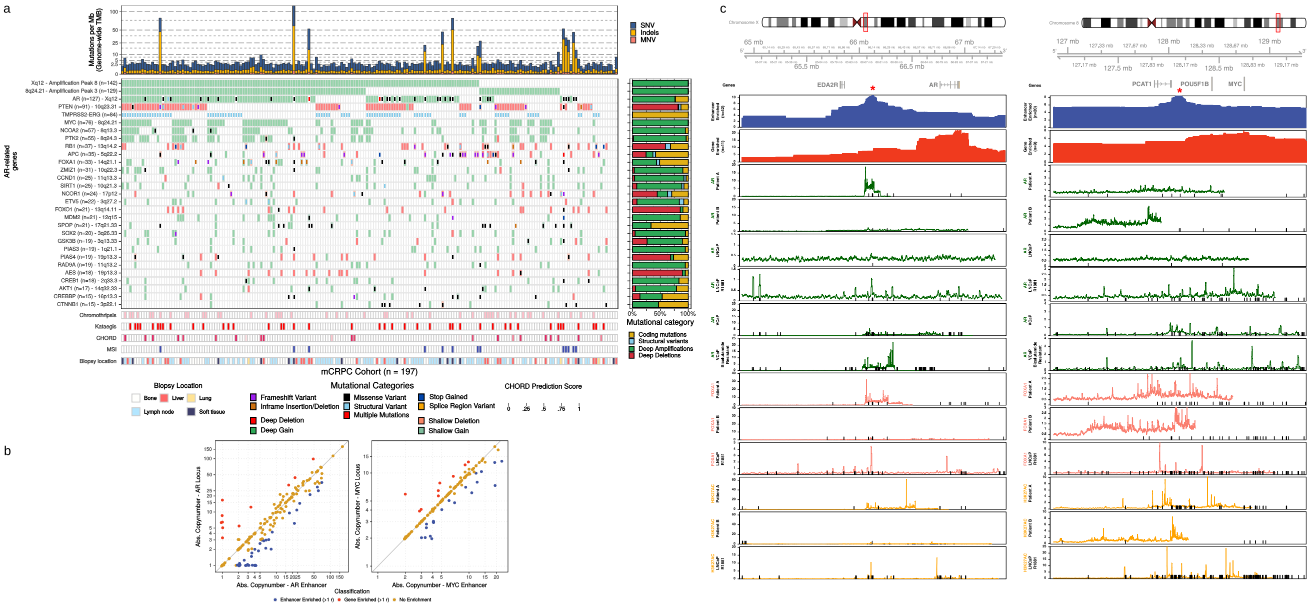


Figure 4.4: **WGS reveals novel insight into the various (non-coding) aberrations affecting *AR* regulation.**
(a): Mutational overview of top recurrently mutated genes affecting *AR* regulation and their putative enhancer foci (as detected by GISTIC2). The first track represents the TMB per SNV (blue), InDels (yellow), and MNV (orange) category genome-wide (square-root scale). Samples are sorted based on mutual-exclusivity of the depicted genes and foci. The heatmap displays the type of mutation(s) per sample. Relative proportions of mutational categories per gene and foci are shown in the bar plot next to the heatmap. The presence of chromothripsis (light pink), kataegis (red), CHORD prediction score (HR-deficiency) (pink gradient), MSI status (dark blue), and biopsy location are shown as bottom tracks. **(b)**: Overview of the copy-number deviations between putative enhancer and gene regions for *AR* and *MYC*. Samples were categorized as enhancer- (blue) or gene- (red) enriched if enhancer-to-gene ratio deviated >1 studentized residual (residual in standard deviation units) from a 1:1 ratio. **(c)**: Copy number and ChIP-seq profiles surrounding the *AR* and *PCAT1*/*MYC* gene loci (with 1.25 additional Mbp up-/downstream). The upper panel displays the selected genomic window and the overlapping genes. The 1st and 2nd track display the aggregated mean copy number (per 0.1Mbp window) of the enhancer- and gene-enriched samples, respectively. These profiles identify distinct amplified regions (indicated by red *) in proximity to the respective gene bodies. The 3th-8th tracks represent AR ChIP-seq profiles (median read-coverage per 0.1Mbp windows) in two mCRPC patients (#3 and 4), LNCaP (#5) and LNCaP with R1881 treatment (#6), VCaP (#7) and bicalutamide-resistant VCaP (#8). The 9th-11th tracks represent FOXA1 ChIP-seq profiles (median read-coverage per 0.1Mbp windows) in two mCRPC patients (#9 and 10) and LNCaP with R1881 treatment (#11). The 12th-14th tracks represent H3K27ac ChIP-seq profiles (median read-coverage per 0.1Mbp windows) in two mCRPC patients (#12 and 13) and LNCaP with R1881 treatment (#14) reflecting active enhancer regions. ChIP-seq peaks (MACS/MACS2; $q < 0.01$) are shown as black lines per respective sample.

tary Figures S4.8, S4.9, S4.10 and S4.11): (A) MSI signature with high TMB and association with mismatch repair deficiency; (B) tandem duplication (>100 kbp) phenotype associated with biallelic *CDK12* inactivation; (D) HRD features with many deletions (>100 kbp) and association with (somatic) mutations in BRCAness-associated genes; this was supported by high HR-deficiency scores (CHORD; Supplementary Figures S4.8 and S4.9); (F) chromothripsis; C, E, G, H); non-significant genomic signature without any currently known biological association. 4.1 summarizes the key features of each subgroup.

Clusters A and B represent previously identified genomic subgroups (MSI and *CDK12*$^{-/-}$).[6,19,21,34] In cluster B, only two patients were allocated to this subgroup without a specific somatic mutation in the identifying gene. The well-known mismatch repair genes: *MLH1*, *MSH2*, and *MSH6* are among the cluster-specific-mutated genes in cluster A (Figure 4.6a). Twelve out of these thirteen patients had at least one inactivating alteration in one of these genes (Figure 4.6b). Interestingly, cluster B (*CDK12*$^{-/-}$) harbors two patients without non-synonymous *CDK12* mutation or copy-number alteration; the cause of their tandem duplication phenotype is currently unknown (Figure 4.6b). Cluster D shows significant features of HRD, specifically biallelic *BRCA2* inactivation (Supplementary Figure /S4.12), mainly mutational signature 3, enrichment of deletions (<100 kbp) and is supported by high HR-deficiency scores (CHORD) (Supplementary Figures S4.8 and S4.9).[22,35] Remarkably, seven out of twenty-two patients did not have a biallelic *BRCA2* inactivation. However, four of these patients showed at least one (deleterious) aberration in other BRCAness-related genes (Figure 4.6b).[36] Cluster F was enriched for chromothripsis events, however we could not reproduce a previous finding, suggesting chromothripsis was associated with inversions and p53 inactivation in prostate cancer.[21] Apart from the chromothripsis events, no clear gene aberration was associated with this cluster (Figure 4.6b). In the remaining patients, there were no distinct genomic signatures or biologic rationale for patient clustering (cluster C, E, G, H). In cluster C, conjoint aberrations of *BRCA1* and *TP53* were observed in one patient with a high HR-deficiency prediction score (CHORD), which is known to lead to a small tandem duplication phenotype (<100 kbp).[32] Two other patients within cluster C displayed a weak CHORD scoring associated with HR-deficiency, however no additional definitive evidence was found for a *BRCA1* loss-of-function mutation within these patients.

In addition to our unsupervised clustering approach, we clustered our samples

	Number of patients (*n*; % of cohort)	TMB (CDS)	SNV/InDels ratio	Number of SV	Main SV category or differentiating category	Ploidy status	Main mutational signature	Top 3 cluster-specific aberrations (% of cluster)	ETS-fusions (*n*)	Chromothripsis (*n*)	Kataegis (*n*)
Cluster A	13 (6.6)	36,88	0,99	149	None	1,92	MSI	MSH6 (69.2), JAK1 (69.2), CIC (58.3)	3	1	6
Cluster B	13 (6.6)	2,44	7,07	669	Tandem duplications (>100 kb)	2,39	N/A	CDK12 (84.6), FGF3 (69.2), FGF4 (69.2)	2	0	1
Cluster C	15 (7.6)	3,00	6,73	237	Tandem duplications (<100 kb)	3,19	N/A	None	7	1	2
Cluster D	22 (11.2)	4,39	7,28	323	Deletions (>100 kb)	2,16	BRCA	BRCA2 (68.2)	7	5	5
Cluster E	55 (27.9)	2,12	7,13	178	None	3,24	N/A	None	25	8	13
Cluster F	20 (10.2)	2,51	6,15	400	None	3,35	N/A	Chromothripsis (80,0)	10	16	7
Cluster G	34 (17.3)	2,12	6,13	222	None	2,98	N/A	None	23	8	5
Cluster H	25 (12.7)	2,30	5,81	201	Insertions	1,97	N/A	None	16	7	3

Table 4.1: **Overview of the distinctive characteristics for each cluster (A-H).** All numbers are median of the cluster, unless otherwise indicated.

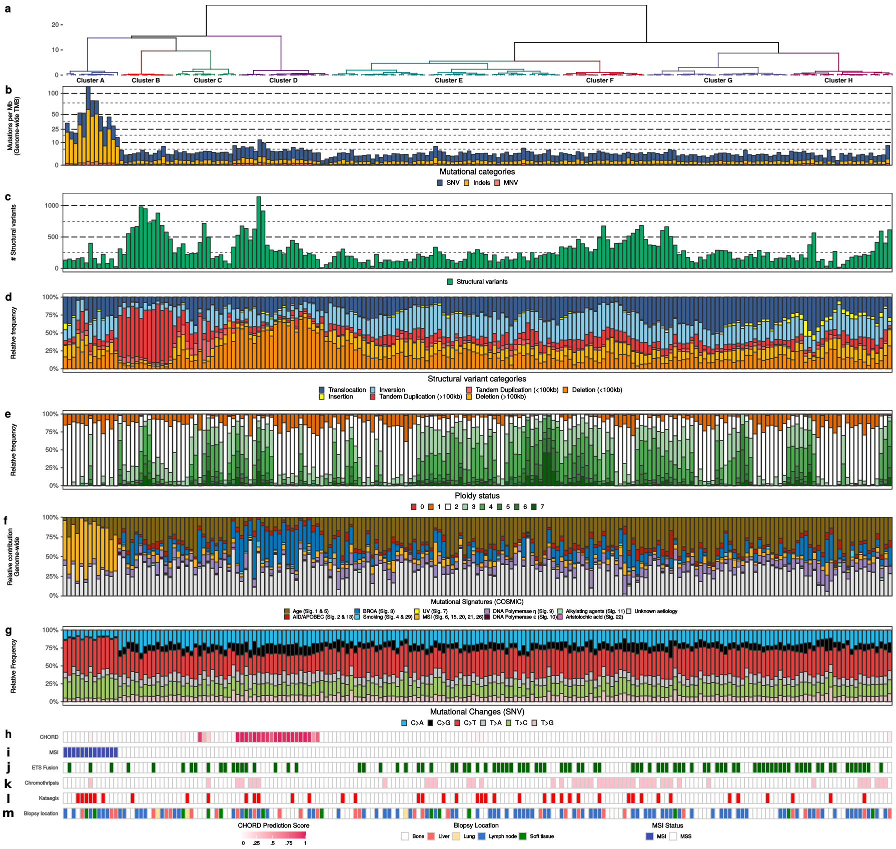


Figure 4.5: **Unsupervised clustering of mCRPC reveals distinct genomic phenotypes.** **(a)**: Dendrogram of unsupervised clustering with optimal leaf ordering (OLO). Top eight clusters are highlighted and denoted based on order of appearance (left to right): A to H. y-axis displays clustering distance (Pearson correlation; ward.D). **(b)**: Number of genomic TMB per SNV (blue), InDels (yellow), and MNV (orange) category. All genome-wide somatic mutations were taken into consideration (square-root scale). **(c)**: Absolute number of unique structural variants per sample. **(d)**: Relative frequency per structural variant category (translocations, inversions, insertions, tandem duplications, and deletions). Tandem Duplications and Deletions are subdivided into >100 kbp and <100 kbp categories. This track shows if an enrichment for particular category of (somatic) structural variant can be detected, which in turn, can be indicative for a specific mutational aberration. **(e)**: Relative genome-wide ploidy status, ranging from 0 to ≥7 copies. This track shows the relative percentage of the entire genome, which is (partially) deleted (ploidy <2 per diploid genome) or amplified (ploidy >2 per diploid genome). **(f)**: Relative contribution to mutational signatures (COSMIC) summarized per proposed etiology. This track displays the proposed etiology of each SNV based on their mutational contexts. **(g)**: Relative frequency of different SNV mutational changes. **(h)**: HR-deficient prediction score as assessed by CHORD. The binary prediction score of CHORD (ranging from 0 to 1) is shown, in which higher scores reflect more evidence for HR-deficiency in a given sample. **(i)**: MSI status as determined using a stringent threshold of MSI characteristics.[33] **(j)**: Presence of a fusion with a member of the ETS family. Green color indicates a possible fusion. **(k)**: Presence of chromothripsis. Pink color indicates presence of chromothripsis as estimated by ShatterSeek. **(l)**: Presence of kataegis. Red color indicates presence of one or more regions showing kataegis. **(m)**: General biopsy location.

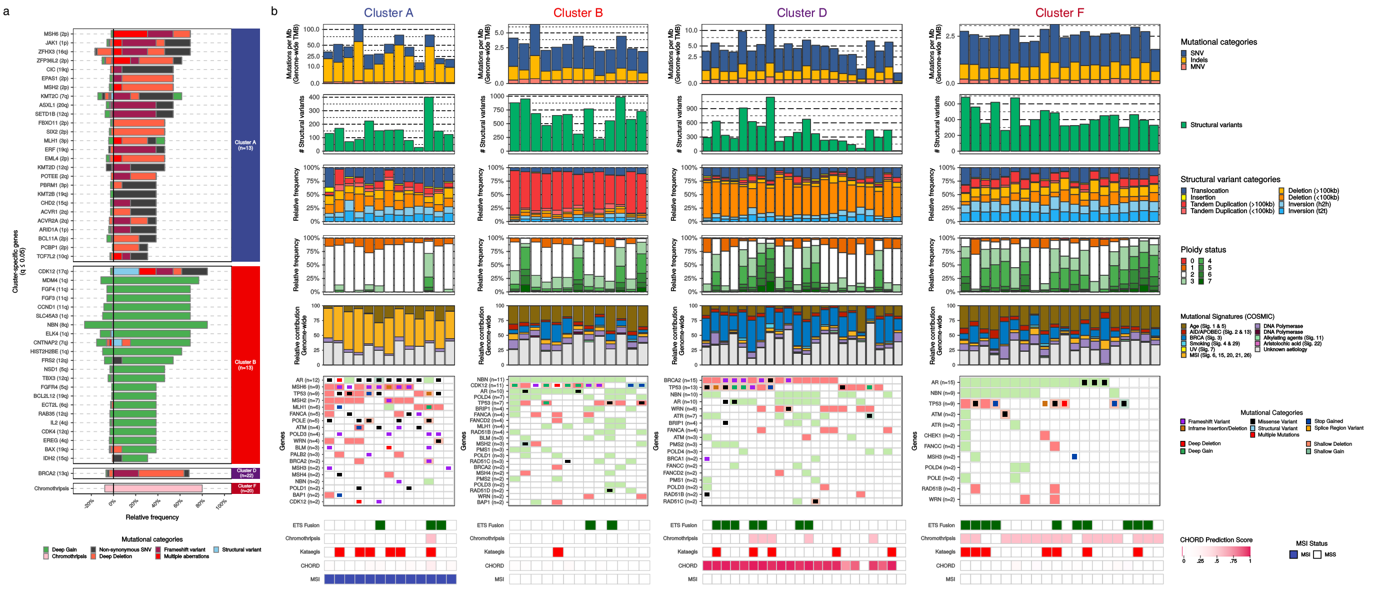


Figure 4.6: **Distinct genomic phenotypes in mCRPC are enriched by mutually exclusive aberrations in key pathways.** **(a)**: Cluster-specific enrichment of mutated genes (multiple colors), chromothripsis (light pink), and structural variants (light blue) (Fisher's Exact Test with BH correction; $q \leq 0.05$). Percentages to the left of the black line represent the relative mutational frequency in mCRPC samples, which are not present in the respective cluster, while the percentages to the right of the black line represent the relative mutational frequency present in the samples from the tested cluster. **(b)**: Genomic overview with biologically relevant genes in the clusters A, B, D, and F with mutational enrichment of genes or large-scale events. The first track represents the number of genomic TMB per SNV (blue), InDels (yellow), and MNV (orange) category genome-wide (square-root scale). The second track represents the absolute number of unique structural variants (green) per sample. The third track represents the relative frequency per structural variant category. Tandem duplications and deletions are subdivided into >100 kbp and <100 kbp categories. The fourth track represents relative genome-wide ploidy status, ranging from 0 to ≥7 copies. The fifth track represents the relative contribution to mutational signatures (Catalogue Of Somatic Mutations In Cancer (COSMIC)) summarized per proposed etiology. The sixth track displays somatic mutations in the relevant genes found in at least one cluster. The lower tracks represent presence of ETS fusions (green), chromothripsis (pink), kataegis (red), CHORD prediction scores (HR-deficiency) (pink gradient), and MSI status (blue) based on a threshold of MSI characteristics.

using the clustering scheme proposed by The Cancer Genome Atlas (TCGA) (Supplementary Figure S4.13a), which defines seven clusters based on coding mutations and copy-number aberrations in *SPOP*, *FOXA1*, *IDH1*, and ETS family gene fusions (and overexpression) per promiscuous partner (*ERG*, *ETV1*, *ETV4*, and *FLI1*).[13] Unfortunately, we currently lack matched messenger RNA (mRNA)-sequencing data in our cohort and therefore cannot observe overexpression of fused ETS family members, which restricted us to only characterize the genomic breaks of these promiscuous partners. Without incorporation of ETS family overexpression, this proposed clustering scheme categorizes 61% of mCRPC into these seven groups versus 68% of the original cohort containing primary prostate cancer described by TCGA (Supplementary Figure S4.13b).[13] There was no significant correlation between the TCGA clustering scheme and our defined genomic subtypes such as MSI, BRCAness or *CDK12*$^{-/-}$. In addition, we did not detect statistical enrichment or depletion ($q \leq 0.05$) between these supervised clusters and additional-mutated genes, kataegis and chromothripsis, only the known enrichment of homozygous *CHD1* deletions in the *SPOP*-cluster was observed.[13]

Performing unsupervised clustering and principal component analysis on the primary prostate cancer and metastatic cohorts revealed no striking primary-only genomic subgroup nor did we detect the presence of the mCRPC-derived genomic subgroups in the primary prostate cancer cohort (Supplementary Figure S4.14). This could reflect the absence of *CDK12* mutations and the presence of only three sporadic *BRCA2*-mutated samples (1%) in the primary prostate cancer cohort. Furthermore, only one sample (1%) with MSI-like and high TMB (>10), respectively, was observed in the primary cancer cohort. Indeed, there is a striking difference in the mutational load between both disease settings.

Discussion

We performed WGS of metastatic tumor biopsies and matched-normal blood obtained from 197 patients with mCRPC to provide an overview of the genomic landscape of mCRPC. The size of our cohort enables classification of patients into distinct disease subgroups using unsupervised clustering. Our data suggest that classification of patients using genomic events, as detected by WGS, improves patient stratification, specifically for clinically actionable subgroups such as BRCA-deficient and MSI patients. Furthermore, we confirm the central role of AR signaling in mCRPC that mediates its effect through regulators located in non-coding regions and the apparent difference in primary versus metastatic prostate cancers.

The classification of patients using WGS has the advantage of being, in theory, more precise in determining genomically defined subgroups in prostate cancer compared to analyses using targeted panels consisting of a limited number of genes, or exome sequencing. The identification of subgroups based on predominant phenotypic characteristics encompassing genomic signatures may be clinically relevant and our clustering analysis refines patient classification. In cluster A, we observed a high TMB, which has been associated in other tumor types with a high sensitivity to immune check-point inhibitors.[9,11,12] Clinical trials using pembrolizumab in selected mCRPC patients are underway (KEYNOTE-028, KEYNOTE-199).[37,38] Interestingly, in both cluster B and cluster D, we identified patients that did not have the defining biallelic *CDK12* or *BRCA2* (somatic) mutation. Such patients might be deemed false-negatives when using FDA-approved assays (BRCAnalysis™ and FoundationFocus™), currently used in breast cancer diagnosis and based on the presence of *BRCA1/2* mutations, to predict response to poly(ADP-ribose) polymerase (PARP) inhibitors and/or platinum compounds. The first clinical trials combining PARP inhibitors with AR-targeted therapies in mCRPC show promising results.[8] Thus, WGS-based stratification may improve the patient classification of DNA repair-deficient tumors as it uses the genome-wide scars caused by defective DNA repair to identify tumors that have these deficiencies.

The use of WGS also allowed us to gain more insight into the role of non-coding regions of the genome in prostate cancer. We confirmed the amplification of a recently reported *AR*-enhancer.[20,21,30] In line with the cell line-based observations, we show AR binding at these mCRPC-specific enhancer regions, providing the first clinical indication that AR-enhancer amplification also increases AR signaling in mCRPC tumors. These findings are supported by previous studies demonstrating that this amplification ultimately resulted in significantly elevated expression of *AR* itself.[20,21,30] Furthermore, we confirm a recurrent focal amplification near *PCAT1*, which shows robust chromatin binding for AR in mCRPC samples, providing clinical proof-of-concept of a functional enhancer that is also active and AR-bound in cell line models. Recent research elucidated to the functional importance of this region in regulating *MYC* expression in prostate cancer, which could highlight a putative role of this somatically acquired amplification.[31] However, the WGS and ChIP-seq data presented here are not conclusive in elucidating the definitive role of this amplified region in regulating *MYC* expression and further mechanistic studies are needed to establish a potential link to *MYC* regulation.

In addition, *PCAT1* is a long non-coding RNA, which is known to be upregulated in prostate cancer and negatively regulates *BRCA2* expression while positively affecting *MYC* expression.[39,40] Combining our WGS approach with AR, FOXA1, and H3K27ac ChIP-seq data, we identify non-coding regions affecting both *AR* itself, and possibly *MYC*, through *AR*-enhancer amplification as a potential mechanism contributing to castration resistance.

A potential pitfall of our clustering analysis is the selection of features used; for this we made a number of assumptions based on the literature and distribution of the structural variants within our cohort.[19–21,32] As the input of features and weights for clustering analysis are inherent to the clustering outcome, we performed additional clustering analyses using various combinations of these features and applied alternative approaches but did not detect striking differences compared to the current approach. Another potential pitfall of the employed hierarchical clustering scheme is that patients are only attributed to a single cluster. An example of this can be seen in cluster A where a patient is grouped based on its predominant genotype (MSI) and associated mutations in mismatch repair (MMR)-related genes (*MLH1*, *POLE*, *POLD3*, and *BLM*), but this sample also displays an increased number of structural variants and increased ploidy status and harbors a pathogenic *BRCA2* mutation. However, it is missing the characteristic number of genomic deletions (<100 kbp) and BRCA mutational signature associated with *BRCA2*$^{-/-}$ samples that define cluster D. Despite these pitfalls we conclude that unbiased clustering contributes towards improved classification of patients.

The CPCT-02 study was designed to examine the correlation of genomic data with treatment outcome after biopsy at varying stages of disease. Our cohort contains patients with highly variable pre-treatment history and since the treatments for mCRPC patients nowadays significantly impacts overall survival, the prognosis of patients differs greatly. Therefore, correlation between genomic data and clinical endpoints, such as survival is inherently flawed due to the very heterogeneous nature of the patient population. Moreover, our analysis comparing primary and metastatic samples shows a significant increase in the number of genomic aberrations with advancing disease, meaning that the difference in timing of the biopsies may bias the prognostic value of the data. In future studies, we plan to gather all known clinically defined prognostic information and determine whether the genomic subtypes increase the ability to predict outcome. Unfortunately, some clinical parameters with prognostic importance such as ethnicity will not be available due to ethical regulations. Moreover, we will increase the sample size, in order to correlate genomic features to clinical parameters to better determine whether the subtypes we identified are stable over time. Therefore, we are currently unable to present meaningful correlations between clinical endpoints and the clusters we identified.

Overall, we show the added value of WGS-based unsupervised clustering in identifying patients with genomic scars who are eligible for specific therapies. Since our clustering method does not rely on one specific genetic mutation we are able to classify patients even when WGS (or our methodology) does not find conclusive evidence for (biallelic) mutations in the proposed gene-of-interest. Further research should validate clinical response and outcome on specific therapies in matched subgroups. This study also shows that a large population of mCRPC patients do not fall into an as-of-yet clinically relevant or biologically clear genotype and further research can help elucidate the oncogenic driver events and provide new therapeutic options.

Material and Methods

Patient cohort and study procedures

Patients with metastatic prostate cancer were recruited under the study protocol (NCT01855477) of the Center for Personalized Cancer Treatment (CPCT). This consortium consists of 41 hospitals in The Netherlands (Supplementary Data 1 (available online)). This CPCT-02 protocol was approved by the medical ethical committee (METC) of the University Medical Center Utrecht and was conducted in accordance with the Declaration of Helsinki. Patients were eligible for inclusion if the following criteria were met: (1) age ≥ 18 years; (2) locally advanced or metastatic solid tumor; (3) indication for new line of systemic treatment with registered anti-cancer agents; (4) safe biopsy according to the intervening physician. For the current study, patients were included for biopsy between 03 May 2016 and 28 May 2018. Data were excluded of patients with the following characteristics: (1) hormone-sensitive prostate cancer; (2) neuro-endocrine prostate cancer (as assessed by routine diagnostics); (3) unknown disease status; (4) prostate biopsy (Figure 4.1a). All patients provided written informed consent before any study procedure. The study procedures consisted of the collection of matched peripheral blood samples for reference DNA and image-guided percutaneous biopsy of a single metastatic lesion. Soft tissue lesions were biopsied preferentially over bone lesions. The clinical data provided by CPCT have been locked at 1st of July 2018.

Collection and sequencing of samples

Blood samples were collected in CellSave preservative tubes (Menarini-Silicon Biosystems, Huntington Valley, PA, USA) and shipped by room temperature to the central sequencing facility at the Hartwig Medical Foundation.[33] Tumor samples were fresh-frozen in liquid nitrogen directly after the procedure and send to a central pathology tissue facility. Tumor cellularity was estimated by assessing a hematoxylin-eosin (HE) stained 6 micron thick section. Subsequently, 25 sections of 20 micron were collected for DNA isolation. DNA was isolated with an automated workflow (QiaSymphony) using the DSP DNA Midi kit for blood and QiaSymphony DSP DNA Mini kit for tumor samples according to the manufacturer's protocol (Qiagen). DNA concentration was measured by Qubit™ fluorometric quantitation (Invitrogen, Life Technologies, Carlsbad, CA, USA). DNA libraries for Illumina sequencing were generated from 50 to100 ng of genomic DNA using standard protocols (Illumina, San Diego, CA, USA) and subsequently whole-genome sequenced in a HiSeq X Ten system using the paired-end sequencing protocol (2 × 150 base pair (bp)). Whole-genome alignment (GRCh37), somatic variants (SNV, InDels (max. 50 bp), MNV), structural variant and copy number calling and *in silico* tumor cell percentage estimation were performed in a uniform manner as detailed by Priestley *et al.*[33]. Mean read coverages of reference and tumor Binary Alignment Map (BAM) were calculated using Picard Tools (v1.141; CollectWgsMetrics) based on GRCh37.[41]

Additional annotation of somatic variants and heuristic filtering

In addition, heuristic filtering removed somatic SNV, InDels, and MNV variants based on the following criteria: (1) minimal alternative reads observations ≤ 3; (2) gnomAD exome (ALL) allele frequency ≥ 0.001 (corresponding to 62 gnomAD individuals); and (3) gnomAD genome (ALL) ≥0.005 (75 gnomAD individuals).[42] gnomAD database v2.0.2 was used. Per gene overlapping a genomic variant, the most deleterious mutation was used to annotate the overlapping gene. Structural variants, with B-Allele Frequency (BAF) ≥0.1, were further annotated by retrieving overlapping and nearest up- and downstream annotations using custom R scripts based on GRCh37 canonical University of California, Santa Cruz (UCSC) promoter and gene annotations with respect to their respective up- or downstream orientation (if known).[43] Only potential fusions with only two different gene-partners were considered (e.g., *TMPRSS2-ERG*); structural variants with both breakpoints falling within the same gene were simply annotated as structural variant mutations. Fusion annotation from the COSMIC (v85), Cancer Genome Interpreter (CGI) and Clinical Interpretation of Variants in Cancer (CIVIC) databases were used to assess known fusions.[44–46] The COSMIC (v85), OncoKB (July 12, 2018), CIVIC (July 26, 2018), CGI (July 26, 2018) and the list from Martincorena *et al.*[26] (dN/dS) were used to classify known oncogenic or cancer-associated genes[44–46].

Ploidy and copy-number analysis

Ploidy and copy-number (CN) analysis was performed by a custom pipeline as detailed by Priestley *et al.*.[33] Briefly, this pipeline combines BAF, read depth, and structural variants to estimate the purity and CN profile of a tumor sample. Recurrent focal and broad CN alterations were identified by GISTIC2.0 (v2.0.23).[27] GISTIC2.0 was run with the following parameters: (a) genegistic 1; (b) gcm extreme; (c) maxseg 4000; (d) broad 1; (e) brlen 0.98; (f) conf 0.95; (g) rx 0; (h) cap 3; (i) saveseg 0; (j) armpeel 1; (k) smallmem 0; (l) res 0.01; (m) ta 0.1; (n) td 0.1; (o) savedata 0; (p) savegene 1; (q) gvt 0.1.

Categorization of shallow and deep CN aberration per gene was based on thresholded GISTIC2 calls. Focal peaks detected by GISTIC2 were re-annotated, based on overlapping genomic coordinates, using custom R scripts and UCSC gene annotations. GISTIC2 peaks were annotated with all overlapping canonical UCSC genes within the wide peak limits. If a GISTIC2 peak overlapped with ≤3 genes, the most-likely targeted gene was selected based on oncogenic or tumor-suppressor annotation in the COSMIC (v85), OncoKB (July 12, 2018), CIVIC (July 26, 2018), and CGI (July 26, 2018) lists.[26,44–46] Peaks in gene deserts were annotated with their nearest gene.

Estimation of tumor mutational burden

The mutation rate per Mbp of genomic DNA was calculated as the total genome-wide amount of SNV, MNV, and InDels divided over the total amount of callable nucleotides (ACTG) in the human reference genome (hg19) FASTA sequence file:

$$TMB_{genomic} = \frac{(SNV_g + MNV_g + InDels_g)}{(2858674662/10^6)} \tag{4.1}$$

The mutation rate per Mbp of coding mutations was calculated as the amount of coding SNV, MNV, and InDels divided over the summed lengths of distinct non-overlapping coding regions, as determined on the subset of protein-coding and fully supported (TSL = 21) transcripts in GenCode v28 (hg19)[47]:

$$TMB_{coding} = \frac{(SNV_c + MNV_c + InDels_c)}{(28711682/10^6)} \tag{4.2}$$

MSI and HR-deficiency prediction

HR-deficiency/BRCAness was estimated using the CHORD classifier (Nguyen, van Hoeck and Cuppen, manuscript in preparation). This classifier was based on the HRDetect[48] algorithm, however, redesigned to improve its performance beyond primary breast cancer. The binary prediction score (ranging from 0 to 1) was used to indicate BRCAness level within a sample. To elucidate the potential target gene(s) in the HR-deficient samples (Figure 4.4), we used the list of BRCAness genes from Lord *et al.*[36].

MSI status was determined based on the following criteria: if a sample contained >11,436 genomic InDels (max. 50 bp, with repeat-stretches of ≥4 bases, repeat length sequence between 2 and 4, or if these InDels consist of a single repeat sequence, which repeats ≥5 times), the sample was designated as MSI.[33]

Detection of (onco-)genes under selective pressure

To detect (onco-)genes under tumor-evolutionary mutational selection, we employed a Poisson-based dN/dS model (192 rate parameters; under the full trinucleotide model) by the R package dndscv (v0.0.0.9).[26] Briefly, this model tests the normalized ratio of non-synonymous (missense, nonsense, and splicing) over background (synonymous) mutations while correcting for sequence composition and mutational signatures. A global *q*-value ≤ 0.1 (with and without taking InDels into consideration) was used to identify statistically significant (novel) driver genes.

Identification of hypermutated foci (kataegis)

Putative kataegis events were detected using a dynamic programming algorithm, which determines a globally optimal fit of a piecewise constant expression profile along genomic coordinates as described by Huber *et al.*[49] and implemented in the tilingarray R package (v1.56.0). Only SNVs were used in detecting kataegis. Each chromosome was assessed separately and the maximum number of segmental breakpoints was based on a maximum of five consecutive SNVs (max. 5000 segments per chromosome). Fitting was performed on $\log_{10}$-transformed intermutational distances. Per segment, it was assessed if the mean intermutational distance was ≤2000 bp and at least five SNVs were used in the generation of the segment. A single sample with >200 distinct observed events was set to zero observed events as this sample was found to be hypermutated throughout the entire genome rather than locally. Kataegis was visualized using the R package karyoploteR (v1.4.1).[50]

Mutational signatures analysis

Mutational signatures analysis was performed using the MutationalPatterns R package (v1.4.2).[51] The 30 consensus mutational signatures, as established by Alexandrov et. al, (matrix S_{ij}; $i = 96$; number of trinucleotide motifs; $j = 30$; number of signatures) were downloaded from COSMIC (as visited on 23-05-2018)[44]. Mutations (SNVs) were categorized according to their respective trinucleotide context (hg19) into a mutational spectrum matrix M_{ij} ($i = 96$; number of trinucleotide contexts; $j = 196$; number of samples) and subsequently, per sample a constrained linear combination of the thirty consensus mutational signatures was constructed using non-negative least squares regression implemented in the R package pracma (v1.9.3).

Between two and 15 custom signatures were assessed using the NMF package (v0.21.0) with 1000 iterations.[52] By comparing the cophenetic correlation coefficient, residual sum of squares and silhouette, we opted to generate five custom signatures. Custom signatures were correlated to existing (COSMIC) signatures using cosine similarity.

Detection of chromothripsis-like events

Rounded absolute copy number (excluded Y chromosome) and structural variants (BAF ≥ 0.1) were used in the detection of chromothripsis-like events by the Shatterseek software (v0.4) using default parameters.[53] As a precise standardized definition of chromothripsis has not yet been fully established, and as per the author's instruction, we performed visual inspection of reported chromothripsis-like events after dynamically adapting criteria thresholds (taking the recommended thresholds into consideration). We opted to use the following criteria: (a) Total number of intrachromosomal structural variants involved in the event ≥25; (b) max. number of oscillating CN segments (two states) ≥7 or max. number of oscillating CN segments (three states) ≥14; (c) Total size of chromothripsis event ≥20 Mbp; (d) Satisfying the test of equal distribution of SV types ($p > 0.05$); and (e) Satisfying the test of non-random SV distribution within the cluster region or chromosome ($p \leq 0.05$).

Unsupervised clustering of mCRPC WGS characteristics

Samples were clustered using the Euclidian distance of the Pearson correlation coefficient ($1 - r$) and Ward.D hierarchical clustering based on five basic whole-genome characteristics; number of mutations per genomic Mbp (SNV, InDels, and MNV), mean genome-wide ploidy, number of structural variants and the relative frequencies of structural variant categories (inversions, tandem duplications (larger and smaller than 100 kbp), deletions (larger and smaller than 100 kbp), insertions and interchromosomal translocations). Data was scaled but not centered (root mean square) prior to calculating Pearson correlation coefficients. After clustering, OLO was performed using the seriation package (v1.2.3).[54] The elbow method was employed to determine optimal number of discriminating clusters (Supplementary Figure S4.10) using the factoextra package (v1.0.5). Bootstrapping was performed using the pvclust package (v2.0) with 5000 iterations.

Cluster-specific enrichment of aberrant genes (either through SV, deep copy-number alteration, or coding SNV/InDels/MNV), kataegis, chromothripsis, GISTIC2 peaks, and predicted fusions between clusters was tested using a two-sided Fisher's Exact Test and Benjamini–Hochberg correction.

A principal component analysis (with scaling and centering) using the prcomp R package[55] was performed on the chosen genomic features and $\cos^2$ values for each feature per principal component were retrieved to determine the importance of each feature per respective principal component.

To test the robustness of our clustering, we performed unsupervised clustering, and also other techniques, using various combinations of structural variants and clustering mechanisms as a surrogate for different genome-instability metrics but this analysis did not reveal any striking new clusters.

Supervised clustering based on mutually exclusive aberrations

Samples were sorted on mutual-exclusivity of *SPOP*, *FOXA1*, and *IDH1* coding mutations and copy-number aberrations and ETS family gene fusions (and overexpression) per promiscuous partner (*ERG*, *ETV1*, *ETV4*, and *FLI1*) as defined in primary prostate cancer.[13] Supplementary Table S1A of the article "The Molecular Taxonomy of Primary Prostate Cancer"[13] was used to determine the relative frequency and mutational types of each of the respective primary prostate cancer within the TCGA cohort. In addition, as the TCGA cohort did not denote high-level/deep amplifications, we did not incorporate these either in this analysis.

Correlation of the detection rate of genomic aberrations versus tumor cell percentages

Absolute counts of SNV, InDels, MNV and SV were correlated to the *in silico* estimated tumor cell percentage using Spearman's correlation coefficient.

Correlation of pre-treatment history with detected aberrations and WGS characteristics

Pre-treatment history of patients was summarized into ten groups:

- Only chemo-treatment (with radio-nucleotides).
- Only chemo-treatment (without radio-nucleotides).
- Only radio-nucleotides.
- Only secondary anti-hormonal therapy (with radio-nucleotides).
- Only secondary anti-hormonal therapy (without radio-nucleotides).
- Secondary anti-hormonal therapy + one chemo-treatment (with radio-nucleotides)
- Secondary anti-hormonal therapy + two chemo-treatment (with radio-nucleotides)
- Secondary anti-hormonal therapy + one chemo-treatment (without radio-nucleotides)
- Secondary anti-hormonal therapy + two chemo-treatment (without radio-nucleotides)

- No additional treatment after androgen deprivation therapy.

Association with mutated genes, presence of chromothripsis, presence of kataegis, MSI-status, and genomic subtypes was tested with a two-sided Fisher's exact test with Benjamini–Hochberg correction.

ChIP-seq experimental set-up and analysis

ChIP-seq cell culturing: VCaP cells were incubated in RPMI medium in additional with 10% fetal bovine serum (FBS). Bicalutamide-resistant VCaP cells (VCaP-Bic) were cultured in RPMI medium supplemented with 10% dextran charcoal-stripped bovine serum (DCC) and 10-6M bicalutamide. VCaP cells were hormone deprived in RPMI medium with 10% DCC for 3 days before the ChIP-seq experiment.

ChIP-seq and peak calling analysis: For both cell and tissue ChIPs, 5 µg of antibody and 50 µg of magnetic protein A or G beads (10008D or 10009D, Thermo Fisher Scientific) were used per immunoprecipitation (IP). The following antibodies were used: Foxa1/2 (M-20, sc-6554 Santa Cruz Biotechnology), AR (N-20, sc-816 Santa Cruz Biotechnology), and H3K27ac (39133, Active Motif). ChIP-seq was performed as described previously.[56] In brief, fresh-frozen tissue was cryosectioned into 30 micron thick slices and stored at −80 °C till processing. Samples were fixed using 2 mM DSG (20593; Thermo Fisher Scientific) in solution A (50 mM Hepes-KOH, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) while rotating for 25 min at room temperature, followed by the addition of 1% formaldehyde and another 20 min incubation at room temperature. The reaction was quenched by adding a surplus of glycine. Subsequently, tissue sections were pelleted and washed with cold PBS. Tissue was disrupted using a motorized pellet pestle (Sigma-Aldrich) to disrupt the tissue in cold PBS and obtain a cell suspension, after which the nuclei were isolated and the chromatin was sheared. During immunoprecipitation, human control RNA (4307281; Thermo Fisher Scientific) and recombinant Histone 2B (M2505S; New England Biolabs) were added as carriers, as described previously.[57]

Immunoprecipitated DNA was processed for sequencing using standard protocols and sequenced on an Illumina HiSeq 2500 with 65 bp single end reads. Sequenced samples were aligned to the reference human genome (Ensembl release 55: Homo sapiens GRCh 37.55) using Burrows-Wheeler Aligner (BWA, v0.5.10)[58], reads with a mapping quality >20 were used for further downstream analysis.

For the tissues, peak calling was performed using MACS2[59] with option –nomodel. In addition, peaks were called against matched input using DFilter[60] in the refine setting with a bandwidth of 50 and a kernel size of 30. Only peaks that were shared between the two algorithms were considered.

For the cell lines, peaks were obtained with MACS (v1.4; $p \leq 10-7$).

The AR and FOXA1 ChIP-seq data for LNCAP with/-out R1881 was obtained from GSE94682[61]. The H3K27ac ChIP-seq data for LNCAP was obtained from GSE114737[56].

Determining enrichment of enhancer to gene ratios: Absolute copy-numbers segments overlapping the gene loci and putative enhancer region (as detected by GISTIC2; focal amplification peaks with a width <5000 bp) were retrieved per sample. If regions overlapped multiple distinct copy-number segments, the maximum copy-number value of the overlapping segments was used to represent the region. Samples with gene-to-enhancer ratios deviating >1 studentized residual from equal 1:1 gene-to-enhancer ratios (linear model: $\log_2$(copy number of enhancer) − $\log_2$(copy number of gene locus) 0) were categorized as gene or enhancer enriched. Based on the direction of the ratio, samples were either denoted as enhancer (if positive ratio) or gene (if negative ratio) enriched.

Comparison of unmatched primary prostate cancer and mCRPC

Mutational frequencies of the drivers (dN/dS and or GISTIC2) and subtype-specific genes were compared to a separate (unmatched) cohort of primary prostate cancer ($n = 210$) focusing on Gleason score (GS) of 3 + 3, 3 + 4, or 4 + 3, as described by Fraser *et al.*[15] and Espiritu *et al.*[25]. Briefly, whole-genome sequencing reads were mapped to the human reference genome (GRCh37) using BWA[58] (v0.5.7) and downstream analysis was performed using Strelka[62] (v.1.0.12) for mutational calling using a matched-normal design (SNVs and InDels), copy-number alterations were estimated with TITAN[63] (v1.11.0), and single-nucleotide polymorphism (SNP) array data as described in Espiritu *et al.*[25] with Delly[64] (v0.5.5 and v0.7.8) was used for detecting structural variants (translocations, inversions, tandem duplications, and deletions). Large insertion calls and overall ploidy was not available for the primary prostate cancer cohort.

TMB was calculated by dividing the number of SNVs and InDels by the total amount of callable bases in the human reference genome (GRCh37), identical to 4.1. MNV calls were not available for the primary prostate cancer cohort.

Multiple aberrations per gene within a sample were summarized as a single mutational event, e.g., a deletion and mutation in *PTEN* would only count for a single mutation in the sample. Only non-synonymous mutations and gains/deletions overlapping with coding regions were used. Statistically significant differences in mutational frequencies were calculated using a two-sided Fisher's Exact test with Benjamini–Hochberg correction.

The primary prostate cancer dataset was clustered together with the mCRPC cohort using the Euclidian distance of the Pearson correlation coefficient ($1-r$) and Ward.D hierarchical clustering based on three basic whole-genome characteristics, which were available for all samples; number of mutations per genomic Mbp (SNVs and InDels), number of structural variants and the relative frequencies of structural variant categories (inversions, tandem duplications (larger and smaller than 100 kbp), deletions (larger and smaller than 100 kbp), and interchromosomal translocations).

Data availability

The data that support the findings of this study are available from Hartwig Medical Foundation, which were used under data request number DR-011 for the current study. Both WGS and clinical data are freely available for academic use from the Hartwig Medical Foundation through standardized procedures and request forms can be found at https://www.hartwigmedicalfoundation.nl.[33] The ChIP-seq profiles (aligned reads and MACS/MACS2 peaks) as analyzed and shown in this manuscript have been deposited on Gene Expression Omnibus (GEO) under accession number: GSE138168.

Code availability

All tools and scripts used for processing of the WGS data are available at https://github.com/hartwigmedical/ and/or can be provided by authors upon request.

References

[1] L. Boyd, X. Mao, and Y.-J. Lu, *The complexity of prostate cancer: Genomic alterations and heterogeneity,* Nature Reviews Urology **9**, 652 (2012).

[2] L. Wei, J. Wang, E. Lampert, S. Schlanger, A. DePriest, *et al.*, *Intratumoral and intertumoral genomic heterogeneity of multifocal localized prostate cancer impacts molecular classifications and genomic prognosticators,* European Urology **71**, 183 (2017).

[3] S. Mullane and E. Van Allen, *Precision medicine for advanced prostate cancer,* Current Opinion in Urology **26**, 231 (2016).

[4] C. Ciccarese, F. Massari, R. Iacovelli, M. Fiorentino, R. Montironi, *et al.*, *Prostate cancer heterogeneity: Discovering novel molecular targets for therapy,* Cancer Treatment Reviews **54**, 68 (2017).

[5] E. Shtivelman, T. Beer, and C. Evans, *Molecular pathways and targets in prostate cancer,* Oncotarget **5**, 7217 (2014).

[6] D. Robinson, E. Van Allen, Y.-M. Wu, N. Schultz, R. Lonigro, *et al.*, *Integrative clinical genomics of advanced prostate cancer,* Cell **161**, 1215 (2015).

[7] H. Chow, P. Ghosh, R. Devere White, C. Evans, M. Dall'Era, *et al.*, *A phase 2 clinical trial of everolimus plus bicalutamide for castration-resistant prostate cancer,* Cancer **122**, 1897 (2016).

[8] N. Clarke, P. Wiechno, B. Alekseev, N. Sala, R. Jones, *et al.*, *Olaparib combined with abiraterone in patients with metastatic castration-resistant prostate cancer: a randomised, double-blind, placebo-controlled, phase 2 trial,* The Lancet Oncology **19**, 975 (2018).

[9] M. Yarchoan, A. Hopkins, and E. Jaffee, *Tumor mutational burden and response rate to pd-1 inhibition,* New England Journal of Medicine **377**, 2500 (2017).

[10] T. Chan, M. Yarchoan, E. Jaffee, C. Swanton, S. Quezada, *et al.*, *Development of tumor mutation burden as an immunotherapy biomarker: Utility for the oncology clinic,* Annals of Oncology **30**, 44 (2019).

[11] N. Rizvi, M. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, *et al.*, *Mutational landscape determines sensitivity to pd-1 blockade in non-small cell lung cancer,* Science **348**, 124 (2015).

[12] R. Samstein, C.-H. Lee, A. Shoushtari, M. Hellmann, R. Shen, *et al.*, *Tumor mutational load predicts survival after immunotherapy across multiple cancer types,* Nature Genetics **51**, 202 (2019).

[13] A. Abeshouse, J. Ahn, R. Akbani, A. Ally, S. Amin, *et al.*, *The molecular taxonomy of primary prostate cancer,* Cell **163**, 1011 (2015).

[14] A. Angeles, S. Bauer, L. Ratz, S. Klauck, and H. Sultmann, *Genome-based classification and therapy of prostate cancer,* Diagnostics **8** (2018).

[15] M. Fraser, V. Sabelnykova, T. Yamaguchi, L. Heisler, J. Livingstone, *et al.*, *Genomic hallmarks of localized, non-indolent prostate cancer,* Nature **541**, 359 (2017).

[16] J. Schoenborn, P. Nelson, and M. Fang, *Genomic profiling defines subtypes of prostate cancer with the potential for therapeutic stratification,* Clinical Cancer Research **19**, 4058 (2013).

[17] R. Nam, L. Sugar, Z. Wang, W. Yang, R. Kitching, *et al.*, *Expression of tmprss2 erg gene fusion in prostate cancer cells is an important prognostic factor for cancer progression,* Cancer Biology and Therapy **6**, 40 (2007).

[18] S. Tomlins, B. Laxman, S. Varambally, X. Cao, J. Yu, *et al.*, *Role of the tmprss2-erg gene fusion in prostate cancer,* Neoplasia **10**, 177 (2008).

[19] Y.-M. Wu, M. Cieślik, R. Lonigro, P. Vats, M. Reimers, *et al.*, *Inactivation of cdk12 delineates a distinct immunogenic class of advanced prostate cancer,* Cell **173**, 1770 (2018).

[20] S. Viswanathan, G. Ha, A. Hoff, J. Wala, J. Carrot-Zhang, *et al.*, *Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing,* Cell **174**, 433 (2018).

[21] D. Quigley, H. Dang, S. Zhao, P. Lloyd, R. Aggarwal, *et al.*, *Genomic hallmarks and structural variation in metastatic prostate cancer,* Cell **174**, 758 (2018).

[22] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, *et al.*, *Landscape of somatic mutations in 560 breast cancer whole-genome sequences,* Nature **534**, 47 (2016).

[23] R. Taylor, M. Fraser, J. Livingstone, S. Espiritu, H. Thorne, *et al.*, *Germline brca2 mutations drive prostate cancers with distinct evolutionary trajectories,* Nature Communications **8** (2017), 10.1038/ncomms13671.

[24] H. Davies, S. Morganella, C. Purdie, S. Jang, E. Borgen, *et al.*, *Whole-genome sequencing reveals breast cancers with mismatch repair deficiency,* Cancer Research **77**, 4755 (2017).

[25] S. Espiritu, L. Liu, Y. Rubanova, V. Bhandari, E. Holgersen, *et al.*, *The evolutionary landscape of localized prostate cancers drives clinical aggression,* Cell **173**, 1003 (2018).

[26] I. Martincorena, K. Raine, M. Gerstung, K. Dawson, K. Haase, *et al.*, *Universal patterns of selection in cancer and somatic tissues,* Cell **171**, 1029 (2017).

[27] C. Mermel, S. Schumacher, B. Hill, M. Meyerson, R. Beroukhim, *et al.*, *Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers,* Genome Biology **12** (2011), 10.1186/gb-2011-12-4-r41.

[28] J. Armenia, S. Wankowicz, D. Liu, J. Gao, R. Kundra, *et al.*, *The long tail of oncogenic drivers in prostate cancer,* Nature Genetics **50**, 645 (2018).

[29] L. Alexandrov, S. Nik-Zainal, D. Wedge, S. Aparicio, S. Behjati, *et al.*, *Signatures of mutational processes in human cancer,* Nature **500**, 415 (2013).

[30] D. Takeda, S. Spisák, J.-H. Seo, C. Bell, E. O'Connor, *et al.*, *A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer,* Cell **174**, 422 (2018).

[31] P. Mazrooei, *Somatic mutations and risk-variants converge on cis-regulatory elements to reveal the cancer driver transcription regulators in primary prostate tumors,* SSRN Electron. J. (2018).

[32] F. Menghi, F. Barthel, V. Yadav, M. Tang, B. Ji, *et al.*, *The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations,* Cancer Cell **34**, 197 (2018).

[33] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, *et al.*, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[34] C. Pritchard, C. Morrissey, A. Kumar, X. Zhang, C. Smith, *et al.*, *Complex msh2 and msh6 mutations in hypermutated microsatellite unstable advanced prostate cancer,* Nature Communications **5** (2014), 10.1038/ncomms5988.

[35] P. Polak, J. Kim, L. Braunstein, R. Karlic, N. Haradhavala, *et al.*, *A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer,* Nature Genetics **49**, 1476 (2017).

[36] C. Lord and A. Ashworth, *Brcaness revisited,* Nature Reviews Cancer **16**, 110 (2016).

[37] A. Hansen, C. Massard, P. Ott, N. Haas, J. Lopez, *et al.*, *Pembrolizumab for advanced prostate adenocarcinoma: Findings of the keynote-028 study,* Annals of Oncology **29**, 1807 (2018).

[38] J. De Bono, J. Goh, and K. Ojamaa, *Keynote-199: Pembrolizumab (pembro) for docetaxel-refractory metastatic castration-resistant prostate cancer (mcrpc),* J Clin Oncol **36** (2018).

[39] J. Prensner, W. Chen, S. Han, M. Iyer, Q. Cao, *et al.*, *The long non-coding rna pcat-1 promotes prostate cancer cell proliferation through cmyc,* Neoplasia **16**, 900 (2014).

[40] J. Prensner, W. Chen, M. Iyer, Q. Cao, T. Ma, *et al.*, *Pcat-1, a long noncoding rna, regulates brca2 and controls homologous recombination in cancer,* Cancer Research **74**, 1651 (2014).

[41] A. Wysoker, K. Tibbetts, and T. Fennell, Picard Tools (2016).

[42] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, *et al.*, *Analysis of protein-coding genetic variation in 60,706 humans,* Nature **536**, 285 (2016).

[43] J. Casper, A. Zweig, C. Villarreal, C. Tyner, M. Speir, *et al.*, *The ucsc genome browser database: 2018 update,* Nucleic Acids Research **46**, D762 (2018).

[44] S. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, *et al.*, *Cosmic: Somatic cancer genetics at high-resolution,* Nucleic Acids Research **45**, D777 (2017).

[45] D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. Schroeder, A. Vivancos, *et al.*, *Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations,* Genome Medicine **10** (2018), 10.1186/s13073-018-0531-8.

[46] M. Griffith, N. Spies, K. Krysiak, J. McMichael, A. Coffman, *et al.*, *Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer,* Nature Genetics **49**, 170 (2017).

[47] J. Harrow, A. Frankish, J. Gonzalez, E. Tapanari, M. Diekhans, *et al.*, *Gencode: The reference human genome annotation for the encode project,* Genome Research **22**, 1760 (2012).

[48] H. Davies, D. Glodzik, S. Morganella, L. Yates, J. Staaf, *et al.*, *Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures,* Nature Medicine **23**, 517 (2017).

[49] W. Huber, J. Toedling, and L. Steinmetz, *Transcript mapping with high-density oligonucleotide tiling arrays,* Bioinformatics **22**, 1963 (2006).

[50] B. Gel and E. Serra, *Karyoploter: An r/bioconductor package to plot customizable genomes displaying arbitrary data,* Bioinformatics **33**, 3088 (2017).

[51] F. Blokzijl, R. Janssen, R. van Boxtel, and E. Cuppen, *Mutationalpatterns: Comprehensive genome-wide analysis of mutational processes,* Genome Medicine **10** (2018), 10.1186/s13073-018-0539-0.

[52] R. Gaujoux and C. Seoighe, *A flexible r package for nonnegative matrix factorization,* BMC Bioinformatics **11** (2010), 10.1186/1471-2105-11-367.

[53] I. Cortes-Ciriano, J.-K. Lee, R. Xi, D. Jain, Y. Jung, *et al.*, *Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing,* BioRxiv (2018).

[54] M. Hahsler, K. Hornik, and C. Buchta, *Getting things in order: An introduction to the r package seriation,* Journal of Statistical Software **25**, 1 (2008).

[55] W. Venables and B. Ripley, Modern Applied Statistics with S-Plus (1994).

[56] A. Singh, K. Schuurman, E. Nevedomskaya, S. Stelloo, S. Linder, *et al.*, *Optimized chip-seq method facilitates transcription factor profiling in human tumors,* Life Science Alliance **2** (2019), 10.26508/lsa.201800115.

[57] W. Zwart, R. Koornstra, J. Wesseling, E. Rutgers, S. Linn, *et al.*, *A carrier-assisted chip-seq method for estrogen receptor-chromatin interactions from breast cancer core needle biopsy samples,* BMC Genomics **14** (2013), 10.1186/1471-2164-14-232.

[58] H. Li and R. Durbin, *Fast and accurate short read alignment with burrows-wheeler transform,* Bioinformatics **25**, 1754 (2009).

[59] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, *et al.*, *Model-based analysis of chip-seq (macs),* Genome Biology **9** (2008), 10.1186/gb-2008-9-9-r137.

[60] V. Kumar, M. Muratani, N. Rayan, P. Kraus, T. Lufkin, *et al.*, *Uniform, optimal signal processing of mapped deep-sequencing data,* Nature Biotechnology **31**, 615 (2013).

[61] S. Stelloo, E. Nevedomskaya, Y. Kim, L. Hoekman, O. Bleijerveld, *et al.*, *Endogenous androgen receptor proteomic profiling reveals genomic subcomplex involved in prostate tumorigenesis,* Oncogene **37**, 313 (2018).

[62] S. Kim, *Strelka2: Fast and accurate variant calling for clinical sequencing applications,* bioRxiv (2017).

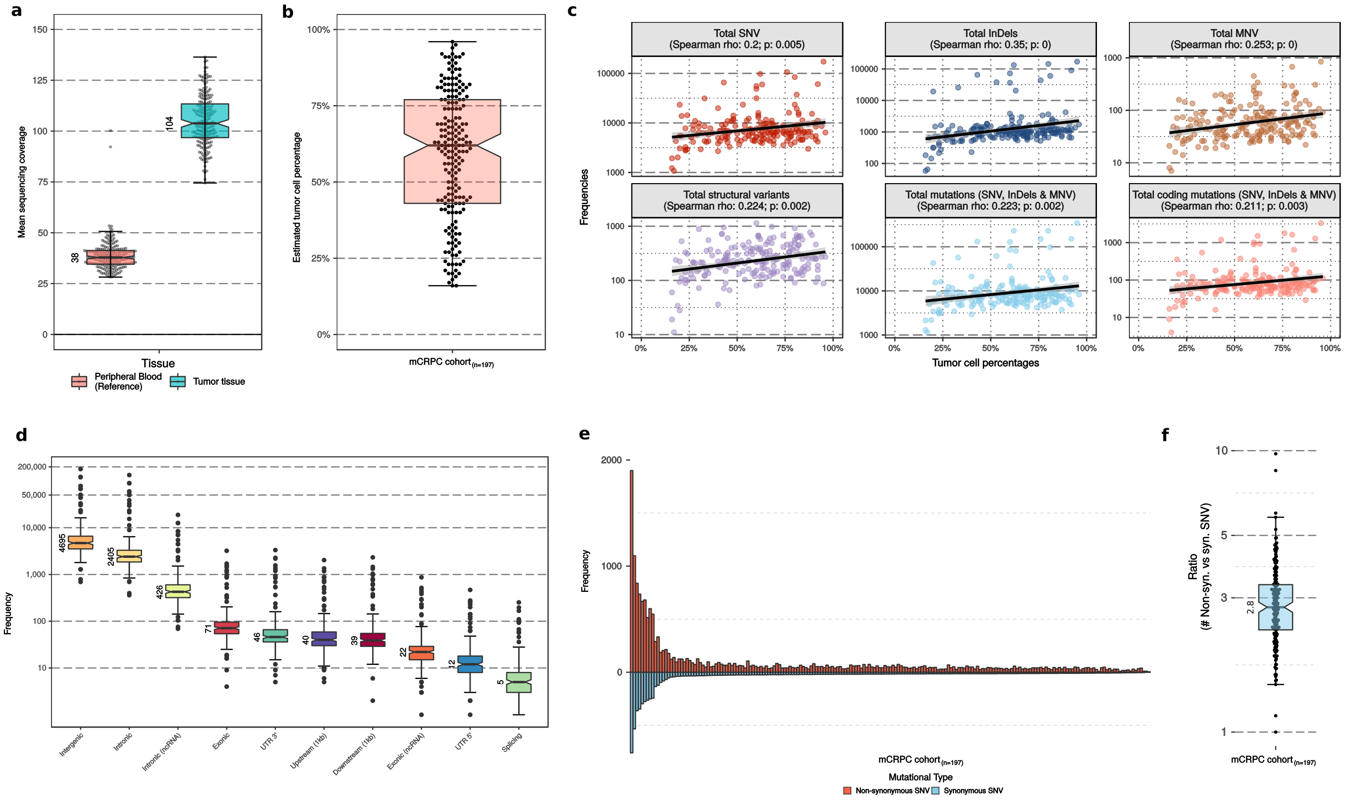
[63] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, *et al.*, *Titan: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data,* Genome Research **24**, 1881 (2014).

[64] T. Rausch, T. Zichner, A. Schlattl, A. Stütz, V. Benes, *et al.*, *Delly: Structural variant discovery by integrated paired-end and split-read analysis,* Bioinformatics **28**, i333 (2012).

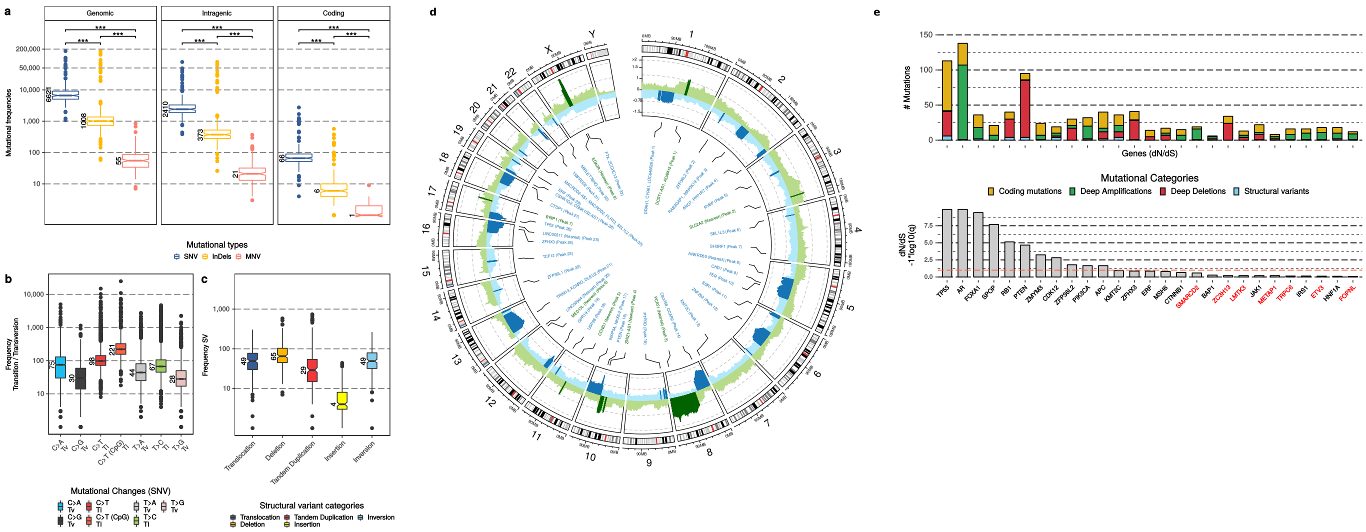
Supplemental Data

Supplementary data and figures accompanying the chapter:

"The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact"

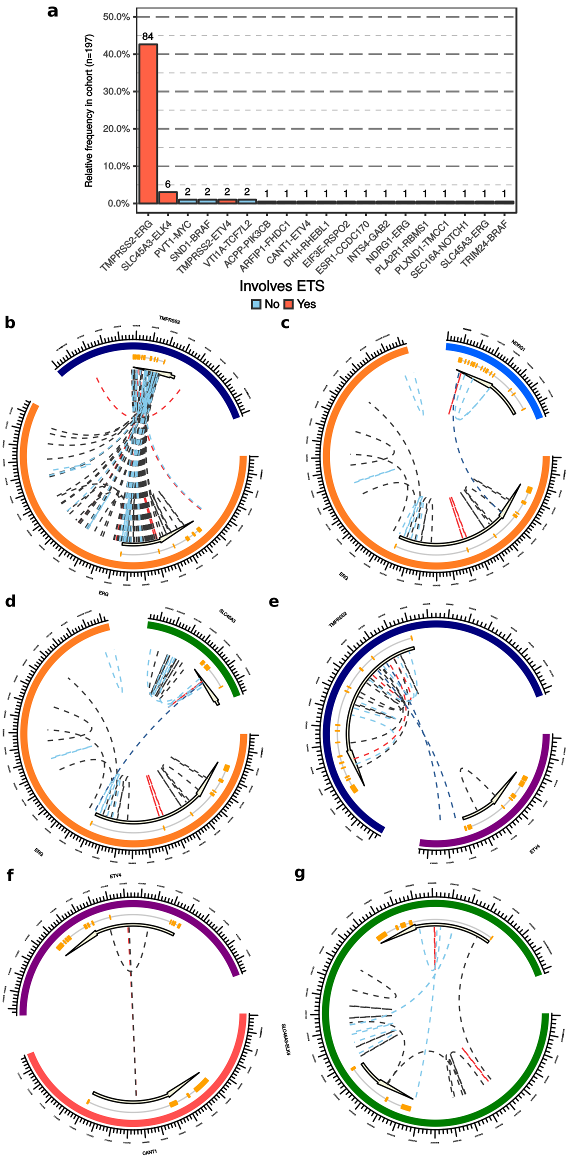


Supplementary figure S4.1: **Overview of sequencing quality metrics.**
(a): Bee-swarm boxplot with notch of the mean read coverage per sample of reference and tumor tissues. Boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. **(b):** Bee-swarm boxplot with notch of the estimated (*in silico*) cohort-wide tumor cell percentages. Boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. **(c):** Correlation (spearman) of estimated tumor cell percentages with observed aberrations per mutational category. Based on the low rank correlation coefficients (Spearman rho) we did not find high correlation with TC% and detected events, however a minor correlation could indeed be seen. **(d):** Overview of the locations of variants (SNV / InDels / MNV) in respect to UCSC gene-models. Boxplot with notch depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. **(e):** Frequency of non-synonymous (red) and synonymous (blue) SNV per mCRPC sample. **(f):** Ratio of non-synonymous over synonymous SNV for the entire mCRPC cohort. Bee-swarm boxplot with notch of the ratio. Boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown.

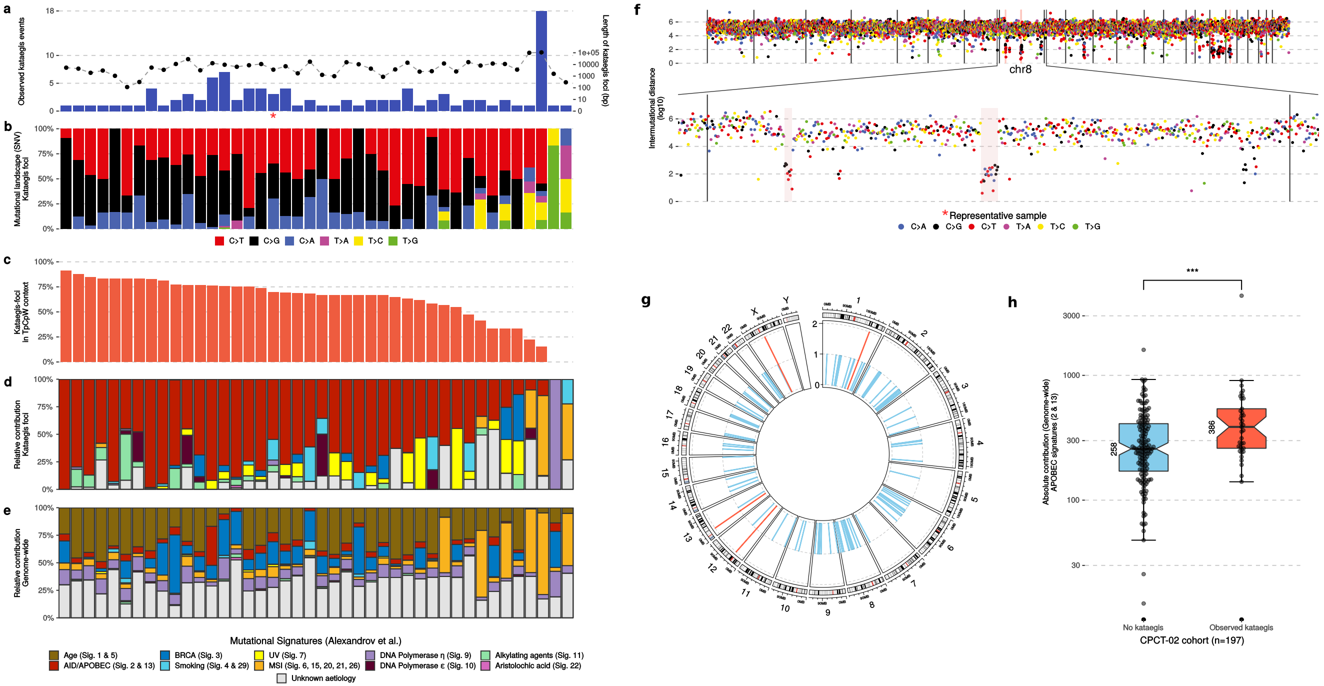


Supplementary figure S4.2: **Overview of cohort-wide mCRPC somatic characteristics**.

(a): Number of SNV (blue), InDels (yellow) and MNV (orange) per whole-genome sequenced sample over three resolutions; genome-wide, within intragenic regions and within coding regions. Boxplot with notch depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. Statistical significance (Wilcoxon rank-sum test) is denoted per comparison. **(b):** Type of genome-wide SNVs. transition (Ti) and transversion (Tv), with a special attention for C to T Ti in CpG context, are indicated per sample. Boxplot with notch depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. **(c):** Frequency of Tandem Duplications (DUP), Insertions (INS), Inversions (INV), Deletions (DEL) and interchromosomal translocations(BND) are indicated per sample. Boxplot with notch depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. **(d):** Overview of recurrent copy number aberrations as detected by GISTIC2. G-scores are depicted on the y-axis ranging from 0 to ≥ 2. Regions with amplifications (G-score > 0) are depicted in green and deletions (G-score < 0) in blue. Regions with significant (and recurring) copy number aberrations ($q \leq 0.1$) are denoted with a darker shade of green or blue, respective of amplification or deletion. Per region, the foci of maximal amplification or deletion (focal peaks; $q \leq 0.1$) are denoted in the inner track; the peak identifier is also denoted as presented in supplementary Data 1 (available online). **(e):** Overview of genes detected by the dN/dS algorithm and corresponding mutational categories. Genes not present in one of our lists of known (onco)genes are colored red; (COSMIC v85, CGI, CIVIC and the list from Martincorena *et al.*). The upper figure displays absolute frequencies per mutational category in the detected genes and the lower figure displays the respective q-value (-1*$\log_{10}(q)$). The red line in the bottom figure indicates the threshold for statistical significance ($q = 0.01$).

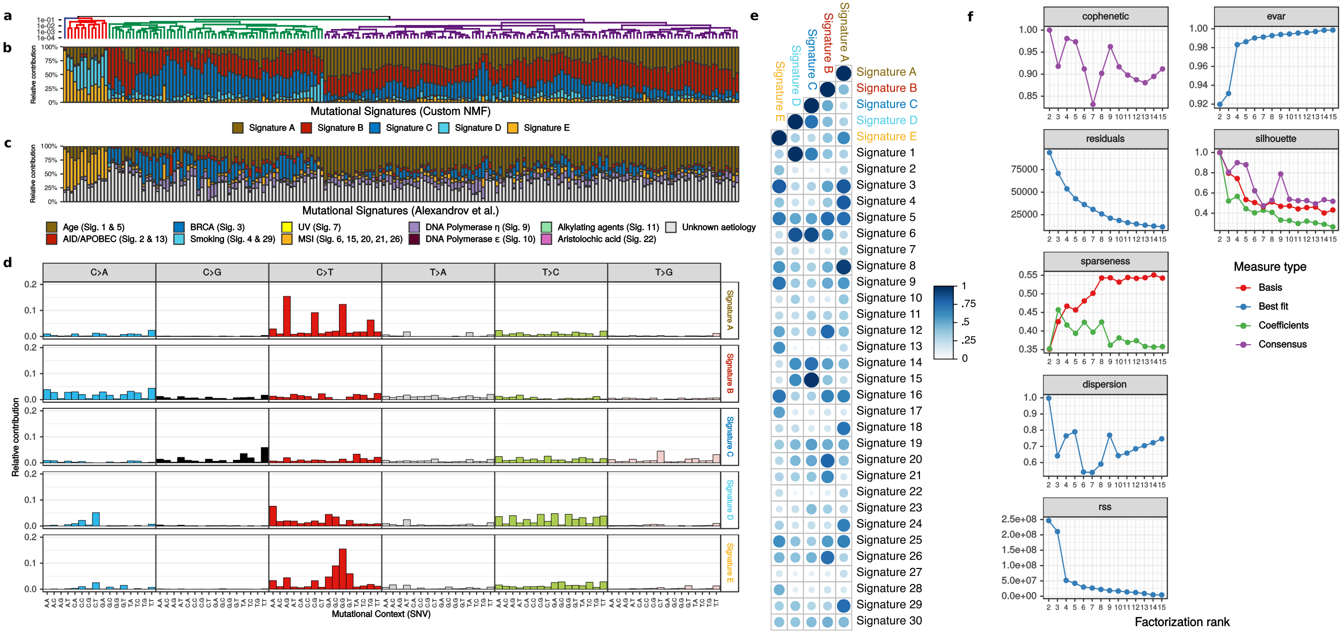


Supplementary figure S4.3: **Overview of genomic (ETS) fusions.**
(a): Relative frequency of observed genomic aberrations resulting in potential fusion products. Fusions involving ETS genes are depicted in red, other potential fusion partners in blue. Numbers above the bars indicate theabsolute number of patients with the genomic aberration in the mCRPC cohort.(**b-g**) Overview of structural variants involving the *TMPRSS2* and *ERG* loci **(b)**, *NDRG1* and *ERG* loci **(c)**, *SLC45A3* and *ERG* loci **(d)**, *TMPRSS2* and *ETV4* loci **(e)**, *CANT1* and *ETV4* loci **(f)** and the *SLC45A3* and *ELK4* loci **(g)** in the mCRPC cohort. Interchromosomal translocations are colored in dark blue, deletions in black, insertions in yellow, inversion in light blue and tandem duplications in red. Orange boxes indicate exons; black line con-necting the boxes are introns.

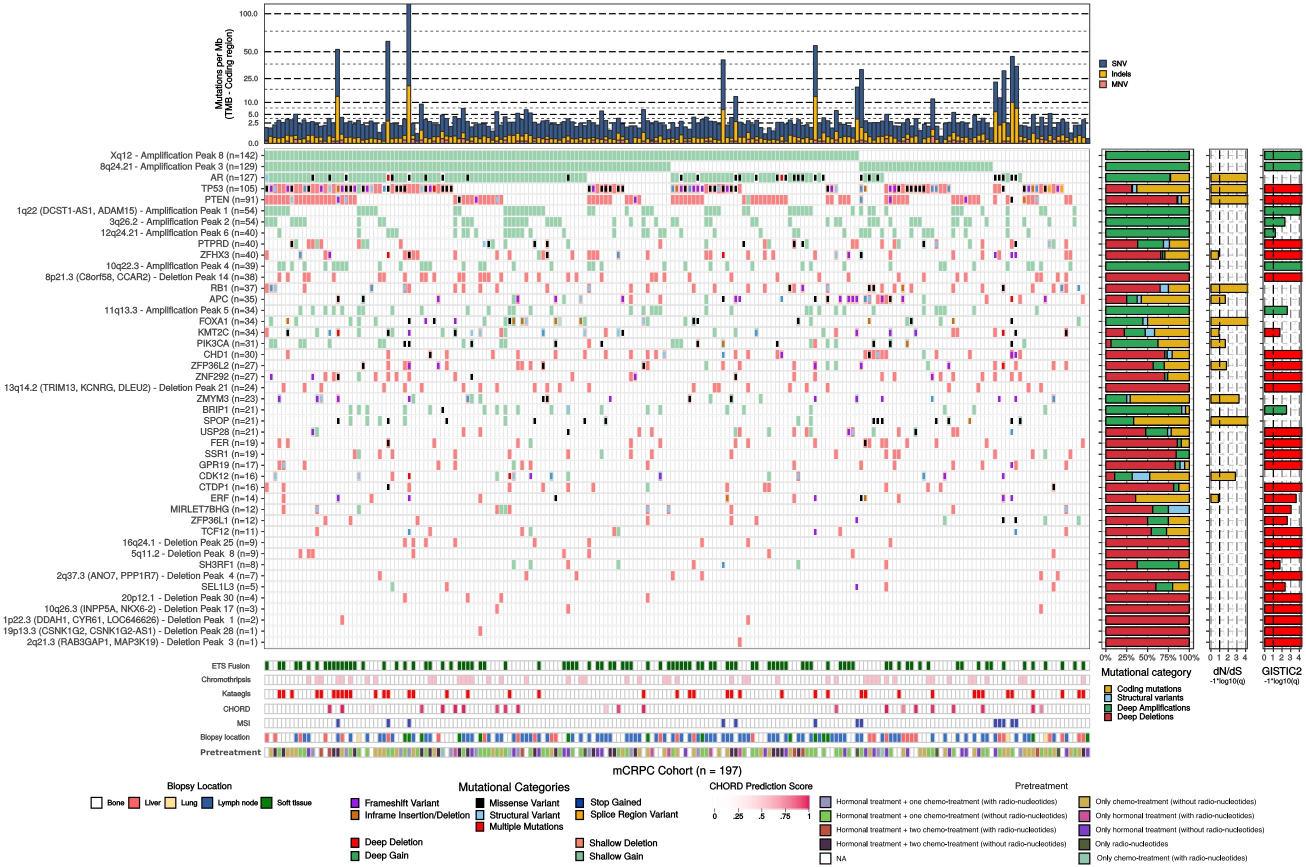


Supplementary figure S4.4: **Kataegis prevalence in mCRPC.**

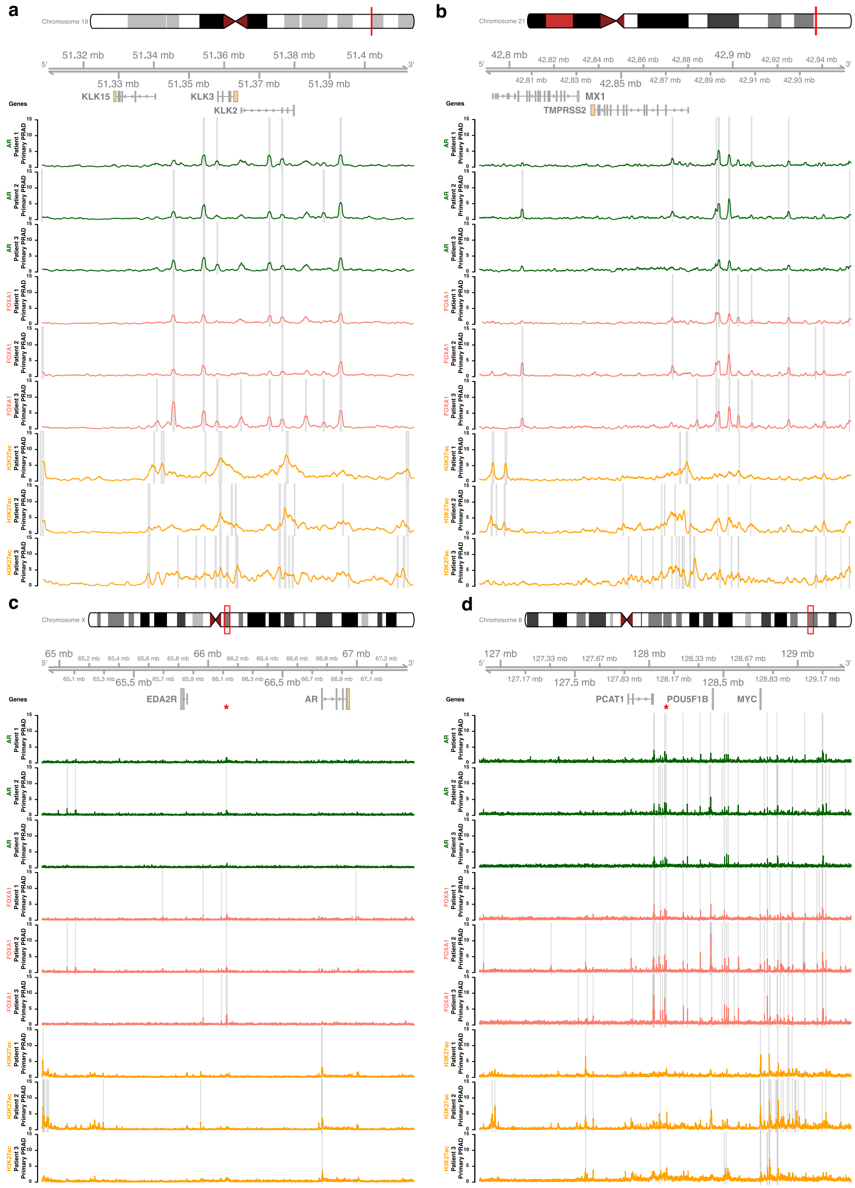
(a): Number of observed kataegis events in mCRPC cohort samples (n = 42, blue bars) and the respective genomic width of all observed kataegis foci per sample (right y-axis; black points). **(b)**: Relative frequency of mutational contexts (of SNV) found in all observed kataegis foci per sample. **(c)**: Relative frequency of SNV in observed kataegis foci in *APOBEC*-related TpCpW mutational context. W stands for T or A. **(d)**: Relative contribution to mutational signatures (COSMIC) within the kataegis foci. **(e)**: Relative contribution to mutational signatures (COSMIC) of all genome-wide events of the sample. **(f)**: Representation of two distinct kataegis foci on chromosome 8 within a single respective sample (highlighted with * in **a**). SNV (colored on Ti/Tv type) are shown with relative genomic distances (in $\log_{10}$) to neighboring SNV. Observed kataegis foci are highlighted with a transparent red background. **(g)**: Frequency and locations of cohort-wide observed kataegis foci, binned per 1 Mbp. Bins with 2 kataegis events in distinct samples are colored red, else blue. **(h)**: Absolute contribution of *APOBEC* signatures (2 & 13) in samples without (n = 155) and with observed kataegis (n = 42). Bee-swarm boxplot with notch of the mean absolute contribution of *APOBEC* signatures (2 & 13). Boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. Statistical significance was tested with Wilcoxon rank-sum test.



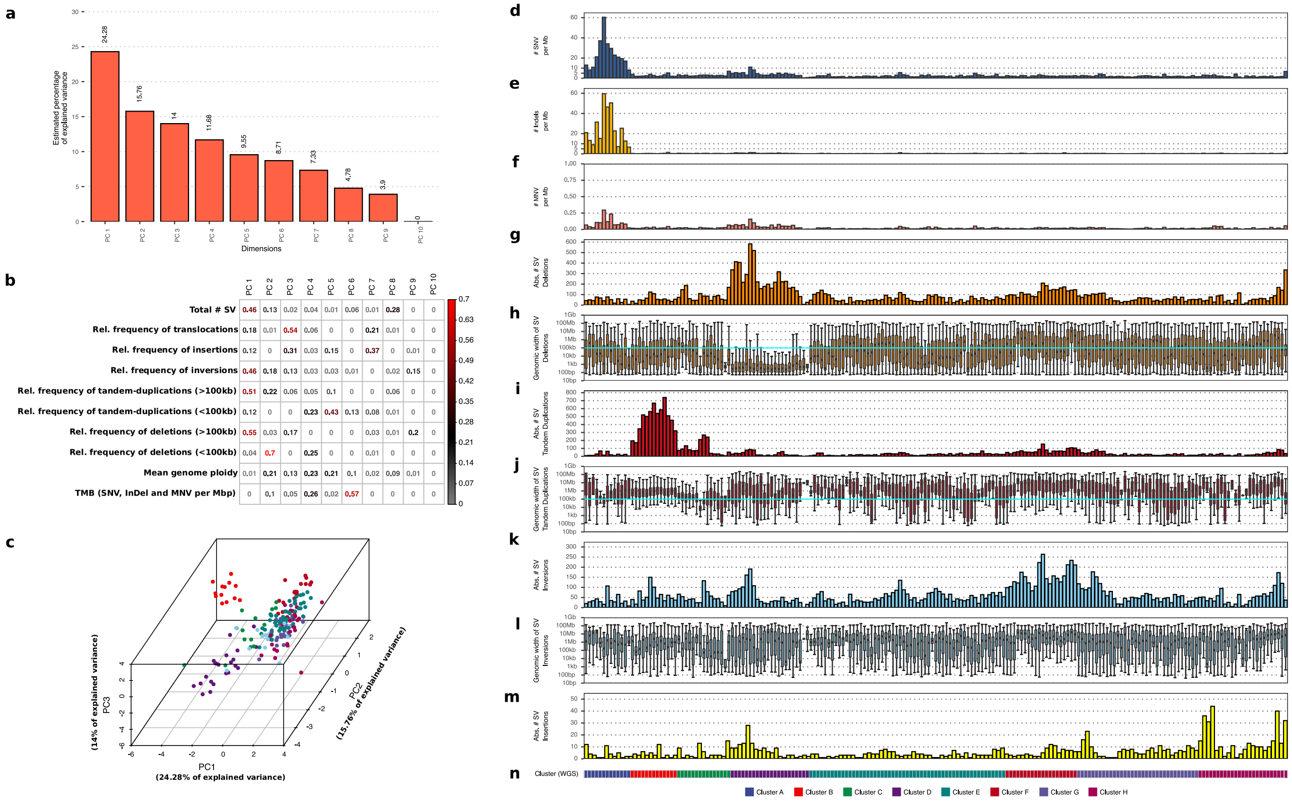
Supplementary figure S4.5: **Mutational signature analysis.**
(a): Dendrogram of unsupervised clustering (Euclidean distance; Ward.D method) on absolute contributions of SNVs in custom signatures A-E. **(b)**: Relative contribution to the five custom mutational signatures A-E. **(c)**: Relative contribution to COSMIC mutational signatures. **(d)**: Relative distribution of the 96 mutational contexts present in the custom signatures. **(e)**: Correlation (cosine similarity) of novel signatures with COSMIC signatures. The size of the dot reflects the cosine similarity, with higher cosine similarity values shown as larger dots. The color gradient indicates the level of cosine similarity. **(f)**: Quality metrics of non-negative matrix factorization (NMF) between two to fifteen ranks.



Supplementary figure S4.6: **The mutational landscape of mCRPC seems unrelated to treatment history.**
The upper track displays the number of genomic TMB) of SNV (blue), InDels (yellow) and MNV (orange) categories. The heatmap displays the type of mutation(s) per sample; (light-)green or (light-)red backgrounds depict copy number aberrations whilst the inner square depicts the type of (coding) mutation(s). Relative proportions of mutational categories (coding mutations [SNV, InDels and MNV] (yellow), SV (blue), deep amplifications [high-level amplifications resulting in many additional copies] (green) and deep deletions [high-level losses resulting in (near) homozygous losses] (red)) per gene and foci are shown in the bar plot next to the heatmap. Narrow GISTIC2 peaks covering ≤ 3 genes were reduced to gene-level rows if one of these genes is present in the dN/dS ($q \leq 0.1$) analysis or is a known oncogene or tumor-suppressor. For GISTIC2 peaks covering multiple genes, only deep amplifications and deep deletions are shown. Recurrent aberrant focal genomic foci in gene deserts are annotated with their nearest gene. Significance scores (-1*$\log_{10}(q)$) of the dN/dS and GISTIC2 analysis are shown on the outer-right bar plots; bars in the GISTIC2 significance plot are colored red if these foci were detected as a recurrent focal deletion and green if detected as a recurrent focal gain. Per sample, the presence of (predicted) ETS fusions (green), chromothripsis (light pink), kataegis (red), CHORD prediction score (HR-deficiency) (pink gradient), MSI status (dark blue), biopsy location and treatment history are shown as bottom tracks.

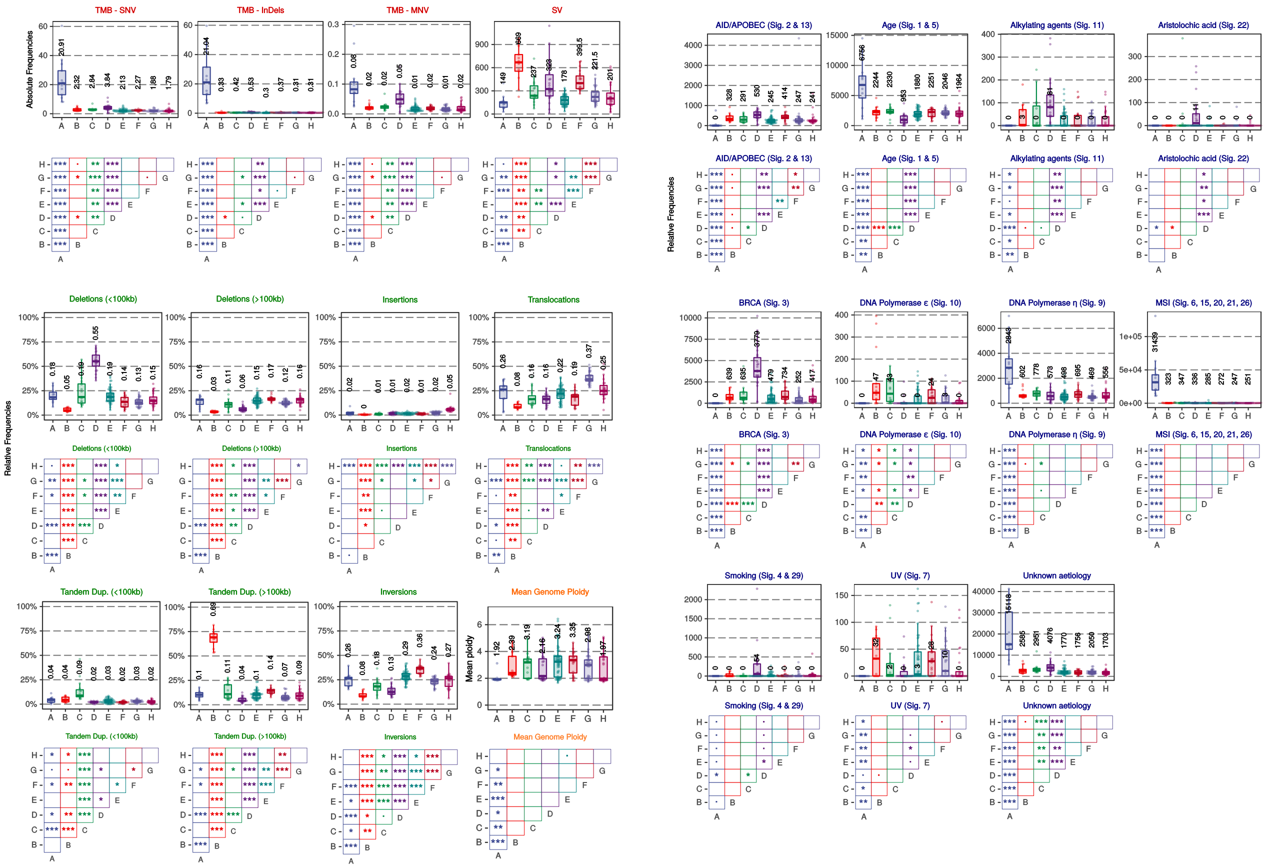


Supplementary figure S4.7: **ChIP-seq profiles in primary prostate cancer for known driver genes.**
ChIP-seq profiles from three independent primary prostate cancer patients surrounding the *AR* and *PCAT1*/*MYC* gene loci (with 1.25 additional Mbp up-/down-stream) and two known AR-regulated positive controls (*KLK3* and *TMPRSS2* with additional 0.5 Mbp up-/downstream). Per subplot, the upper panel displays the selected genomic window and the overlapping genes. The 1th to 3th tracks represent AR ChIP-seq profiles (median read-coverage per 1000bp windows) in the three primary prostate cancer patients. The 4th to 6th tracks represent FOXA1 ChIP-seq profiles (median read-coverage per 1000bp windows) in the three primary prostate cancer patients. Finally, the 7th to 9th track represent H3K27ac ChIP-seq profiles (median read-coverage per 1000bp windows) in the three primary prostate cancer patients. ChIP-seq peaks (MACS/MACS2; q < 0.01) are shown as grey transparent lines per respective sample.(a) ChIP-seq profiles surrounding the positive control *KLK3* region.(b) ChIP-seq profiles surrounding the positive control *TMPRSS2* region.(c) ChIP-seq profiles surrounding the *AR* region. The red asterisk denotes the location of the amplified region within the mCRPC setting.(d) ChIP-seq profiles surrounding the *PCAT1*/*MYC* region. The red asterisk denotes the location of the amplified region within the mCRPC setting.



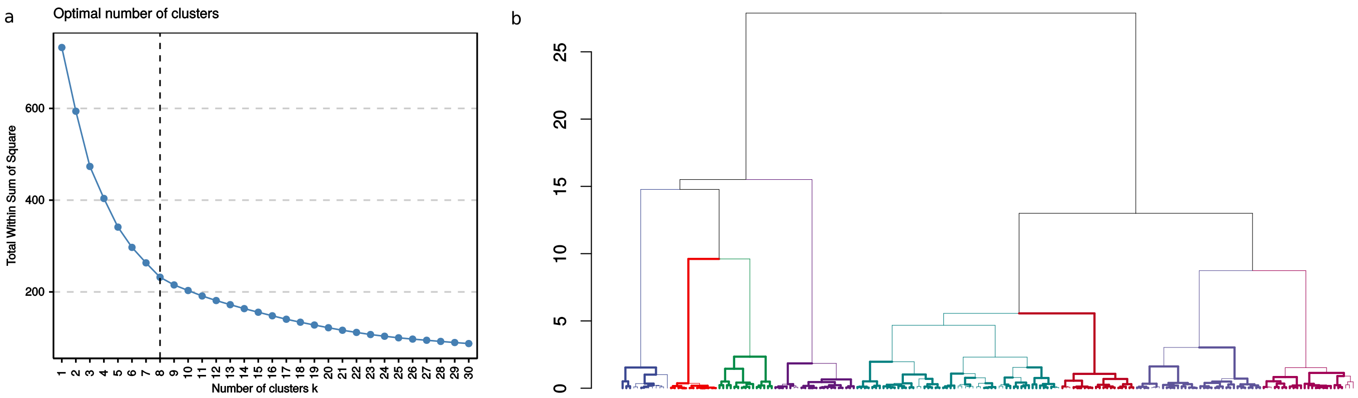
	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
Total # SV	0.46	0.13	0.02	0.04	0.01	0.06	0.01	0.28	0	0
Rel. frequency of translocations	0.18	0.01	0.54	0.06	0	0	0.21	0.01	0	0
Rel. frequency of insertions	0.12	0	0.31	0.03	0.15	0	0.37	0	0.01	0
Rel. frequency of inversions	0.46	0.18	0.13	0.03	0.03	0.01	0	0.02	0.15	0
Rel. frequency of tandem-duplications (>100kb)	0.51	0.22	0.06	0.05	0.1	0	0	0.06	0	0
Rel. frequency of tandem-duplications (<100kb)	0.12	0	0	0.23	0.43	0.13	0.08	0.01	0	0
Rel. frequency of deletions (>100kb)	0.55	0.03	0.17	0	0	0	0.03	0.01	0.2	0
Rel. frequency of deletions (<100kb)	0.04	0.7	0	0.25	0	0	0.01	0	0	0
Mean genome ploidy	0.01	0.21	0.13	0.23	0.21	0.1	0.02	0.09	0.01	0
TMB (SNV, InDel and MNV per Mbp)	0	0.1	0.05	0.26	0.02	0.57	0	0	0	0

Supplementary figure S4.8: **Rationale of the chosen genomic features for unsupervised clustering.**
Principal component analysis (PCA) and overview of the genomic features included in the unsupervised clustering analysis highlighting the chosen size cut-offs and striking differences between samples. **(a)**: Overview of the explained variance per principal component (PC) in PCA. **(b)**: The quality of representation for each feature per PC ($\cos^2$), this ranges from 0 (no importance / representation in PC) to 1 (absolute importance / representation in PC). Color gradient (0 to 0.7) denotes $\cos^2$, red values denote important / representation of feature within PC. Numbers shown are the $\cos^2$ values. **(c)**: Visualization of the first three principal components of PCA, each sample is colored based on their assigned cluster (depicted in **n**) after unsupervised clustering on their genomic features. **(d)**: All genome-wide somatic SNVs per Mbp (square root scale). **(e)**: All genome-wide somatic InDels per Mbp (square root scale). **(f)**: All genome-wide somatic MNV per Mbp (square root scale). **(g)**: Absolute number of deletions (SV) per sample. **(h)**: Distribution of the genomic width of deletions (SV) per sample. Cyan line indicates the chosen size cut-offs (< 100 kbp and ≥ 100 kbp). **(i)**: Absolute number of tandem duplications (SV) per sample. **(j)**: Distribution of the genomic width of tandem duplications (SV) per sample. Cyan line indicates the chosen size cut-offs (< 100 kbp and ≥ 100 kbp). **(k)**: Absolute number of inversions (SV) per sample. **(l)**: Distribution of the genomic width of inversions (SV) per sample. **(m)**: Absolute number of insertions (SV) per sample. Genomic width of insertions could not be estimated accurately due to repeat-like sequences. **(n)**: Assigned clusters based on unsupervised clustering of genomic features.



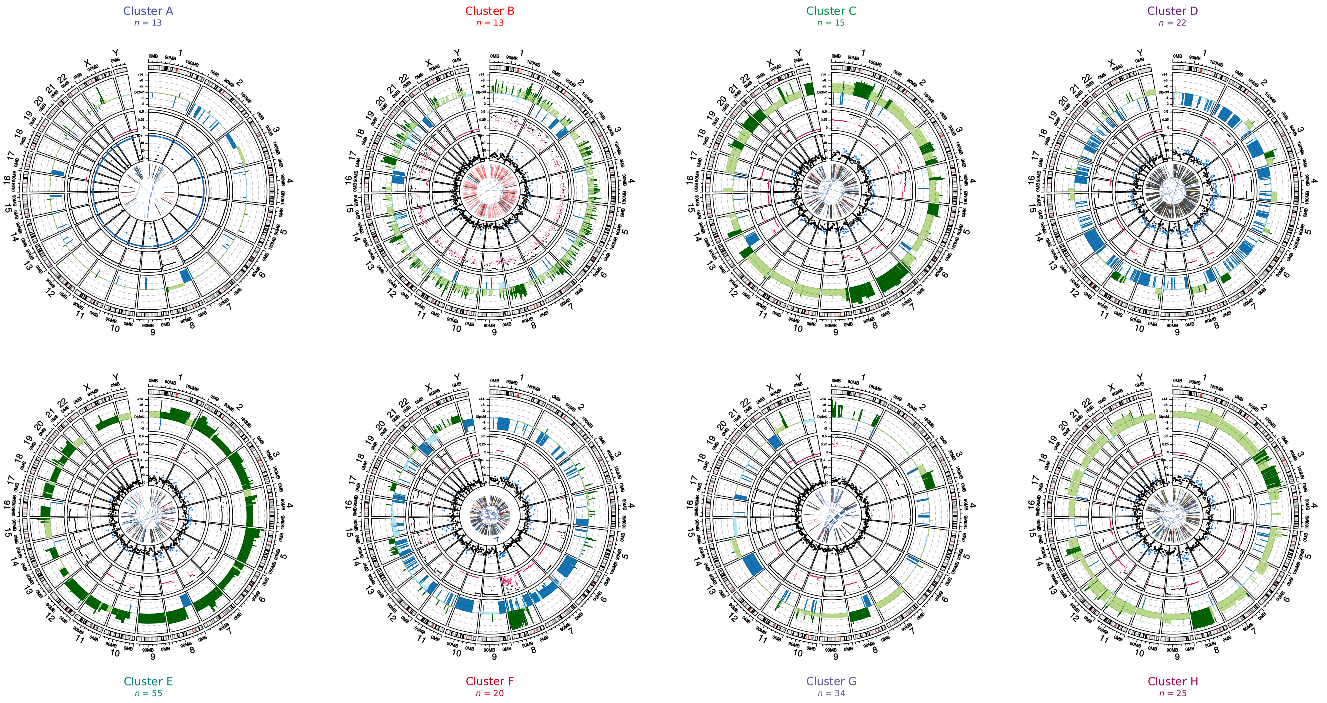
Supplementary figure S4.9: **Cluster characteristics.**

Overview of genomic characteristics and COSMIC mutational signatures per cluster (A-H) derived from unsupervised clustering of the mCRPC cohort using basic WGS characteristics. Bee-swarm boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. A pairwise Wilcoxon rank-sum test (BH correction) was performed to detect statistically significant differences between clusters; * denotes $p \leq 0.05$, ** denotes $p \leq 0.01$, *** denotes $p \leq 0.001$. Significant differences of events were also found in clusters without a clear biological association (C, E and G-H), such as increased numbers of translocations in cluster G and insertions in cluster H.

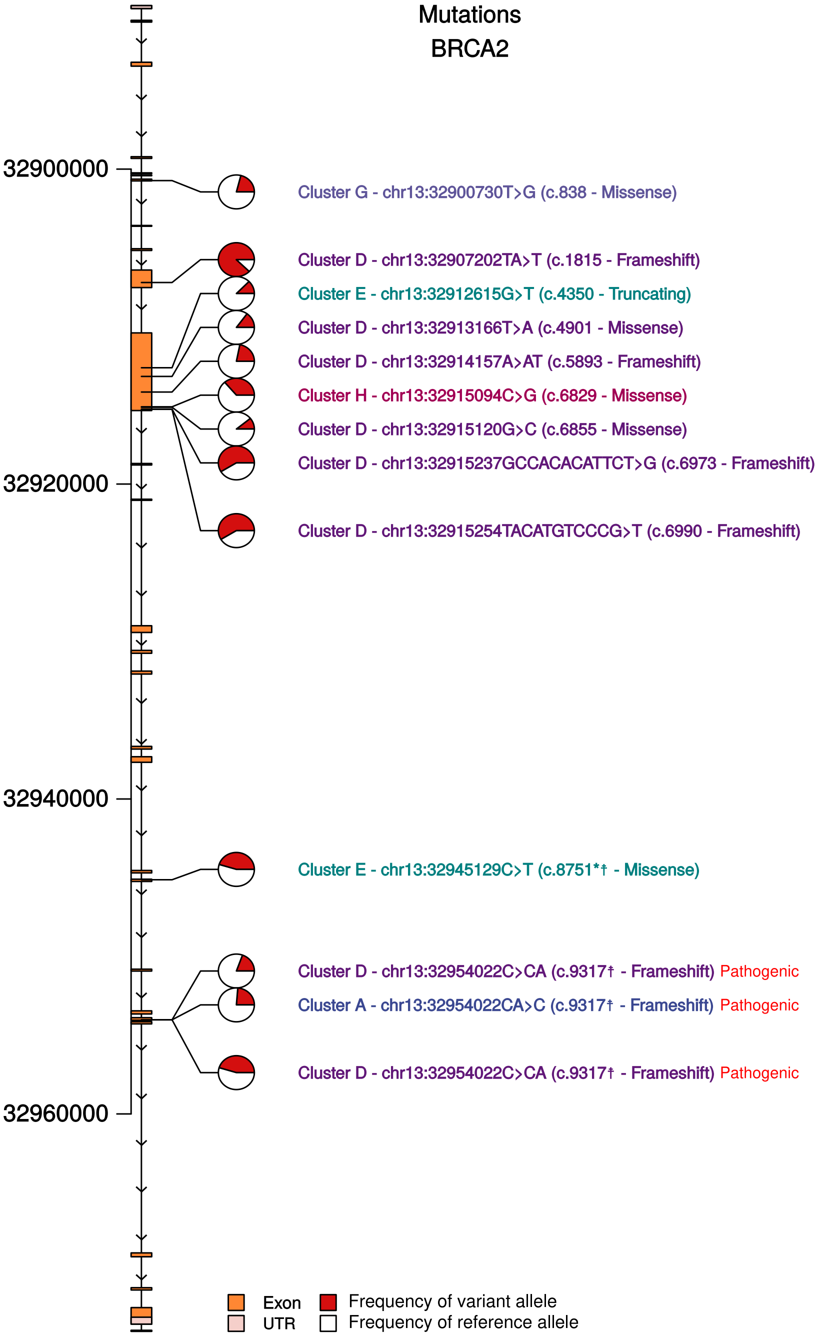


Supplementary figure S4.10: **Clustering QC.**

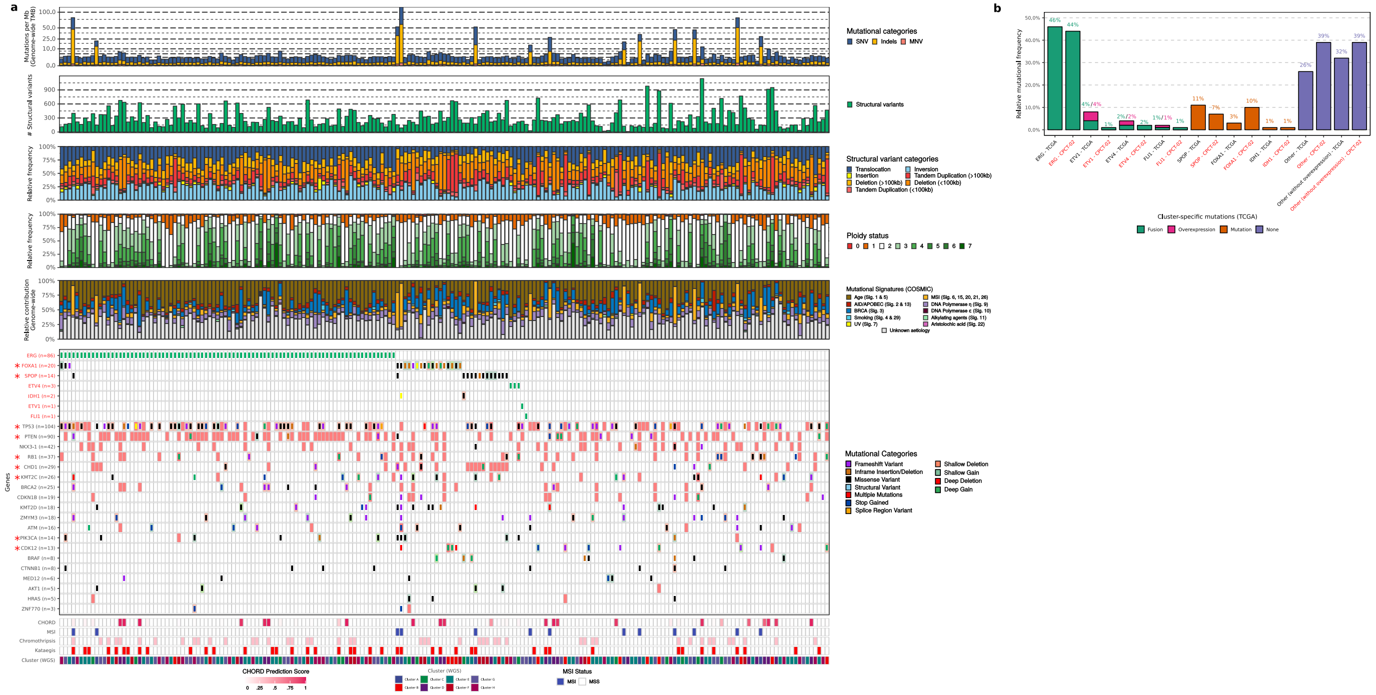
(a): Clustering estimation using optimum total within-cluster sum of square (wss). The final selection of the most optimal number of clusters was based on the knee in the blue line, e.g. the moment when increasing the number of clusters does not dramatically decrease wss. **(b)**: Bootstrapping results (5000 iterations) of the unsupervised hierarchical clustering with additional coloring of the eight defined clusters as used in this manuscript. Branches with Approximately Unbiased (AU) *p*-values ≤ 0.05 are highlighted with bolder lines and reflect significantly robust groups of samples based on similar characteristics derived from WGS. Y-axis displays clustering distance (Pearson correlation; ward.D).



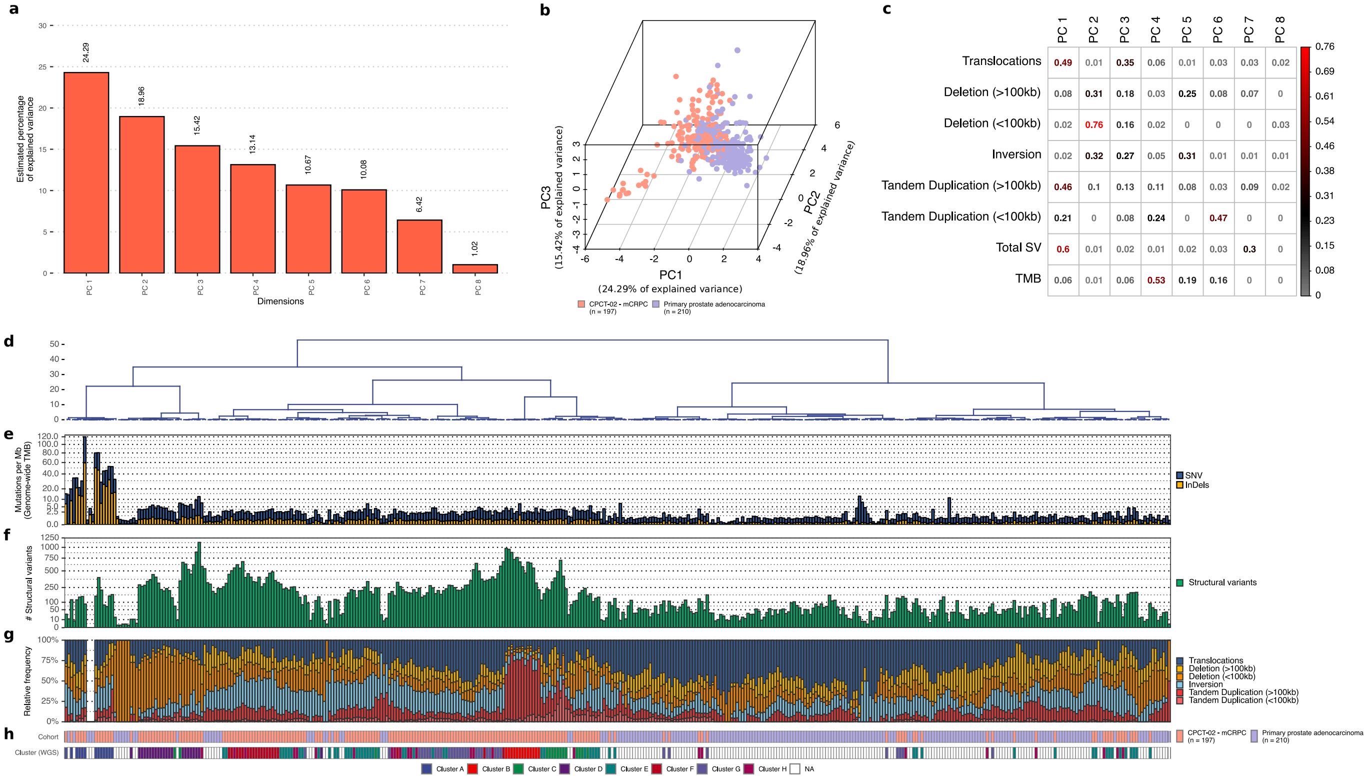
Supplementary figure S4.11: **Representative mCRPC sample per cluster.**
Genomic overviews of a single representative sample per (unsupervised) cluster. The outer track displays the genomic ideogram, the second-outer track displays copy number profiles (amplification in light green; deep amplification beyond sample-specific threshold (GISTIC2) in dark green, deletions in blue; deep deletions beyond sample-specific threshold (GISTIC2) in dark blue). The third track displays tumor cell percentage (TC)corrected lower allele-frequency (LAF) values of individual copy number segments (LAF ≤ 0.33 in pink; LAF ≥ 0.33 in black). The fourth track displays the number of mutations per 5 Mbp, ranging from 0 to 60+; bins with ≥ 20 mutations are highlighted in blue. The innermost track displays structural variants; interchromosomal translocations in dark blue, deletions in grey, insertions in yellow, inversion in light blue and tandem duplications in red.



Supplementary figure S4.12: **Overview of *BRCA2* mutations.**
Genomic distribution of non-synonymous *BRCA2* mutations found within our mCRPC cohort. Mutations are displayed as pie-charts depicting the variant allele frequency (red portion of the pie-chart) and reference allele frequency (white portion of the pie-chart). Samples are colored based on their respective cluster after unsupervised clustering (figure 4.4). Known COSMIC mutations are annotated with ** and/or known dbSNP variants with a cross. Alleles with known pathogenicity within ClinVar are highlighted, mutations without ClinVar annotation could be considered as variants with as-of-yet uncertain significance.



Supplementary figure S4.13: **Supervised clustering of mCRPC based on TCGA criteria.**
(a): Samples are sorted based on mutual-exclusivity of the same genes and aberrations as the TCGA clustering, depicted in red colors in the heatmap. In addition, all genes which the TCGA defined as recurrent alterations in primary prostate cancer are shown, genes also discovered in the mCRPC cohort as enriched in non-synonymous mutations or copy-number alterations (dN/dS and/or GISTIC2) are depicted with a red asterisk. The upper track displays the number of genomic TMB of SNV (blue), InDels (yellow) and MNV (orange) categories. Second track displays the absolute number of unique structural variants per sample. Third track displays the relative frequency per structural variant category, Tandem Duplications and Deletions are subdivided into > 100 kbp and <100 kbp categories. The fourth track displaysthe relative genome-wide ploidy status, ranging from 0 to ≥ 7 copies and the fifth track displays the relative contribution to mutational signatures (COSMIC) summarized per proposed etiology. The heatmap displays the type of mutation(s) per sample; (light-)green or (light-)red backgrounds depict copy number aberrations whilst the inner square depicts the type of (coding) mutation(s). In addition, the lower tracks display CHORD prediction score (HR-deficiency) (pink gradient), MSI status (blue), chromothripsis (pink), presence of kataegis (red) and in which of the eight genomic cluster, as defined by this manuscript, each sample falls. **(b)**: Overview of the relative frequency of samples captured per mutually-exclusive group for both the TCGA and mCRPC cohort. Promiscuous ETS family fusions (*ETV1*, *ETV4* and *FLI1*) which were captured in the TCGA cohort using mRNA overexpression were split as the mCRPC cohort did not have accompanying mRNA sequencing data to perform a similar capturing.



Supplementary figure S4.14: **Overview of clustering scheme on primary prostate cancer and mCRPC.** **(a)**: Overview of the explained variance per PC in PCA of the combined dataset of primary prostate cancer (n = 210) and mCRPC (n = 197). PCA was performed in the following features: Total number of SV, genome-wide TMB (SNV and InDels) and relative frequency ofstructural variants (except insertions). **(b)**: Visualization of the first three principal components of PCA, each sample is colored based on their respective disease-setting, primary prostate cancer is colored as violet whilst mCRPC is colored as light-red. **(c)**: The quality of representation for each feature per principal component (cos^2), this ranges from 0 (no importance / representation in PC) to 1 (absolute importance / representation in PC). Color gradient (0 to 0.7) denotes cos^2, red values denote important / representation of feature within PC. Numbers shows are the cos^2 values. **(d)**: Dendrogram of unsupervised clustering with OLO on genomic features. Y-axis displays clustering distance (Pearson correlation; ward.D). **(e)**: All genome-wide somatic SNVs (blue) and InDels (yellow) per Mbp (square root scale). **(f)**: Absolute frequency of SV per sample (square root scale). **(g)**: Relative frequency per SV category, Tandem Duplications and Deletions are subdivided into > 100 kbp and < 100 kbp categories. **(h)**: Respective cohort of the samples, primary prostate cancer is colored as violet whilst mCRPC is colored as light-red. **(i)**: Assigned clusters of mCRPC samples based on unsupervised clustering of genomic features as described in figure 4.4.

Chapter 5

The genomic landscape of 85 advanced neuroendocrine neoplasms reveals subtype-heterogeneity and potential therapeutic targets

J. van Riet[a,b,c*], H.J.G. van de Werken[a,b*], E. Cuppen[d], F. Eskens[c], M. Tesselaar[e], L.M. van Veenendaal[e], HJ. Klümpen[f], W. Dercksen[g], G.D. Valk[h], M.P.J.K. Lolkema[c,i], S. Sleijfer[c,i], B. Mostert[c]

a Cancer Computational Biology Center, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.

b Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.

c Department of Medical Oncology, Erasmus Medical Center Cancer Institute, Rotterdam, the Netherlands; Cancer Genomics Netherlands, Erasmus Medical Center Cancer Institute, Rotterdam, the Netherlands.

d Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands.

e Department of Medical Oncology, Cancer Institute, Amsterdam, The Netherlands.

f Department of Medical Oncology, Amsterdam University Medical Centers, Cancer Center Amsterdam, Amsterdam, The Netherlands.

g Department of Internal Medicine, Maxima Medisch Centrum, Veldhoven, The Netherlands.

h Department of Endocrine Oncology, University Medical Center Utrecht, Utrecht, The Netherlands.

i Center for Personalized Cancer Treatment, Rotterdam, the Netherlands.

* These authors contributed equally.

This chapter has been published in **Nature Communications** Volume 12, July 2021, Article number: #4612

Abstract

Metastatic and locally-advanced neuroendocrine neoplasms (aNEN) form clinically and genetically heterogeneous malignancies, characterized by distinct prognoses based upon primary tumor localization, functionality, grade, proliferation index and diverse outcomes to treatment. Here, we report the mutational landscape of 85 whole-genome sequenced aNEN. This landscape reveals distinct genomic subpopulations of aNEN based on primary localization and differentiation grade; we observe relatively high tumor mutational burdens (TMB) in neuroendocrine carcinoma (average 5.45 somatic mutations per megabase) with *TP53*, *KRAS*, *RB1*, *CSMD3*, *APC*, *CSMD1*, *LRATD2*, *TRRAP* and *MYC* as major drivers versus an overall low TMB in neuroendocrine tumors (1.09). Furthermore, we observe distinct drivers which are enriched in somatic aberrations in pancreatic (*MEN1*, *ATRX*, *DAXX*, *DMD* and *CREBBP*) and midgut-derived neuroendocrine tumors (*CDKN1B*). Finally, 49% of aNEN patients reveal potential therapeutic targets based upon actionable (and responsive) somatic aberrations within their genome; potentially directing improvements in aNEN treatment strategies.

Introduction

Neuroendocrine neoplasms (NEN) are a heterogeneous and uncommon tumor type. It can arise from any of the neuroendocrine cells distributed widely throughout the body. As outlaid by the International Agency for Research on Cancer and World Health Organization, a clinical distinction is made between the poorly differentiated neuroendocrine carcinomas (NEC) and the more differentiated neuroendocrine tumors (NET)[1,2], the latter are further subdivided based on their primary site in pancreas, gastro-intestinal tract or lung. Further distinctions are made based upon grade (as assessed by Ki-67 or MIB-1 staining as a measure of proliferation index), differentiation, histology (small-cell vs. large-cell) and functionality (the presence or absence of hormone secretion resulting in typical clinical syndromes dependent upon the predominant hormone that is secreted). Tumor grade and differentiation are associated with prognosis, and all the aforementioned factors affect the choice of treatment. However, also in small subgroups of Neuroendocrine neoplasms (NEN), such as well-differentiated low-proliferating pancreatic neuroendocrine tumors (pNET), marked clinical and genetic heterogeneity occur, as well as vastly different responses to treatment with only few mutant genes such as *DAXX*, *ATRX*, and *MEN1* serving as prognostic markers[3–6]. Thus, the parameters by which NEN are currently classified do not sufficiently separate patients and tumors according to prognosis and response to therapy. Nonetheless, certain antitumor therapies (i.e., sunitinib and everolimus) have been registered for distinct NEN-subtypes. Hence, there is a high unmet need to better classify and understand these diverse tumors, ultimately leading to more tumor- or patient-tailored therapeutic strategies.

Thus far, limited whole-exome sequencing (WES) and whole-genome sequencing (WGS) data are available for NEN, probably reflecting the rarity of this disease. Currently, pNET have been characterized most extensively; 81 primary tumors were subjected to WGS as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project[7] and another set of primary pNET ($n = 102$) was described by Scarpa *et al.*[5]. In addition, smaller series using diverse sequencing approaches of varying resolution on primary NET subtypes have been published; which include genomic studies on pNET (WES and targeted sequencing; $n = 10$ and 58, respectively)[4], DNA methylation and RNA-sequencing of pNET ($n = 32$ and $n = 33$, respectively)[3], well-differentiated carcinoid (SNP-array; $n = 29$)[8], NEC (targeted sequencing; $n = 63$)[9] and two studies on a multi-institution cohort of small intestine NET (SI-NET) using combined approaches of targeted sequencing ($n = 81$), WES ($n = 48$; $n = 29$) and WGS ($n = 15$)[10,11]. These studies have shown that NET have a relatively stable genome and only few commonly observed driver mutations and allelic imbalances, often associated with their primary tissue of origin. Previously associated genetic drivers of NET include the cell-cycle regulator *CDKN1B* in SI-NET[10–13], chromatin-remodeling genes (*DAXX*, *ATRX*, *MEN1*, and *SETD2*), DNA-repair genes (*CHEK2*, *BRCA2*, and *MUTYH*), mTOR-related genes (*TSC2*, *PTEN*, and *PIK3CA*) and the oxygen-sensing modulator *VHL*[14] in pNET[3–5,7,15] whilst NEC is associated with aberrations in *TP53*, *RB1*, *MYC*, *CCND1*, *KRAS*, *PIK3CA*/*PTEN* and *BRAF*[9,16–18]. However, these studies were all performed on primary tumor specimens, whilst a patient generally dies from the consequences of metastatic disease. In addition, we know from other tumor types that marked heterogeneity can occur between primary and metastatic tumor cells[19–22], due to inherent genomic instability and/or the influence of targeted or cytotoxic treatment on the tumor genome. These discrepancies should be taken into account when assessing a patient's prognosis and possible treatment options, and can be better understood through thorough genomic characterization of metastases. To date, analysis of metastatic NET is limited to two studies describing series of five patients with NET originating in the pancreas and the small intestine (or midgut), respectively[23,24]. These studies have shown focal amplification of *MYCN* concomitant with loss of *APC* and *TP53* in one sample as important metastatic genetic aberrations. For NECs, only two series of WGS of the primary tumors of (1) five cervical and (2) 12 genitourinary NECs have been published[25,26].

In this work, WGS was performed on 85 biopsies from patients with locally advanced or metastatic (advanced) neuroendocrine neoplasm (aNEN); a single biopsy per patient was selected for analysis. The vast majority of these biopsies are taken from metastatic lesions ($n = 70$ out of 85) whilst for 15 patients suffering from metastatic or incurable locally advanced disease, their treating physician judged a biopsy of a metastasis too high-risk or not feasible, and instead had a biopsy taken from their primary lesion at the time of locally advanced or metastatic disease. All aNEN patients underwent these biopsies as part of their participation in the Dutch CPCT-02 and DRUP studies[27,28]. We report on the presence of genomic alterations, mutational and rearrangement signatures for the whole aNEN cohort and reveal genomic characteristics and alterations distinguishing aNEC from aNET. Furthermore, we make a genomic distinction between pancreas- and midgut-derived aNET. In addition, we investigate the presence of actionable genetic alteration within aNEN patients, which might render them eligible for off-label or experimental systemic treatments to extend therapy options.

Results

Overview of included patients within the CPCT-02 aNEN cohort and whole-genome sequencing

A total of 108 patients, originally classified as having a neuroendocrine neoplasm, were included in the CPCT-02 and DRUP studies and had a primary or metastatic tumor biopsy taken in parallel with a blood control (Figure 5.1). Five patients were excluded because of missing or withdrawn informed consent, and another five had non-evaluable biopsies due to low (<20%) tumor cell percentage or low DNA yield. Thirteen biopsies were excluded because of incomplete clinical records, misclassifications of the tumor (based on additional checks of the medical records), or were duplicate biopsies from the same patient. An overview of the aNEN patient inclusion per participating Dutch center (*n* = 13) can be found in Supplementary Figure S5.1a.

The aNEN cohort is represented by 37 females and 48 males with a median age of 62 (Q(uartile)1-Q3: 57-68) and 61 (Q1-Q3: 56-68) years, at time of biopsy, respectively (Figure 5.1c). In total, 69 NET and 16 NEC were included. The primary tumor location in the midgut was most common (*n* = 41, 48%), followed by pancreas (*n* = 23, 27%) and unknown (*n* = 12, 14%) (Figure 5.1b). Most of the tumor biopsies were taken from liver metastases, and a minority from relapses at the primary site (Figure 5.1d).

To gain more in-depth knowledge of the pathological information of this cohort, we requested pathological reports of primary tumor and/or metastatic tumor tissue as available in the nationwide (Dutch) PALGA registry. Of note, these tissues were often not acquired at the time of biopsy for the CPCT-02 study. For the majority of patients, pathology reports on metastases and/or the primary tumor were available (**Online Suppl. Data 1**). In the minority, the pathological record of a previous primary biopsy or resection specimen was assessed.

We also characterized our cohort with regard to previously administered systemic anti-tumor treatment. Sixty-nine percent of patients had not undergone any previous anti-tumor treatment, 31% had undergone a large variety of previous

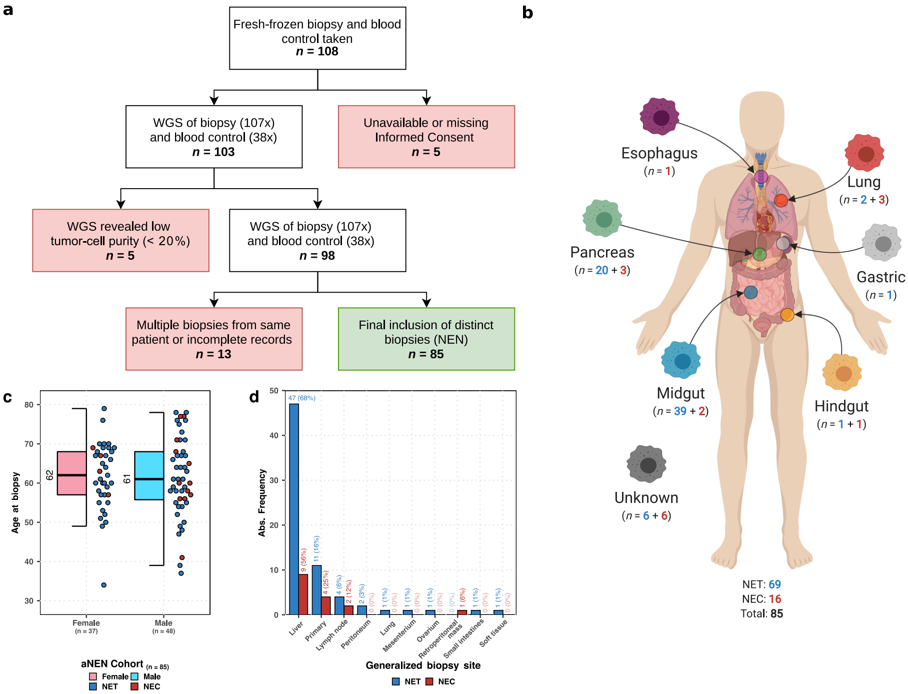


Figure 5.1: **Overview of patient inclusion and subclassification of biopsies.** **(a):** Flowchart of patient inclusion. From the CPCT-02 cohort, single biopsies from 85 distinct patients with aNEN were selected. From the total pool of available whole-genome sequenced aNET samples. If multiple derived aNET biopsies from the same patient were available, we selected the aNET biopsy with the highest tumor cell purity. The tumor and matching blood sample (reference) were whole-genome sequenced to a median read coverage of 107x and 38x (paired-end) reads per base, respectively. Filtering criteria in which patients were excluded are highlighted in red. The final inclusion of aNEN patients is depicted in green. **(b):** Subclassification of aNEN based on primary localization. The 85 aNEN were subclassified, based on their primary localization, into six major categories; gastric, hindgut, lung, esophagus, pancreas, and midgut; whilst samples with indeterminable localization were categorized as unknown. The number of aNET (in blue) and aNEC (in red) are shown per category. **(c):** Age distribution stratified by gender of the aNEN cohort. Observed median per variable displayed in a boxplot with individual data points (aNET and aNEC are depicted as blue and red points respectively). The median, interquartile range (IQR), and 1.5× the IQR are represented by a solid black line, box, and whiskers, respectively. **(d):** Barplot of generalized location of the tumor biopsy. Absolute and relative (in brackets) frequency of aNET (blue) and aNEC (red) biopsy locations.

treatments, mainly consisting of somatostatin analogs, radioisotopes, chemotherapy, and targeted therapy (Figure S5.1d).

The tumor biopsies and corresponding peripheral blood controls from the 85 distinct patients were whole-genome sequenced using paired-end protocols, to a median mean read coverage of 107× (Q(uartile)1-Q3: 99×-116×) and 38x (Q1-Q3: 35-42×), respectively to a median in silico estimated tumor cell purity of 0.7 (Q1-Q3: 0.5-0.82).

The mutational landscape of advanced neuroendocrine neoplasms reveals differences related to primary localization and degree of differentiation

The overall mutational landscape of aNEN (*n* = 85; Figure 5.2) reveals two strikingly distinct genomic populations of neuroendocrine neoplasms, i.e., the aNEC and aNET populations. The aNEC (*n* = 16) reveals diploid to triploid genomes and a median tumor mutational burden (TMB) of 5.45 somatic mutations per Mb (Q1-Q3: 3.84-8.85), which is in the mid-range of TMB known for human primary cancers[29]. However, the aNET (*n* = 69) are hallmarked by a relatively stable diploid tumor genome with only few, but specific, chromosomal arm aberrations and harbors the lowest overall TMB of only 1.09 (Q1-Q3: 0.79-1.52) of all metastatic cohorts within the CPCT-02 study[27].

The somatically acquired and whole-genomic mutational landscape of aNEC (*n* = 16) revealed a median of 13,996 single-nucleotide variants (SNVs; Q1-Q3: 9465-22,830), 1.756 small insertions and deletions (InDels; Q1-Q3: 752-2,245), 114 multiple-nucleotide variants (MNVs; Q1-Q3: 49-198), 150 structural variants (SVs; Q1-Q3: 82-264) and an overall diploid to triploid genome (Q1-Q3: 1.9-3.1; Supplementary Figure S5.2). Concordant with the lower TMB of the aNET (n = 69), the aNET revealed a median of 2870 SNVs (Q1-Q3: 1995-3904), 254 InDels (Q1-Q3: 185-325), 19 MNVs (Q1-Q3: 12-27), 17 SVs (Q1-Q3: 7-53) and an overall diploid genome (Q1-Q3: 1.9-2.19). The discrepancy in mutational load between aNEC and aNET also held true when inspecting only the coding regions, in which aNEC revealed a higher number of SNVs, InDels, MNV compared to aNET (Supplementary Figure S5.2a). Similarly, aNEC displayed elevated numbers of all SV classes (translocations, deletions, tandem duplications, insertions and inversions; Supplementary Figure S5.2d).

The majority of somatic coding mutations for all aNEC and all aNET (*n* = 3333

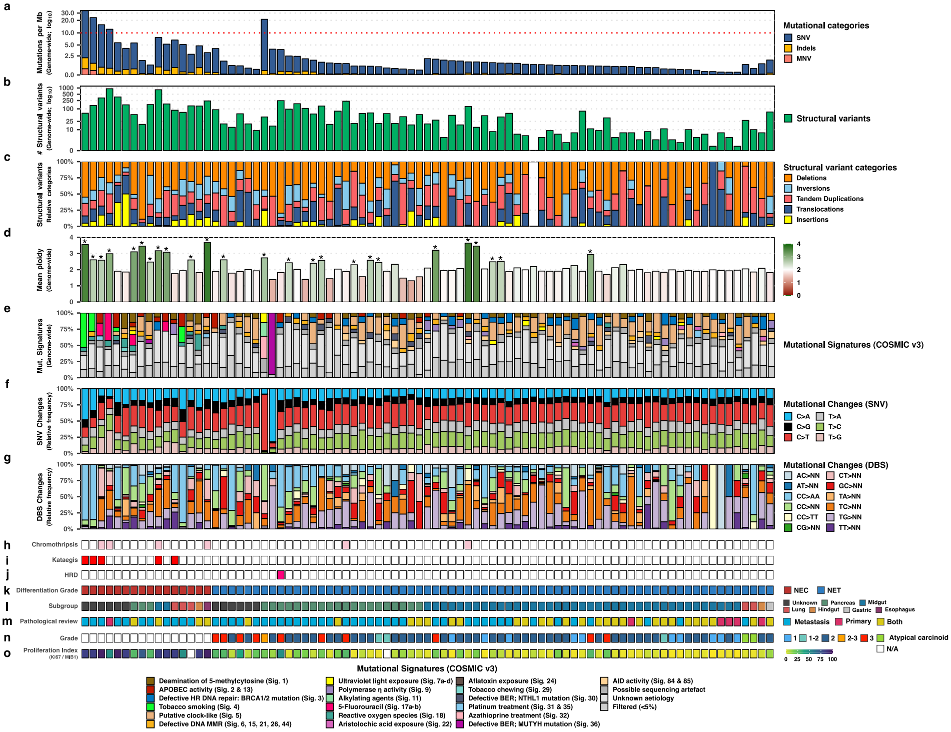


Figure 5.2: **Landscape of large-scale genomic alterations detected in aNEN, ordered by differentiation grade (NEC/NET) and primary localization.** Overview of genome-wide characteristics of the aNEN cohort ordered by aNEC/aNET and primary localization on decreasing median tumor mutational burden. For each aNEN ($n = 85$), the following tracks are shown: **(a):** Number of genomic mutations per megabase over the entire genome (TMB). Threshold for high TMB (≥10) is shown by a horizontal red dotted line. Y-axis is shown in $\log_{10}$-scale. **(b):** Total number of structural variants (green) including deletions, tandem duplications, translocations, inversions, and insertions as detected by GRIDSS. Y-axis is shown in $\log_{10}$-scale. **(c):** Relative frequency of each of the structural variant categories; deletions in orange, tandem duplications in red, translocations in blue, inversions in light-blue, and insertions in yellow. **(d):** Mean genome-wide ploidy, ranging from 0 (red) to 4 (green; tetraploid). Common diploid status is shown in white. Suspected whole-genome duplication (WGD) events have been marked by an asterisk (*). **(e-g):** Relative contribution of the COSMIC SBS (**e**), SNV mutational changes (**f**) and DBS (**g**) mutational signatures (v3; $n = 67$). Proposed etiology of signatures is denoted below. **(h-j):** Presence of chromothripsis, kataegis and homologous recombination deficiency (HRD).; aNEN with chromothripsis are shown in pink. **(k-o):** Molecular and pathological characteristics per aNEN sample.

and 3663; SNV, InDel, and MNV) were found to be predicted missense variants (52% in aNEC vs. 52% in aNET), followed by synonymous variants (18% vs. 21%). The number of genes harboring somatic mutations within their coding regions differed between aNEC and aNET. Over the entire aNEC cohort ($n = 16$), 2845 distinct mutant genes were observed, versus 3112 distinct genes within the entire aNET cohort ($n = 69$). Per sample, a median of 150 (Q1-Q3: 127-270) versus 37 (median; Q1-Q3: 26-51) genes harboring mutations within coding regions were observed for aNEC and aNET samples, respectively; revealing that aNEC harbor greater numbers of mutant genes compared to aNET.

The median genome-wide ratio of transitions (Ti; A ⇔ G or T ⇔ C) to transversions (Tv; C ⇔ A, C ⇔ G, T ⇔ A or T ⇔ G) within aNEC was found to be 0.78 Ti \TV (Q1-Q3: 0.72-1.02) vs. 1.52 Ti \Tv (Q1-Q3: 1.12-2.20) in the coding regions. For aNET the median genome-wide and coding Ti \TV were found to be 1.09 (Q1-Q3: 0.98-1.32) and 1.42 (Q1-Q3: 1-1.96), respectively (Supplementary Figure S5.2f).

High-TMB (≥10) are often associated with DNA-repair deficiency and/or tumors with sensitivity for immune therapy, e.g., checkpoint inhibitors. Four aNEC samples, all from unknown origin, and a single pancreatic aNET showed this high-TMB genotype (Figure 5.2a). One aNET displayed signs of *BRCA2*-associated HRD, as determined using the CHORD classifier which is mainly based on deletions with flanking microhomology and 1100 kb structural duplications (Figure 5.2j; Supplementary Figure S5.3). Further inspection revealed that this aNET harbored a somatic frameshift mutation within *RAD51C*, a known HRD-associated gene[30–33].

Regional hypermutation (kataegis)

Regional hypermutation (kataegis) was detected in five aNEC; Figure 5.2i; Supplementary Figure S5.4). Canonically, kataegis is associated with APOBEC activity and indeed, four out of five (80%) of these kataegis events predominantly showed the canonical TpCpW context associated with APOBEC alteration[34]. In addition, in the five samples harboring kataegis, the absolute contribution of APOBEC single-base substitution (SBS) mutational signatures (2 & 13) was significantly higher (median 45 vs. 533; $p < 0.01$; Wilcoxon rank-sum test) compared to aNEN without kataegis ($n = 80$).

Chromothripsis

Multiple distinct aNEN (four aNEC and two aNET; 7%) revealed the presence of chromothripsis, a catastrophic phenomenon of the shattering and interchromosomal recombination of one or more chromosomes (Figure 5.2h; Supplementary Figure S5.5). Strikingly, four of the six observed chromothripsis events from distinct aNEN (two aNEC and two aNET) involved the same chromosome, namely chromosome 12. Within these four aNEN, we observed possible evidence for extrachromosomal deoxyribonucleic acid (DNA) due to copy-number oscillations between one low ($\text{CN} \leq 4$) and one very high ($\text{CN} \geq 10$) states, consistent with the presence of double minutes[35–37].

Catalog of the cohort-wide mutational signatures provide biological insights into treatment effect

Different mutational processes, such as exposure to exogenous or endogenous mutagens and defective DNA-repair mechanisms generate unique combinations of mutational trinucleotide contexts which are reflected in mutational signatures[38,39]. To determine these mutational signatures within aNEN, we performed *de novo* mutational signature analysis and determined the contribution of previously described SBS mutational signatures (COSMIC v3). The *de novo* mutational signature assessment revealed seven signatures, denoted as Sig. A to Sig. G, (Supplementary Figure S5.6b, h, i) which all strongly correlated to previously known mutational signatures (Supplementary Figure S5.6a-f). In particular, we observed samples with large relative contributions (>20%) of *de novo* signatures similar to the known signatures associated with aging (SBS1 & 5; Sig A and D), APOBEC activity (SBS2 & 13; Sig B.), tobacco smoking (SBS4; Sig F.), alkylating agents exposure (SBS11; Sig E.), 5-Fluorouracil exposure (SBS17a-b; Sig. C.) and *MUTYH* mutations (SBS36; Sig. G.).

Overall, the mutational signature profiles do not differ greatly within the aNEN cohort. SBS5 ($n = 48$; putative clock-like), SBS8 ($n = 45$; possibly late-replication errors[40]), SBS40 ($n = 22$; Unknown), SBS3 ($n = 16$; HRD-like), SBS1 ($n = 10$; clock-like), SBS39 ($n = 7$; Unknown), and SBS9 ($n = 5$; polymerase η (POLH) activity) were classified as dominant signatures (i.e., contributed at least 10% of total contribution within ≥5 aNEN; Figure 5.2e). When comparing between our major subgroups (aNEC, midgut- and pancreas-derived aNET), we observed significant

($q \leq 0.05$) differences for five previously described SBS mutational signatures (Supplementary Figure S5.6g). The relative contribution of SBS3 (HRD-like) and SBS5 (clock-like) was lower in aNEC compared to midgut- and/or pancreas-derived aNET whilst conversely, SBS18 (reactive oxygen species) was elevated in aNEC. In addition, SBS8 (possibly late-replication errors) was elevated in midgut-derived aNET compared to the others. Finally, the relative presence SBS40 (unknown) was higher in pancreas-derived aNET compared to others.

Two included aNEC of unknown primary localization are characterized by high-TMB (≥10) and SBS4, which is associated with smoking; likely due to tobacco mutagens. This could reflect that these metastases could be primary lung non-small-cell lung cancer. However, as no somatic coding mutations in canonical lung cancer-associated genes were observed and the clinicopathological data of these patients did not point to any different primary tumor other than a NEC, it seems unlikely that these could be primary non-small-cell lung cancers. Smoking has also been implicated as a risk factor for pulmonary and extrapulmonary NEC such as those of the urinary bladder and the esophagus[41].

Strikingly, the only high-TMB (pancreatic) NET was strongly characterized by SBS11, which exhibits a mutational pattern resembling that of alkylating agents, with a strong enrichment for C/T (G > A) transitions. Previously, an association between treatment with the alkylating agent temozolomide and SBS11 mutations has been found[38,42]. This same patient showed the highest TMB with a TMB of 21.3 (median TMB of NET: 1.09) and was treated with a combination of 5-fluorouracil and streptozocin before undergoing a biopsy for the CPCT-02 study. Streptozocin is a capable of DNA alkylation and inhibition of DNA synthesis, and its mechanism of action closely resembles that of temozolomide.

One aNET was strongly characterized by SBS36, associated with base excision repair (BER) deficiency due to MUTYH alterations, C > A mutations and previously also seen in pancreatic NET[42–44]. Strikingly, this tumor did not harbor specific somatic alterations within *MUTYH* but possessed a heterozygous germline pathogenic missense mutation within *MUTYH* (c.527A>G / p.Tyr176Cys; rs34612342) coupled with a complete loss of a single chromosome 1, resulting in subsequent loss of heterozygosity.

Driver catalog of aNEN

We next performed an unbiased driver gene discovery analysis by performing GISTIC2[45] to detect recurrent somatic copy-number alterations and dN/dS[46] to detect genes under positive (or negative) selection pressure on the entire aNEN cohort and separately on all aNET and aNEC samples. With this analysis, we detected eighteen focal deletion peaks and two focal copy-number amplifications peaks throughout the genome ($q \leq 0.1$) and ten genes enriched with non-synonymous mutations ($q \leq 0.1$; Fig. 3 and Supplementary Figure S5.7). Within these focal peaks, several oncogenes and tumor suppressors were present which could be the potential target of the copy-number alteration. These genes, which have been previously associated as driver genes in NET and/or pan-cancer cohorts[5,11,27] are shown in Figure 5.3 for all aNEN with a distinction between aNEC and aNET. We detected several previously known tumor suppressors and oncogenes such as *TP53*, *KRAS*, *MEN1*, *RB1*, *CDKN1B*, *DAXX*, and *APC* enriched with non-synonymous mutations ($q \leq 0.05$) as well three additional genes (*LPCAT2*, *SETD2*, and *CREBBP*) just above the statistical threshold value ($q \leq 0.1$). By overlapping known drivers within the observed focal amplification and deletion peaks, we detected a plethora of putative drivers with copy-number alteration; such as deletions of *TP53*, *CDKN2A*, *CDKN2B*, *CDKN1B*, *PTPRD*, *DR1*, *CBFA2T3*, *PLCG2*, *ANKDR11*, *IRF8*, *LINC01881*, *PRKN*, *ZNF407*, common fragile sites such as *DMD*, *FHIT* and *MACROD2*, and amplifications of genes such as *PCAT1/MYC* and *MDM2*. Furthermore, focal deletions of additional genes such as *CAMTA1*, *DLUE1/2*, *TRIM13*, *KCNRG*, *FXD1* were found in ≤2 samples (**Online Suppl. Data 1**). Large perturbations on chromosome 12q15 (*MDM2*) were observed within aNEN harboring chromothripsis (Supplementary Figure S5.5). Furthermore, we could detect a single in-frame fusion of the common fusion-partner *EWSR1* seen in *pNET*[5]. Moreover, we observed only two genes harboring hotspot coding mutations (on base-level) which were shared between three samples (*ZNF829* and *KRAS*) and seven genes between two samples (*UHRF1BP1L*, *CDKN1B*, *MEN1*, *LEKR1*, *OR5L1*, *CTNNB1*, and *GNAS*; **Online Suppl. Data 1**).

We observed an overall heterogeneous pattern of putative drivers, the most frequently putative driver was found to be *CDKN2A/B* ($n = 17$; 14), followed by *TP53* ($n = 17$), *CDKN1B* ($n = 11$), *PTPRD* ($n = 11$), *KRAS* ($n = 11$), *MEN1* ($n = 11$) and *RB1* ($n = 11$). Strikingly, a significant portion of the total aNEN cohort had no mutual putative driver(s) (9 out of 85; 11%) and only contained patient-specific putative drivers.

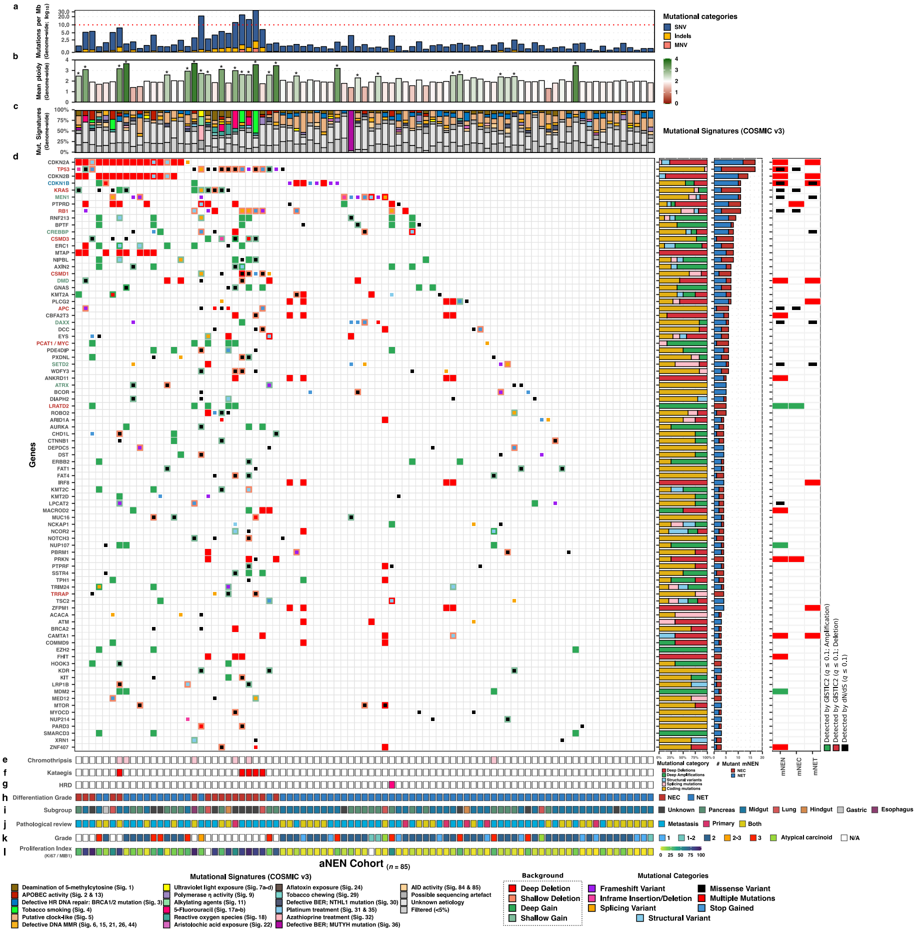


Figure 5.3: **Putative drivers and NEN-associated genes within the aNEN cohort as detected by unbiased discovery (dN/dS, GISTIC2) and literature.**
Overview of putative drivers harboring coding mutations within ≥3 aNEN. We show putative drivers as detected by dN/dS and/or GISTIC2 and supplemented this list with additional NEN-associated drivers. aNEN and genes are sorted based on mutually exclusivity of the depicted putative drivers. In addition, genes found to be mutually exclusive between our major subgroups are highlighted in the respective color of the enriched subgroup; Supplementary Figure S5.8e). This overview depicts the genomic features and the somatic inventory for the entire aNEN cohort ($n = 85$). **(a):** Number of genomic mutations per megabase over the entire genome (TMB). Threshold for high TMB (≥10) is shown by a horizontal red dotted line. Y-axis is shown in $\log_{10}$-scale. **(b):** Mean genome-wide ploidy, ranging from 0 (red) to 4 (green; tetraploid). Common diploid status is shown in white. Suspected WGD events have been marked by an asterisk (*). **(c):** Relative contribution of the COSMIC single-base substitution mutational signatures (v3; $n = 67$). Proposed etiology of signatures is denoted below. **(d):** Overview of coding mutation(s) per aNEN, (light-)green or (light-)red backgrounds depict copy-number aberrations whilst the inner square depicts the type of (coding) mutation(s). The adjacent bar plots represent the relative proportions of mutational categories per gene, the percentage of aNEC (in red) and aNET in blue harboring mutation and the dN/dS and/or GISTIC2 support, per analysis. **(e-g):** Relative contribution of the COSMIC SBS (**e**), SNV mutational changes (**f**) and DBS (**g**) mutational signatures (v3; $n = 67$). Proposed etiology of signatures is denoted below. **(h-j):** Presence of chromothripsis, kataegis and HRD.; aNEN with chromothripsis are shown in pink. **(k-o):** Molecular and pathological characteristics per aNEN sample.

We next investigated whether any form of mutational enrichment, such as somatic alterations within certain genes (mutations and/or copy-number alterations) or evidence of large-scale events (kataegis and chromothripsis), could be related to one of our three major subgroups relating to subtype or primary localization; being aNEC ($n = 16$), pancreas- ($n = 20$), and midgut-derived aNET ($n = 39$). Using a one-sided Fisher's exact test (with Benjamini-Hochberg correction) on relevant genes ($n = 20$) captured within either within our dN/dS and GISTIC2 analysis or present as mutant genes with either mutations or copy-number alterations within 20% of each major subgroup, we detected the enrichment of at least one such event(s) within these subgroups (Supplementary Figure S5.8e).

Within aNEC, an enrichment of alterations within *TP53* (88% of aNEC), *KRAS* (50%), *RB1* (50%), *CSMD3* (44%), *APC* (31%), *CSMD1* (31%), *LRATD2* (31%), *PCAT1/MYC* (31%), *TRRAP* (25%), and presence of kataegis (31%) and chromothripsis (25%) could be appreciated ($q \leq 0.05$). Likewise, within pancreas-derived aNET, an enrichment was seen for *MEN1* (40% of pancreas-derived aNET), *DAXX* (25%), *DMD* (25%), *SETD2* (25%), *ATRX* (20%) and *CREBBP* (20%) whilst midgut-derived aNET revealed enrichment of *CDKN1B* alterations (23% of midgut-derived aNET).

Genomic differences relating to primary localization of aNET

Due to distinct prognosis and previous genetic associations, we investigated genome-wide differences in regards to primary localization within the aNET population ($n = 69$). We observed several genome-wide differences relating to primary localization (Figure 5.2, Supplementary Figure S5.8), such as the median genome-wide TMB; ranging from 1.05 (aNET-Midgut; Q1-Q3: 0.75-1.4) and 1.07 (aNET - Unknown; Q1-Q3: 0.84-1.53) to 1.27 (aNET - Other; Q1-Q3: 1.10-1.44) and 1.35 (aNET-Pancreas; Q1-Q3: 0.9-2.12). A similar pattern was detected regarding the number of distinct genes with coding mutations. Midgut-derived aNET also presented a surprisingly low number of SVs compared to the other aNET subpopulations.

Next, we investigated possible differences in putative drivers between our major aNET subpopulations, being midgut- ($n = 39$) and pancreas-derived ($n = 20$) aNET (Figure 5.4, Supplementary Figure S5.8). The copy-number profiles (GISTIC2) of both populations differed, in which midgut-derived aNET presented focal deletion peaks at 9p21 (*CDKN2A/B*), 11q23 (7 possible driver genes), 12p13 (*CDKN1B*), 13q14 (17 genes), 14q24 (20 genes) and 16q23 (5 possible driver genes; common fragile site) coupled with an overall flat diploid profile. Pancreas-derived aNET presented a different profile harboring focal deletion peaks at 2q37 (*LINC01881*), 9p21 (*CDKN2A/B*) and Xp21 (*DMD*; common fragile site gene) couples with a more instable genomic profile, including several samples with large-scale chromosomal losses (Supplementary Figures S5.7 and S5.8c). When investigating the statistically significant large-scale copy-number alterations of the chromosomal arms, we also detect striking differences between the major subgroups (Supplementary Figure S5.9). Within aNEC, we detected a large number of samples (69%) harboring a loss of 22q. Midgut-derived aNET revealed amplifications of chromosome 4p/q, 5p/q, 7p/q, 10p/q, 14p/q, 20p/q and loss of 9p/q in various samples (30%) and a loss of 18p/q in 66% of samples. This re-confirms the high frequency of chromosome 18 loss in midgut-derived NET and the association with *DDC*[47], as *DCC* (18q21.2) is the most recurrently mutated gene on chromosome 18 in our cohort also ($n = 6$) together with *CDH7* (n = 6; 18q22.1). Finally, over half of pancreas-derived aNET revealed amplifications of chromosome 5p/q, 7p/q, 9q, 12p/q, 13q, 14p/q, 17p/q, 18p/q, 19p/q, 20p/q and loss of 22q.

Unbiased driver gene analysis (dN/dS) on midgut-derived aNET presented *CDKN1B* whilst pancreas-derived aNET revealed *MEN1*, *DAXX*, and *SETD2*. Several genes (present in ≥2 samples) were found only, or predominately, within midgut-derived aNET: *CDKN1B*, *KMT2A*, *PSIP1*, and *PTPRD* (Figure 5.4; Supplementary Figure S5.8e). Conversely, *MEN1*, *DAXX*, *DMD*, *SETD2*, *ATRX*, *CREBBP*, *DST*, *KDR*, *PTPRC*, and *TSC2* were found to be mutated only within pancreas-derived aNET. Several midgut-derived aNET ($n = 9$; 23%) did not readily present a shared mutual driver and only harbored somatic mutations in private or as-of-yet unassociated cancer driver genes.

Clinically actionable mutations

We observed forty-two aNEN (49%) harboring one or more target-specific or general somatic aberrations which are known as possible (and responsive) druggable targets against currently available (or under development) treatment agents. Twenty-one aNEN (24%) harbored somatic aberrations corresponding to a treatment that is currently registered for NEN or specifically for the NEN subtype of that particular patient (Figure 5.5, **Online Suppl. Data 1**). In addition, 14 patients (16%) could benefit from therapies that are off-label, but are commonly considered best

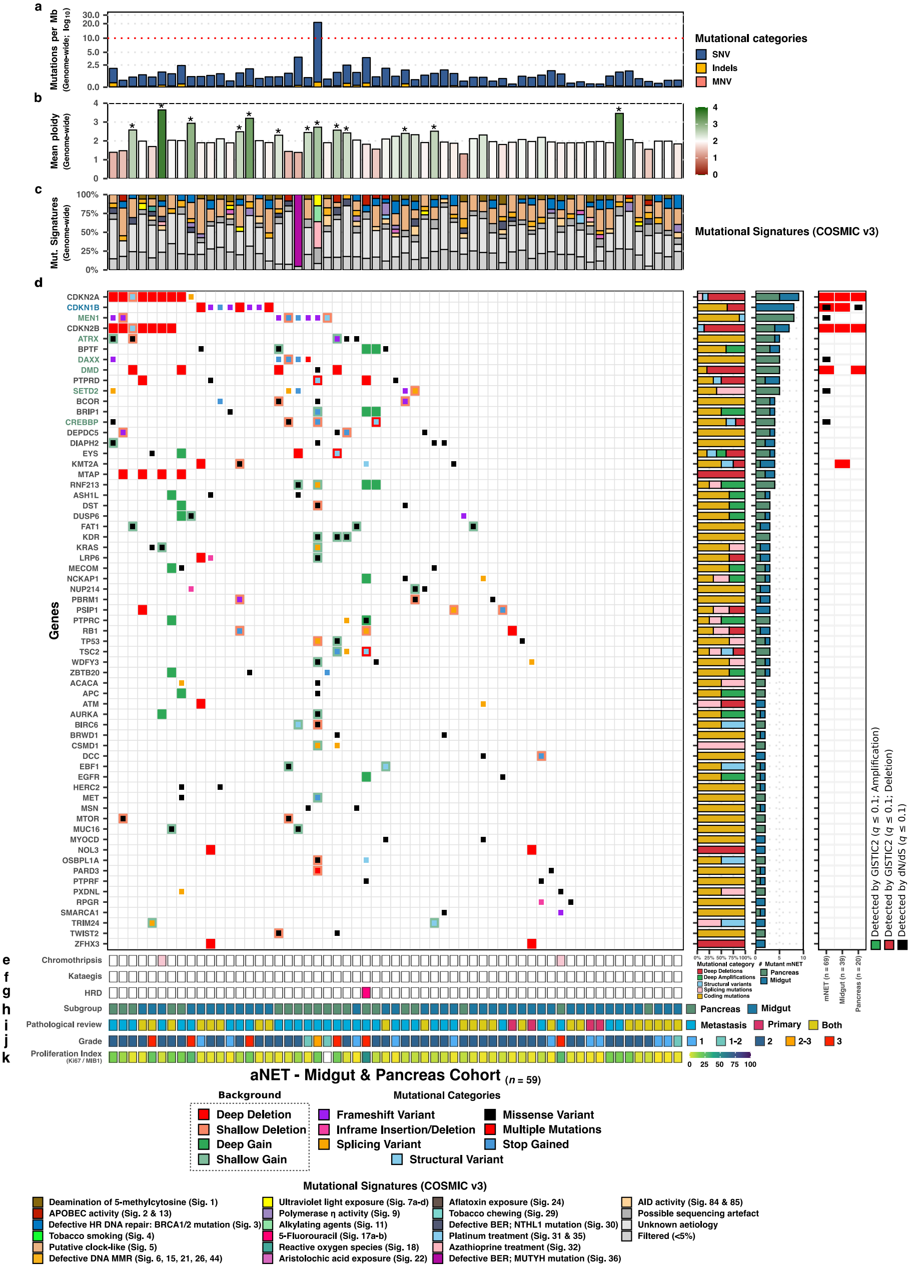


Figure 5.4: **Putative drivers and NEN-associated genes within the pancreas- and midgut-derived aNET as detected by unbiased discovery (dN/dS, GISTIC2) and literature.** Overview of putative drivers harboring coding mutations within at least two pancreas- and/or midgut-derived aNET. We show putative drivers as detected by subgroup-specific dN/dS and/or GISTIC2 and supplemented this list with additional NEN-associated drivers. aNET and genes are sorted based on mutually exclusivity of the depicted putative drivers. Same layout as Figure 5.3, except the adjacent middle-outer bar (in **d**) depicts the percentage of pancreas-derived m(NET) in green and midgut-derived aNET in blue. In addition, genes found to be mutually exclusive between our major subgroups are highlighted in the respective color of the enriched subgroup (aNET-Pancreas (green and aNET-Midgut (blue); Supplementary Figure S5.8e).

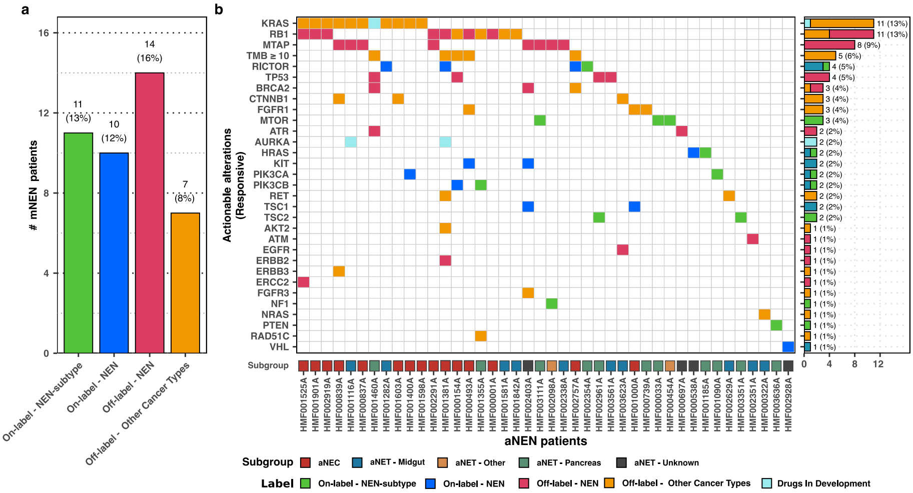


Figure 5.5: **Clinically actionable somatic alterations observed within aNEN.** **(a):** Overview of distinct aNEN harboring current clinically actionable alterations for on- and off-label NEN therapies. The highest NEN-therapy option (ranked as on-label NEN subtype (green), on-label NEN (dark blue), off-label for NEN (pink), off-label for other cancer types but currently available (orange) and drugs in development (turquoise) per distinct aNEN is shown. **(b):** aNEN harboring current clinically actionable alterations, per gene. Full description: aNEN harboring current clinically actionable alterations, per gene. The highest NET-therapy option per aNEN and gene is shown. Bottom track represents the categorized primary localization of the aNEN (aNEC in red, midgut-derived aNET in blue, non-midgut/pancreas-derived aNET in orange, pancreas-derived aNET in green and aNET of unknown origin in black) whilst the right-hand side figure shown the number of samples harboring a somatic alteration within the given gene and the proposed level of therapy.

practice for NEN. Another seven patients (8%) could benefit from drugs which are registered for another indication but not currently administered in NEN treatment. We found *RB1* ($n = 11$), *KRAS* ($n = 11$), *MTAP* ($n = 8$), high-TMB (≥10; $n = 5$), *RICTOR* ($n = 4$), and *TP53* ($n = 4$) to be the most frequently observed (target-specific or general) somatic aberrations which granted eligibility to various possible treatment options. In total, 10 midgut-derived aNET (26%) and 11 pancreas-derived aNET (55%) revealed potentially responsive alterations in various genes and most strikingly, almost all aNEC (94%) revealed potential responsive targets due to *RB1* and/or *KRAS* mutations or toward checkpoint inhibitors due to high TMB (≥10).

Discussion

Historically, NEN has long been considered as a difficult malignancy to diagnose, monitor, and treat due to presentation of an inherently wide spectrum of disease progression, cellular differentiation and low mutational burden, resulting in few targetable mutations and a relatively stable tumor genome. Indeed, aNET is characterized by the lowest TMB of all metastatic cohorts sequenced in the CPCT-02 study[27]. This study is the first to have an in-depth look into the whole genome and mutations of a large cohort of 85 advanced NEN from various primary localizations and differentiation grades. The relatively large number of unknown primary tumor localizations in this aNEN cohort ($n = 12$; 14%) reflects the difficulties in daily clinical practice to determine the site of origin for aNEN. Recently, we have become more aware of the phenomena of trans-differentiation, in which a NEC arises within a pre-existing adenocarcinoma of for instance the lung or prostate. However, in the six aNEC patients with an unknown primary tumor, no molecular clues, such as *TMPRSS2-ERG* fusions were found pointing to a specific tissue of origin.

In our aNEN cohort, it is apparent that the molecular landscape of aNEC is markedly dissimilar from that of the more differentiated aNET, in terms of mutational burden (median TMB of 5.45 vs. 1.09, respectively), genomic stability, and distinct mutant (driver) genes. With respect to TMB, four aNEC and a single aNET presented a high-TMB genotype (TMB ≥10) which could render these patients eligible for immune-based therapies such as checkpoint inhibitors[48,49].

The single high-TMB pancreas-derived aNET presented a striking contribution of the mutational signature associated with alkylating agents (temozolomide) and was previously treated with a combination of 5-fluorouracil and the alkylating antineoplastic agent streptozotocin. The mechanism of action for streptozocin closely resembles that of temozolomide as both react with DNA by undergoing substitution reactions forming a methyldiazonium ion, resulting in methylation of primarily N7 guanine (67). They both induce high levels of DNA methylation, and recognition and repair of this methylation results in single- and double-strand DNA breaks[50]. To the best of our knowledge, no data have been published on a correlation between hypermutation and streptozocin treatment, but as streptozocin and temozolomide so closely resemble each other in their mechanism of action, one can hypothesize the same mechanism to occur in streptozocin-treated patients. It would be interesting to investigate whether prior treatment with streptozocin or temozolomide indeed induces high-TMB in aNEN, and if so, whether pre-treatment with streptozocin or temozolomide could render these tumors more sensitive to checkpoint inhibition. Likewise, temozolomide (with capecitabine) for advanced pancreatic NETs has shown to be an effective therapy for these patients[51]. Similarly, we observed a large contribution of the mutational signature associated with BER deficiency due to *MUTYH* aberrations in the second highest-TMB aNET, and indeed this patient harbored a pathogenic germline *MUTYH* allele coupled with a complete somatic loss of the respective chromosomal arm. *MUTYH* abnormalities have also previously described to occur in pancreatic NET[5]. A single aNET presented a *BRCA2*-genotype associated with HRD but did not harbor (somatic) mutations within *BRCA2*. It did harbor a somatic mutation in *RAD51C*, a gene known to be involved with homologous recombination and repair of DNA.

Concerning genomic stability, we observed evidence of chromothripsis, a large-scale and catastrophic chromosomal rearrangement, within six aNEN (four aNEC, two aNET). Strikingly, four out of six chromothripsis events occurred on chromosome 12. In addition, we observe the first occurrence of localized hypermutation (kataegis) in five aNEC. Kataegis encompasses a pattern of localized hypermutations, which has been identified in various, but not all and to a varying degree, cancer types[52,53]. These regions of kataegis often co-localize with regions of genetic rearrangements. Kataegis is thought to arise from frequent genomic C-to-U deamination events as a result of APOBEC-family enzyme activity, a DNA cytosine deaminase which was recently identified as an internal and thus far unrecognized source of DNA damage and mutagenesis in various cancer types[54]. More recently, kataegis, rather than TMB, microsatellite instability or mismatch repair deficiency, was found to independently correlate with PD-L1/PD-L2 expression, and could thus be a marker in response to immune checkpoint inhibition[55].

Using unbiased driver gene analysis (dN/dS and GISTIC2) on the aNEN cohort, and on aNEC/aNEC separately to explore putative driver genes, we (re-)discovered 10 genes to be enriched with non-synonymous mutations (*TP53*, *CDKN1B*, *KRAS*, *MEN1*, *RB1*, *CREBBP*, *APC*, *DAXX*, *LPCAT2*, and *SETD2*) and detected 18 focal deletion and 2 focal amplification peaks overlapping with a plethora of (driver) genes, including deletions of *TP53*, *CDKN2A*, *CDKN2B*, *CDKN1B*, *PTPRD*, *CBFA2T3*, *CAMTA1*, *ANKDR11*, *LINC00881*, *PRKN*, *ZNF407* and fragile site genes *FHIT*, *DMD* and *MACROD2*, and amplifications of *PCAT1/MYC* and *MDM2*. Investigation of mutational enrichment within our major subgroups revealed that somatic alterations in *TP53*, *KRAS*, *RB1*, *CSMD3*, *CSMD1*, *MYC*, *APC*, *LRATD2*, and *TRRAP*, as well as the presence of chromothripsis and kataegis was enriched within aNEC. Within pancreas-derived aNET, we report the enriched presence of mutant *MEN1*, *DAXX*, *DMD*, *SETD2*, *ATRX*, and *CREBBP*, whilst midgut-derived aNET showed preference for *CDKN1B* alterations.

As previously mentioned, the majority of these detected somatic aberrations have been previously associated to primary NEN in regards to their tissue of origin. These include the associations with midgut-derived NET (*CDKN1B*)[10,12], lung NET (*FHIT*)[56–58], pNET (*TP53*, *MEN1*, *ATRX*, *DAXX* and *SETD2*)[3–5,7,15] and NEC (*TP53*, *KRAS*, *MYC*, *APC* and *RB1*, and chromothripsis)[17,18,59,60]. Aberrations within *CDKN2A* and *CDKN2B* have been associated to gastro-intestinal NETs[61,62] and have been observed with increased mutational frequency within metastatic pNET compared to primary pNET and is associated to poor prognosis.[63]

A recent large-scale study utilizing organoids derived from gastroenteropancreatic (GEP) neuroendocrine neoplasms also revealed similar genomic landscapes and (mutually exclusive) enrichment for drivers such as *TP53*, *RB1*, *APC*, and *MYC* within NEC for GEP-NEN organoids and chromosome-wide loss of heterozygosity within both NET and NEC tissues[64]. Concordantly, mutational enrichment of drivers within one population (i.e. pNET) does not imply exclusivity; e.g., *MEN1* aberrations were also found to be (sporadically) present within GEP-NECs and within a single NEC of our cohort.

Other frequently altered genes within our aNEN cohort are associated with various other malignancies (*PTPRD*[56], *CBFA2T3*[65], *ANKRD11*[66–68], and *MDM2*[69]) or genomic instability (*DMD* and *PRKN*[57], *MACROD2*)[70]. In particular, *CSMD1* and *CSMD3* (CUB And Sushi Multiple Domains 1 and 3) were found almost exclusively mutated within aNEC (31% and 44% of aNEC, respectively) yet have not previously obtained much attention in context to aNEC. *CSMD1*, a regulator of complement activation and inflammation, has been proposed as a tumor suppressor gene in advanced oral, gastric, prostate and breast cancer and subsequent loss of *CSMD1* functionality is associated to poor prognosis and enhanced proliferation, migration and invasion[71–74]. Moreover, *CSMD3* is reported as frequently mutated in lung cancers and associated with proliferation of airway epithelial cells[75] and has been recently also reported as enriched within NEC compared to NET[76]. Taken together, this prompts further investigation for *CSMD1* and *CSMD3* as aNEC-related drivers.

Currently, the choice of treatment in an individual aNEN patient is, apart from factors such as comorbidity and patient preference, determined by primary tumor localization, proliferation index (as determined by Ki-67 or MIB-1 staining), and somatostatin expression. The distinction based on primary tumor localization stems from the different embryologic structures the tumor can originate from, e.g. foregut, midgut or hindgut. When comparing the various origins of the aNEN at a genomic level, we conclude that aNEN harbors a strikingly low TMB compared to cancers[29], yet do observe slight deviations on total TMB; ranging from 1.05 (aNET-midgut; Q1–Q3: 0.75–1.4) and 1.07 (aNET—unknown; Q1-Q3: 0.84–1.53) to 1.27 (aNET—other; Q1–Q3: 1.10–1.44) and 1.35 (aNET-Pancreas; Q1–Q3: 0.9–2.12) to 5.45 (aNEC; Q–Q3: 3.8–8.85). In addition, when we compared the two largest groups of aNET per primary localization (midgut and pancreas), we can readily distinguish between the two subtypes based on somatic mutation and copy-number profiles. Yet strikingly, many midgut-derived aNET (*n* = 9; 23%) did not present a mutual driver gene but each was characterized by distinct sets of mutated genes reflecting the heterogenous nature of the malignancy.

Almost half of aNEN (*n* = 42; 49%) harbored a specific genomic alteration or genotype for which an FDA-approved drug is currently available, either on (registered for that indication) or off label. Thus, WGS revealed 49% of aNEN patients harboring clinically relevant and potentially targetable somatic aberrations which could possibly extend their treatment repertoire. It should be noted that we do not yet know whether these identified associations between genomic alterations and specific drugs indeed translate into clinical response in these patients. However, for instance, when looking at TMB as a predictive factor for checkpoint inhibitors, it was recently shown that TMB-high aNEC can respond to pembrolizumab[77]. These drugs are currently not readily available for these patients, but could provide new treatment options in the future. When deciding upon a new line of systemic treatment, a metastatic biopsy could always be considered, preferably in the context of a study, as this could shed light upon additional and effective treatment options for these late-stage patients with otherwise few remaining treatment options. In the Netherlands, we have the DRUP study active, a study in which patients for whom no standard treatments are currently available and whom might be treated with anticancer treatments outside of their approved label based on the presence of actionable mutations in their tumors[28].

In this current study exploring the largest whole-genome sequenced aNEN repository to date (*n* = 85), we focused on the genetic aberrations driving aNEN and analyzed several additional aspects of genomic instability, such as SVs, kataegis, chromothripsis, and HRD. This study improves our understanding of the complex molecular makeup of (m)NEN and reveals that the underlying genomic alterations could be exploited for better distinction of tumor subgroups and new treatment options. This study furthermore underscores that whilst the number of genetic aberrations is increased[27], the inventory of somatic drivers does not significantly change between primary and metastatic NEN. The major advantages of characterizing the genomic landscape of metastatic NEN lie within the identification of potentially actionable targets and treatment-induced (resistance-)mechanisms within the late-stage disease.

In addition, the recent major collaborative efforts in acquiring, (whole-genome) sequencing and releasing several large-scale pan-cancer datasets comprising both primary and metastatic malignancies, such as the PCAWG[78] and CPCT-02[27], could spark insights and the development of methods on how to fully interrogate and map the whole tumor genome, including the still relatively unexplored non-coding regions. This could deduce new shared oncogenic mechanisms but also, by contrast, reveal driving forces unique to (m)NEN. Within this presented aNEN repository, the full range of the somatic principles driving this enigmatic disease are likely still hidden from us but ever-present.

Material and Methods

Patient cohort and study procedures

Patients with aNEN were recruited under the study protocol (CPCT-02 Biopsy Protocol, ClinicalTrial.gov no. NCT01855477; **Online Suppl. Note 1**) of the Center for Personalized Cancer Treatment (CPCT) within the CPCT-02 and the DRUP (The Drug Rediscovery Protocol (DRUP Trial), ClinicalTrial.gov no. NCT02925234) studies. All analyzed biopsies were taken prior to treatment within the DRUP trial. The CPCT-02 (NCT01855477) and DRUP (NCT02925234) clinical studies were approved by the medical ethical committees of the University Medical Center Utrecht and the Netherlands Cancer Institute, respectively. Patients were eligible for inclusion if the following criteria were met: (1) age ≥18 years; (2) locally advanced or metastatic solid tumor; (3) indication for new line of systemic treatment with registered anti-cancer agents; (4) safe biopsy according to the intervening physician. All patients have given explicit consent for WEG and data sharing for cancer research purposes. The study procedures consisted of the collection of matched peripheral blood samples for reference DNA and image-guided percutaneous biopsy of the preferred metastatic site or, if no high-quality metastatic biopsy was available, a biopsy of the primary tumor site was collected. For the current study, patients were included for biopsy between May 10, 2016 and July 17, 2018 resulting in a cohort of 85 distinct patients from 13 Dutch hospitals (**Online Suppl. Data 1**).

Collection of the pathological records and generalization of pre-treatment(s)

Primary tumor characteristics of the 85 included aNEN patients were checked within the nationwide network and registry of histo- and cytopathology in the Netherlands (PALGA)[79].

From PALGA, we collected the differentiation grade and proliferation index (Ki67/MIB-1) based on the pathological records of the patient-specific primary and/or any metastatic lesion. If more than one pathological report was available, we chose to include the report most close in date, but always prior to, the biopsy for the CPCT-02 study.

The pre-treatment(s) of aNEN patients prior to the collection and sequencing of the tumor biopsy has been collected and generalized on treatment classification. Out of all included aNEN patients ($n = 85$), 26 patients received pre-treatment according to our clinical records.

Collection, sequencing, and processing of aNEN biopsies

Blood samples were collected in CellSave preservative tubes (Menarini-Silicon Biosystems, Huntington Valley, PA, USA) and shipped by room temperature to the central sequencing facility at the Hartwig Medical Foundation. Tumor samples were fresh-frozen in liquid nitrogen directly after the procedure and send to a central pathology tissue facility. Tumor cellularity was estimated by assessing a hematoxylin-eosin stained 6-micron section. Subsequently, 25 sections of 20 microns were collected for DNA isolation. DNA was isolated with an automated workflow (QiaSymphony) using the DSP DNA Midi kit for blood and QiaSymphony DSP DNA Mini kit for tumor samples according to the manufacturer's protocol (Qiagen). DNA concentration was measured by Qubit™ fluorometric quantitation (Invitrogen, Life Technologies, Carlsbad, CA, USA). DNA libraries for Illumina sequencing were generated from 50–100 ng of genomic DNA using standard protocols (Illumina, San Diego, CA, USA) and subsequently whole-genome sequenced in a HiSeq X Ten system using the paired-end sequencing protocol (2 × 150bp) for both the biopsy and matched blood sample.

Subsequent alignment, somatic mutation detection, and in silico tumor cell percentage estimation were performed in a uniform manner as detailed by Priestley *et al.*[27]. Briefly, paired-end sequencing reads were aligned against the human reference genome (GRCh37) using BWA-mem (v0.7.5a)[80]. Duplicate reads were marked and small insertion and deletions (InDels) were realigned using GATK IndelRealigner (v3.4.46). Prior to somatic SNV and InDel variant calling, base qualities were recalibrated using GATK BQSR (v3.4.46)[81]. Somatic SNV, InDels, and MNV were called by Strelka (v1.0.14) using the matched peripheral blood WGS sample for matched-normal variant calling[82].

Additional in-depth settings and optimizations of the HMF pipeline are described by Priestley *et al.*[27].

The somatic mutations (SNV, InDels, and MNV) were further annotated with Ensembl Variant Effect Predictor[83] (VEP, version 99, cache 99_GRCh37) using GENCODE (v33) annotations in tandem with the dbNSFP[84] plugin (version 3.5, hg19) for gnomAD[85] population frequencies. SIFT[86] and PolyPhen-2[87] scoring was applied for additional functional effect prediction.

During downstream analysis, we only retained SNV, InDels, and MNV which passed all of the following heuristic filters; default Strelka filters (PASS-only), gnomAD exome (ALL) allele frequency <0.001, gnomAD genome (ALL) <0.005, not present in ≥5 samples from the Hartwig Medical Foundation germline panel-of-normals (GATK Haplotyper) and not present in ≥3 samples from the Hartwig Medical Foundation Strelka-specific somatic blacklist.

Putative protein-altering (coding) or high-impact (e.g., splicing) mutations were aggregated per sample and gene by selecting the most deleterious annotated effect (from VEP) on any known overlapping gene-wise transcript (except those transcripts flagged as retained intron and nonsense-mediated decay). In addition, SVs with a Tumor Allele Frequency (TAF) ≥ 0.1, as calculated by PURPLE and GRIDSS[88], that overlapped only partly with the respective coding sequences (i.e., not all exons of the respective gene), were annotated as 'SV' mutations. Multiple coding mutations and/or SV per gene were annotated as 'multiple mutations'.

Discovery of somatic SVs, copy-number alterations, and in-frame fusions of *EWSR1* was performed using the GRIDDS (v2.9.3), PURPLE (v2.47) and LINX (v2.47) suite[88]. During the downstream analyses, we only retained somatic SVs passing all default QC filters (PASS-only) and with an upstream and/or downstream TAF ≥ 0.1.

Mean read coverages of the reference and tumor samples were calculated using Picard Tools (v1.141; CollectWgsMetrics) based on GRCh3789. Genomic and coding TMB was calculated as previously described by van Dessel *et al.* (2019)[89]. Briefly, the number of somatic mutations (SNVs, InDels and MNVs) was divided over the total mappable bases and the superset of coding sequences, respectively.

Discovery of genes under evolutionary selection

We performed a dN/dS analysis on somatic mutations (SNV and InDels) using dndscv[46] (v0.0.1.0) on respective genome sequences and transcript annotations using a custom transcript database based on ENSEMBL[90] Genes (v99)/GENCODE (v33) annotations. We performed a dN/dS analysis over the entire NEN cohort ($n = 85$) and four separate dN/dS analysis on the major subgroups (aNEC; n = 16, aNET; n = 69, aNET-midgut; n = 39 and aNET-pancreas; n = 20). Genes-of-interest were selected based on the statistical significance, corrected for multiple hypothesis testing (Benjamini-Hochberg), which integrated all mutation types (missense, nonsense, essential splice-site mutations and InDels; qglobal_cv ≤ 0.1) and/or without InDels (qallsubs_cv ≤ 0.1).

Detection and annotation of recurrent copy-number alterations

To detect recurrent copy-number alterations, we performed a GISTIC2[45] (v2.0.23) analysis over the entire aNEN cohort and, again, four separate GISTIC2 analysis on the major subgroups (aNEC, aNET and pancreas- and midgut-derived aNET).

GISTIC2 was performed using the following settings:

```
gistic2 -b $inputFolder-seg $inputSegmentation-refgene hg19.UCSC.
    add_miR.140312.refgene.mat -genegistic 1 -gcm extreme -maxseg
    4000 -broad 1 -brlen 0.98 -conf 0.95 -rx 0 -cap 3 -saveseg 0 -
    armpeel 1 -smallmem 0 -res 0.01 -ta 0.1 -td 0.1 -savedata 0 -
    savegene 1 -qvt 0.1.
```

Genes were annotated to GISTIC2 peaks ($q \leq 0.1$) based on the following strategy:

1. GISTIC2 focal peaks (all_lesions.conf_95.txt) were overlapped to genes (from

verified and manually annotated loci, no pseudogenes or read-throughs and from standard chromosomes; $n = 36574$) from GENCODE (GRCh37; v33), taking into consideration only the genes overlapping with at least 100 base pairs within the detected GISTIC2 peak.

2. If a GISTIC2 focal peak overlapped with multiple GENCODE genes, a combined database containing known drivers detected in a metastatic pan-cancer dataset (CPCT-02)[27], COSMIC Cancer Gene Census (v85)[91], OncoKB Cancer Gene Census (June 2019)[92] Martincorena et al.[46], and Priestley et al.[27] were used to further pinpoint the possible target gene(s) ($n = 1272$), e.g., if a GISTIC2 peak overlapped both *PTEN* and near-adjacent non-driver gene, only *PTEN* would be chosen as possible gene. The list of all overlapping GENCODE[93] (v33) genes per GISTIC2 peak can be found in **Online Suppl. Data 1**.

3. If no overlapping genes were found, GISTIC2 peaks were annotated with the nearest GENCODE (v33) protein-coding gene ($n = 19{,}988$).

Genes detected as deep amplifications or deep deletions within GISTIC2 focal peaks were considered as GISTIC2-derived driver genes in this cohort.

Mutational signature analysis

Mutational signatures based on the trinucleotide contexts of SNVs were performed, using the MutationalPatterns package (1.10.0)[94] and as previously described[89]. The 96 SBS mutational signatures (COSMIC v3) as established by Alexandrov *et al.* (2019)[42], (matrix S_{ij}; $i = 96$; number of trinucleotide motifs; $j =$ number of signatures) were downloaded from COSMIC (as deposited on May 2019). The proposed etiology of each SBS signature was derived from Alexandrov *et al.* (2019)[29], Petljak *et al.*[42], Angus *et al.*[19] and Christensen *et al.* (2019)[95].

In addition, *de novo* mutational signature analysis by MutationalPatterns was performed based on the max. number of relevant signatures as assessed using the NMF R package[96] (v0.21.0) with 1000 iterations (Supplementary Figure S5.6d). By comparing the cophenetic correlation coefficient, residual sum of squares and silhouette, we opted to generate seven custom *de novo* signatures. Custom signatures were correlated to existing (COSMIC v3) mutational signatures using cosine similarity.

Per sample, mutational signatures with less than five percent relative contribution were categorized into the "Filtered (<5%)" category.

Detection of chromothripsis

Shatterseek[97] (v0.4) using default parameters was used to detect chromothripsis-like events. As input, we used the rounded absolute copy numbers (as derived by PURPLE) and SVs with an TAF ≥ 0.1 at either end of the breakpoint. The male sex chromosome (chrY) was excluded. The criteria for a chromothripsis-like event were based on the following criteria: (a) total number of intra-chromosomal SVs involved in the event ≥25; (b) max. number of oscillating CN segments (2 states) ≥7 or max. number of oscillating CN segments (3 states) ≥14; (c) total size of chromothripsis event ≥20 megabase pairs (Mbp); (d) satisfying the test of equal distribution of SV types ($p > 0.05$); and (e) satisfying the test of non-random SV distribution within the cluster region or chromosome ($p \leq 0.05$).

Classification of homologous recombination deficiency genotypes

To determine HRD due to possible loss of function of *BRCA1* and/or *BRCA2* (amongst others), we utilized the Classifier for HRD with default settings (CHORD; v2.0). CHORD uses a random-forest approach to classify samples into HR-deficient/HR-proficient categories[31]. Briefly, we make use of CHORD;[31] a random-forest-based classifier designed to classify samples with evidence of HRD (*BRCA1*-type, *BRCA2*-type or otherwise) by using all the information captured within all the somatic small mutations and somatic SVs of whole-genome sequenced samples. If a sample contains sufficient HRD-related genomic scars (SVs) and additional markers for HRD, that sample will be classified as HR-deficient (HRD).

Detecting enrichment of mutant genes within major subgroups

To determine the enrichment of mutant genes within our major subgroups (aNEC, midgut- and pancreas-derived aNET), we generated a list of potential driver genes based on captured genes through our dN/dS ($q \leq 0.1$) analysis and/or present within the focal amplification and deletion peaks captured by GISTIC2. We extended this list by selecting genes which contained a coding mutation in ≥20% of a respective subgroup or which harbored a deep amplification or deletion in ≥20% of the respective subgroup (i.e., 20% of the respective subgroup contained coding mutations and/or ≥20% contained a copy-number alteration, irrespective of coding mutation).

Using this list of genes ($n = 20$), we performed a one-sided (enrichment) Fisher's exact test with Benjamini–Hochberg correction between each pairwise comparison per major subgroup against the remaining major subgroups (e.g., aNEC vs. the combined group of midgut- and pancreas-derived aNET).

Inventory of clinically actionable somatic alterations and putative therapeutic targets

Current clinical relevance of somatic alterations in relation to putative treatment options or resistance mechanisms and trial eligibility was determined based upon the following databases; CiViC[98] (Nov. 2018), OncoKB[92] (Nov. 2018), CGI[99] (Nov. 2018) and the iClusion (Dutch) clinical-trial database (Dec. 2020) from iClusion (Rotterdam, the Netherlands). The databases were aggregated and harmonized using the HMF knowledgebase-importer (v1.7). This list was manually corrected for discrepancies and subsequently, we curated the linked putative treatments for current on- and off-label aNEN and aNEN-subtype treatment options, as defined within the Netherlands by the Dutch Medicines Evaluation Board ("College ter Beoordeling van Geneesmiddelen; CBG)[100].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article: https://www.nature.com/articles/s41467-021-24812-3#MOESM5

Data availability

The WGS and corresponding clinical data used in this study was made available by the Hartwig Medical Foundation (Dutch nonprofit biobank organization) after signing a license agreement stating data cannot be made publicly available via third-party organizations. Therefore, the data are available under restricted access and can be requested upon by contacting the Hartwig Medical Foundation (https://www.hartwigmedicalfoundation.nl/applying-for-data/) under the accession code DR-036. Within this manuscript, we furthermore made use of the actionable gene-variant and associated drug databases of CiViC (01-Nov-2018; https://civicdb.org), OncoKB (Nov. 2018; https://www.oncokb.org), CGI (Nov. 2018; https://www.cancergenomeinterpreter.org) and the iClusion (Dutch) clinical trial database (Dec. 2020) from iClusion (Rotterdam, the

Netherlands; **Online Suppl. Data 1**). The remaining data are available within the Article, Supplementary Information or available from the authors upon request.

Code availability

Next to the initial processing workflows and software which are available at `https://github.com/hartwigmedical/`, any additional custom code and scripts used within this study (processing, analysis, and visualization) have been deposited on Bitbucket under the GPL-3.0 License: `https://bitbucket.org/ccbc/dr-036_anen/`.

References

[1] I. D. Nagtegaal, R. D. Odze, D. Klimstra, V. Paradis, M. Rugge, *et al.*, *The 2019 who classification of tumours of the digestive system,* Histopathology **76**, 182 (2020).

[2] G. Rindi, D. S. Klimstra, B. Abedi-Ardekani, S. L. Asa, F. T. Bosman, *et al.*, *A common classification framework for neuroendocrine neoplasms: an international agency for research on cancer (iarc) and world health organization (who) expert consensus proposal,* Modern Pathology **31**, 1770 (2018).

[3] C. S. Chan, S. V. Laddha, P. W. Lewis, M. S. Koletsky, K. Robzyk, *et al.*, *Atrx, daxx or men1 mutant pancreatic neuroendocrine tumors are a distinct alpha-cell signature subgroup,* Nature Communications **9**, 1 (2018).

[4] Y. Jiao, C. Shi, B. H. Edil, R. F. D. Wilde, D. S. Klimstra, *et al.*, *Daxx/atrx, men1, and mtor pathway genes are frequently altered in pancreatic neuroendocrine tumors,* Science **331**, 1199 (2011).

[5] A. Scarpa, D. K. Chang, K. Nones, V. Corbo, A. M. Patch, *et al.*, *Whole-genome landscape of pancreatic neuroendocrine tumours,* Nature **543**, 65 (2017).

[6] J. K. Park, W. H. Paik, K. Lee, J. K. Ryu, S. H. Lee, *et al.*, *Daxx/atrx and men1 genes are strong prognostic markers in pancreatic neuroendocrine tumors,* Oncotarget **8**, 49796 (2017).

[7] J. Cheng, J. Demeulemeester, D. C. Wedge, H. K. M. Vollan, J. J. Pitt, *et al.*, *Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors,* Nature Communications **8**, 1 (2017).

[8] H. K. Do, Y. Nagano, I. S. Choi, J. A. White, J. C. Yao, *et al.*, *Allelic alterations in well-differentiated neuroendocrine tumors (carcinoid tumors) identified by genome-wide single nucleotide polymorphism analysis and comparison with pancreatic endocrine tumors,* Genes Chromosomes and Cancer **47**, 84 (2008).

[9] N. Vijayvergia, P. M. Boland, E. Handorf, K. S. Gustafson, Y. Gong, *et al.*, *Molecular profiling of neuroendocrine malignancies to identify prognostic and therapeutic markers: A fox chase cancer center pilot study,* British Journal of Cancer **115**, 564 (2016).

[10] J. M. Francis, A. Kiezun, A. H. Ramos, S. Serra, C. S. Pedamallu, *et al.*, *Somatic mutation of cdkn1b in small intestine neuroendocrine tumors,* Nature genetics **45**, 1483 (2013).

[11] M. S. Banck, R. Kanwar, A. A. Kulkarni, G. K. Boora, F. Metge, *et al.*, *The genomic landscape of small intestine neuroendocrine tumors,* Journal of Clinical Investigation **123**, 2502 (2013).

[12] J. E. Maxwell, S. K. Sherman, G. Li, A. B. Choi, A. M. Bellizzi, *et al.*, *Somatic alterations of cdkn1b are associated with small bowel neuroendocrine tumors,* Cancer Genetics **208**, 564 (2015).

[13] R. Alrezk, F. Hannah-Shmouni, and C. A. Stratakis, *Men4 and cdkn1b mutations: The latest of the men syndromes,* Endocrine-Related Cancer **24**, T195 (2017).

[14] W. Y. Kim and W. G. Kaelin, *Role of vhl gene mutation in human cancer,* Journal of Clinical Oncology **22**, 4991 (2004).

[15] J. Zhang, R. Francois, R. Iyer, M. Seshadri, M. Zajac-Kaye, *et al.*, *Current understanding of the molecular biology of pancreatic neuroendocrine tumors,* Journal of the National Cancer Institute **105**, 1005 (2013).

[16] J. George, J. S. Lim, S. J. Jang, Y. Cun, L. Ozretia, *et al.*, *Comprehensive genomic profiles of small cell lung cancer,* Nature **524**, 47 (2015).

[17] L. H. Tang, B. R. Untch, D. L. Reidy, E. O'Reilly, D. Dhall, *et al.*, *Well-differentiated neuroendocrine tumors with a morphologically apparent high-grade component: A pathway distinct from poorly differentiated neuroendocrine carcinomas,* Clinical Cancer Research **22**, 1011 (2016).

[18] G. Mollaoglu, M. R. Guthrie, S. Böhm, J. Brägelmann, I. Can, *et al.*, *Myc drives progression of small cell lung cancer to a variant neuroendocrine subtype with vulnerability to aurora kinase inhibition,* Cancer Cell **31**, 270 (2017).

[19] L. Angus, M. Smid, S. M. Wilting, J. van Riet, A. V. Hoeck, *et al.*, *The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies,* Nature genetics , 1 (2019).

[20] D. Brown, D. Smeets, B. Székely, D. Larsimont, A. M. Szász, *et al.*, *Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations,* Nature Communications **8**, 14944 (2017).

[21] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, *et al.*, *Tumour evolution inferred by single-cell sequencing,* Nature **472**, 90 (2011).

[22] K. W. Hunter, R. Amin, S. Deasy, N. H. Ha, and L. Wakefield, *Genetic insights into the morass of metastatic heterogeneity,* Nature Reviews Cancer **18**, 211 (2018).

[23] D. Walter, P. N. Harter, F. Battke, R. Winkelmann, M. Schneider, *et al.*, *Genetic heterogeneity of primary lesion and metastasis in small intestine neuroendocrine tumors,* Scientific Reports **8**, 3811 (2018).

[24] H. L. Wong, K. C. Yang, Y. Shen, E. Y. Zhao, J. M. Loree, *et al.*, *Molecular characterization of metastatic pancreatic neuroendocrine tumors (pnets) using whole-genome and transcriptome sequencing,* Cold Spring Harbor Molecular Case Studies **4**, a002329 (2018).

[25] S. Y. Cho, M. Choi, H. J. Ban, C. H. Lee, S. Park, *et al.*, *Cervical small cell neuroendocrine tumor mutation profiles via whole exome sequencing,* Oncotarget **8**, 8095 (2017).

[26] P. Shen, Y. Jing, R. Zhang, M. C. Cai, P. Ma, *et al.*, *Comprehensive genomic profiling of neuroendocrine bladder cancer pinpoints molecular origin and potential therapeutics,* Oncogene **37**, 3039 (2018).

[27] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, *et al.*, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[28] D. Van der Velden, L. Hoes, H. van der Wijngaart, J. van Berge Henegouwen, E. van Werkhoven, *et al.*, *The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs,* Nature **574**, 127 (2019).

[29] L. Alexandrov, S. Nik-Zainal, D. Wedge, S. Aparicio, S. Behjati, *et al.*, *Signatures of mutational processes in human cancer,* Nature **500**, 415 (2013).

[30] K. Somyajit, S. Subramanya, and G. Nagaraju, *Rad51c: A novel cancer susceptibility gene is linked to fanconi anemia and breast cancer,* Carcinogenesis **31**, 2031 (2010).

[31] L. Nguyen, J. W. M. Martens, A. V. Hoeck, and E. Cuppen, *Pan-cancer landscape of homologous recombination deficiency,* Nature Communications **11**, 5584 (2020).

[32] F. Vaz, H. Hanenberg, B. Schuster, K. Barker, C. Wiek, *et al.*, *Mutation of the rad51c gene in a fanconi anemia-like disorder,* Nature Genetics **42**, 406 (2010).

[33] A. Min, S. A. Im, Y. K. Yoon, S. H. Song, H. J. Nam, *et al.*, *Rad51c-deficient cancer cells are highly sensitive to the parp inhibitor olaparib,* Molecular Cancer Therapeutics **12**, 865 (2013).

[34] K. Chan, S. A. Roberts, L. J. Klimczak, J. F. Sterling, N. Saini, *et al.*, *An apobec3a hypermutation signature is distinguishable from the signature of background mutagenesis by apobec3b in human cancers,* Nature Genetics **47**, 1067 (2015).

[35] I. Cortés-Ciriano, J. J.-K. Lee, R. Xi, D. Jain, Y. L. Jung, *et al.*, *Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing,* Nature Genetics **52**, 331 (2020).

[36] T. Rausch, D. T. Jones, M. Zapatka, A. M. Stütz, T. Zichner, *et al.*, *Genome sequencing of pediatric medulloblastoma links catastrophic dna rearrangements with tp53 mutations,* Cell **148**, 59 (2012).

[37] J. Z. Sanborn, S. R. Salama, M. Grifford, C. W. Brennan, T. Mikkelsen, *et al.*, *Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons,* Cancer Research **73**, 6036 (2013).

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

[39] L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. T. Ng, *et al.*, *The repertoire of mutational signatures in human cancer,* Nature **578**, 94 (2020).

[40] V. K. Singh, A. Rastogi, X. Hu, Y. Wang, and S. De, *Mutational signature sbs8 predominantly arises due to late replication errors in cancer,* Communications Biology **3**, 421 (2020).

[41] A. Dasari, K. Mehta, L. A. Byers, H. Sorbye, and J. C. Yao, *Comparative study of lung and extrapulmonary poorly differentiated neuroendocrine carcinomas: A seer database analysis of 162,983 cases,* Cancer **124**, 807 (2018).

[42] M. Petljak, L. B. Alexandrov, J. S. Brammeld, S. Price, D. C. Wedge, *et al.*, *Characterizing mutational signatures in human cancer cell lines reveals episodic apobec mutagenesis,* Cell **176**, 1282 (2019).

[43] S. S. David, V. L. O'Shea, and S. Kundu, *Base-excision repair of oxidative dna damage,* Nature **447**, 941 (2007).

[44] A. Pea, J. Yu, L. Marchionni, M. Noe, C. Luchini, *et al.*, *Genetic analysis of small well-differentiated pancreatic neuroendocrine tumors identifies subgroups with differing risks of liver metastases,* Annals of Surgery **271**, 566 (2020).

[45] C. Mermel, S. Schumacher, B. Hill, M. Meyerson, R. Beroukhim, *et al.*, *Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers,* Genome Biology **12** (2011), 10.1186/gb-2011-12-4-r41.

[46] I. Martincorena, K. Raine, M. Gerstung, K. Dawson, K. Haase, *et al.*, *Universal patterns of selection in cancer and somatic tissues,* Cell **171**, 1029 (2017).

[47] M. Nieser, T. Henopp, J. Brix, L. Stoß, B. Sitek, *et al.*, *Loss of chromosome 18 in neuroendocrine tumors of the small intestine: The enigma remains,* Neuroendocrinology **104**, 302 (2017).

[48] R. M. Samstein, C. H. Lee, A. N. Shoushtari, M. D. Hellmann, R. Shen, *et al.*, *Tumor mutational load predicts survival after immunotherapy across multiple cancer types,* Nature Genetics **51**, 202 (2019).

[49] T. A. Chan, M. Yarchoan, E. Jaffee, C. Swanton, S. A. Quezada, *et al.*, *Development of tumor mutation burden as an immunotherapy biomarker: Utility for the oncology clinic,* Annals of Oncology **30**, 44 (2019).

[50] G. M. Li, *Mechanisms and functions of dna mismatch repair,* Cell Research **18**, 85 (2008).

[51] P. L. Kunz, P. J. Catalano, H. Nimeiri, G. A. Fisher, T. A. Longacre, *et al.*, *A randomized study of temozolomide or temozolomide and capecitabine in patients with advanced pancreatic neuroendocrine tumors: A trial of the ecog-acrin cancer research group (e2211).* Journal of Clinical Oncology **36**, 4004 (2018).

[52] F. Yousif, S. Prokopec, R. Sun, F. Fan, C. Lalansingh, *et al.*, *The origins and consequences of localized and global somatic hypermutation,* bioRxiv , 287839 (2018).

[53] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. V. Loo, C. D. Greenman, *et al.*, *Mutational processes molding the genomes of 21 breast cancers,* Cell **149**, 979 (2012).

[54] R. S. Harris, *Molecular mechanism and clinical impact of apobec3b-catalyzed mutagenesis in breast cancer,* Breast Cancer Research **17**, 8 (2015).

[55] A. Boichard, I. F. Tsigelny, and R. Kurzrock, *High expression of pd-1 ligands is associated with kataegis mutational signature and apobec3 alterations,* OncoImmunology **6**, e1284719 (2017).

[56] S. Veeriah, C. Brennan, S. Meng, B. Singh, J. A. Fagin, *et al.*, *The tyrosine phosphatase ptprd is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers,* Proceedings of the National Academy of Sciences of the United States of America **106**, 9435 (2009).

[57] J. Mitsui, Y. Takahashi, J. Goto, H. Tomiyama, S. Ishikawa, *et al.*, *Mechanisms of genomic instabilities underlying two common fragile-site-associated loci, park2 and dmd, in germ cell and cancer cell lines,* American Journal of Human Genetics **87**, 75 (2010).

[58] H. Duan, Z. Lei, F. Xu, T. Pan, D. Lu, *et al.*, *Park2 suppresses proliferation and tumorigenicity in non-small cell lung cancer,* Frontiers in Oncology **9**, 790 (2019).

[59] D. Jia, A. Augert, D. W. Kim, E. Eastwood, N. Wu, *et al.*, *Crebbp loss drives small cell lung cancer and increases sensitivity to hdac inhibition,* Cancer Discovery **8**, 1422 (2018).

[60] C. N. Arnold, A. Sosnowski, and H. E. Blum, *Analysis of molecular pathways in neuroendocrine cancers of the gastroenteropancreatic system,* Annals of the New York Academy of Sciences **1014**, 218 (2004).

[61] J. Serrano, S. U. Goebel, P. L. Peghini, I. A. Lubensky, F. Gibril, *et al.*, *Alterations in the p16ink4a/cdkn2a tumor suppressor gene in gastrinomas,* Journal of Clinical Endocrinology and Metabolism **85**, 4146 (2000).

[62] N. Lubomierski, M. Kersting, T. Bert, K. Muench, U. Wulbrand, *et al.*, *Tumor suppressor genes in the 9p21 gene cluster are selective targets of inactivation in neuroendocrine gastroenteropancreatic tumors,* Cancer Research **61**, 5905 (2001).

[63] S. Roy, W. A. LaFramboise, T. C. Liu, D. Cao, A. Luvison, *et al.*, *Loss of chromatin-remodeling proteins and/or cdkn2a associates with metastasis of pancreatic neuroendocrine tumors and reduced patient survival times,* Gastroenterology **154**, 2060 (2018).

[64] K. Kawasaki, K. Toshimitsu, M. Matano, M. Fujita, M. Fujii, *et al.*, *An organoid biobank of neuroendocrine neoplasms enables genotype-phenotype mapping,* Cell **183**, 1420 (2020).

[65] M. Kochetkova, O. L. McKenzie, A. J. Bais, J. M. Martin, G. A. Secker, *et al.*, *Cbfa2t3 (mtg16) is a putative breast tumor suppressor gene from the breast cancer loss of heterozygosity region at 16q24.3,* Cancer Research **62**, 4599 (2002).

[66] P. M. Neilsen, K. M. Cheney, C. W. Li, J. D. Chen, J. E. Cawrse, *et al.*, *Identification of ankrd11 as a p53 coactivator,* Journal of Cell Science **121**, 3541 (2008).

[67] S. P. Lim, N. C. Wong, R. J. Suetani, K. Ho, J. L. Ng, *et al.*, *Specific-site methylation of tumour suppressor ankrd11 in breast cancer,* European Journal of Cancer **48**, 3300 (2012).

[68] J. E. Noll, J. Jeffery, F. Al-Ejeh, R. Kumar, K. K. Khanna, *et al.*, *Mutant p53 drives multinucleation and invasion through a process that is suppressed by ankrd11,* Oncogene **31**, 2836 (2012).

[69] J. D. Oliner, A. Y. Saiki, and S. Caenepeel, *The role of mdm2 amplification and overexpression in tumorigenesis,* Cold Spring Harbor Perspectives in Medicine **6**, 6 (2016).

[70] A. Sakthianandeswaren, M. J. Parsons, D. Mouradov, R. N. Mackinnon, B. Catimel, *et al.*, *Macrod2 haploinsufficiency impairs catalytic activity of parp1 and promotes chromosome instability and growth of intestinal tumors,* Cancer Discovery **8**, 988 (2018).

[71] C. L. Farrell, H. Crimm, P. Meeh, R. Croshaw, T. D. Barbar, *et al.*, *Somatic mutations to csmd1 in colorectal adenocarcinomas,* Cancer Biology and Therapy **7**, 609 (2008).

[72] X. L. Chen, L. L. Hong, K. L. Wang, X. Liu, J. L. Wang, *et al.*, *Deregulation of csmd1 targeted by microrna-10b drives gastric cancer progression through the nf-κb pathway,* International Journal of Biological Sciences **15**, 2075 (2019).

[73] D. M. Kraus, G. S. Elliott, H. Chute, T. Horan, K. H. Pfenninger, *et al.*, *Csmd1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues,* The Journal of Immunology **176**, 4419 (2006).

[74] M. Kamal, D. L. Holliday, E. E. Morrison, V. Speirs, C. Toomes, *et al.*, *Loss of csmd1 expression disrupts mammary duct formation while enhancing proliferation, migration and invasion,* Oncology Reports **38**, 283 (2017).

[75] P. Liu, C. Morrison, L. Wang, D. Xiong, P. Vedell, *et al.*, *Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing,* Carcinogenesis **33**, 1270 (2012).

[76] G. Centonze, D. Biganzoli, N. Prinzi, S. Pusceddu, A. Mangogna, *et al.*, *Beyond traditional morphological characterization of lung neuroendocrine neoplasms: In silico study of next-generation sequencing mutations analysis across the four world health organization defined groups,* Cancers **12**, 1 (2020).

[77] A. Marabelle, M. Fakih, J. Lopez, M. Shah, R. Shapira-Frommer, *et al.*, *Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 keynote-158 study,* The Lancet Oncology **21**, 1353 (2020).

[78] I. The, T. P.-C. A. of Whole, G. Consortium, *et al.*, *Pan-cancer analysis of whole genomes,* Nature **578**, 82 (2020).

[79] M. Casparie, A. T. Tiebosch, G. Burger, H. Blauwgeers, A. V. D. Pol, *et al.*, *Pathology databanking and biobanking in the netherlands, a central role for palga, the nationwide histopathology and cytopathology data network and archive,* Cellular Oncology **29**, 19 (2007).

[80] H. Li and R. Durbin, *Fast and accurate short read alignment with burrows-wheeler transform,* Bioinformatics **25**, 1754 (2009).

[81] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, *The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data,* Genome Research **20**, 1297 (2010).

[82] S. Kim, *Strelka2: Fast and accurate variant calling for clinical sequencing applications,* bioRxiv (2017).

[83] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, *et al.*, *The ensembl variant effect predictor,* Genome Biology **17**, 122 (2016).

[84] X. Liu, C. Wu, C. Li, and E. Boerwinkle, *dbnsfp v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site snvs,* Human Mutation **37**, 235 (2016).

[85] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, *et al.*, *Analysis of protein-coding genetic variation in 60,706 humans,* Nature **536**, 285 (2016).

[86] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, *et al.*, *A method and server for predicting damaging missense mutations,* Nature Methods **7**, 248 (2010).

[87] P. C. Ng and S. Henikoff, *Sift: Predicting amino acid changes that affect protein function,* Nucleic Acids Research **31**, 3812 (2003).

[88] D. Cameron, J. Baber, C. Shale, A. Papenfuss, J. E. Valle-Inclan, *et al.*, *Gridss, purple, linx: Unscrambling the tumor genome via integrated analysis of structural variation and copy number,* bioRxiv (2019), 10.1101/781013.

[89] L. F. van Dessel, J. van Riet, M. Smits, Y. Zhu, P. Hamberg, *et al.*, *The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact,* Nature communications **10**, 1 (2019).

[90] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, *et al.*, *Ensembl 2018,* Nucleic Acids Research **46**, D754 (2018).

[91] S. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, *et al.*, *Cosmic: Somatic cancer genetics at high-resolution,* Nucleic Acids Research **45**, D777 (2017).

[92] D. Chakravarty, J. Gao, S. Phillips, R. Kundra, H. Zhang, *et al.*, *Oncokb: A precision oncology knowledge base,* JCO Precision Oncology **1**, 1 (2017).

[93] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, *et al.*, *Gencode: The reference human genome annotation for the encode project,* Genome Research **22**, 1760 (2012).

[94] F. Blokzijl, R. Janssen, R. van Boxtel, and E. Cuppen, *Mutationalpatterns: Comprehensive genome-wide analysis of mutational processes,* Genome Medicine **10** (2018), 10.1186/s13073-018-0539-0.

[95] S. Christensen, B. V. der Roest, N. Besselink, R. Janssen, S. Boymans, *et al.*, *5-fluorouracil treatment induces characteristic t>g mutations in human cancer,* Nature Communications **10**, 4571 (2019).

[96] R. Gaujoux and C. Seoighe, *A flexible r package for nonnegative matrix factorization,* BMC Bioinformatics **11** (2010), 10.1186/1471-2105-11-367.

[97] I. Cortes-Ciriano, J.-K. Lee, R. Xi, D. Jain, Y. Jung, *et al.*, *Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing,* BioRxiv (2018).

[98] M. Griffith, N. Spies, K. Krysiak, J. McMichael, A. Coffman, *et al.*, *Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer,* Nature Genetics **49**, 170 (2017).

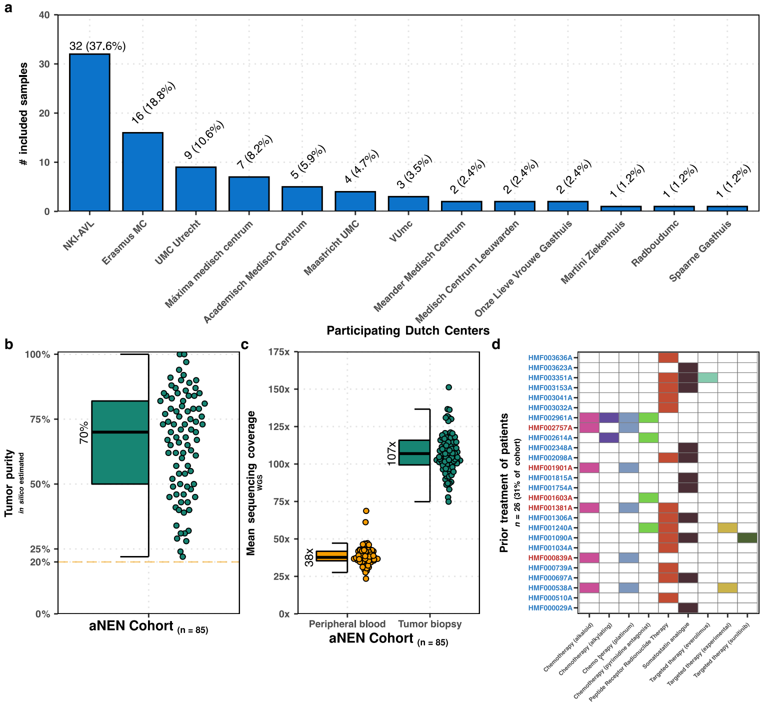
[99] D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. Schroeder, A. Vivancos, *et al.*, *Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations,* Genome Medicine **10** (2018), 10.1186/s13073-018-0531-8.

[100] M. Pavel, D. O'Toole, F. Costa, J. Capdevila, D. Gross, *et al.*, *Enets consensus guidelines update for the management of distant metastatic disease of intestinal, pancreatic, bronchial neuroendocrine neoplasms (nen) and nen of unknown primary site,* (2016) pp. 172–185.

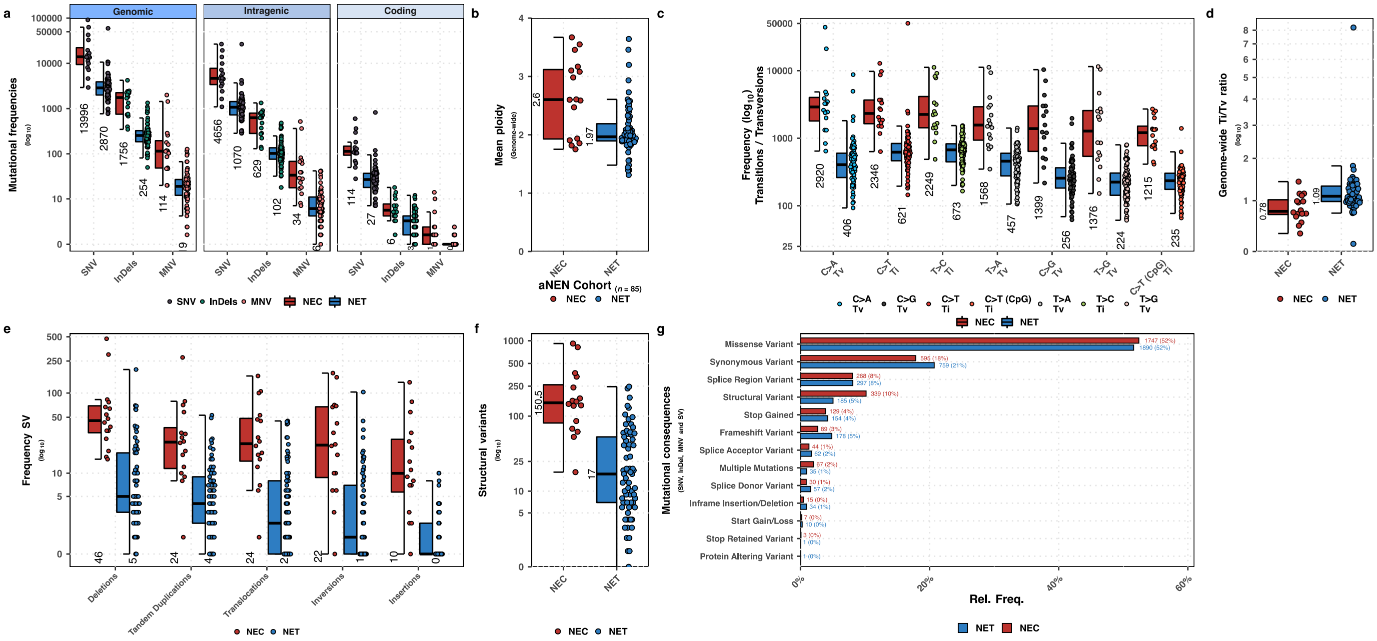
Supplemental Data

Supplementary data and figures accompanying the chapter:

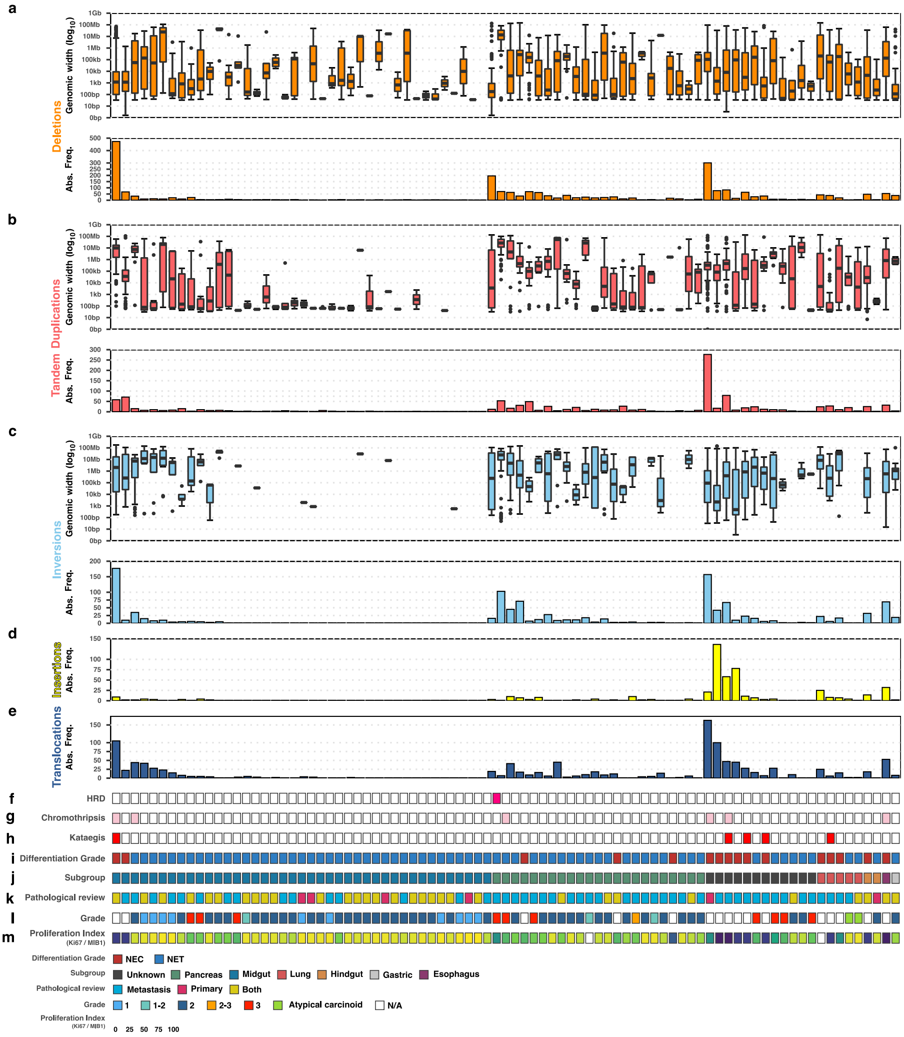
"The genomic landscape of 85 advanced neuroendocrine neoplasms reveals subtype-heterogeneity and potential therapeutic targets."



Supplementary figure S5.1: **Overview of sequencing quality metrics.**
(a): Absolute and relative frequencies of distinct included patients in the CPCT-02 aNEN cohort per participating center within the Netherlands. **(b):** Boxplot with individual data points of the estimated (*in silico*) tumor cell percentages based on the whole genome sequencing data with observed median displayed. **(c):** Boxplot with individual data points of the mean read-coverages (WGS) of the peripheral blood (reference; blue) and biopsy tissues (red) with observed median per variable displayed. **(c):** Age distribution stratified by gender of the aNEN cohort with observed median per variable displayed in a boxplot with individual data points. **(d):** Type of prior treatment per aNEN, if applicable (*n* = 26 out of 85 patients). Patients with aNET are highlighted in blue on the y-axis whulst patients with aNEC are highlighted in red on the y-axis.

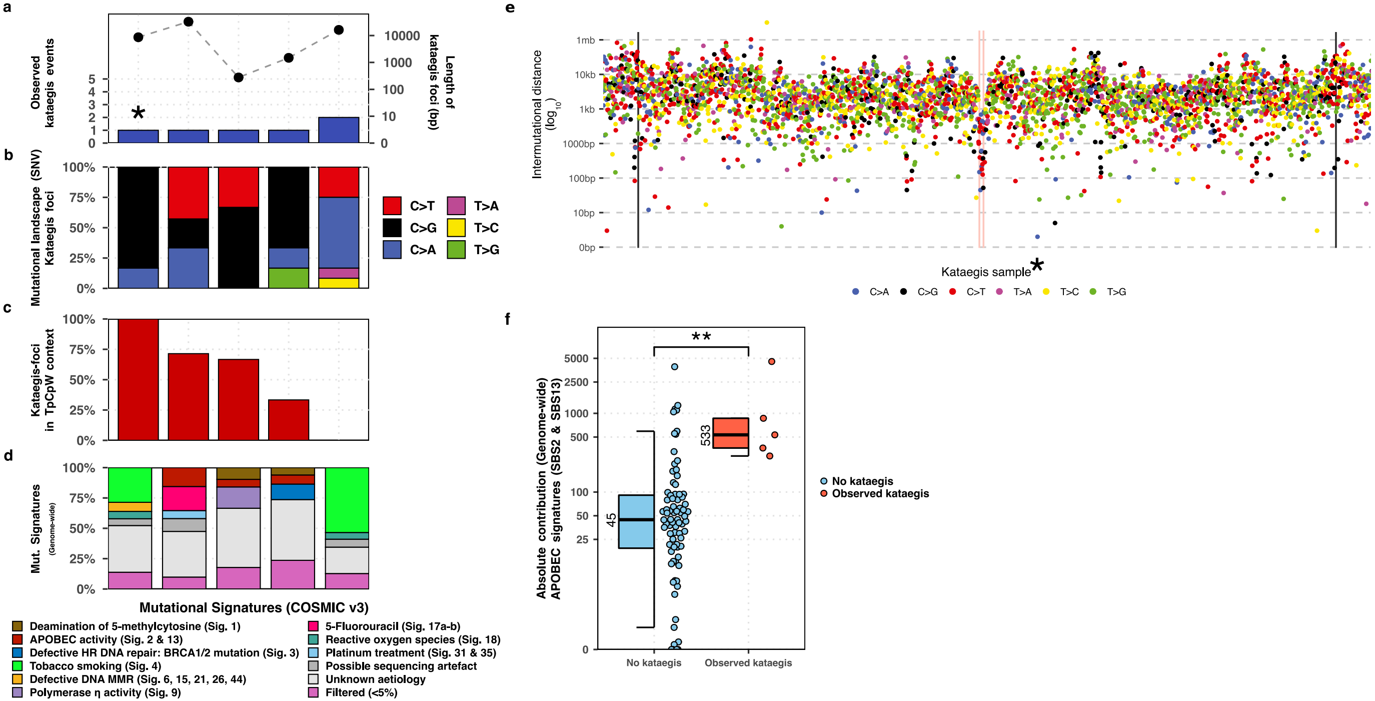


Supplementary figure S5.2: **Overview of mutational landscape, categorized per differentiation grade.**
(a): Bee-swarm boxplot with notch of the mean read coverage per sample of reference and tumor tissues. Boxplot depicts the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the IQR. Data points outside the IQR are shown. **(b):** Mean genome-wide tumor ploidy based on all autosomal chromosomes with observed median displayed. Data is categorized on aNEC / aNET status. **(c):** Type of genome-wide SNVs. Transition (Ti) and transversion (Tv), with a special attention for C to T (Ti) in CpG context, are indicated per sample with observed median per variable displayed. Data is categorized on aNEC / aNET status. **(d):** Genome-wide ratio of transitions (Ti) over transversion (Tv) with observed median displayed. Data is categorized on aNEC / aNET status. **(e):** Frequency of Tandem Duplications, Insertions, Inversions, Deletions and interchromosomal translocations are indicated per sample with observed median per variable displayed. Data is categorized on aNEC / aNET status. **(f):** Frequency of structural variants (SV) per sample with observed median per variable displayed. Data is categorized on aNEC / aNET status. **(g):** Mutational consequences of genomic variants overlapping genes using Ensembl Variant Effect Predictor (VEP). Data is categorized on aNEC / aNET status.

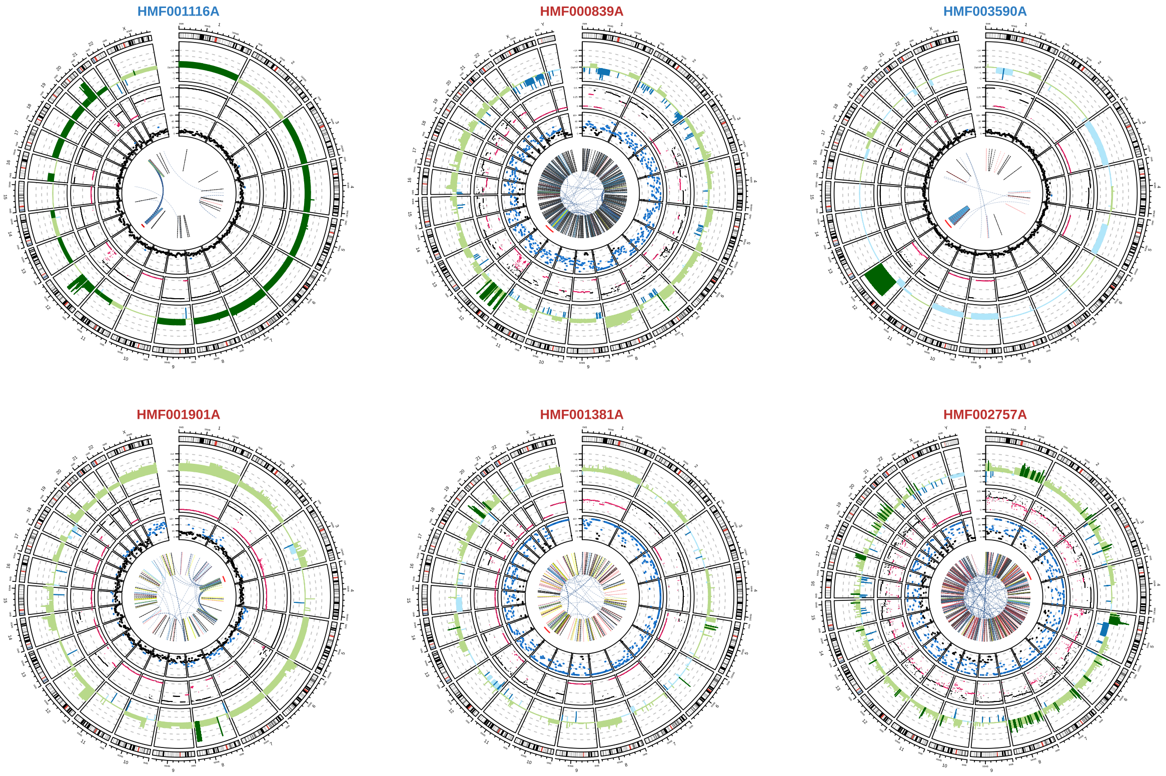


Supplementary figure S5.3: **Overview of the distribution of somatically-acquired structural variants.**

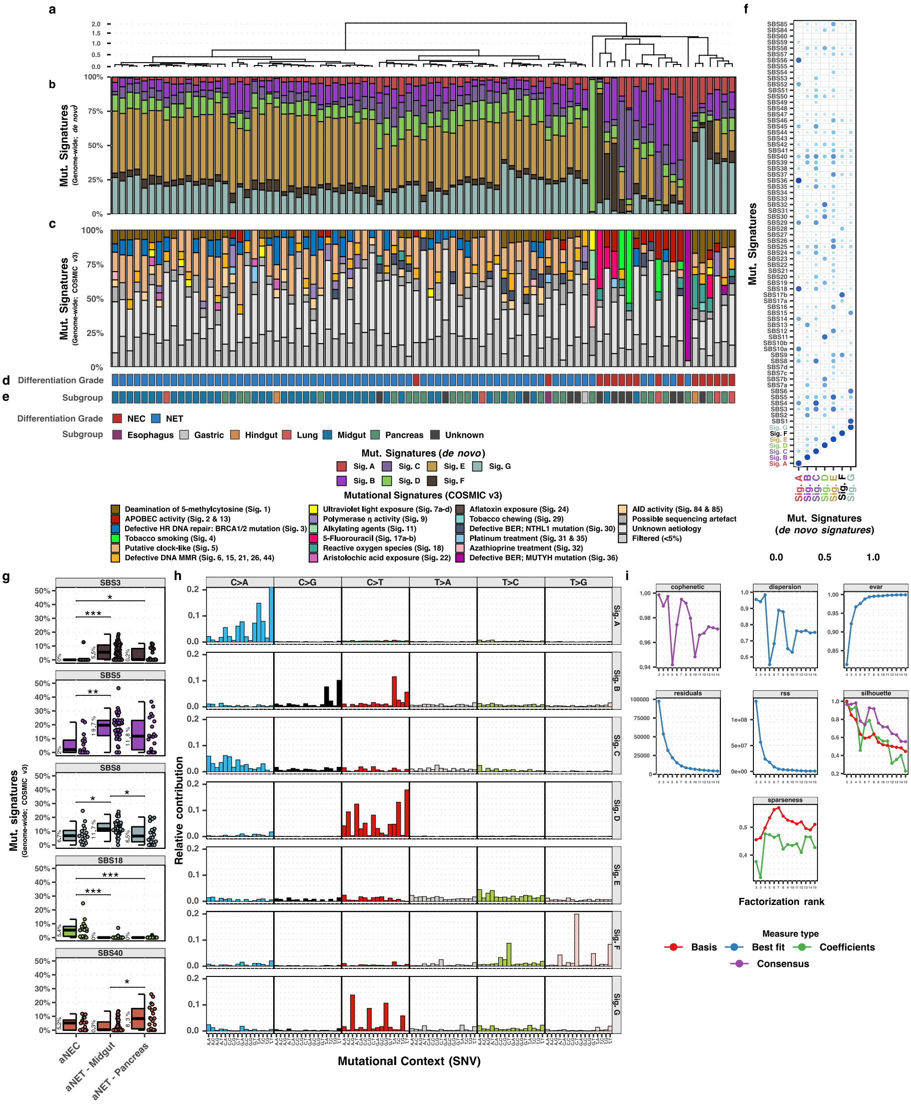
Overview of the genomic sizes and numbers of structural variants present in the aNEN cohort. Samples are sorted based on primary localization and decreasing number of total observed structural variants over all categories (deletions, tandem duplications, inversions, translocations and insertions).



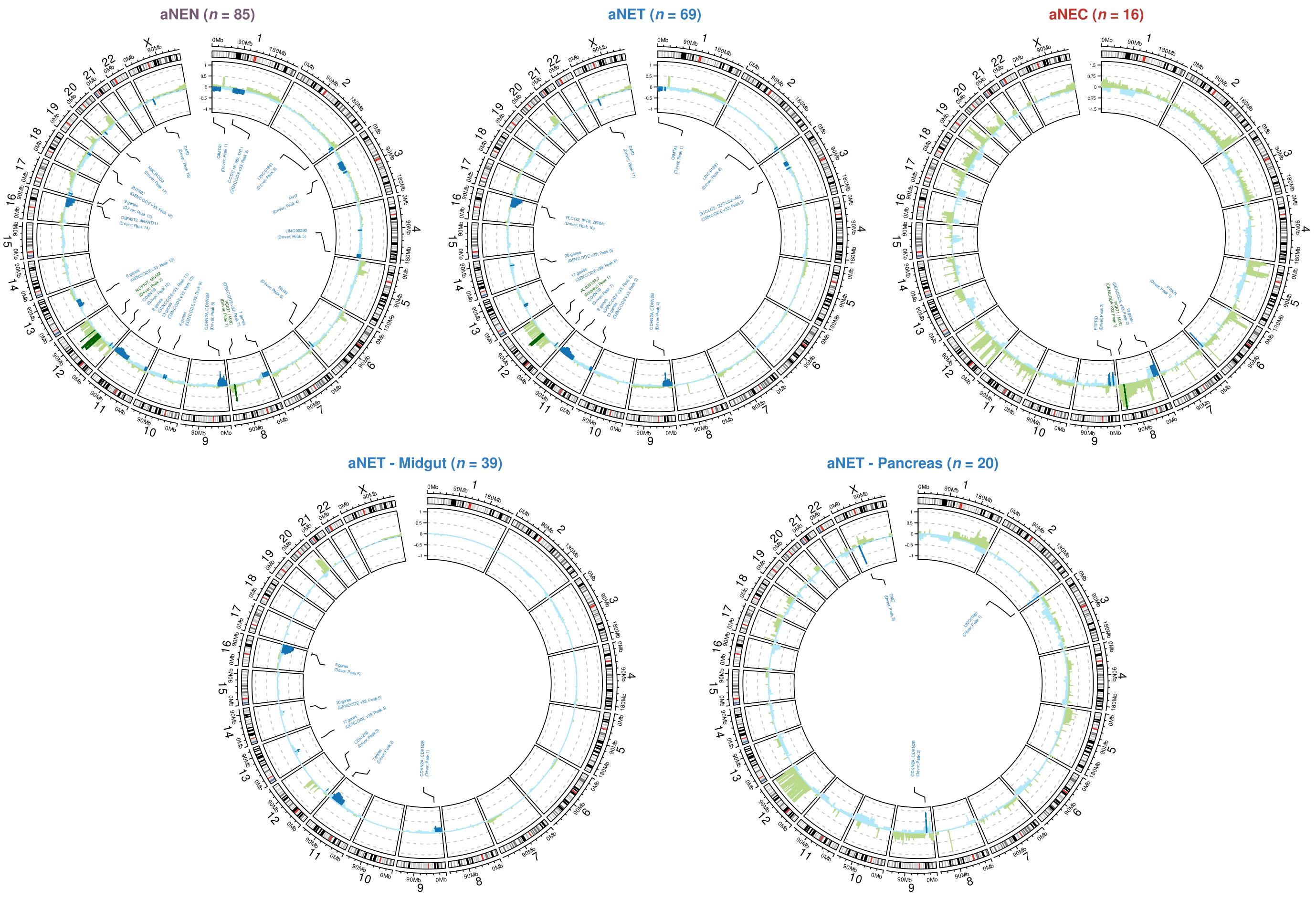
Supplementary figure S5.4: **Observed kataegis events within the aNEN cohort.**
(a): Number of observed kataegis foci in the aNEN cohort (found in 6 distinct samples, blue bars) and the respective cumulative genomic width of all observed kataegis foci per sample (right y-axis; black points). **(b):** Relative frequency of SNV categories found in all observed kataegis foci per sample. **(c):** Relative frequency of SNV in observed kataegis foci with APOBEC-related TpCpW mutational context. W stands for T or A changes. **(d):** Genome-wide relative contribution to mutational signatures (COSMIC v3) for the respective aNEN sample. **(e):** Representation of a single kataegis foci on chromosome 8 within a single respective sample (highlighted with * in **a**). SNV (colored on pyrimidine mutations) are shown with relative genomic distances (in $\log_{10}$) to neighboring SNV. Observed kataegis focus on chromosome 8 is highlighted with a transparent red background. **(f):** Absolute mutational contribution of APOBEC COSMIC (v3) signatures (2 & 13) for samples without (n = 79) and with observed kataegis foci (n = 6). Statistical significance was tested with Wilcoxon rank-sum test and is denoted with * ≤ 0.05, ** ≤ 0.01 and *** ≤ 0.001.



Supplementary figure S5.5: **Genomic overview of aNEN displaying chromothripsis-like events.**
Genomic representations of the chromothripsis-harboring aNEN (*n* = 6). The outer track displays the genomic ideogram, the second-outer track displays copy number profiles (amplification in light green; deep amplification beyond sample-specific threshold (GISTIC2) in dark green, deletions in blue; deep deletions beyond sample-specific threshold (GISTIC2) in dark blue). The third track displays TC%-corrected lower allele-frequency (LAF) values of individual copy number segments (LAF ≤ 0.33 in pink; LAF ≥ 0.33 in black). The fourth track displays the number of mutations per 5 Mbp, ranging from 0 to 60+; bins with ≥ 20 mutations are highlighted in blue. The fifth track highlights the regions harboring chromothripsis in a red line. The innermost track displays the breakpoints of the structural variants; interchromosomal translocations in dark blue, deletions in gray, insertions in yellow, inversion in light blue and tandem duplications in red. Samples are colored per aNEC (in red) and aNET (in blue) status.



Supplementary figure S5.6: ***de novo* mutational signatures assessment on aNEN.** Assessment and comparison of extracted *de novo* single base substitution mutational signatures (n = 7; Sig. A - I) using non-negative matrix factorization (NMF) within the aNEN cohort against the known COSMIC (v3; n = 67) signatures. **(a-e):** Overview of extracted *de novo* single base substitution mutational signatures (n = 7; Sig. A – I; upper track) vs. COSMIC signatures (v3; n = 67; lower track), per aNEN. aNEN are sorted based on unsupervised clustering (Ward.D; Euclidean distance; distances plotted in log$_{10}$-scale) of the relative contribution of the seven *de novo* mutational signatures. **(f):** Cosine similarity of the de novo mutational signatures against the known COSMIC v3 signatures (n = 67). **(g):** Overview of the statistically significant mutational signatures between the major subgroups (Wilcoxon Rank-Sum test with Benjamini-Hochberg correction);; significance is denoted by *** ($q \leq$ 0.001), ** ($q \leq 0.01$), * ($q \leq 0.05$) and . ($q \leq 0.1$). **(h):** Trinucleotide mutational contexts of the seven extracted *de novo* signatures. **(i):** NMF quality metrics using between two to fifteen ranks over 1000 iterations.



aNEN (*n* = 85)

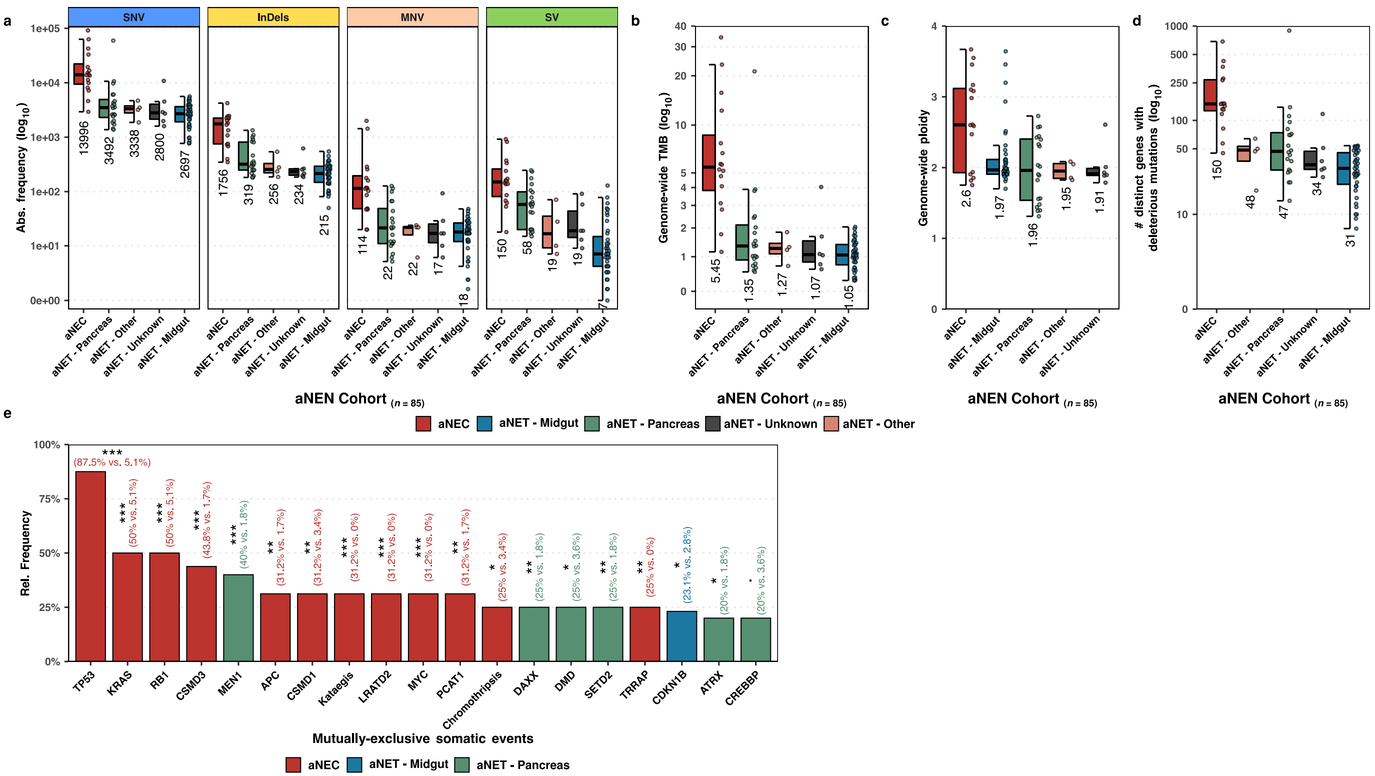
aNET (*n* = 69)

aNEC (*n* = 16)

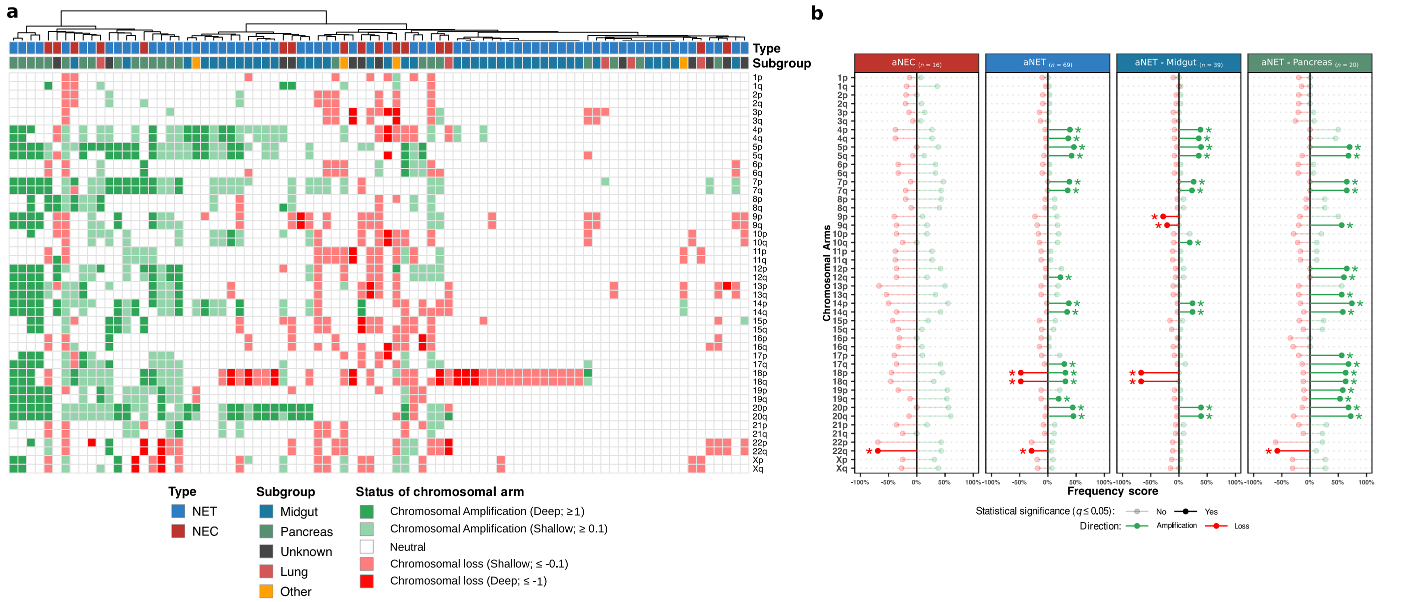
aNET - Midgut (*n* = 39)

aNET - Pancreas (*n* = 20)

Supplementary figure S5.7: **Circosplots with ideogram of recurrent copy-number aberrations as detected by GISTIC2 per sub-population (as shown above each circosplot)**.
G-scores are depicted on the y-axis. Regions with amplifications (G-score >0) are depicted in green and deletions (G-score <0) in blue. Regions with significant (and recurring) copy-number aberrations ($q \leq 0.1$) are denoted with a darker shade of green or blue, respective of amplification or deletion. Per region, the foci of maximal amplification or deletion (focal peaks; q ≤ 0.1) are denoted in the inner track; the peak identifiers with associated genes are also denoted and presented in **Online Suppl. Data 1**.



Supplementary figure S5.8: **Genomic characteristics per differentiation grade (aNEC/aNET) and primary localization within aNEN.** **(a):** Number of SNV, InDels, MNV and SV per whole-genome aNEN with observed median per variable displayed. Data is categorized on aNEC and distinct aNET subgroups based on primary localization. **(b):** Tumor mutational burdens (genome-wide; $\log_{10}$), with observed median per variable displayed. Data is categorized on aNEC and distinct aNET subgroups based on primary localization. **(c):** Mean genome-wide ploidy, with observed median per variable displayed. Data is categorized on aNEC and distinct aNET subgroups based on primary localization. **(d):** Number of genes harboring somatic coding mutations, with observed median per variable displayed. Data is categorized on aNEC and distinct aNET subgroups based on primary localization. **(a):** Mutational enrichment of mutant genes (mutations and copy-number alterations) and large-scale events (kataegis and chromothripsis) between our three major subgroups; aNEC, pancreas- and midgut-derived aNET. Statistical significance was tested using a one-sided Fisher's Exact Test with BH correction; significance is denoted by *** ($q \leq 0.001$), ** ($q \leq 0.01$), * ($q \leq 0.05$) and . ($q \leq 0.1$).



Supplementary figure S5.9: **Copy-number aberrations of chromosomal arms per differentiation grade (NEC/NET) and primary localization within aNEN.**

(a): Unsupervised clustering (Euclidean distances, Ward.D2 method) of the aNEN samples based on the categorization of chromosomal arm copy-number aberrations (based on GISTIC2 value per arm). Top color-bars depict the differentiation grade of the aNEN (aNEC in red, aNET in blue) and the primary localization. **(b):** Overview of the relative frequency of samples with amplifications (green) and losses (red) per arm within the given subgroup. Statistically significant ($q \leq 0.05$) arm-level copy-number aberrations are depicted with an asterisk whilst the non-significant events are shown as transparent.

Chapter 6

ERG$^+$ PRAD shows enhanced frequencies of innate immune cells and elevated gene expression of *TDO2* when compared to *ERG*$^-$ PRAD

J. van Riet[a,b,c*], D.M. Hammerl[d*], M.A. Komor[e,f], Y. Hoogstrate[b], B. Janssen[g], R.J.A. Fijneman[e], G.J.L.M. van Leendersh, H.J.G. van de Werken[a,b,i¶], G.W. Jenster[b¶], R. Debets[d¶]

a Cancer Computational Biology Center, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.
b Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.
c Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.
d Laboratory of Tumor Immunology, Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands.
e Translational Gastrointestinal Oncology, Department of Pathology, Netherlands Cancer Institute, Amsterdam, the Netherlands.
f Oncoproteomics Laboratory, Department of Medical Oncology, VU University Medical Center, Amsterdam, the Netherlands.
g GenomeScan, Leiden, the Netherlands.
h Department of Pathology, Erasmus MC Cancer Institute, University Medical Center, Rotterdam, the Netherlands.
i Department of Immunology, Erasmus MC Cancer Institute, University Medical Center, Rotterdam, the Netherlands

* These authors contributed equally.
¶ These authors jointly supervised this work.

This chapter has not yet been peer-reviewed and is a preliminary draft-version.

Abstract

The *TMPRSS2-ERG* fusion-event induces abundant expression of *ERG* (*ERG*$^{+}$) and is present within half of all prostate adenocarcinomas (PRAD), with dissimilar expression profiles compared to *ERG*$^{-}$ PRAD. Here, we questioned whether these PRAD subgroups could harbor unique immune determinants which might expand our understanding into tumor evolution and overall poor response to immune-therapies. Utilizing whole-transcriptomics from a discovery and two validation data-sets constituting early-stage, treatment-naïve PRAD (n = 539) and normal adjacent prostate tissue (NAP; n = 148), we investigated differential expression of immune-related genes and gene-sets, immune cell composition, T cell receptor repertoire and antigenicity using immunogenomics algorithms.

We revealed a distinct yet considerable subgroup of *ERG*$^{+}$ PRAD with abundant expression of *TDO2* and dissimilar expression of immune-regulatory gene-sets such as metabolic and lysosome pathways, MHC I complex, IL-2–STAT5, IFNα signaling and IFNγ signaling. We furthermore observed significant differences within the relative immune-cell compositions, including lower frequency of neutrophils within PRAD vs. NAP, whereas higher frequencies of M1 macrophages, dendritic cells, Tregs and NK cells could be attributed to *ERG*$^{+}$ PRAD, with even greater frequencies within *ERG*$^{+}$/*TDO2*$^{+}$ PRAD. This was coupled with increased numbers of coding mutations, TCR repertoire and CTL neo-epitopes within *ERG*$^{+}$/*TDO2*$^{+}$ PRAD compared to NAP, whilst *TDO2*$^{-}$ PRAD harbored similar quantities to NAP. In addition, a subset of Cancer Germline Antigens were differentially expressed between NAP, *ERG*$^{+}$ PRAD and *ERG*$^{-}$ PRAD.

In short, we reveal dissimilar immune-contextures and potential origins of innate immunity based on *ERG* and/or *TDO2* expression within PRAD. The *ERG*$^{+}$/*TDO2*$^{+}$ subgroup of PRAD constitutes a hitherto unidentified and potentially immunosuppressive subgroup. Further validation is required to investigate these potential avenues into these otherwise immunologically-cold tumors.

Introduction

Prostate cancer is the most prevalent and second most frequently diagnosed malignancy in men worldwide, surpassed only by the incidences of lung cancer.[1,2] Several treatment options are available for primary and advanced disease including radical surgery, radiation therapy, chemotherapy and/or androgen deprivation therapy (ADT).[3] Dependent on the disease stage, molecular characteristics and time of diagnosis, curative treatment success varies between patients with primary (localized) prostate adenocarcinoma (PRAD) whilst curative treatment outcomes for progressed metastatic or androgen deprivation-resistant disease are scarce.[3,4]

An alternative and promising option to treat solid tumors are immunotherapies using Immune Checkpoint Inhibitors (ICI), typically against PD-1, PD-L1 and CTLA-4.[5] However, only a small subset of patients with PRAD benefit from ICI-based treatment with PRAD revealing overall lower response rates as compared to various other solid tumors.[6] For example, treatment of metastatic castration-resistant prostate cancer (mCRPC) with anti-CTLA4 (ipilimumab) in a phase III trial showed no significant improvement of survival compared to placebo.[7] Similarly, anti-PD1 treatment with nivolumab in mCRPC did not result in any objective response (OR).[8] Conversely, treatment of mCRPC with another anti-PD1 treatment (pembrolizumab) resulted in 17% OR in PD-L1$^+$ patients.[9] PD-L1-positivity, however, is not a prerequisite for response to pembrolizumab as evaluated by a phase II trial that yielded equivalent response (of 4%) in the PD-L1$^+$ and PD-L1$^-$ treatment arms.[10] In contrast, a recent trial in which ipilimumab and pembrolizumab were combined yielded OR in 25% of mCRPC patients, particularly in patients with high tumor mutational burden (TMB) and PD-L1-positivity >1%. Of note, grade 3-5 adverse events were observed in about 50% of patients, urging for improvements in patient selection.[11] To better understand the variable response to ICI, it is critical to identify immune determinants that are unique for localized PRAD. As localized PRAD is generally considered a poorly immunogenic tumor with a relatively low TMB[12], the identification of immune determinants could greatly improve the selection of patients for ICI treatments.

Generally, PRAD harbors low numbers of tumor infiltrating lymphocytes (TILs), and those present commonly have an immunosuppressive phenotype, including FOXP3$^+$ regulatory CD4+ T cells, tumor-associated macrophages (TAMs) as well as myeloid-derived suppressor cells (MDSC).[13] High frequencies of innate immune cells, such as natural killer (NK) cells and dendritic cells (DC) have been associated with good prognosis in PRAD. Conversely, high frequencies of TAMs and, in contrast to many other tumor types, $CD8^{+}$ T cells have been associated with poor prognosis.[14] In recent years, molecular subclassification of both primary PRAD and mCRPC has been significantly refined and includes genotypes with potential responses to immune-therapies.[4,15,16] The latter genotypes generally harbor a multitude of somatic aberrations, for example, due to microsatellite instability (MSI), BRCAness and CDK12ness that could potentially lead to high loads of neo-antigens.[4,17–19]. In addition, somatic aberrations potentially impact the expression of oncogenic drivers, such as *PTEN*, *PIK3CA*, *TP53*, members of the *RAS* family, *RB1* and *TMPRSS2-ERG* fusions. Recent and seminal reports, mainly in melanoma, have demonstrated that examples of such tumor cell-intrinsic oncogenes are highly-related to immune evasion, such as the cyclin-dependent kinase *CDK4*[20] and the p21-activated kinases *PAK4*[21].

The genomic fusion between *TMPRSS2* and *ERG* (*TMPRSS2-ERG*) occurs in 50% of all PRAD yet remains an enigmatic feature of the malignancy, whilst constituting a highly prevalent and distinct molecular subgroup with significant and large-scale changes in transcriptomic and epigenetic profiles.[17,22] Briefly, *TMPRSS2-ERG* is a genomic re-arrangement, mostly originating from a single 3 megabase genomic deletion, placing the promoter of the androgen-responsive *TMPRSS2* gene in proximity to the ETS-family transcription factor *ERG*; leading to (over-)expression of the *TMPRSS2-ERG* fusion product.[22,23] Beyond the highly-specific diagnostic potential of *TMPRSS2-ERG*, there are conflicting reports regarding the prognostic value of *TMPRSS2-ERG* within PRAD.[24–26] This controversy likely reflects the complex interplay in which expression of *ERG* or even the mere genomic presence of *TMPRSS2-ERG* somehow contributes to the overall lack of proficient immunogenicity within PRAD. In fact, recent investigations suggest that *TMPRSS2-ERG* fusions are associated with lower abundances of TILs and deregulated interactions between innate and adaptive immune cells.[27]

In the current study, we present an integrative *in silico* study of three independent transcriptomics cohorts (NGS-ProToCoL[28], EMC-FFPE-PCa and TCGA-PRAD[17]) consisting of localized PRAD (Gleason Score 6-10) and normal-adjacent prostate (NAP) tissue, comprising in total 593 PRAD and 148 NAP samples. These cohorts were independently interrogated and cross-evaluated by contemporary *in silico* immunogenomics algorithms to investigate the immune cell composition, T cell receptor repertoire, antigenicity and immune evasive pathways between NAP, *ERG*$^+$ and *ERG*$^-$ PRAD (Figure 6.1).

Results

Study design and overview of included prostate cancer cohorts

The publicly available whole-transcriptome (fresh-frozen) sequenced cohort (NGS-ProToCol) containing NAP (n = 40) and localized PRAD (n = 49) tissues (Gleason scores 6-10) was interrogated to investigate immune determinants of *ERG*$^+$ and *ERG*$^-$ PRAD (Figure 6.1, Supplementary Figure S6.1). In addition, two independent whole-transcriptome cohorts containing both NAP and PRAD tissues were used as validation cohorts. These validation cohorts comprised the whole-transcriptome TCGA-PRAD cohort (n = 536) and an additional whole-transcriptome (Formalin-Fixed Paraffin-Embedded (FFPE)) sequenced cohort (EMC-PCa-FFPE; n = 57 paired samples) (Supplementary Figure S6.1).

To investigate possible associations between the PRAD immune-landscape and increased *ERG* mRNA expression, samples were categorized into two groups based on expression profiles (Supplementary Figure S6.2). Using the normalized expression of *ERG* and two known downstream *ERG*-regulated genes (*PCAT5* and *TDRD1*), samples were categorized according to unsupervised clustering (k = 2). Samples were denoted as *ERG*$^+$ and *ERG*$^-$ based on the respective expression of these three genes within the two clusters.

Using these cohorts, we assessed differential expression of genes and gene-sets related to immune evasive mechanisms, the composition of immune-cell populations, the T cell receptor (TCR) repertoire and the expression of CD8 T lymphocytes (CTL) neo-epitopes and cancer germline antigens (CGAs).

A major subpopulation existing within *ERG*$^+$, and absent within *ERG*$^-$ PRAD, harbors abundant expression of *TDO2* and is accompanied by pathways related to antigen presentation and processing, and IFN production

We first investigated the differential gene expression profiles between NAP, *ERG*$^-$ PRAD and *ERG*$^+$ PRAD and performed gene-set enrichment analysis (GSEA) using, among others, gene-sets related to immune evasion.[29] For these analyses and all

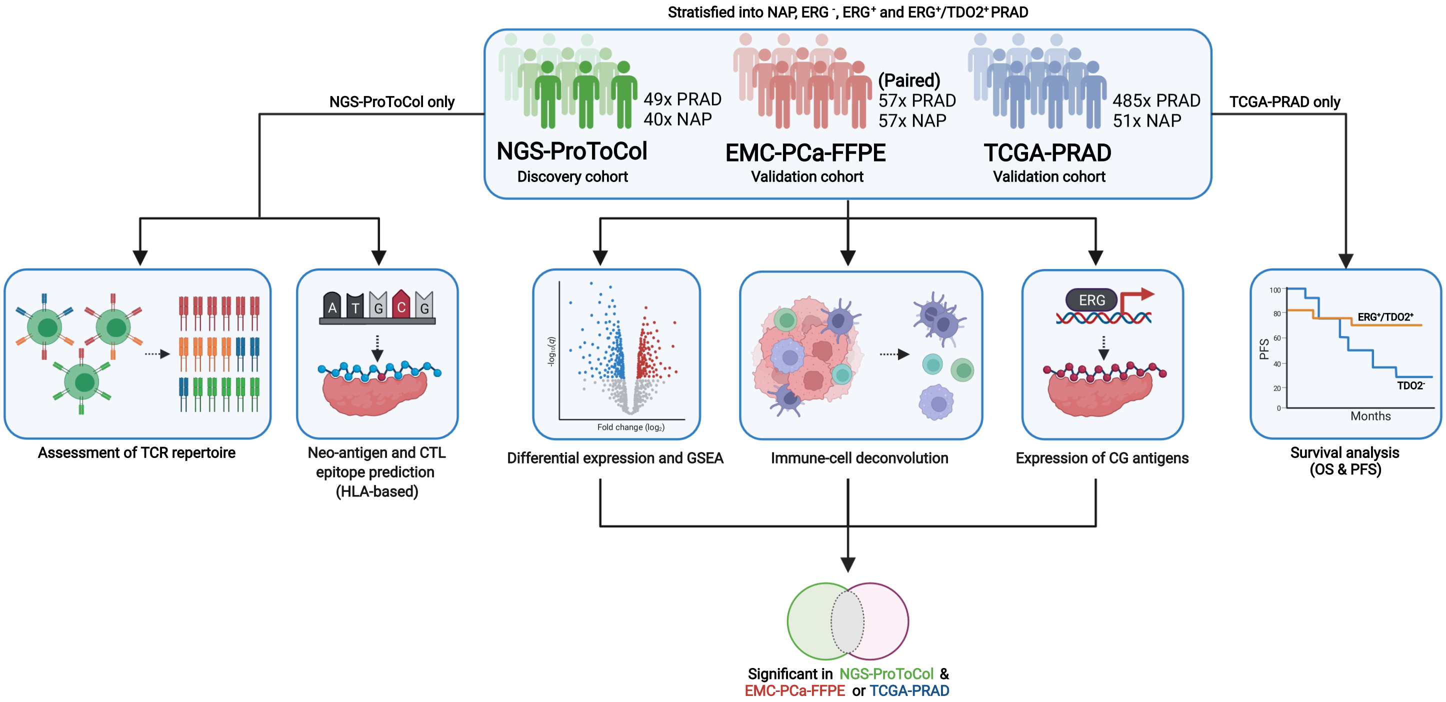

Stratisfied into NAP, ERG -, ERG+ and ERG+/TDO2+ PRAD
NGS-ProToCol only
49x PRAD
40x NAP
NGS-ProToCol
Discovery cohort
(Paired)
57x PRAD
57x NAP
EMC-PCa-FFPE
Validation cohort
485x PRAD
51x NAP
TCGA-PRAD
Validation cohort
TCGA-PRAD only
Assessment of TCR repertoire
A T G C G
Neo-antigen and CTL epitope prediction (HLA-based)
-log10(q)
Fold change (log2)
Differential expression and GSEA
Immune-cell deconvolution
ERG
Expression of CG antigens
PFS
100
80
60
40
20
0
ERG+/TDO2+
TDO2-
Months
Survival analysis (OS & PFS)
Significant in NGS-ProToCol &
EMC-PCa-FFPE or TCGA-PRAD

Figure 6.1: **Flowchart visualizing study design and aims.**
Three independent whole-transcriptome cohorts; NGS-ProToCol as discovery set and TCGA-PRAD and EMC-PCa-FFPE as validation cohorts were investigated for differential gene-expression (incl. cancer germline (CG) antigens) and gene-sets, immune-cell deconvolution (TIL10) based on stratification into NAP, *ERG*$^+$, *ERG*$^-$ and *ERG*$^+$/*TDO2*$^+$ PRAD. The TCR repertoire, neo-antigens and predicted CTL epitopes was assessed within NGS-ProToCol only whilst survival analysis was performed using the TCGA-PRAD cohort only.

subsequent analyses, we only retained statistically significant results which were obtained within the discovery cohort (NGS-ProToCol) and also within at least one of the additional validation cohorts.

These analyses confirmed the different overall expression profile(s) of PRAD in regard to NAP, with canonical PRAD-specific genes within the top results of differentially expressed genes (DEGs), such as *CRISPR3*, *ONECUT2*, *ALOX15*, *FOXB2*, *ANKRD34B*, *AMACR*, *PCAT5*, *PCAT7*, *EPCAM*, *TDRD1* and *ERG* among others (Figure 6.2a). In total, 2.610 DEGs were found between PRAD vs. NAP. In addition, GSEA revealed a myriad of perturbed pathways (Figure 6.2d), such as ribosomal biogenesis and metabolism of pyrimidine, aminoacyl-tRNA, sodium regulation and cytochrome P450. In addition, several intracellular and immune signaling pathways revolving around calcium, endothelial development, MYC, KRAS and WNT were observed.

Concordant with earlier studies, a considerable number of DEGs (n = 747) and perturbed gene-sets could be observed between *ERG*$^+$ vs. *ERG*$^-$ PRAD (Figure 6.2b,d). Amongst these DEGs and perturbed gene-sets are prominent immunological modulators, including higher expression of ribosomal biogenesis and MYC signaling within *ERG*$^+$ PRAD whilst conversely peroxisome proliferator-activated receptor (PPAR) signaling and the catabolism of β-alanine were higher expressed within *ERG*$^-$ PRAD. In addition, *TDO2* was revealed as the gene with the highest $\log_2$ fold-change ($\log_2$FC) within *ERG*$^+$ PRAD compared to *ERG*$^-$ PRAD, yet also harbored a relatively large $\log_2$FC standard error. This suggested two roughly-equally sized separate groups within the *ERG*$^+$ PRAD population; one with abundant expression of *TDO2* and one without or with only limited expression of *TDO2*. As TDO2 is the counterpart to IDO1 as a key rate-limiting factor of tryptophan to kynurenine metabolism, we characterized the differential expression of all members of the tryptophan to kynurenine catabolism pathway by comparing *ERG*$^+$/*TDO2*$^+$ PRAD vs. NAP. We observed additional dysregulation of crucial members , including increased expression of *AFMID* and *KMO* and decreased expression of *KYAT1* and *HAAO* (Supplementary Figure S6.3). Therefore, we added abundant *TDO2* expression within *ERG*$^+$ PRAD as additional stratification within downstream analysis (*ERG*$^+$/*TDO2*$^+$ PRAD).

Indeed, comparing *ERG*$^+$/*TDO2*$^-$ vs. *ERG*$^+$/*TDO2*$^+$ PRAD revealed DEGs and perturbed immuno-regulatory pathways exclusive to the *ERG*$^+$/*TDO2*$^+$ population, which are not readily observed when only comparing either *ERG*$^+$ (without *TDO2* status) vs. *ERG*$^-$ nor PRAD vs. NAP (Figure 6.2b-d). Notably, *ERG*$^+$/*TDO2*$^+$ PRAD showed enrichment of glycolysis and valine leucine and isoleucine metabolism, IL-2–STAT5, IFNγ and IFNα signaling, lysosome, MHC I complex and depletion of MYC and ribosomal biogenesis. In addition, we observed higher expression of known immunoregulatory genes (next to *TDO2*), such as *DAPP1* (Bam32)[30] and the immunogenic peptide *GP2* (glycoprotein 2)[31]. Furthermore, epithelial-mesenchymal transition (EMT) was found to be highly enriched within *ERG*$^+$/*TDO2*$^+$ PRAD.

ERG$^+$ vs. *ERG*$^-$ PRAD demonstrates enhanced frequencies of dendritic cells and M1 macrophages

We next investigated the relative frequency of intra-tumoral immune cell populations based on the established TIL10 transcriptome signatures[32] and compared this between PRAD vs. NAP and subsequently between PRAD based on *ERG* and/or *TDO2* status. Unsupervised clustering revealed clusters with distinct immune cell compositions between PRAD and NAP and additionally between PRAD based on *ERG* status (Figure 6.3). Statistically significant differences between PRAD and NAP (Figure 6.3j) demonstrated a lower frequency of neutrophils within PRAD, whereas frequencies of M1 and M2 macrophages, DC and NK cells were higher within PRAD. Notably, higher frequencies of M1 macrophages, DC, Tregs, NK cells could be attributed to *ERG*$^+$ and *ERG*$^+$/*TDO2*$^+$ status. Similar distinctive patterns of immune cell fractions were observed within the validation cohorts (Supplementary Figure S6.4).

ERG$^+$/*TDO2*$^+$ PRAD reveal increased diversity of TCR repertoire and neo-antigen burden

We next assessed and analyzed measures of tumor antigenicity, size and diversity of TCR repertoire, as well as the potential (neo)antigen load. Briefly, we investigated characteristics of the TCR repertoire using MiXCR[33] and the abundance of putative CTL neo-epitopes as identified by NetCTLPan[34].

We observed higher frequencies ($q < 0.1$) of the number of TCR-Vβ clonotypes, TCR-Vβ diversity and TCR-Vβ skewness between NAP vs. *ERG*$^+$/*TDO2*$^+$ PRAD but not between NAP vs. *TDO2*$^-$ PRAD (Figure 6.4a-c). We did note that the overall composition of the TCR repertoire is relatively limited as compared to ICI-sensitive tumors such as melanoma.[12] In addition, we observed no difference in TILs score[29]

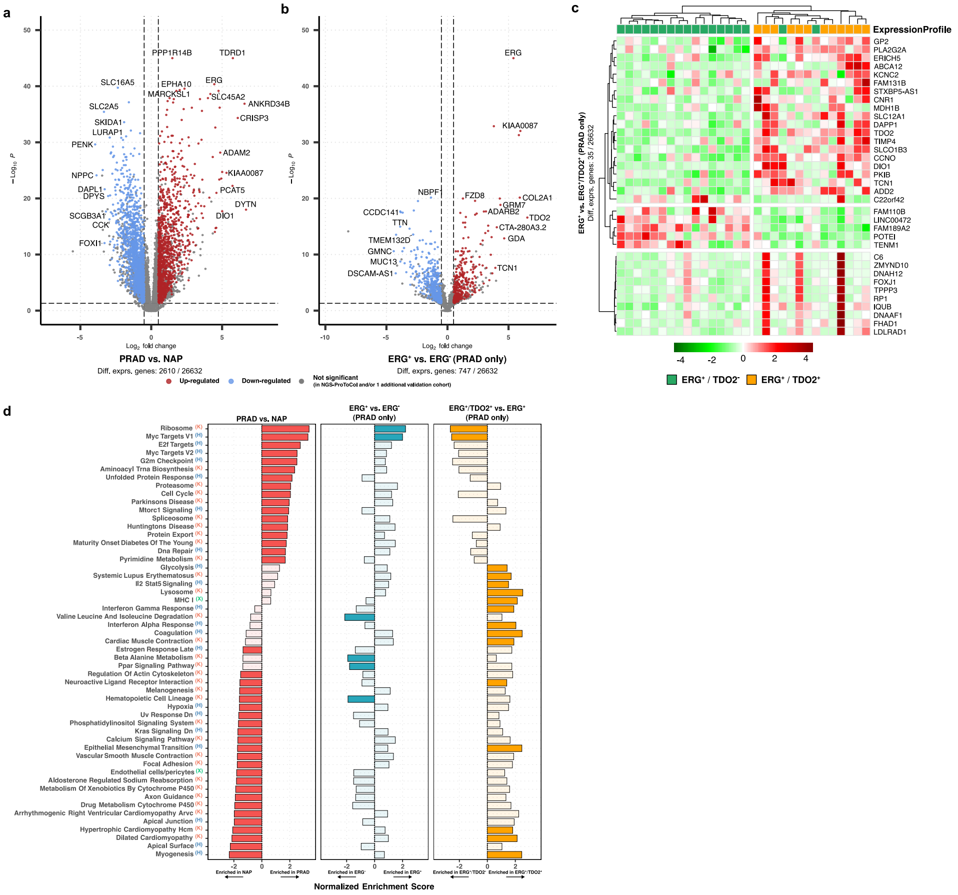


Figure 6.2: ***ERG*+ PRAD demonstrates increased gene expression of *TDO2* compared to *ERG*- PRAD, which is accompanied by pathways of antigen presentation and processing and IFN production.**
(a): Volcano-plot of the differential expression analysis between PRAD vs. NAP, genes significantly ($q \leq 0.05$, $\log_2\mathrm{FC} \geq |0.5|$, avg. read count ≥ 10 and also significant within ≥1 validation set) down-regulated (blue) and up-regulated (red) in PRAD are shown. The x-axis displays the $\log_2$FC and y-axis displays the adjusted p (q) shown on a $\log_{10}$ scale. The total amount of tested elements is shown below. **(b):** Same as **a)** but between *ERG*+ and *ERG*- PRAD. **(c):** Heatmap representing VST-corrected expression of DEGs between *ERG*+/*TDO2*+ vs. *ERG*+/*TDO2*- PRAD, shown as Z-scores. The upper tracks display the respective expression profiles. Columns and rows are clustered based on maximum distance metrics and the Ward.D2 unsupervised hierarchical clustering. **(d):** Significantly perturbed gene-sets ($q \leq 0.05$ and also significant within ≥1 validation set) between each performed analysis (as shown on top). The origin of each gene-set is shown as suffix to the description (K: KEGG, H: Hallmark, X: Hammerl et al.[29]). The Normalized Enrichment Scores (NES) is shown for each analysis and gene-set. Transparent NES were not found significant within the respective analysis, e.g., E2F targets was deemed significantly perturbed only between PRAD vs.NAP.

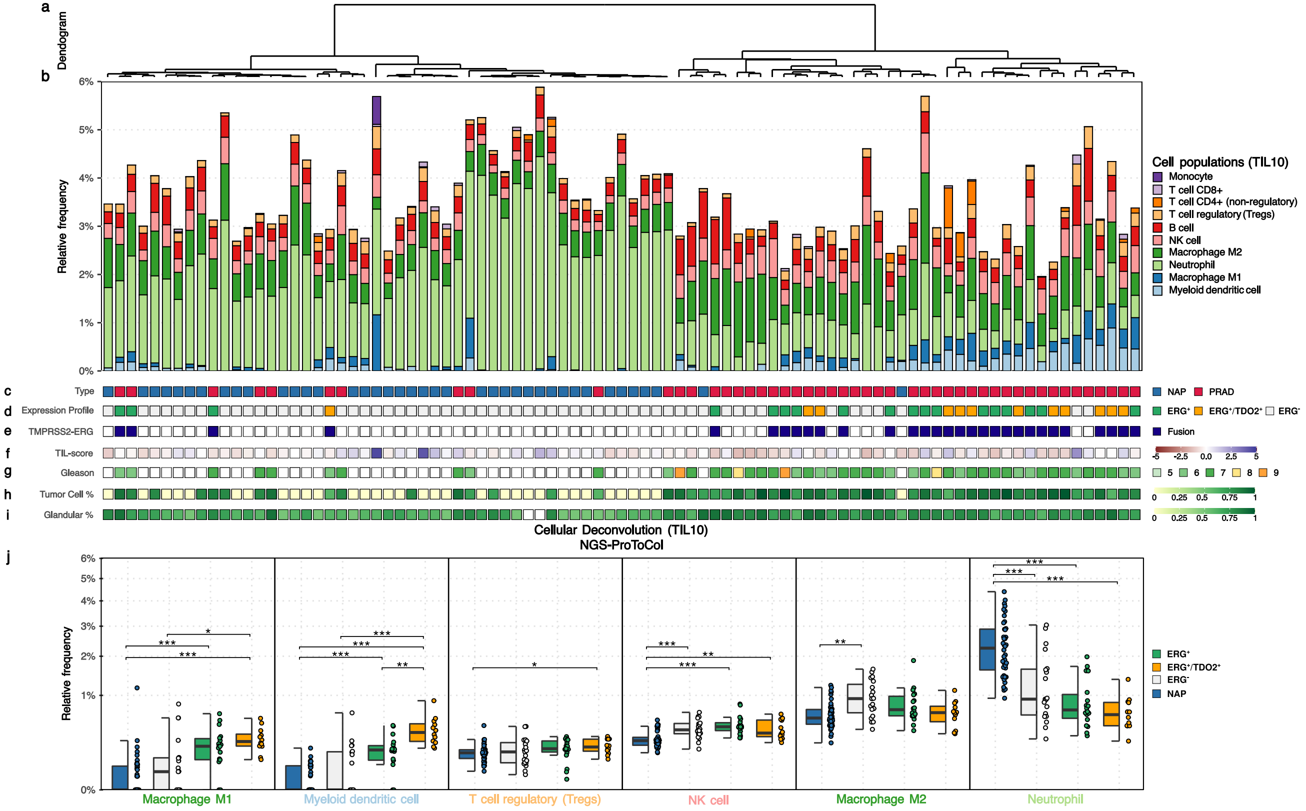


Figure 6.3: ***ERG***$^+$ **PRAD shows lowered frequencies of B cells, and enhanced frequencies of dendritic cells and M1 macrophages compared to** ***ERG***$^-$ **PRAD.**
Unsupervised clustering of immune-cell population for the discovery cohort (NGS-ProToCol; $n = 89$) according to quanTIseq using the TIL10 gene-signatures. **a**) Dendrogram of hierarchical unsupervised clustering (Ward.D2 method on Euclidean distances). **b**) Relative frequency of different immune cell populations (TIL10) per sample. **c**) Respective tissue type (NAP in blue; PRAD in red). **d**) *ERG* expression profiles (*ERG*$^+$ in green, *ERG*$^-$ in white). **e**) Presence of genomic *TMPRSS2-ERG* fusions (dark blue if present). **f**) Sample-wise TILs score. **g**) Gleason scores (Gleason 5 (light-green) to Gleason 9 (orange)). **h**) Tumor cell percentage (0% (yellow) to 100% (dark green)). **i**) Glandular cell percentage (0% (yellow) to 100% (dark green)). **j**) Boxplots with individual data points of the frequency of the TIL10 populations which were found statistically significant between NAP (blue), *ERG*$^-$ PRAD (grey), *ERG*$^+$ PRAD (green) and/or *ERG*$^+$/*TDO2*$^+$ PRAD (orange). Pairwise Wilcoxon Rank-Sum test against NAP with Benjamini-Hochberg correction was used to determine statistical significance with $q < 0.05$ (*); $q < 0.01$ (**); $q < 0.001$ (***).

between NAP and PRAD nor when including *ERG* and/or *TDO2* status (Figure 6.3f).

Further, we evaluated the presence of CTL neo-epitopes derived from variant proteins within PRAD. CTL neo-epitopes, denoted as either the distinct number of variant proteins containing ≥1 CTL neo-epitope(s) bound by HLA-A/B/C or as the total number of CTL neo-epitopes per sample, again revealed higher frequencies within *ERG*$^{+}$/*TDO2*$^{+}$ PRAD compared to *TDO2*$^{-}$ PRAD (Figure 6.4d-f). Concordant with the TCR repertoire assessment, the median quantity of CTL neo-epitopes is again relatively low compared to ICI-sensitive tumors such as melanoma.[12]

Compared to NAP and *ERG*$^{-}$ PRAD, *ERG*$^{+}$ PRAD shows higher expression for a subset of CGAs

We next investigated differential expression of CGAs (*n* = 247), another set of antigens recognized for its ability to elicit anti-tumor T cell responses. We observed several differentially expressed CGAs between NAP vs. PRAD and *ERG*$^{+}$ vs. *ERG*$^{-}$ PRAD (Figure 6.5). The gene expression of 6 CGAs (*PIWIL2*, *RGS22*, *PAGE4*, *LYPD6B*, *POTEG* and *POTEH*) was lower whilst conversely 9 CGAs (*TTK*, *KIF2C*, *CEP55*, *CCDC110*, *ADAM2*, *TDRD1*, *ARMC3*, *NOL4* and *TMEFF2*) were higher in PRAD vs. NAP. Between *ERG*$^{+}$ vs. *ERG*$^{-}$ PRAD, 3 CGAs (*POTEE*, *LYP6B* and *POTEH*) had lower expression whilst conversely *ADAM2* and *TDRD1* were up-regulated. Of note, only *POTEE* was found exclusively when taking *ERG* status into account; signifying the observed changes to CGAs in regard to PRAD *ERG* expression. No additional differences based on *TDO2* status was observed.

Stratification of PRAD based on abundant *ERG* and *TDO2* expression suggests an overall greater progression-free survival

Stratification of patients from the TCGA-PRAD cohort, based on abundant *ERG* and *TDO2* expression status (*ERG*$^{+}$/*TDO2*$^{+}$) vs. those without abundant *TDO2* expression (*TDO2*$^{-}$) revealed no significant difference in Overall survival (OS) yet does suggests a potentially more stable Progression-free survival (PFS), with log-rank p-values of 0.96 and 0.079, respectively (Supplementary Figure S6.5).

Discussion

Vaccine-based strategies, such as Sipuleucel-T (dendritic cells mediated) and Prostavac (T cell mediated) are currently the only FDA-approved immune therapies for (metastatic)

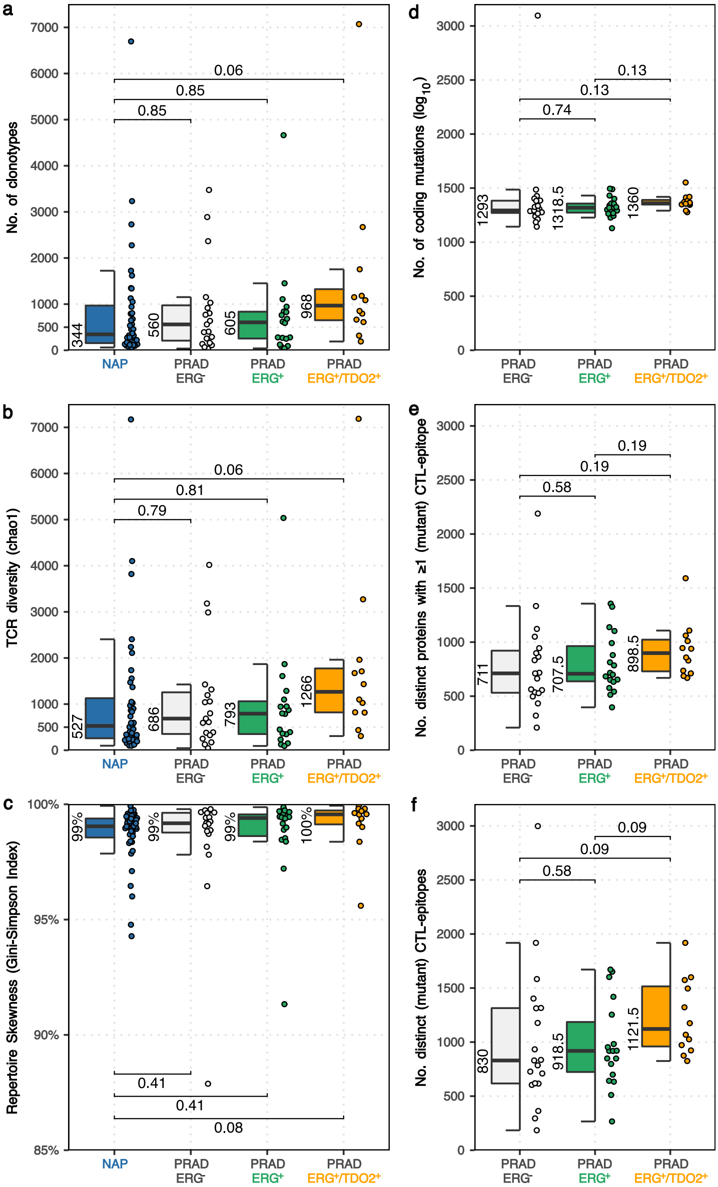


Figure 6.4: ***ERG*$^+$ / *TDO2*$^+$ PRAD reveals enrichment of size and diversity of TCR repertoire as well as expression of HLA-A/B/C-restricted neo-antigens.**
(a): Boxplot with individual data points showing the number of distinct TCR clonotypes (as TCR-Vb reads) between NAP and PRAD based on *ERG* and/or *TDO2* status. Pairwise Wilcoxon Rank-Sum tests vs. NAP with Benjamini-Hochberg correction were used to determine statistical significance, *q*-values are shown. Median values per group are indicated beside their respective boxplot. **(b):** Same as **a)** but reporting the TCR diversity (chao1 index). **(c):** Same as **a**) but reporting the repertoire skewness (Gini-Simpson index). **(d):** Boxplot with individual data points showing the number of the number of protein-coding mutations (incl. silent mutations) between PRAD based on *ERG* and/or *TDO2* status. Pairwise Wilcoxon Rank-Sum tests with Benjamini-Hochberg correction were used to determine statistical significance, *q*-values are shown. Median values per group are indicated beside their respective boxplot. **(e):** Same as **d**) but reporting the number of unique variant proteins harboring at least one predicted HLA-A2-binding epitope. **(f):** Same as **d**) but reporting the number of unique HLA-A2-binding epitopes from variant proteins.

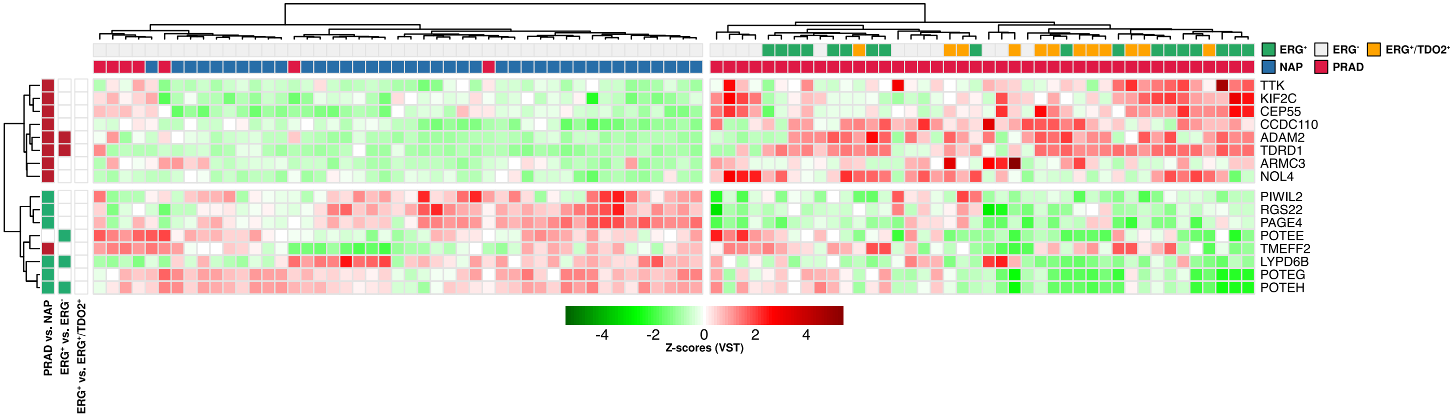


Figure 6.5: **A subset of CGAs reveal aberrant expression within PRAD and can attributed to *ERG* status**
Heatmap representing the variance-stabilizing transformation (VST)-corrected expression of differentially expressed CGAs, shown as Z-scores, between NAP vs. PRAD and between PRAD based on *ERG* and/or *TDO2* status. The upper tracks display the respective expression profiles and tissue type. Columns and rows are clustered based on maximum distance and the Ward.D2 unsupervised hierarchical clustering.

prostate cancer, of which presently a significant survival benefit is only observed for a small fraction of treated patients (5% patients, 10-13% increase in overall survival).[35] Interestingly, early-stage prostate cancer patients (compared to late stage mCRPC) show better clinical response to these therapies; perhaps caused by an increased number of immunosuppressive cells and genomic alterations in later stage tumors.[36]

In this current study, we interrogated *ERG*$^+$ PRAD for the presence of selective immune determinants to better understand therapy unresponsiveness and to potentially improve stratification of patients for therapies. Employing three independent large whole-transcriptome cohorts capturing PRAD and NAP, we investigated DEGs and perturbed gene-sets for immune evasive mechanisms, composition of immune-cell populations, the TCR repertoire, neo-antigen burden and their putative role as CTL epitopes, and aberrant expression of CGAs.

Concordant with previous studies[17–19,37,38], we observed major differences in the overall expression profiles of genes and gene-sets between NAP vs. PRAD and we could further attribute immunologically-related genes and gene-sets to *ERG*$^-$ and *ERG*$^+$ PRAD. Compared to *ERG*$^-$ PRAD and NAP, *ERG*$^+$ PRAD revealed distinctly lower expression of PPAR-signaling with higher expression of Wnt[39] and MYC signaling, coupled with ribosomal biogenesis among the perturbed gene-sets. Likewise, prominent differential expression distinctive to *ERG*$^+$ PRAD was found for a multitude of immunologically-related genes, including abundant expression of *FZD8*[40], *MYCL*[41], *FOXD3*[42] and *F5* coupled with lower expression of genes such as *WNT11*[43], *WIF1*[43] and *FABP5*[44].

Our findings extend the previously reported association between the deregulation of Wnt/PPAR-signaling and the tolerization (and induction) of DCs[39] as well as T-cell exclusion[45]. Furthermore, we observed differential expression of 15 CGAs between NAP and PRAD (regardless of *ERG* expression) and 5 CGAs between *ERG*$^-$ and *ERG*$^+$ PRAD. These differentially expressed CGAs in *ERG*$^+$ PRAD include *TDRD1*, a known downstream target of ERG.[46] As a CGA, *TDRD1* has predominant expression in tumorous and immune-privileged tissues and very low expression in healthy tissues; suggesting a possible role as target antigen for immune-based therapies.

Of particular interest was the discovery of a considerable subset of *ERG*$^+$ PRAD with abundant expression of *TDO2* (n = 12 out of 30; 40%). As critical rate-limiting factor (similar to IDO1) within the tryptophan to kynurenine catabolism, TDO2 functions to convert (L-)tryptophan into the immunosuppressive catabolite L-kynurenine (Supplementary Figure S6.3) which harbors considerable and remarkable immunosuppressive qualities. This includes counter-regulatory and tolerogenic effects such as activation of the aryl hydrocarbon receptor (AhR) and promoting the differentiation of tolerogenic DCs, TAMs and MDSCs.[47–52] Further investigation into *ERG*$^+$/*TDO2*$^+$ PRAD indeed revealed distinct high expression of known immunogenic gene-sets and genes including metabolic pathways, IL-2–STAT5 and IFN□signaling, which are pathways typically present or associated with activated T lymphocytes, as well as lysosome, MHC I complex, and IFNα signaling, which are pathways typically linked to antigen presentation and processing and activation of innate immune cells.[53–55] In addition, the MYC and ribosomal biogenesis pathways were both distinctly depleted in *ERG*$^+$/*TDO2*$^+$ PRAD whilst other forms of PRAD revealed enrichment of these pathways compared to NAP. MYC has been previously been reported as master regulator of ribosomal biogenesis[56] and regulator of T-cell activation[57]. In addition, EMT was found to be enriched within *ERG*$^+$/*TDO2*$^+$ PRAD and is known to be effected by the tumor microenvironment (TME).[58] Taken together, these pathways may hint that T cells, whilst also present at low to negligible numbers, are in an exhaustive, non-functional state within *ERG*$^+$/*TDO2*$^+$ PRAD.

Concordantly, we observed statistically significant differences within the relative immune-cell compositions. This included lower frequency of neutrophils within PRAD, whereas frequencies of M1 and M2 macrophages, DC and NK cells were higher within PRAD. Notably, higher frequencies of M1 macrophages, DC, Tregs, NK cells could be attributed to *ERG* expression and even more so to *ERG* coupled with *TDO2* expression. We also observed notable differences in the number of coding mutations, TCR repertoire and CTL neo-epitopes within *ERG*$^+$/*TDO2*$^+$ PRAD compared to NAP, whilst *TDO2*$^-$ PRAD revealed similar quantities.

The enhanced frequencies of M1 macrophages, Tregs, NK cells and specifically DCs, together with enrichment of antigen-processing and presentation, IL-2–STAT5 and IFN signaling within *ERG*$^+$/*TDO2*$^+$ PRAD suggest that these patients may benefit from targeted immunotherapies or T cell therapies. Within this line of reasoning, it is interesting to note that human monocyte-derived DCs are natural and potent producers of L-kynurenine, thereby reducing local tryptophan levels, limiting T-cell proliferation and functions and consequently promoting localized immune tolerance.[50–52]

Stratification of PRAD patients with and without abundant *TDO2* expression suggest that *TDO2*$^{+}$ PRAD patients suffer from a more stable disease compared to those lacking *TDO2* expression; further investigation into this sub-population might herald new findings relating to clinical benefit.

With our current study, we have revealed several promising new aspects of PRAD which should be validated using follow-up experiments. These findings are based on bulk whole-transcriptome sequencing data which obfuscates the cellular origin(s) of the observed DEGs and perturbed gene-sets. It would therefore be crucial to validate these findings using single-cell transcriptome and spatial sequencing or multiplex immunofluorescence staining to deduce whether these findings are wholly tumor-intrinsic or are facilitated by external factors within the TME.

In conclusion, our data points to significant differences regarding immune determinants between *ERG*$^{-}$ PRAD, *ERG*$^{+}$ PRAD and *ERG*$^{+}$/*TDO2*$^{+}$ PRAD. The revelation of a hitherto unidentified subgroup of PRAD with abundant and potentially immunosuppressive properties, facilitated by *TDO2* and the tryptophan to kynurenine catabolism, could spark new insights into the lackluster response rates of immune-therapies within PRAD. This could imply that inhibition of the kynurenine pathway enhances the effectiveness of ICI-based therapies within patients with *ERG*$^{+}$/*TDO2*$^{+}$ PRAD, a suggestion that warrants further research and verification.

Material and Methods

Sample acquisition and sequencing

NGS-ProToCol

For the CTMM NGS-ProToCol study[28], 50 localized PRAD and 40 NAP tissues from treatment-naïve patients within the Erasmus MC were snap-frozen and stored in liquid nitrogen as previously described by Hendriksen and colleagues.[59] The use of these samples for research purposes was approved by the Erasmus MC Medical Ethics Committee according to the Medical Research Involving Human Subjects Act (MEC-2004-261; MEC-2010-176). The RNA extraction and paired-end sequencing protocol has been described previously by Chen et al.[60] and the respective sequencing data (whole-transcriptome) and clinical information were retrieved from the European Genome-phenome Archive (EGA)[61] under accession EGAS00001002816.

We performed an additional screening of the in-house clinical records and retained 49 PRAD and 40 NAP tissues after resolving conflicting pathological records. This cohort has been used as discovery cohort and significant findings were validated in two additional validation cohorts (EMC-PCa-FFPE and TCGA-PRAD).

EMC-PCa-FFPE

As validation cohort, we have utilized PRAD and matching NAP tissues (n = 57 paired samples) from treatment-naïve patients within the EMC-PCa-FFPE cohort. Use of samples for research purposes was approved by the Erasmus MC Medical Ethics Committee (MEC-2004-261). These patient samples represent two groups with matching of age and histopathological characteristics, but having different outcomes after radical prostatectomy with long event-free follow-up (n = 38) and poor outcome, defined as time to metastatic disease ≤ 7.6 years and/or time to death from prostate cancer ≤ 10 years (n = 19). For this cohort, RNA was extracted from FFPE radical prostatectomy tissue blocks (the Erasmus MC, the Netherlands) followed by rRNA-depletion and random-priming for Illumina paired-end sequencing to an min. depth of 60 million reads per sample (performed by Aros Applied Biotechnology A/S, Aarhus, Denmark).

TCGA-PRAD

As additional validation cohort, we have used the TCGA-PRAD cohort from which transcriptome profiling data (GRCh38.p12; HTSeq-counts and HTSeq-FPKM) was collected (in verbatim) from the Genomic Data Commons (GDC) using TCGAbiolinks (v2.20.0)[62]. Only primary PRAD (n = 485) and NAP tissues (n = 51) were retained and used in all subsequent analysis. In addition, clinical data capturing overall survival and progression-free survival was *in verbatim* collected from the Genomic Data Commons (GDC).

Pre-processing and alignment of NGS-ProToCol and EMC-PCa-FFPE

Raw paired-end sequencing data of the NGS-ProToCol and EMC-PCa-FFPE cohorts were pre-processed using fastp[63] (v0.20.0) to remove leftover sequencing adapters (TruSeq3) and perform low-quality filtering using the following command and parameters:

```
fastp --detect_adapter_for_pe -L --html --thread 5 --in1 <R1> --in2
     <R2> --out1 <R1.gz> --out2 <R2.gz>
```

The trimmed reads were aligned to the human reference (GRCh38.p13) using STAR[64] (2.7.9a) with genomic annotations from GENCODE release 38[65]. Per sample, all lanes were aligned simultaneously and annotated with read groups using the following command and parameters:

```
STAR --genomeDir <genome> --readFilesIn <R1.gz> <R2.gz> --
    readFilesCommand zcat --outFileNamePrefix <sampleID_> --
    outSAMtype BAM SortedByCoordinate --outSAMunmapped Within --
    outSAMattributes NH HI AS nM NM MD jM jI MC ch XS --
    outSAMstrandField intronMotif --outFilterMultimapNmax 10 --
    outFilterMismatchNmax 3 --limitOutSJcollapsed 3000000 --
    chimSegmentMin 10 --chimOutType WithinBAM SoftClip --
    chimJunctionOverhangMin 10 --chimSegmentReadGapMax 3 --
    chimScoreMin 1 --chimScoreDropMax 30 --chimScoreJunctionNonGTAG
     0 --chimScoreSeparation 1 --outFilterScoreMinOverLread 0.33 --
    outFilterMatchNminOverLread 0.33 --outFilterMatchNmin 35 --
    alignSplicedMateMapLminOverLmate 0.33 --alignSplicedMateMapLmin
     35 --alignSJstitchMismatchNmax 5 -1 5 5 --twopassMode Basic --
    twopass1readsN -1 --runThreadN 10 --limitBAMsortRAM 25000000000
     --quantMode TranscriptomeSAM --outSAMattrRGline <read group>
```

Generation of alignment quality metrics (flagstat) and duplicate reads marking was performed by Sambamba[66] (v0.7.0). FeatureCounts[67] (v2.0.2) was used to generate raw read count tables using annotations from GENCODE release 38[65]; only primary (uniquely mapped) reads were counted per exon and summarized per gene using paired-end modus and with respective strand-specific modus.

RSEM[68] (v1.3.2) was used to quantify, with respective strand-specific modus, RNA expression into transcripts per million (TPM) values for use in downstream immune-population deconvolution analysis.

For the TCGA-PRAD, HTSeq-FPKM values were converted to TPM values for use in downstream analysis.

Detection and heuristic filtering of somatic variants within NGS-ProToCol.

As matched NAP data for the NGS-ProToCol dataset was not available, we performed variant calling on the PRAD and NAP samples separately and employed a panel of normals (PON) design to filter possible germline variants and artifacts. Following the suggested best practices and workflow for variant detection for RNA-Seq using Genome Analysis Toolkit (GATK) (v4.2.2.0)[69], we performed the SplitNCigarReads and base-quality recalibration as suggested for all samples. Subsequently, we performed germline variant calling (exons-only) utilizing Haplotyper for all NAP samples using only the primary-aligned and non-duplicated reads. In addition, we performed MuTect2[70] (v4.2.2.0) on the PRAD samples to detect somatic variants (exons-only) using only primary-aligned and non-duplicated reads. Possible sequencing artifacts, as detected by Mutect2, were removed prior to downstream analysis (PASS-only). Additional annotation was performed by Variant Effect Predictor (VEP) (v104)[71] using the default picking scheme.

Using the germline variant calling on the NAP (n = 40), we generated a PON of potential germline variants by employing the following heuristic filtering to retain variants; present in ≥5 NAP, a total read depth of ≥10 reads and with ≥3 supporting reads for the alternate allele. Using the PON, we subsequently filtered the sample-specific somatic variants within the PRAD samples (n = 49) by only retaining the somatic variants satisfying the following thresholds; not present within the PON, a total read depth of ≥10 reads with ≥3 supporting reads for the alternate allele, an allele frequency ≥0.1 and they should not be harbor a gnomAD (v2.0.1)[72] exome and genome allele frequency of ≥0.0008 (100 out of 125748 exomes) and/or ≥0.006 (100 out of 15708 genomes). In addition, we only retained protein-coding somatic variants.

Assessment of TMPRSS2-ERG fusion status.

Arriba[73] (v2.1.0)-mediated analysis of the aligned Binary Alignment Map (BAM) files with respective reference genome and annotations (GRCh38.p13 and GENCODE v38), default blacklist and known fusion-gene lists, enabled assessment of the presence of known *TMPRSS2-ERG* fusion-gene transcripts within the NGS-ProToCol and EMC-PCa-FFPE cohorts. In addition, genomic *TMPRSS2-ERG* fusions for the NGS-ProToCol cohort were also known from previous exome-based experiments.[59] For the TCGA-PRAD cohort, the existing annotations for the *TMPRSS2-ERG* fusion-gene was used as denoted on GDC.

Determining expression profiles of samples.

Using the DESeq2-normalized counts with additional VST[74] for expression of *ERG* and two canonical downstream *ERG*-regulated genes (*PCAT5* and *TDRD1*), we employed unsupervised hierarchical clustering (Euclidean distance and Ward.D2 method) and split the discovery and validation cohorts into two groups ($k = 2$) based on the earliest branchpoint. Respective to the expression of these *ERG*-markers, we then labeled these group as *ERG*$^{+}$ and *ERG*$^{-}$; NAP samples were denoted as NAP regardless of clustering. In addition, we performed an identical approach using only the normalized and VST-transformed expression of *TDO2* to denote *TDO2*$^{+}$ and *TDO2*$^{-}$ samples. We then combined both *ERG* and *TDO2* status to denote *ERG*$^{+}$/*TDO2*$^{+}$, *ERG*$^{+}$/*TDO2*$^{-}$ and *ERG*$^{-}$/*TDO2*$^{-}$ samples.

Assessment of frequencies of immune cell populations

QuanTIseq[32], as implemented in the immunedeconv R package (v2.0.3)[75], was performed with default TIL10 gene-signatures (136 out of 138 TIL10 genes were matched) using a constrained Least Squares with Equality and Inequality (lsei) with mRNA scaling and filtering of signature genes which are known to be highly expressed in tumor samples (*NUPR1*, *CD36*, *CSTA*, *HPGD*, *CFB*, *ECM1*, *FCGBP*, *PLTP*, *FXYD6*, *HOPX*, *SERPING1*, *ENPP2*, *GATM*, *PDPN*, *ADAM6*, *FCRLA*, *SLC1A3*). The TIL10 gene-signature is capable of identifying B cells, classically-activated (M1) macrophages, alternatively-activated (M2) macrophages, monocytes, neutrophils, natural killer (NK) cells, non-regulatory (helper) CD4+ T cells, cytotoxic CD8+ T cells, regulatory CD4+ T (Treg) cells, myeloid dendritic cells and other uncharacterized cells.[32] Samples were subsequently clustered using unsupervised hierarchical clustering (Euclidean distance and Ward.D2 method) according to the frequencies of all immune-cell populations after scaling based on the standard deviation.

In addition, we calculated the TIL score as previously detailed by Hammerl et al. (2020)[29]. Briefly, we calculated a per-sample average of a list of 119 genes using their respective gene-wise TPM values.

Assessment of T cell receptor repertoire

MiXCR[33] (v3.0.13) was performed to estimate the T cell receptor repertoires using default settings and following the best-practices for RNA-Seq using the Fastp-trimmed reads. Subsequent analysis was performed with the immunarch (v0.6.6) package in R.[76] We quantified the number of clonotypes using the 'volume' method, TCR diversity using the 'chao1' method and repertoire skewness using the 'gini.simp' methods within the immunarch package.

Assessment of neo-antigens in NGS-ProToCol.

Single nucleotide variations (SNV), insertions and deletions (InDels), as well as multi-nucleotide variants (MNV) detected in the NGS-ProToCol samples were *in silico* incorporated into their respective mRNA transcripts and translated to variant protein sequences with ProteoDisco (v1.1.3))[77] in a sample-specific manner. Briefly, per overlapping transcript(s), mutations were incorporated into their respective coding sequence (GRCh38.p13) based on GENCODE (v38) annotations and were subsequently *in silico* translated into their corresponding protein variant sequence(s). All mutations were incorporated simultaneously per transcript, i.e., multiple mutations were incorporated within the same mutant protein sequence.

Using the default workflow of arcasHLA[78] (v0.2.5), the *HLA-A*, *HLA-B* and *HLA-C* alleles of each sample was determined from the STAR-aligned BAM files. Briefly, potential *HLA*-related reads were retrieved from known *HLA* loci and supplemented with all unmapped reads and used to determine the most likely HLA-genotype after pseudo-alignment with Kallisto (v0.46.1)[79]. Only *HLA*-alleles observed with ≥10 reads were taken along.

Per sample, netCTLpan[34] (v1.1) was used to determine CTL epitopes (9mers) of the *in silico* derived protein variants sequences (from ProteoDisco) using their sample-respective *HLA-A*, *HLA-B* and *HLA-C* alleles. From the netCTLpan output, we only retained the CTL epitopes which had a rank <1 for any of the three sample-specific *HLA*-alleles. CTL epitopes originating from canonical protein sequences (wild-type UniProtKB / Swiss-Prot)[80] were removed by overlapping the CTL peptides to all wild-type protein sequences and removing any CTL peptides which fully matched within any canonical protein sequence. In turn, this generated a list of potential CTL neo-epitopes.

Differential gene-expression analysis.

Differential gene-analysis of whole-transcriptome data (NGS-ProToCol, TCGA-PRAD and EMC-PCa-FFPE) was performed with DESeq2[74] (v1.32.0). To correct for multiple hypothesis testing after DESeq2 analysis, we employed independent hypothesis weighting (IHW) (v1.20.0)[81]. Fold-changes (log2) were shrunken using their respective coefficient using adaptive shrinkage estimator (ashr)[82]. Per cohort, differential genes were selected based on the following criteria: adjusted p (q) ≤ 0.05; $\log_2$FC $\geq |0.5|$; $\log_2$FCstandard error ≤ 1 (unless $\log_2$FC $\geq |3|$); and an average read count ≥ 25 over all samples with the respective analysis. Only genes found to be differentially expressed (using the above threshold) within the NGS-ProToCol cohort and at least one additional validation cohort (TCGA-PRAD, EMC-PCa-FFPE) were included into the final list of DEGs.

Gene-set enrichment analysis.

Canonical pathways from KEGG[83] (C2) and hallmark gene-sets (H) were obtained from MSigDB (v7.4; accessed on 25-08-2021)[84]. In addition, immune-related gene-sets (1,474 genes in 37 distinct gene-sets, which were based on literature and in-house data) were selected from Hammerl et al. (2020)[29] and further annotated with gene-identifiers and screened for erroneous misclassifications. For all the gene-sets, we only retained gene-sets with at least 15 genes and fewer than 300 genes resulting in a total number of 255 gene-sets (KEGG, Hallmark and Hammerl et al.). GSEA on Wald statistics per cohort (derived from DESeq2) were used as input within the fgsea package[85] (v1.19.2). We only retained Wald statistics from genes with an average read count ≥ 10 over all samples within the respective analysis. The final list of significantly enriched gene-sets was selected based on the following criteria: adjusted $p \leq 0.05$ within the NGS-ProToCol cohort; and at least one additional validation cohort (TCGA-PRAD, EMC-PCa-FFPE).

Overview of the tryptophan and kynurenine metabolism pathway

The tryptophan metabolism pathway relating to kynurenine, was downloaded from KEGG86 (map00380) using the pathview R package[86] (v1.13.1) and the underlying genes were annotated with their respective $\log_2$FC (derived from the aforementioned DESeq2 analysis) between *ERG*$^+$/*TDO2*$^+$ PRAD vs. NAP.

Data and code availability

All processed data can be obtained from authors upon request. The raw data for the discovery cohort (NGS-ProToCol) is available from the EGA[59] with accession number EGAS00001002816. Processed data from the EMC-PCa-FFPE cohort can be offered upon request. Processed transcriptome profiling data (GRCh38.p12; HTSeq-counts and HTSeq-FPKM) from the TCGA-PRAD cohort was collected from the Genomic Data Commons (GDC) using TCGAbiolinks (v2.20.0)[62].

Statistical analysis

All analysis was performed with the statistical language platform R (v4.1.1)[87]. Unless stated otherwise, pairwise Wilcoxon Rank-Sum tests with additional Benjamini-Hochberg multiple-testing correction was used to test the statistical significance of differences between conditions. An adjusted p (q) < 0.05 was used to considered statistically significant differences. When q-values shown in figures, we have used the following annotations: $q < 0.05$ (*); $q < 0.01$ (**); $q < 0.001$ (***). Kaplan-Meijer univariate analysis (log-rank test) was performed using the survminer R package (v0.4.9).

Acknowledgments

This study was supported by the framework of CTMM (The Netherlands) NGS ProToCol (grant 03O-402) and we thank the NGS-ProToCol consortium for the sample series, performing the sequencing and critical discussions. The EMC-PCa-FFPE study was funded by Janssen Pharmaceutica (ICD446026; 2012) and we particularly thank Dr. Hugo Ceulemans as the Janssen Project Manager. The results [published or shown] here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. Figure 1 has been created with BioRender.com.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, *Cancer statistics, 2020,* CA: A Cancer Journal for Clinicians **70**, 7 (2020).

[2] J. Ferlay, M. Colombet, I. Soerjomataram, D. M. Parkin, M. Piñeros, *et al.*, *Cancer statistics for the year 2020: An overview,* International Journal of Cancer (2021).

[3] J. L. Mohler, E. S. Antonarakis, A. J. Armstrong, A. V. D'Amico, B. J. Davis, *et al.*, *Prostate cancer, version 2.2019, nccn clinical practice guidelines in oncology,* Journal of the National Comprehensive Cancer Network **17**, 479 (2019).

[4] L. F. van Dessel, J. van Riet, M. Smits, Y. Zhu, P. Hamberg, *et al.*, *The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact,* Nature communications **10**, 1 (2019).

[5] S. Menon, S. Shin, and G. Dy, *Advances in cancer immunotherapy in solid tumors,* Cancers **8**, 106 (2016).

[6] P. Isaacsson Velho and E. S. Antonarakis, *Pd-1/pd-l1 pathway inhibitors in advanced prostate cancer,* Expert review of clinical pharmacology **11**, 475 (2018).

[7] E. D. Kwon, C. G. Drake, H. I. Scher, K. Fizazi, A. Bossi, *et al.*, *Ipilimumab versus placebo after radiotherapy in patients with metastatic castration-resistant prostate cancer that had progressed after docetaxel chemotherapy (ca184-043): a multicentre, randomised, double-blind, phase 3 trial,* The lancet oncology **15**, 700 (2014).

[8] S. L. Topalian, F. S. Hodi, J. R. Brahmer, S. N. Gettinger, D. C. Smith, *et al.*, *Safety, activity, and immune correlates of anti–pd-1 antibody in cancer,* New England Journal of Medicine **366**, 2443 (2012).

[9] A. Hansen, C. Massard, P. Ott, N. Haas, J. Lopez, *et al.*, *Pembrolizumab for advanced prostate adenocarcinoma: findings of the keynote-028 study,* Annals of Oncology **29**, 1807 (2018).

[10] J. De Bono, J. Goh, and K. Ojamaa, *Keynote-199: Pembrolizumab (pembro) for docetaxel-refractory metastatic castration-resistant prostate cancer (mcrpc),* J Clin Oncol **36** (2018).

[11] C. Caruso, *Anti–pd-1–ctla4 combo hits prostate cancer,* (2019).

[12] L. Alexandrov, S. Nik-Zainal, D. Wedge, S. Aparicio, S. Behjati, *et al.*, *Signatures of mutational processes in human cancer,* Nature **500**, 415 (2013).

[13] N. Vitkin, S. Nersesian, D. R. Siemens, and M. Koti, *The tumor immune contexture of prostate cancer,* Frontiers in immunology **10**, 603 (2019).

[14] S. G. Zhao, J. Lehrer, S. L. Chang, R. Das, N. Erho, *et al.*, *The immune landscape of prostate cancer and nomination of pd-l2 as a potential therapeutic target,* JNCI: Journal of the National Cancer Institute **111**, 301 (2019).

[15] C. S. Grasso, Y.-M. Wu, D. R. Robinson, X. Cao, S. M. Dhanasekaran, *et al.*, *The mutational landscape of lethal castration-resistant prostate cancer,* Nature **487**, 239 (2012).

[16] Y.-M. Wu, M. Cieślik, R. Lonigro, P. Vats, M. Reimers, *et al.*, *Inactivation of cdk12 delineates a distinct immunogenic class of advanced prostate cancer,* Cell **173**, 1770 (2018).

[17] A. Abeshouse, J. Ahn, R. Akbani, A. Ally, S. Amin, *et al.*, *The molecular taxonomy of primary prostate cancer,* Cell **163**, 1011 (2015).

[18] M. Fraser, V. Sabelnykova, T. Yamaguchi, L. Heisler, J. Livingstone, *et al.*, *Genomic hallmarks of localized, non-indolent prostate cancer,* Nature **541**, 359 (2017).

[19] S. Viswanathan, G. Ha, A. Hoff, J. Wala, J. Carrot-Zhang, *et al.*, *Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing,* Cell **174**, 433 (2018).

[20] L. Jerby-Arnon, P. Shah, M. S. Cuoco, C. Rodman, M.-J. Su, *et al.*, *A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade,* Cell **175**, 984 (2018).

[21] G. Abril-Rodriguez, D. Y. Torrejon, W. Liu, J. M. Zaretsky, T. S. Nowicki, *et al.*, *Pak4 inhibition improves pd-1 blockade immunotherapy,* Nature Cancer **1**, 46 (2020).

[22] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, *et al.*, *Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer,* science **310**, 644 (2005).

[23] S. A. Tomlins, B. Laxman, S. Varambally, X. Cao, J. Yu, *et al.*, *Role of the tmprss2-erg gene fusion in prostate cancer,* Neoplasia (New York, NY) **10**, 177 (2008).

[24] J. Rubio-Briones, A. Fernández-Serra, A. Calatrava, Z. García-Casado, L. Rubio, *et al.*, *Clinical implications of tmprss2-erg gene fusion expression in patients with prostate cancer treated with radical prostatectomy,* The Journal of urology **183**, 2054 (2010).

[25] C. Hägglöf, P. Hammarsten, K. Strömvall, L. Egevad, A. Josefsson, *et al.*, *Tmprss2-erg expression predicts prostate cancer survival and associates with stromal biomarkers,* PloS one **9**, e86824 (2014).

[26] R. K. Nam, L. Sugar, Z. Wang, W. Yang, R. Kitching, *et al.*, *Expression of tmprss2: Erg gene fusion in prostate cancer cells is an important prognostic factor for cancer progression,* Cancer biology & therapy **6**, 40 (2007).

[27] S. R. Rao, N. K. Alham, E. Upton, S. McIntyre, R. J. Bryant, *et al.*, *Detailed molecular and immune marker profiling of archival prostate cancer samples reveals an inverse association between tmprss2: Erg fusion status and immune cell infiltration,* The Journal of Molecular Diagnostics **22**, 652 (2020).

[28] CETMM, *Next generation sequencing from prostate to colorectal cancer - center for translational molecular medicine,* (2019).

[29] D. Hammerl, M. P. Massink, M. Smid, C. H. van Deurzen, H. E. Meijers-Heijboer, *et al.*, *Clonality, antigen recognition, and suppression of cd8+ t cells differentially affect prognosis of breast cancer subtypes,* Clinical Cancer Research **26**, 505 (2020).

[30] L. Hao, A. J. Marshall, and L. Liu, *Suppressive role of bam32/dapp1 in chemokine-induced neutrophil recruitment,* International journal of molecular sciences **22**, 1825 (2021).

[31] L. Werner, D. Paclik, C. Fritz, D. Reinhold, D. Roggenbuck, *et al.*, *Identification of pancreatic glycoprotein 2 as an endogenous immunomodulator of innate and adaptive immune responses,* The Journal of Immunology **189**, 2774 (2012).

[32] F. Finotello, C. Mayer, C. Plattner, G. Laschober, D. Rieder, *et al.*, *Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of rna-seq data,* Genome medicine **11**, 1 (2019).

[33] D. A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, *et al.*, *Mixcr: software for comprehensive adaptive immunity profiling,* Nature methods **12**, 380 (2015).

[34] T. Stranzl, M. V. Larsen, C. Lundegaard, and M. Nielsen, *Netctlpan: pan-specific mhc class i pathway epitope predictions,* Immunogenetics **62**, 357 (2010).

[35] M. L. Cheng and L. Fong, *Beyond sipuleucel-t: immune approaches to treating prostate cancer,* Current treatment options in oncology **15**, 115 (2014).

[36] G. Bryant, L. Wang, and D. J. Mulholland, *Overcoming oncogenic mediated tumor immunity in prostate cancer,* International journal of molecular sciences **18**, 1542 (2017).

[37] O. R. Saramäki, A. E. Harjula, P. M. Martikainen, R. L. Vessella, T. L. Tammela, *et al.*, *Tmprss2: Erg fusion identifies a subgroup of prostate cancers with a favorable prognosis,* Clinical cancer research **14**, 3395 (2008).

[38] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, *et al.*, *Integrative genomic profiling of human prostate cancer,* Cancer cell **18**, 11 (2010).

[39] F. Zhao, C. Xiao, K. S. Evans, T. Theivanthiran, N. DeVito, *et al.*, *Paracrine wnt5a-β-catenin signaling triggers a metabolic program that drives dendritic cell tolerization,* Immunity **48**, 147 (2018).

[40] V. Murillo-Garzón, I. Gorroño-Etxebarria, M. Åkerfelt, M. C. Puustinen, L. Sistonen, *et al.*, *Frizzled-8 integrates wnt-11 and transforming growth factor-β signaling in prostate cancer,* Nature communications **9**, 1 (2018).

[41] C. V. Dang, *Myc on the path to cancer,* Cell **149**, 22 (2012).

[42] Y. Zhang, Z. Wang, H. Xiao, X. Liu, G. Zhu, *et al.*, *Foxd3 suppresses interleukin-10 expression in b cells,* Immunology **150**, 478 (2017).

[43] I. Ramachandran, V. Ganapathy, E. Gillies, I. Fonseca, S. Sureban, *et al.*, *Wnt inhibitory factor 1 suppresses cancer stemness and induces cellular senescence,* Cell death & disease **5**, e1246 (2014).

[44] S. Kobayashi, T. Wannakul, K. Sekino, Y. Takahashi, Y. Kagawa, *et al.*, *Fatty acid-binding protein 5 limits the generation of foxp3+ regulatory t cells through regulating plasmacytoid dendritic cell function in the tumor microenvironment,* International Journal of Cancer **150**, 152 (2022).

[45] S. Spranger, R. Bao, and T. F. Gajewski, *Melanoma-intrinsic β-catenin signalling prevents anti-tumour immunity,* Nature **523**, 231 (2015).

[46] J. L. Boormans, H. Korsten, A. J. Ziel-van der Made, G. J. van Leenders, C. V. de Vos, *et al.*, *Identification of tdrd1 as a direct target gene of erg in primary prostate cancer,* International journal of cancer **133**, 335 (2013).

[47] J.-P. Routy, B. Routy, G. M. Graziani, and V. Mehraj, *The kynurenine pathway is a double-edged sword in immune-privileged sites and in cancer: implications for immunotherapy,* International Journal of Tryptophan Research **9**, IJTR (2016).

[48] T. A. Triplett, K. C. Garrison, N. Marshall, M. Donkor, J. Blazeck, *et al.*, *Reversal of indoleamine 2, 3-dioxygenase–mediated cancer immune suppression by systemic kynurenine depletion with a therapeutic enzyme,* Nature biotechnology **36**, 758 (2018).

[49] P. Terness, T. M. Bauer, L. Röse, C. Dufter, A. Watzlik, *et al.*, *Inhibition of allogeneic t cell proliferation by indoleamine 2, 3-dioxygenase–expressing dendritic cells: mediation of suppression by tryptophan metabolites,* The Journal of experimental medicine **196**, 447 (2002).

[50] C. A. Opitz, U. M. Litzenburger, F. Sahm, M. Ott, I. Tritschler, *et al.*, *An endogenous tumour-promoting ligand of the human aryl hydrocarbon receptor,* Nature **478**, 197 (2011).

[51] M. Platten, W. Wick, and B. J. Van den Eynde, *Tryptophan catabolism in cancer: beyond ido and tryptophan depletion,* Cancer research **72**, 5435 (2012).

[52] U. Grohmann, F. Fallarino, and P. Puccetti, *Tolerance, dcs and tryptophan: much ado about ido,* Trends in immunology **24**, 242 (2003).

[53] N. Ron-Harel, J. M. Ghergurovich, G. Notarangelo, M. W. LaFleur, Y. Tsubosaka, *et al.*, *T cell activation depends on extracellular alanine,* Cell reports **28**, 3011 (2019).

[54] F. Castro, A. P. Cardoso, R. M. Gonçalves, K. Serre, and M. J. Oliveira, *Interferon-gamma at the crossroads of tumor immune surveillance or evasion,* Frontiers in immunology **9**, 847 (2018).

[55] T. Welte, D. Leitenberg, B. N. Dittel, B. K. Al-Ramadi, B. Xie, *et al.*, *Stat5 interaction with the t cell receptor complex and stimulation of t cell proliferation,* Science **283**, 222 (1999).

[56] J. Van Riggelen, A. Yetil, and D. W. Felsher, *Myc as a regulator of ribosome biogenesis and protein synthesis,* Nature Reviews Cancer **10**, 301 (2010).

[57] J. M. Marchingo, L. V. Sinclair, A. J. Howden, and D. A. Cantrell, *Quantitative analysis of how myc controls t cell proteomes and metabolic pathways during t cell activation,* Elife **9**, e53725 (2020).

[58] P. J. Chockley and V. G. Keshamouni, *Immunological consequences of epithelial–mesenchymal transition in tumor progression,* The Journal of Immunology **197**, 691 (2016).

[59] P. J. Hendriksen, N. F. Dits, K. Kokame, A. Veldhoven, W. M. van Weerden, *et al.*, *Evolution of the androgen receptor pathway during progression of prostate cancer,* Cancer research **66**, 5012 (2006).

[60] S. Chen, V. Huang, X. Xu, J. Livingstone, F. Soares, *et al.*, *Widespread and functional rna circularization in localized prostate cancer,* Cell **176**, 831 (2019).

[61] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, *et al.*, *The european nucleotide archive,* Nucleic acids research **39**, D28 (2010).

[62] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, *et al.*, *Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data,* Nucleic acids research **44**, e71 (2016).

[63] S. Chen, Y. Zhou, Y. Chen, and J. Gu, *fastp: an ultra-fast all-in-one fastq preprocessor,* Bioinformatics **34**, i884 (2018).

[64] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, *Star: Ultrafast universal rna-seq aligner,* Bioinformatics **29**, 15 (2013).

[65] A. Frankish, M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, *et al.*, *Gencode reference annotation for the human and mouse genomes,* Nucleic Acids Research **47**, D766 (2019).

[66] A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins, *Sambamba: fast processing of ngs alignment formats,* Bioinformatics **31**, 2032 (2015).

[67] Y. Liao, G. K. Smyth, and W. Shi, *featurecounts: an efficient general purpose program for assigning sequence reads to genomic features,* Bioinformatics **30**, 923 (2014).

[68] B. Li and C. N. Dewey, *Rsem: accurate transcript quantification from rna-seq data with or without a reference genome,* BMC bioinformatics **12**, 1 (2011).

[69] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, *The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data,* Genome Research **20**, 1297 (2010).

[70] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, *et al.*, *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples,* Nature biotechnology **31**, 213 (2013).

[71] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, *et al.*, *The ensembl variant effect predictor,* Genome Biology **17**, 122 (2016).

[72] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, *et al.*, *Analysis of protein-coding genetic variation in 60,706 humans,* Nature **536**, 285 (2016).

[73] S. Uhrig, J. Ellermann, T. Walther, P. Burkhardt, M. Fröhlich, *et al.*, *Accurate and efficient detection of gene fusions from rna sequencing data,* Genome research **31**, 448 (2021).

[74] M. I. Love, W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for rna-seq data with deseq2,* Genome biology **15**, 1 (2014).

[75] G. Sturm, F. Finotello, F. Petitprez, J. D. Zhang, J. Baumbach, *et al.*, *Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology,* Bioinformatics **35**, i436 (2019).

[76] I. Team, *immunarch: An r package for painless bioinformatics analysis of t-cell and b-cell immune repertoires,* (2019).

[77] W. S. van de Geer, J. van Riet, and H. J. G. van de Werken, *Proteodisco: A flexible r approach to generate customized protein databases for extended search space of novel and variant proteins in proteogenomic studies,* Bioinformatics (2021), 10.1093/bioinformatics/btab809, btab809.

[78] R. Orenbuch, I. Filip, D. Comito, J. Shaman, I. Pe'er, *et al.*, *arcashla: high-resolution hla typing from rnaseq,* Bioinformatics **36**, 33 (2020).

[79] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, *Near-optimal probabilistic rna-seq quantification,* Nature biotechnology **34**, 525 (2016).

[80] *Uniprot: the universal protein knowledgebase in 2021,* Nucleic Acids Research **49**, D480 (2021).

[81] N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber, *Data-driven hypothesis weighting increases detection power in genome-scale multiple testing,* Nature methods **13**, 577 (2016).

[82] M. Stephens, *False discovery rates: a new deal,* Biostatistics **18**, 275 (2017).

[83] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, *Kegg as a reference resource for gene and protein annotation,* Nucleic acids research **44**, D457 (2016).

[84] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, *et al.*, *The molecular signatures database hallmark gene set collection,* Cell systems **1**, 417 (2015).

[85] G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, *et al.*, *Fast gene set enrichment analysis,* BioRxiv , 060012 (2021).

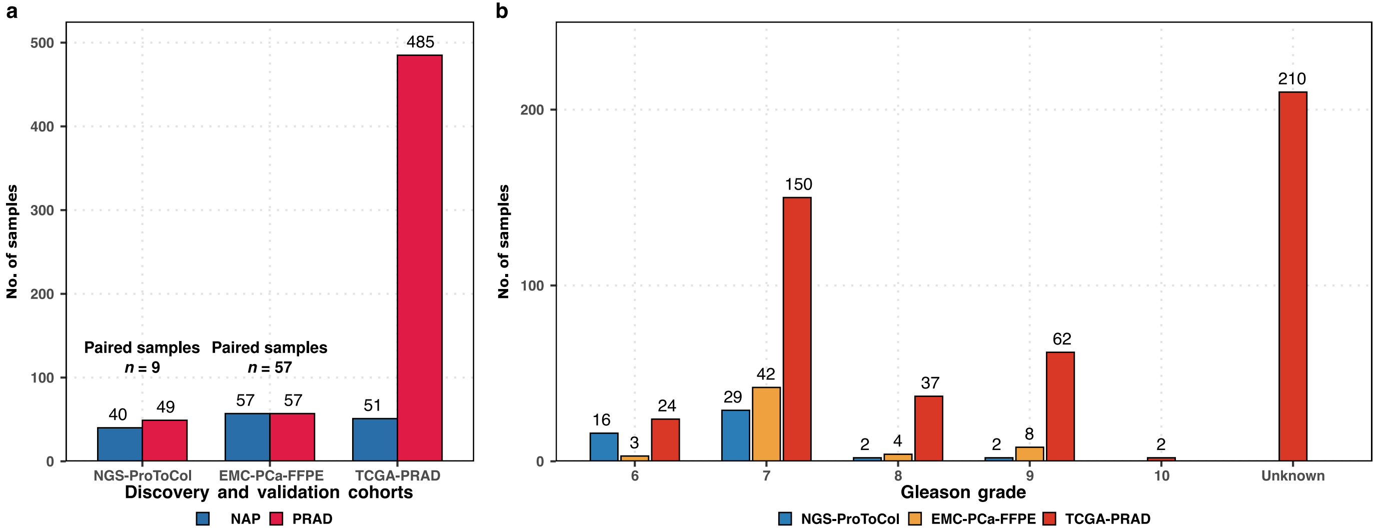
[86] W. Luo and C. Brouwer, *Pathview: an r/bioconductor package for pathway-based data integration and visualization,* Bioinformatics **29**, 1830 (2013), https://academic.oup.com/bioinformatics/article-pdf/29/14/1830/16916584/btt285.pdf .

[87] R. Core Team, *R: A language and environment for statistical computing,* R: A Language and Environment for Statistical Computing (2013).

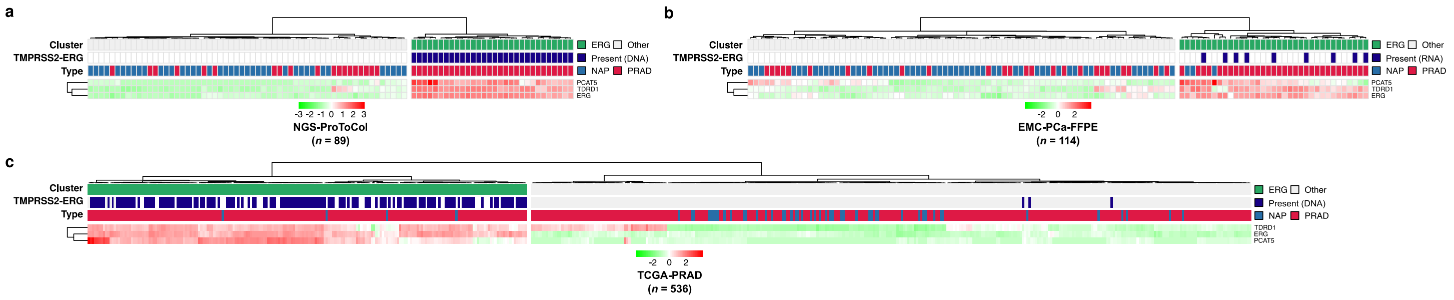
Supplemental Data

Supplementary data and figures accompanying the chapter:

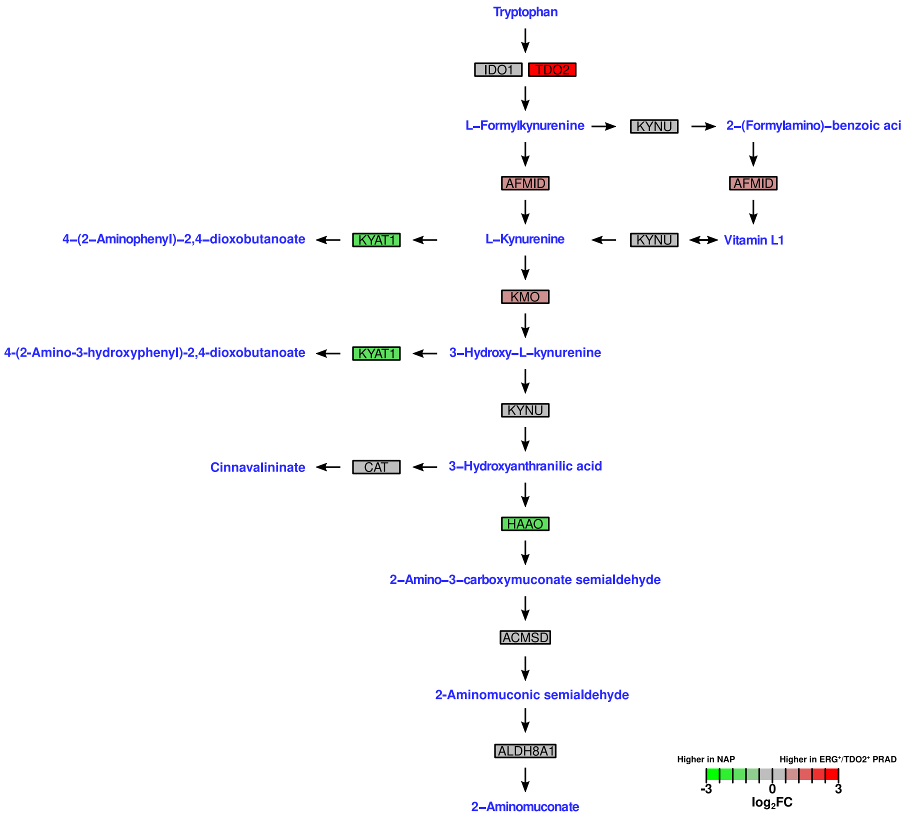
"ERG$^+$ PRAD shows enhanced frequencies of innate immune cells and elevated gene expression of TDO2 when compared to ERG$^-$ PRAD"



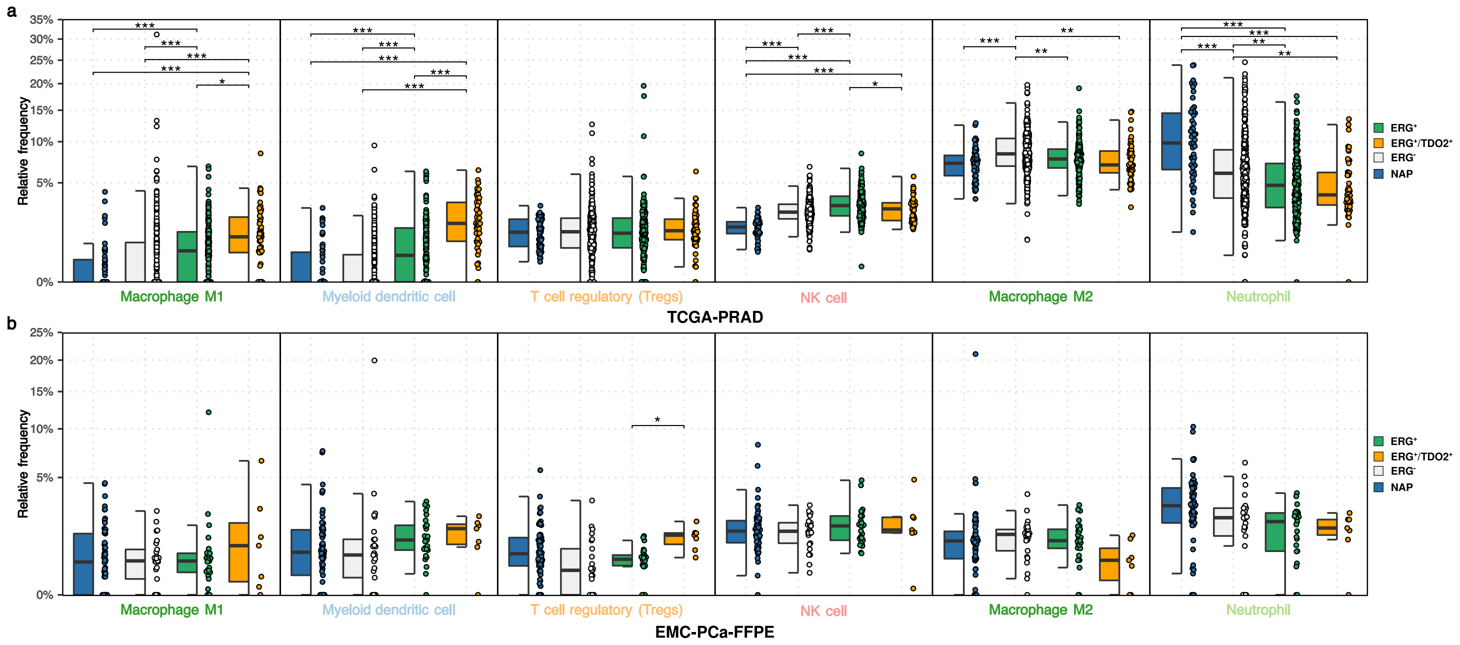
Supplementary figure S6.1: **Overview of included PRAD and NAP tissues, per included cohort.**
(a): Overview of the discovery and validation cohorts depicting the total number of samples per tissue type, respectively NAP and PRAD. NGS-ProToCol and EMC-PCa-FFPE contain matched non-malignant and malignant pairs of the same patient as shown for each respective cohort. **(b):** Distribution of the Gleason scores (x-axis) of the malignant prostate tissues in the discovery and validation cohorts. Within the TCGA-PRAD, 160 samples had unknown Gleason scores.



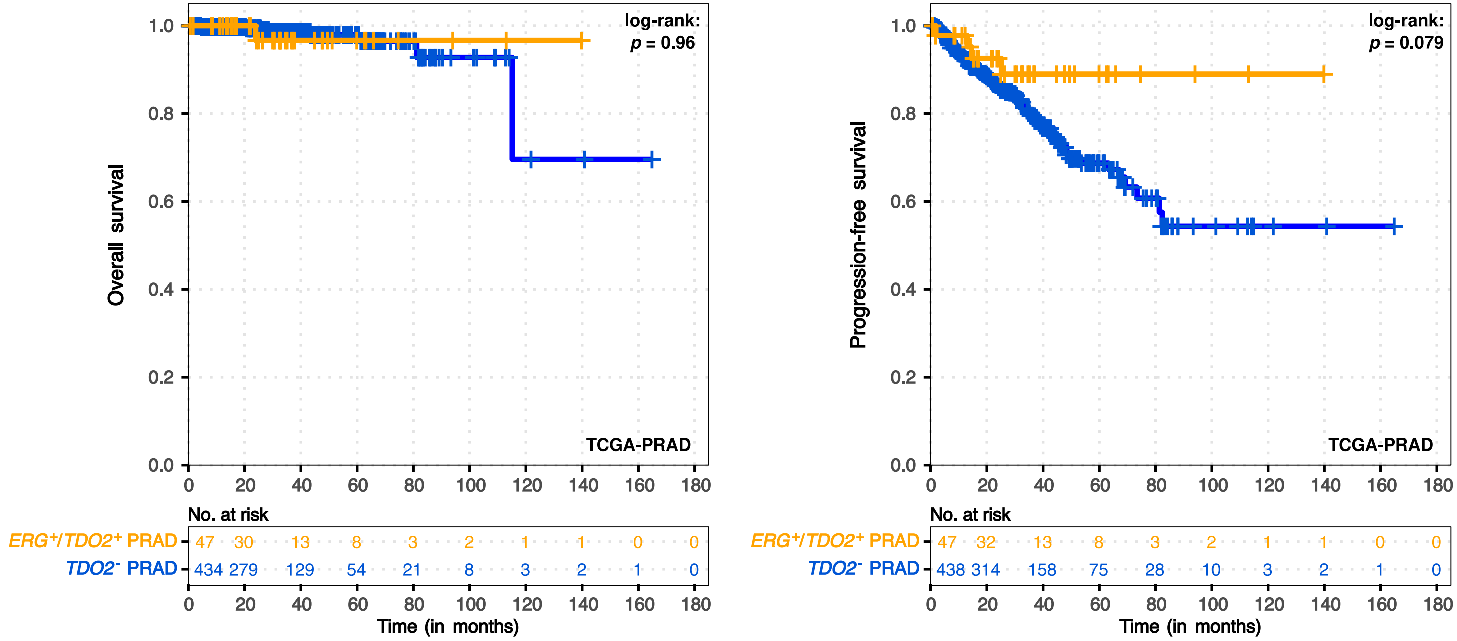
Supplementary figure S6.2: **Categorization based on aberrant ERG expression profiles.**
To determine possible aberrant *ERG* expression profiles, normalized and VST-transformed read counts for ERG and two downstream targets (*PCAT5* and *TDRD1*) were used in unsupervised hierarchical clustering (Euclidean distance and Ward.D2 metrics) for each of the included cohorts and are shown within the heatmap as Z-scores. Subsequently, the upper two clusters were used to distinguish*ERG*$^-$ from *ERG*$^+$ samples. The upper tracks display the respective cluster, presence of genomic *TMPRSS2-ERG* fusions and tissue type. **(a):** Overview of NGS-ProToCoL. **(b):** Overview of EMC-PCA-FFPE. **(c):** Overview of TCGA-PRAD.


Tryptophan
IDO1
TDO2
L–Formylkynurenine
KYNU
2–(Formylamino)–benzoic acid
AFMID
AFMID
4–(2–Aminophenyl)–2,4–dioxobutanoate
KYAT1
L–Kynurenine
KYNU
Vitamin L1
KMO
4-(2-Amino-3-hydroxyphenyl)-2,4-dioxobutanoate
KYAT1
3–Hydroxy–L–kynurenine
KYNU
Cinnavalininate
CAT
3–Hydroxyanthranilic acid
HAAO
2–Amino–3–carboxymuconate semialdehyde
ACMSD
2-Aminomuconic semialdehyde
ALDH8A1
2–Aminomuconate
Higher in NAP
Higher in ERG⁺/TDO2⁺ PRAD
-3
0
3
log₂FC

Supplementary figure S6.3: **Differential expression for members within the tryptophan to kynurenine catabolism pathway within *ERG*$^+$ / *TDO2*$^+$ PRAD compared to NAP.**
Overview of the tryptophan to kynurenine catabolism pathway (as derived from KEGG; map00380) displaying the $\log_2$FC from comparing *ERG*$^+$ / *TDO2*$^+$ PRAD vs. NAP within the NGS-ProToCol cohort.



Supplementary figure S6.4: **Immune cell populations (TIL10) within TCGA-PRAD and EMC-PCa-FFPE.**
(a): Boxplots with individual data points of the frequency of the TIL10 populations within the TCGA-PRAD which were found statistically significant between NAP (blue), *ERG*$^-$ PRAD (grey),*ERG*$^+$ PRAD (green) and/or *ERG*$^+$/*TDO2*$^+$ PRAD (orange) within the discovery cohort. Pairwise Wilcoxon Rank-Sum test against NAP with Benjamini-Hochberg correction was used to determine statistical significance with $q < 0.05$ (*); $q < 0.01$ (**); $q < 0.001$ (***). **b**) Same as **a**), but for the EMC-PCa-FFPE cohort.



Supplementary figure S6.5: **Overall and progression-free survival within the TCGA-PRAD, stratified on *ERG* and *TDO2* status.**
(a): Survival probability (OS) using univariate analysis of all TCGA-PRAD patients (*n* = 481; y-axis), stratified and colored on abundant expression of *ERG* and *TDO2* status and those without abundant *TDO2* expression, depicted in months (x-axis); censoring is shown by crosses (+). The bottom table represent the total number of remaining cases per depicted time-point. The log-rank p-value (between all groups) is shown on the right-hand top-side. **(b):** Same as **a)** but reporting on PFS.

Chapter 7

General discussion and future perspectives

Our current understanding of the complex and inter-connective mechanisms which underlie the intricate nature of genetics, and its malignant perversion into cancer, has come a long way from the early theories and concepts proposed by Hippocrates, Gregor Johann Mendel, Hugo de Vries, Thomas Hunt Morgan and fellow colleagues. With this expanded knowledge, so too has the need for collaborative efforts increased. The design, execution and analysis of the multitude of biological experiments and data warranted for current-day high-impact research demands an interdisciplinary and collaborative approach and mindset. Likewise, the demands on processing time and the multitude of molecular analysis from Next-Generation Sequencing (NGS) experiments cannot be performed without open-source software and community-driven development. To prevent re-inventing the wheel with less-than-optimal solutions, the possibility to turn to peer-reviewed and contemporary open-source software allows scores of researchers to address a wide range of biological questions from NGS experiments. This allows numerous research groups and institutes to robustly study and catalog the inventories of aberrations between multiple cellular states, perform extensive molecular classification and more; whilst ensuring the quality and uniformity of methods and analysis.

The necessity of transparent computational biology and open science

One of the pillars of scientific reporting is the ability to reproduce and extend upon published results. As detailed within the introduction of this thesis, the dependency on robust and reproducible computational methods to process and analyze current and future biological data-sets is increasing. As such, major journals have implemented renewed policies stating that custom software and biological data needs to be accessible or deposited upon publication which in turn, also leads to higher citations and greater impact.[1,2]

To underscore the importance of this inherent principle of science, all work in this thesis has been performed using peer-reviewed open-source software and crucial validations have been made possible due to public data accessible within the The Cancer Genome Atlas (TCGA) or (portions of the) Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium and through exchange of data between research-groups. Similarly, open-source software from the Bioconductor[3] or from other sources (e.g., GitHub) have made the majority of this thesis possible. We therefore made conscious efforts to publish *SNPitty*[4] and *ProteoDisco*[5] as open-source software for use in similar external research and open for other institutes.

To promote *open science*, several major Dutch and international grant initiatives such as the Koningin Wilhelmina Fonds voor de Nederlandse Kankerbestrijding (KWF) and Horizon Europe 2020 (H2020) have started to request the adoption of Findability, Accessibility, Interoperability, and Reusability (FAIR) principles[6] in current and future proposals. A similar trend has emerged within major publishing groups such as Springer Nature by embracing the FAIR principles as requirement (or advantage) for publication. Whilst a worthy goal to pursue for open science, a practical (and proper) implementation of the FAIR principles is still lacking in most research institutes and practical guidelines are still evolving.[7,8]

The merits of open science will likely outweigh practical concerns and will be adopted by current researchers as external validation, the need for increased sample-sizes and secondary analysis (i.e., re-analysis on public data) have become commonplace. The CPCT-02, Drug Rediscovery Protocol (DRUP) and WGS Implementation in the standard Diagnostics for Every cancer patient (WIDE) studies, as maintained by the Hartwig Medical Foundation (HMF), are also leading forces of this principle by facilitating the (re-)analysis of this unique and large cohort of metastatic malignancies (>4000 whole-genome sequencing (WGS) samples).[9–11] Despite the immense dependency on open-source software and computational biology in high-impact research, due credit is not always given as employed methodologies, software or concepts are not always properly cited.[12] Likewise, the ”antiquated” academic metrics such as the major focus on primary and last authorship rather than collective metrics, leads to challenges in correctly rewarding collaborative yet crucial roles within current and future collaborative and interdisciplinary efforts. As consequence, this leads to diminished academic progression and hinders software development and long-term maintenance thereof.[13]

Regardless, current and future enormous batches of sequencing data warrants the use of transparent and robust computational methodologies and will secure computational biology (bioinformatics) as a pillar of high-throughput science in the foreseeable future.

The myriad role of sequencing in the research, diagnosis, monitoring and treatment of malignancies

Major investments, technical advances and innovations have dramatically reduced the cost of routine sequencing and have led to the use of whole-exome sequencing (WES) or WGS as suitable alternatives to assess a wide range of distinct geno- and phenotypes from rare and common genetic diseases. This has the crucial advantage of employing only a single standardized technique and workflow, rather than utilizing multiple standalone molecular screenings to achieve similar results.[9,11,14] Within this thesis, we have interrogated the CPCT-02 WGS cohorts of metastatic castration-resistant prostate cancer (mCRPC)[15] and locally advanced or metastatic (advanced) neuroendocrine neoplasm (aNEN)[16] to illustrate that utilizing WGS can expand our molecular knowledge and provide clinically-relevant information. From these two cohorts, we have identified distinct treatment-relevant molecular subtypes, putative personalized treatment targets and captured the entire landscape of somatic alterations driving these complex malignancies. These large-scale WGS efforts were also able to detect relevant and recurrent genetic drivers from the total somatic landscape and were able to reproduce and reconfirm findings from multiple previous publications and research efforts within a single study-design.

One additional often-underrepresented benefit of utilizing a single standardized technique for a diverse range of genetic diseases is the ability to retrospectively quantify or investigate previously unknown genetic features or perform (retrospective) re-analysis on large numbers of samples, even more so in the case of WGS. Even as post-processing steps such as sequence alignment and downstream analysis might differ between institutes and research-groups, uniform re-analysis can always be performed if the original unaligned reads were retained or were captured within the Binary Alignment Map (BAM) file.

This raises the question as to whether more effort should be spent in promoting large international and interdisciplinary collaborations with pooled resources. This could improve the overall sample-sizes and increase population and ancestral diversity, robustness of analysis and enable supplementary techniques (e.g., transcriptomic and epigenetic) of matched samples. This would also consolidate expert knowledge and discussion, and ease research into different avenues by using the same uniformly-processed and greater data-set. In addition, this would also help prevent reinventing methodologies, duplicate research efforts and improves overall accuracy and open science principles. As previously alluded towards, this would also benefit from a rigorous overhaul of the antiquated system promoting the non-collaborative nature of academic achievements.

The work presented in this thesis was performed on solid biopsies obtained from invasive collection methods such as radical prostatectomy or fine-needle biopsy of metastatic sites. Due to these collection methods, we were likely incapable of capturing the full repertoire of clonal populations from locally(-advanced) or metastatic disease, as only the major subclones are often detected due to lack of sequencing depth. As these methods are invasive and complex, only few biopsies can be captured throughout the disease trajectory and subsequent progression. This leads to a potential under-representation of the full repertoire of treatment-induced mechanisms at play such as dynamic resistance mechanisms and clonal evolution. However, current-day feasibility for the sequencing of liquid biopsies such as circulating tumor DNA (ctDNA) or exosomal content, acquired from minimally-invasive blood-sampling, has the potential to improve our understanding of these dynamic mechanisms.[17–19] Coupled with the advantages of single-cell sequencing (from solid biopsies), this will likely spur new research and insights into to the dynamic interplay of malignant cells with the tumor microenvironment (TME), clonal trajectories, cellular trans-differentiation (plasticity) and treatment-induced changes. As this field of research is still in its infancy, robust clinical and experimental guidelines are still under debate and will require further evaluations to be implemented in daily practice but will nevertheless herald a groundbreaking extension in utilizing high-throughput NGS for the monitoring of disease progression and timely adjustments to treatment-regiments.

Future perspectives of computational biology in oncology research

With the increasing volumes of NGS-data for primary and metastatic malignancies from varied sequencing and experimental setups, we now stand for the challenge of incorporating and simplifying these multitude of observations into generalized and clinically-beneficial insights. This warrants an inter-disciplinary approach to bridge the various and interrelated molecular apparatus promoting and sustaining cancer. The recent advancements and implementation of ctDNA-based and single-cell sequencing approaches provide an unheralded understanding of the clonal evolution and interplay with the TME, which underlies the initiation and progression of malignant cellular states. This can also shed light onto the treatment-induced clonal evolution regarding cellular plasticity such as the treatment-induced neuroendocrine prostate cancer (t-NEPC) seen ever more often due the use of more potent androgen receptor pathway inhibition agents.[20,21] Research into this complex interplay using single-cell sequencing could elucidate novel treatment-targets or strategies to counteract or even steer these clonal developments. This could also shed further light upon the role of the TME regarding the induction of an adequate immune responses in otherwise (immunologically) cold tumors such as prostate cancer.[22]

Utilizing the full potential of WGS also necessitates the investigation of the non-coding portions of the malignant genome(s); a somewhat undervalued region of the potentially-exploitable genome due to its more complicated nature of analysis and functional consequences. This could yet identify further sets of crucial somatic aberrations which drive the progression and treatment-resistance(s) of cancer and which are simply not easily detectable using other sequencing approaches such as WES. These non-coding genetic elements could include perturbations such as the amplification of enhancers directly driving the expression of critical onco-genes or, vice-versa, via the suppression of enhancers and subsequent silencing of tumor-suppressors.[15,23–25] However, *in vitro* validation regarding the proposed functional consequences of *bona fide* enhancers is more convoluted than their coding counterparts and require extensive screening experiments.[26] Beyond the functional effects hidden within the non-coding genome, the non-coding regions also harbors additional information exploitable to reveal the likely cell-of-origin for cancers of unknown primary (CUP) or used in estimating intratumoral heterogeneity and the likely moment when certain somatic events were acquired or lost again.[9,27,28]

The central dogma of molecular biology also alludes to potentially detectable and exploitable perturbations within the transcriptome or chromatin states of malignancies. Using general NGS experiments or those tailored towards singular queries (such as circular deoxyribonucleic acid (DNA) or B cell receptor sequencing), there still exists many opportunities to delve into biological markers or prognostic metrics for the detection or monitoring of malignancies. Overall, it can be concluded that NGS coupled with their recent successful implementations on single-cell and ctDNA-level will remain crucial in identifying new avenues of treatment-strategies and to expand our current hypotheses regarding this dreadful malady.

References

[1] G. Colavizza, I. Hrynaszkiewicz, I. Staden, K. Whitaker, and B. McGillivray, *The citation advantage of linking publications to research data,* PloS one **15**, e0230416 (2020).

[2] J. B. Byrd, A. C. Greene, D. V. Prasad, X. Jiang, and C. S. Greene, *Responsible, practical genomic data sharing that accelerates research,* Nature Reviews Genetics **21**, 615 (2020).

[3] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, *et al.*, *Bioconductor: open software development for computational biology and bioinformatics.* Genome biology **5**, R80 (2004).

[4] J. van Riet, N. M. Krol, P. N. Atmodimedjo, E. Brosens, W. F. van Ijcken, *et al.*, *Snpitty: an intuitive web application for interactive b-allele frequency and copy number visualization of next-generation sequencing data,* The Journal of Molecular Diagnostics **20**, 166 (2018).

[5] W. S. van de Geer, J. van Riet, and H. J. G. van de Werken, *Proteodisco: A flexible r approach to generate customized protein databases for extended search space of novel and variant proteins in proteogenomic studies,* Bioinformatics (2021), 10.1093/bioinformatics/btab809, btab809.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, *et al.*, *The fair guiding principles for scientific data management and stewardship,* Scientific data **3**, 1 (2016).

[7] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, *et al.*, *Fair principles: interpretations and implementation considerations,* (2020).

[8] M. Boeckhout, G. A. Zielhuis, and A. L. Bredenoord, *The fair guiding principles for data stewardship: fair enough?* European journal of human genetics **26**, 931 (2018).

[9] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, *et al.*, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[10] D. Van der Velden, L. Hoes, H. van der Wijngaart, J. van Berge Henegouwen, E. van Werkhoven, *et al.*, *The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs,* Nature **574**, 127 (2019).

[11] K. G. Samsom, L. J. Bosch, L. J. Schipper, P. Roepman, E. de Bruijn, *et al.*, *Study protocol: Whole genome sequencing implementation in standard diagnostics for every cancer patient (wide),* BMC medical genomics **13**, 1 (2020).

[12] J. D. Wren, *Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades,* Bioinformatics **32**, 2686 (2016).

[13] J. Chang, *Core services: Reward bioinformaticians,* Nature **520**, 151 (2015).

[14] C. R. Marshall, S. Chowdhury, R. J. Taft, M. S. Lebo, J. G. Buchan, *et al.*, *Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease,* NPJ genomic medicine **5**, 1 (2020).

[15] L. F. van Dessel, J. van Riet, M. Smits, Y. Zhu, P. Hamberg, *et al.*, *The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact,* Nature communications **10**, 1 (2019).

[16] J. van Riet, H. J. van de Werken, E. Cuppen, F. A. Eskens, M. Tesselaar, *et al.*, *The genomic landscape of 85 advanced neuroendocrine neoplasms reveals subtype-heterogeneity and potential therapeutic targets,* Nature communications **12**, 1 (2021).

[17] E. Crowley, F. Di Nicolantonio, F. Loupakis, and A. Bardelli, *Liquid biopsy: monitoring cancer-genetics in the blood,* Nature reviews Clinical oncology **10**, 472 (2013).

[18] R. A. Mathai, R. V. S. Vidya, B. S. Reddy, L. Thomas, K. Udupa, *et al.*, *Potential utility of liquid biopsy as a diagnostic and prognostic tool for the assessment of solid tumors: implications in the precision oncology,* Journal of clinical medicine **8**, 373 (2019).

[19] M. Russano, A. Napolitano, G. Ribelli, M. Iuliani, S. Simonetti, *et al.*, *Liquid biopsy and tumor heterogeneity in metastatic solid tumors: the potentiality of blood samples,* Journal of Experimental & Clinical Cancer Research **39**, 1 (2020).

[20] S. Akamatsu, T. Inoue, O. Ogawa, and M. E. Gleave, *Clinical and molecular features of treatment-related neuroendocrine prostate cancer,* International Journal of Urology **25**, 345 (2018).

[21] R. Aggarwal, J. Huang, J. J. Alumkal, L. Zhang, F. Y. Feng, *et al.*, *Clinical and genomic characterization of treatment-emergent small-cell neuroendocrine prostate cancer: a multi-institutional prospective study,* Journal of Clinical Oncology **36**, 2492 (2018).

[22] Y. Zhang and Z. Zhang, *The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications,* Cellular & molecular immunology **17**, 807 (2020).

[23] D. Takeda, S. Spisák, J.-H. Seo, C. Bell, E. O'Connor, *et al.*, *A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer,* Cell **174**, 422 (2018).

[24] E. Rheinbay, M. M. Nielsen, F. Abascal, J. A. Wala, O. Shapira, *et al.*, *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes,* Nature **578**, 102 (2020).

[25] Y. Liu, S. Yu, V. K. Dhiman, T. Brunetti, H. Eckart, *et al.*, *Functional assessment of human enhancer activities using whole-genome starr-sequencing,* Genome biology **18**, 1 (2017).

[26] M. Gasperini, J. M. Tome, and J. Shendure, *Towards a comprehensive catalogue of validated and target-linked human enhancers,* Nature Reviews Genetics **21**, 292 (2020).

[27] W. Jiao, G. Atwal, P. Polak, R. Karlic, E. Cuppen, *et al.*, *A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns,* Nature communications **11**, 1 (2020).

[28] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, *Machine learning applications in cancer prognosis and prediction,* Computational and structural biotechnology journal **13**, 8 (2015).

Chapter 8

Summary

Current-day technologies laying bare the whole genome, transcriptome and even epigenome of (metastatic) malignancies had long been out of reach for generations of earlier researchers whom had to make due with only snippets of genetic information at a time. In particular, the recent advent of whole-genome sequencing (WGS) has sparked many insights and novel methods into utilizing genome-wide information to extensively characterize malignancies based on their genomic features, perform unbiased detection of disease-driving genes and to derive potential targets or genotypes for personalized treatment.

The use of computational methods and algorithms to process and analyze these large quantities of data is a non-trivial effort for bioinformaticians, pathologists, biologists and clinicians alike. The sheer quantity and complexity of current-day experiments warrants custom yet intuitive software such that researchers are better able to interrogate and interpret their results. To aid in this daunting effort, we have developed two open-source R-based software packages. **Chapter 2** details *SNPitty* which visualizes Next-Generation Sequencing (NGS)-derived genomic alterations such as somatic mutations, heterozygous markers coupled with loss of heterozygosity (LOH) and copy-number alterations relevant to daily molecular diagnostics. **Chapter 3** describes *ProteoDisco*, an algorithm capable of accurately incorporating a myriad of genomic variants into their overlapping coding sequences to produce their respective protein-variant sequence(s) for use in downstream analysis such as neo-antigen prediction and extending the search space of novel and variant proteins in proteogenomic studies.

Within this thesis, we highlight our results on the analysis of our unique and large cohorts of whole-genome sequenced metastatic castration-resistant prostate cancer (mCRPC) and locally advanced or metastatic (advanced) neuroendocrine neoplasm (aNEN), both performed within the CPCT-02 study. **Chapter 4** details the somatic inventory of mCRPC and the use of genomic features such as tumor mutational burden (TMB) and various classes of mutations and structural variants to stratify patients into distinct treatment-related groups based on the large-scale accumulation of specific somatic alterations due to (minute) changes in underlying drivers. As WGS also captures the non-coding regions of the genome, we were able to identify (among others) somatically-acquired amplifications of upstream enhancers regulating the expression of the master transcription factors *MYC* and the androgen-dependent *AR*, thwarting castration regimens by allowing androgen-independent malignant progression.

Chapter **5** describes the somatic inventory of the enigmatic neuroendocrine neoplasms; a malignancy with an generally uncharacteristic stable somatic genome harboring only few somatic alterations coupled with a prolonged disease trajectory. Using WGS to capture the largest repository of neuroendocrine carcinomas (NEC) and neuroendocrine tumors (NET), we could distinguish NEC and NET based on large-scale genomic features, distinct drivers and overall TMB. Furthermore, ~49% of aNEN patients revealed potential therapeutic targets based upon actionable (and responsive) somatic aberrations within their genome; potentially directing improvements in aNEN treatment strategies.

Finally, **chapter** **6** describes the discovery of a distinct and potentially immune-privileged form of (primary) prostate adenocarcinoma with characteristic and abundant expression of the transcription factor *ERG* coupled with the rate-limiting member of the immunoregulatory kynurine pathway *TDO2* and deregulation of immune-related mechanisms and a defunct tumor microenvironment (TME). This previously-undiscovered subgroup within ~25% of all prostate adenocarcinoma could renew efforts into utilizing immune-based therapies within prostate cancer; a strategy long thought to be unsuited this malignancy.

Chapter 9

Samenvatting

Hedendaagse technologieën waarbij het gehele genoom, transcriptoom en zelfs epigenoom van (metastatische) maligniteiten wordt blootgelegd was lange tijd enkel een droom voor eerdere generaties aan onderzoekers; zij moesten genoegen nemen met enkel minuscule fragmenten aan genetische informatie. In het bijzonder heeft de recente introductie van whole-genome sequencing (WGS) vele nieuwe inzichten en methodieken gebracht betreffende het benutten van genoom-brede informatie voor het karakteriseren van maligniteiten op basis van onderliggende genomische eigenschappen, onbevooroordeelde detectie van ziekmakende genen en om potentiële doelwitten of genotypes te bepalen voor doelgerichte en gepersonaliseerde therapie.

Het gebruik van computer-toegepaste methodieken en algoritmes voor het verwerken en analyseren van deze grote hoeveelheden aan data is een niet-triviale zaak voor zowel bioinformatici, pathologen als clinici. De gigantische hoeveelheid en complexiteit van hedendaagse experimenten vragen om maatwerk en intuïtieve software waarbij onderzoekers hun resultaten op de juiste wijze kunnen verwerken en interpreteren. Om aan deze moeilijke klus mee te helpen hebben wij twee open-source R-gebaseerde applicaties ontworpen. **Hoofdstuk 2** gaat in op *SNPitty*, een applicatie voor de visualisatie van Next-Generation Sequencing (NGS)-ontdekte genomische alteraties zoals somatische mutaties, heterozygote markers samen met loss of heterozygosity (LOH) en alteraties in het aantal kopieën van een gen; welke allen relevant zijn voor de dagelijkse moleculaire diagnostiek. **Hoofdstuk 3** beschrijft *ProteoDisco*, een algoritme waarmee op een correcte wijze een verscheidenheid aan genomische varianten kan worden geïncorporeerd binnen overlappende coderende sequenties en waarbij de resulterende proteïne variant(en) kan worden opgeleverd. Deze proteïne-varianten kunnen vervolgens worden gebruikt in additionele analyses zoals het voorspellen van neo-antigenen en ter verbetering van de inventarisatie van (nieuwe) proteïnen in proteogenomische studies.

Ook beschrijven wij de resultaten van onze unieke en grote cohorten aan whole-genome sequenced metastatische castratieresistente prostaatkanker (mCRPC) en (lokaal-geavanceerde) neuro-endocriene neoplasma (aNEN); beide uitgevoerd binnen de CPCT-02 studie. **Hoofdstuk** 4 beschrijft de somatische inventarisatie van mCRPC en het gebruik van onderliggende genomische eigenschappen zoals de totale mutatielast (TMB) en verscheidende categorieën aan mutaties en structurele varianten om patiënten te stratificeren in therapie-gerelateerde groepen a.d.h.v. de enorme accumulatie specifieke somatische alteraties wegens (kleine) veranderingen in onderliggende ziekmakende genen. Omdat WGS ook de niet-coderende regionen van het genoom blootlegt, zijn wij in staat geweest om (o.a.) somatische amplificaties te ontdekken van omliggende elementen (enhancers) van de regulatoire genen *MYC* en *AR*; waarbij castratie-therapie wordt omzeild wegens castratie-onafhankelijke progressie van de kanker.

Hoofdstuk 5 beschrijft de somatische inventarisatie van de raadselachtige neuro-endocriene neoplasma; een maligniteit welke vaak gekoppeld gaat met een ongebruikelijk stabiel somatisch genoom met hierin weinig somatische alteraties en een langdurig ziektebeeld. Met WGS hebben wij het grootste cohort van neuro-endocriene carcinoma (NEC) en neuro-endocriene tumoren (NET) bestudeerd en waren wij in staat om NEC van NET te onderscheiden a.d.h.v. genomische eigenschappen, exclusieve ziekmakende genen en algehele mutatielast. Gebaseerd op responsieve somatische aberraties binnen het tumor-genoom kon voor ~49% van alle aNEN-patiënten een potentieel therapeutisch doelwit worden gevonden; met mogelijke verbeteringen in therapiekeuzen voor aNEN als gevolge.

Tenslotte wordt in **hoofdstuk** 6 de ontdekking beschreven van een unieke en potentieel immuun-geprivilegieerde vorm van (primaire) prostaatadenocarcinoom met karakteristieke en enorme expressie van de transcriptiefactor *ERG* gekoppeld met *TDO2*, het snelheid-limiterende lid van de immuun-regulatoire kynurine pathway samen met deregulatie van immuun-gerelateerde mechanismes en een verstoorde tumor microenvironment (TME). Deze voorheen verborgen subgroep binnen ~25% van alle (primaire) prostaatadenocarcinoma kan ons hernieuwde hoop opleveren voor immuun-gebaseerde therapieën binnen prostaatkanker; een invalshoek welke lang werd gedacht dat dit niet effectief was voor deze maligniteit.

Appendices

Appendix A

PhD Portfolio

Year	Attended Course	ECTS
2020	MolMed - Workshop presenting skills for PhD students and Post Docs	1.0
2019	MolMed - Basic and Translational Oncology	1.8
2019	Erasmus MC - Research Integrity	0.3
2017	EMBL-EBI - Translational Bioinformatics	1.0
2016	MolMed - Molecular Diagnostics XII	1.0
2015	MolMed - The CLC Workbench / Ingenuity Variant Analysis Workshop	0.5

Year	Lectured Courses
2016 - 2021	MolMed - Basic course on R
2016 - 2020	MolMed - Expression Course

Year	Supervised students	Degree
2021	D. Hazelaar *(Vrije Universiteit Amsterdam)*	MSc. Thesis
2021	Y. Ping *(Vrije Universiteit Amsterdam)*	MSc. Thesis
2020	C. Berns *(Technische Universiteit Delft)*	BSc. Thesis
2019	L. Perdaems *(AVANS Breda)*	BSc. Thesis
2018	N. van der Horst *(AVANS Breda)*	BSc. Thesis
2016	N. Stoker *(Rijksuniverseit Groningen)*	BSc. Thesis
2016	W. van de Geer *(Hogeschool Leiden)*	BSc. Thesis

Year	Scientific Presentations / Poster	Type
2018 - 2022	Dept. of Medical Oncology	Presentation(s)
2015 - 2022	Annual MolMed Day	Poster(s)
2015 - 2022	Josephine Nefkens Institute	Presentation(s)
2015 - 2019	Dept. of Urology	Presentation(s)
2021	MORM	Presentation
2020	AACR 2020	Presentation
2019	The Erasmus MC Cancer Institute Day	Presentation
2019	Tour d'Europe / SANOFI	Presentation
2017	EMBL-EBI	Poster
2016	BioSB 2016	Poster

Year	Conferences, Meetings and Workshops	Type
2017 - 2022	Monthly JC - CCBC	Journal Club
2015 - 2019	Monthly JC - Dept. of Urology	Journal Club
2015 - 2022	Bridge Meeting	Meeting
2015 - 2022	JNI Scientific Meeting	Meeting
2015 - 2022	Annual MolMed Day & Symposium	Conference
2021	AVL-NKI | WGS Symposium	
2016 - 2019	Dutch Techcentre for Life Sciences (DLTS)	Workshop(s)
2020	AACR 2020	Conference
2016	BioSB 2016	Conference
2015	VIB - Revolutionizing Next-Generation Sequencing: Tools and Technologies	Conference

Appendix B

List of Publications

Publications in print or under review; ordered by date of publication.

1. A.C. de Jong*, A. Danyi*, **J. van Riet**, R. de Wit, M. Sjöström, F. Feng, J. de Ridder, M.P.J.K. Lolkema
Predicting response to androgen receptor signaling inhibitors in metastatic castration resistant prostate cancer patients through machine learning-based analysis of whole genome and transcriptome sequencing data
Under review.

2. K.T. Isebia, B. Mostert, B.P.S. Belderbos, T. Deger, J.C.A. Helmijr, J. Kraan, V. de Weerd, M.N. Van, E. Oomen-de Hoop, A.M. Sieuwerts, P. Hamberg, B.C.M. Haberkorn, H.H. Helgason, R. de Wit, S. Sleijfer, R.A.H.J. Mathijssen, J.W.M. Martens, S.M. Wilting, **J. van Riet***, M.P.J.K. Lolkema*
mFAST-SeqS based aneuploidy score is a prognostic biomarker in prostate cancer
Under review.

3. K.T. Isebia, B. Mostert, B.P.S. Belderbos, S.A.J. Buck, J.C.A. Helmijr, J. Kraan, C.M. Beaufort, M.N. Van, E. Oomen-de Hoop, A.M. Sieuwerts, W.F.J. van IJcken, M.C.G.N. van den Hout-van Vroonhoven, R.W.W. Brouwer, E. Oole, P. Hamberg, B.C.M. Haberkorn, H.H. Helgason, R. de Wit, S. Sleijfer, R.A.H.J. Mathijssen, J.W.M. Martens, M.P.H.M. Jansen, **J. van Riet***, M.P.J.K. Lolkema*
CABA-V7: a prospective biomarker selected trial of cabazitaxel treatment in AR-V7 positive prostate cancer patients
Under review.

4. D.M. Hazelaar*, **J. van Riet***, Y. Hoogstrate, M.P.J.K Lolkema, H.J.G. van de Werken
Katdetectr: utilising unsupervised changepoint analysis for robust kataegis detection
bioRxiv; Under review.

5. **J. van Riet***, C. Saha*, N. Strepis, R. Brouwers, W.S. van de Geer, S.M.A. Swagemakers, A. Koning, A. Stubbs, J. Mouton, M.A. Komor, Y. Hoogstrate, B. Janssen, R.J.A. Fijneman, Y.S. Niknafs, A.M. Chinnaiyan, W. van IJcken, P.J. van der Spek, G. Jenster, R. Louwen

CRISPRs in the human genome are differentially expressed between malignant and normal adjacent to tumor tissue
Nature Communications Biology 5, 338 (2022).

6. W. Drabarek*, **J. van Riet***, J.Q.N. Nguyen, K.N. Smit, N. van Poppelen, R. Jansen, E. Medico, J. Vaarwater, F.J. Magielsen, T. Brands, B. Eussen, T.P.P. van den Bosch, R.M. Verdijk, N. Naus, D. Paridaens, A. de Klein, E. Brosens, H.J. G. van de Werken, E. Kilic on behalf of the Rotterdam Ocular Melanoma Study Group
Identification of Early-Onset Metastasis in SF3B1 Mutated Uveal Melanoma
Cancers 14(3), 846; Feb. 2022.

7. A. Nakauma-Gonzalez*, M. Rijnders*, **J. van Riet**, M.S. van der Heijden, J. Voortman, E. Cuppen, N. Mehra, S. van Wilpe, S. Oosting, E.C. Zwarthoff, R. de Wit, A.A.M. van der Veldt, H.J.G. van de Werken, M.P.J.K Lolkema, J.L. Boormans
Comprehensive Molecular Characterization Reveals Genomic and Transcriptomic Subtypes of Metastatic Urothelial Carcinoma
European Urology; Jan. 2022.

8. Y. Hoogstrate, M.A. Komor, R. Böttcher, **J. van Riet**, H.J.G. van de Werken, S. van Lieshout, R. Hoffmann, E. van den Broek, A.S. Bolijn, N. Dits, D. Sie, D. van der Meer, F. Pepers, C.H. Bangma, G.J.L.H. van Leenders, M. Smid, P.J. French, J.W.M. Martens, W. van Workum, P.J. van der Spek, B. Janssen, E. Caldenhoven, C. Rausch, M. de Jong, A.P. Stubbs, G.A. Meijer, R.J.A. Fijneman, G.W. Jenster
Fusion transcripts and their genomic breakpoints in polyadenylated and ribosomal RNA–minus RNA sequencing data
GigaScience; Volume 10, Issue 12, December 2021, giab080.

9. W.S. van de Geer*, **J. van Riet***, H.J.G. van de Werken
ProteoDisco: A flexible R approach to generate customized protein databases for extended search space of novel and variant proteins in proteogenomic studies
Bioinformatics; btab809, Dec. 2021.

10. Y. Hoogstrate, S.A. Ghisai, M. de Wit, I. de Heer, K. Draaisma, **J. van Riet**, H.J.G. van de Werken, V. Bours, J. Buter, Is. van den Bempt, M. Eoli, E. Franceschi, J. Frenel, T. Gorlia, M.C. Hanse, A. Hoeben, M. Kerkhof, J.M. Kros, S. Leenstra, G. Lombardi, S. Lukacova, P.A. Robe, J.M. Sepulveda, W. Taal, M. Taphoorn, R.M. Vernhout, A.M.E. Walenkamp, C. Watts, M. Weller, F.Y.F. de Vos, G.W. Jenster, M. van den Bent, P.J. French
The EGFRvIII transcriptome in glioblastoma, a meta-omics analysis
Neuro-oncology, 2021 Oct. 5;noab231 (2021).

11. **J. van Riet***, H.J.G. van de Werken*, E. Cuppen, F.A.L.M. Eskens, M. Tesselaar, L.M. van Veenendaal, H. Klümpen, M.W. Dercksen, G.D. Valk, M.P.J.K. Lolkema, S. Sleijfer,

B. Mostert
The genomic landscape of 85 advanced neuroendocrine neoplasms reveals subtype-heterogeneity and potential therapeutic targets
Nature Communications, volume 12, Article number: 4612 (2021).

12. G. Snaterse, L.F. van Dessel, **J. van Riet**, A.E. Taylor, M. Van Der Vlugt-Daane, P. Hamberg, R. de Wit, J.A. Visser, W. Arlt, M.P.J.K. Lolkema, J. Hofland
11-Ketotestosterone is the predominant active androgen in prostate cancer patients after castration
JCI Insight, 2021;6(11):e148507.

13. L. Mout*, L.F. van Dessel*, J. Kraan, A.C. de Jong, R.P.L. Neves, S. Erkens-Schulze, C.M. Beaufort, A.M. Sieuwerts, **J. van Riet**, T.L.C. Woo, R. de Wit, S. Sleijfer, P. Hamberg, Y. Sandberg, P.A.W. Te Boekhorst, H.J.G. van de Werken, J.W.M. Martens, N.H. Stoecklein, W.M. van Weerden, M.J.P.K. Lolkema
Generating human prostate cancer organoids from leukapheresis enriched circulating tumour cells
European Journal of Cancer Volume 150, June 2021, Pages 179-189.

14. T. van Doeveren*, J.A. Nakauma-Gonzalez*, A.S. Mason, G.J.L.H. van Leenders, T.C.M. Zuiverloon, E.C. Zwarthoff, I.C. Meijssen, A.C. van der Made, A.G. van Der Heijden, K. Hendricksen, B.W.G. van Rhijn, C.S. Voskuilen, **J. van Riet**, W.N.M. Dinjens, H.J. Dubbink, H.J.G. van de Werken, J.L. Boormans
The clonal relation of primary upper urinary tract urothelial carcinoma and paired urothelial carcinoma of the bladder
Molecular Cancer Biology, Volume 148, Issue 4, 15 February 2021, Pages 981-987.

15. P.A.J. Mendelaar, M. Smid, **J. van Riet**, L. Angus, M. Labots, N. Steeghs, M.P. Hendriks, G.A. Cirkel, J.M. van Rooijen, A.J. Ten Tije, M.P.J.K. Lolkema, E. Cuppen, S. Sleijfer, J.W.M. Martens, S.M. Wilting
Whole genome sequencing of metastatic colorectal cancer reveals prior treatment effects and specific metastasis features
Nature Communications, (2021)12:57.

16. A. Nakauma-Gonzalez*, M. Rijnders*, **J. van Riet**, M.S. van der Heijden, J. Voortman, E. Cuppen, N. Mehra, S. van Wilpe, S. Oosting, E.C. Zwarthoff, R. de Wit, A.A.M. van der Veldt, H.J.G. van de Werken, M.P.J.K Lolkema, J.L. Boormans
Genomic and Transcriptomic Characterization of Metastatic Urothelial Carcinoma
Urologic Oncology: Seminars and Original Investigations, Volume 38, Issue 12, December 2020, Page 910.

17. A.C. de Jong*, M. Smits*, **J. van Riet**, J.J. Fütterer, T. Brabander, P. Hamberg, I.M. van Oort, R. de Wit, M.J.P.K. Lolkema, N. Mehra, M. Segbers, A.A.M. van der Veldt

^{68}Ga-PSMA–Guided Bone Biopsies for Molecular Diagnostics in Patients with Metastatic Prostate Cancer
Journal of Nuclear Medicine November 2020, 61 (11) 1607-1614.

18. T. van Doeveren, H.J.G. van de Werken, **J. van Riet**, K.K.H. Aben, P.J. van Leeuwen, E.C. Zwarthoff, J.L. Boormans
Synchronous and metachronous urothelial carcinoma of the upper urinary tract and the bladder: are they clonally related? A systematic review.
Urologic Oncology: Seminars and Original Investigations, Volume 38, Issue 6, 2020, Pages 590-598.

19. A.T. Kenter, E. Rentmeester, **J. van Riet**, R. Boers, J. Boers, M. Ghazvini, V.J. Xavier, G.J.L.H. van Leenders, P.C.M.S. Verhagen, M.E. van Til, B. Eussen, M. Losekoot, A. de Klein, D.J.M. Peters, W.F.J. van IJcken, H.J.G. van de Werken, R. Zietse, E.J. Hoorn, G. Jansen, J.H. Gribnau
Cystic renal-epithelial derived induced pluripotent stem cells from polycystic kidney disease patients
STEM CELLS Translational Medicine 2020 Mar. 9;4:478-490.

20. L.F. van Dessel*, **J. van Riet***, M. Smits, Y. Zhu, P. Hamberg, M.S. van der Heijden, A.M. Bergman, I.M. van Oort, R. de Wit, E.E. Voest, N. Steeghs, T.N. Yamaguchi, J. Livingstone, P.C. Boutros, J.W.M. Martens, S. Sleijfer, E.P.J.G. Cuppen, W. Zwart, H.J.G. van de Werken, N. Mehra, M.P.J.K. Lolkema
The genomic landscape of metastatic castration-resistant prostate cancers using whole genome sequencing reveals multiple distinct genotypes with potential clinical impact
Nature Communications 2019 Nov 20;10(1):5251.

21. L. Angus, M. Smid, S.M. Wilting, **J. van Riet**, A. Van Hoeck, L. Nguyen, S. Nik-Zainal, T.G. Steenbruggen, V.C.G. Tjan-Heijnen, M. Labots, J.M.G.H. van Riel, H.J. Bloemendal, N. Steeghs, M.P.J.K. Lolkema, E.E. Voest, H.J.G. van de Werken, A. Jager, E. Cuppen, S. Sleijfer, J.W.M. Martens
The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies
Nature Genetics 2019 Oct;51(10):1450-1458

22. D.G. Robbrecht*, F. Atrafi*, **J. van Riet**, F.A.L.M. Eskens, P.J. van Diest, E.P.J.G. Cuppen, G.J.L.H. van Leenders, H.J.G. van de Werken, M.J.P.K. Lolkema
Unique Case of a Rare Mesenchymal Tumor Harboring a Somatic c.119delC VHL Mutation
JCO Precision Oncology - published online (2019).

23. L.C.J. Dorssers, A.J.M. Gillis, H. Stoop, R. van Marion, M.M. Nieboer, **J. van Riet**, H.J.G. van de Werken, J.W. Oosterhuis, J. de Ridder, L.H.J. Looijenga

Molecular heterogeneity and early metastatic clone selection in testicular germ cell cancer development
British Journal of Cancer volume 120, Pages 444–452 (2019).

24. W. Drabarek*, S.Yavuzyigitoglu*, A. Obulkasim**, **J. van Riet****, K. N. Smit, N.M. van Poppelen, J. Vaarwater, T. Brands, B. Eussen, R.M. Verdijk, N.C. Naus, H.W. Mensink, D. Paridaens, E. Boersma, H.J.G. van de Werken, E. Kilic, A. de Klein for the Rotterdam Ocular Melanoma Study Group
Multi-Modality Analysis Improves Survival Prediction in Enucleated Uveal Melanoma Patients
Investigative Ophthalmology & Visual Science 2019 Aug 1;60(10):3595-3605.

25. N. Mehra, **J. van Riet**, M. Smits, H. Westdorp, M. Gorris, T. van Ee, M. van der Doelen, I. van Oort, M. Sedelaar, J. Textor, E. Cuppen, K. Grunberg, M.J.L. Ligtenberg, W. Zwart, A. Bergman, H.J.G. van de Werken, J. Schalken, I.J.M. de Vries, M.P.J.K. Lolkema, W.R. Gerritsen
In-depth assessment of metastatic prostate cancer with high tumour mutational burden
Annals of Oncology (2018) 29.

26. T.G. Meijer, N.S. Verkaik, A.M. Sieuwerts, **J. van Riet**, K.A.T. Naipal, C.H.M. van Deurzen, M.A. den Bakker, H.F.B.M. Sleddens, H.J. Dubbink, T.D. den Toom, W.N.M. Dinjens, E. Lips, P.M. Nederlof, M. Smid, H.J.G. van de Werken, R. Kanaar, J.W.M. Martens, A. Jager, D.C. van Gent
Functional Ex Vivo Assay Reveals Homologous Recombination Deficiency in Breast Cancer Beyond BRCA Gene Defects
Clinical Cancer Research 2018 Dec 15;24(24):6277-6287.

27. M. Smid, R.R.J. Coebergh van den Braak, H.J.G. van de Werken, **J. van Riet**, A, van Galen, V. de Weerd, M. van der Vlugt-Daane, S.I. Bril, Z.S. Lalmahomed, W.P. Kloosterman, S.M. Wilting, J.A. Foekens, J.N.M. IJzermans on behalf of the MATCH study group, J.W.M. Martens, A.M. Sieuwerts
Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons
BMC Bioinformatics volume 19, Article number: 236 (2018).

28. **J. van Riet**, N.M.G. Krol, P.N. Atmodimedjo, E. Brosens, W.F.J. van IJcken, M.P.H.M. Jansen, J.W.M. Martens, L.H. Looijenga, G. W. Jenster, H.J. Dubbink, W.N.M. Dinjens, H.J.G. van de Werken
SNPitty: an intuitive web application for interactive B-allele frequency and copy number

visualization of next-generation sequencing data
The Journal of Molecular Diagnostics 2018 Mar;20(2):166-176.

29. A. Mohd-Sarip, M. Teeuwssen, A.G. Bot, M.J. De Herdt, S.M. Willems, R. Baatenburg de Jong, L.H.J. Looijenga, D. Zatreanu, K Bezstarosti, **J. van Riet**, E. Oole, W.F.J. van IJcken, H.J.G. van de Werken, J.A.Demmers, R. Fodde, C. Peter Verrijzer
DOC1-Dependent Recruitment of NURD Reveals Antagonism with SWI/SNF during Epithelial-Mesenchymal Transition in Oral Cancer Cells
Cell Reports 2017 Jul 5;20(1):61-75..

30. S. Yavuzyigitoglu, W. Drabarek, K.N. Smit, N. van Poppelen, A.E. Koopmans, J. Vaarwater, T. Brands, B. Eussen, H.J. Dubbink, **J. van Riet**, H.J.G. van de Werken, B. Beverloo, R.M. Verdijk, N. Naus, D. Paridaens, E. Kilic, A. de Klein; Rotterdam Ocular Melanoma Study Group
Correlation of gene mutation status with copy number profile in uveal melanoma
Ophthalmology. 2017 Apr;124(4):573-575.

Manuscript(s) in preparation.

31. **J. van Riet***, D.M. Hammerl*, M.A. Komor, Y. Hoogstrate, B. Janssen, R.J.A. Fijneman, H.J.G. van de Werken, G. Jenster, R. Debets
ERG$^+$ PRAD shows distinct immunoregulatory mechanism with enrichment towards innate immune cells and elevated expression of TDO2
Manuscript in preparation.

32. M. Vos, H.J.G. van de Werken, **J. van Riet**, N. Steeghs, M.P.J.K. Lolkema, I.M.E. Desar, J.J. de Haan, H. Gelderblom, J. Nin, E.P.J.G. Cuppen, S. Sleijfer, E.A.C. Wiemer
Genomic landscape of metastatic soft tissue sarcoma reveals new potential actionable targets
Manuscript in preparation.

33. W.S. van de Geer, R.H.J. Mathijssen, **J. van Riet**, Neeltje Steeghs, H.M. Verheul, C. van Herpen, P.O. Witteveen, V.C.G. Tjan-Heijnen, M.P.J.K. Lolkema, E.E. Voest, S. Sleijfer, J.W.M. Martens, E.P.J.G. Cuppen, H.J.G. van de Werken, S. Bins
Landscape of drug transporter gene aberrations in metastatic cancer
Manuscript in preparation.

34. A. Kenter, H. van Willigenburg, F. Schutgens, E. Rentmeester, **J. van Riet**, H.J.G. van de Werken, B. Tan, A. Shankar, M.J. Hoogduijn, G. van Leenders, P. Verhagen, R. de Bruin, P.L.J. de Keizer, W.F.J. van IJcken, J.A.A. Demmers, M.C. Verhaar, G. Jansen, R. Zietse, M.B. Rookmaaker, D. Peters, E. Hoorn, J. Gribnau
Senescence is a hallmark of polycystic kidney disease
Manuscript in preparation.

Shared first-authorship and second-authorship is highlighted by * and **, respectively.

Appendix C

About the author

Job van Riet was born on the 19th of July 1991, Soest, the Netherlands. In 2007 he finished secondary school (MAVO) at the Waldheim, Baarn. He then proceeded to the vocational school (MBO) for IT management and network engineering at ROC ASA, Amersfoort, which he graduated at the top of his class (cum laude), and with one year reduction, in 2010. Subsequently, he pursued his BSc. degree in Bioinformatics at the HAN University of Applied Sciences (Nijmegen) and graduated at the top of his class (cum laude) in 2014.

Thereafter, a short attendance at the MSc. Bioinformatics program at Wageningen University & Research (WUR) was undertaken but was ended prematurely to pursue the unique opportunity as bioinformatician within the Cancer Computational Biology Center (CCBC) whilst pursuing a PhD degree within the Erasmus Medical Center, Rotterdam. This PhD was performed in collaboration between the CCBC and the departments of Urology and Medical Oncology under the combined supervision of Prof. dr. R. de Wit, Prof. dr. ir. G.W. Jenster, Dr. M.P.J.K. Lolkema and Dr. ir. H.J.G. van de Werken. In 2020, Job briefly (re-)visited the group of Prof. dr. ir. E.J.P.G. Cuppen at UMC Utrecht in close collaboration with the group of Dr. M.P.J.K. Lolkema (Erasmus MC) to further our understanding of tumor biology using large-scale genomic analysis from the CPCT-02 study.

Early 2021, Job rejoined the group of Dr. M.P.J.K. Lolkema within the dept. of Medical Oncology to further our understanding of the complex genetics of (metastatic) prostate cancer by interrogating the unique and large whole-genome sequenced metastatic cohorts of the CPCT-02, DRUP and WIDE studies and by taking advantage of liquid biopsies to study heterogeneity and treatment-induced resistance mechanisms. In June 2022, Job joined the group of Prof. dr. Moritz Gerstung (Artificial Intelligence in Oncology) at the Deutsches Krebsforschungszentrum to further our understanding of the complex underlying and interconnected mechanisms of cancer.

Appendix D

Dankwoord

Gedurende mijn PhD heb ik het geluk gehad om vele samenwerkingen aan te gaan en om hierbij veel ontzettend gemotiveerde en fijne collega's te leren kennen.

Als eerste wil ik graag mijn (oud-)collega's van het CCBC bedanken voor de vele discussies, avondjes uit en gezelligheid op het kantoor; Dr. Youri Hoogstrate, Wesley van de Geer, Dr. Alberto Nakauma-González, Rick Jansen, Dr. Arlin Keo, Dr. Lisanne Mout, Niels Krol & Heleen Pruijn. Ook mijn stagiaires door de jaren heen wens ik het allerbeste, ik vond het ontzettend leuk om jullie te mogen begeleiden in allerlei verschillende projecten; Luuk Perdaems, Coen Berns, Yi Ping & Daan Hazelaar (en zelfs ook Wesley eventjes!). Dit proefschrift en de voltallige lijst aan publicaties zouden er uiteraard ook niet zijn zonder de begeleiding en hulp van de *vele* co-auteurs.

Daarnaast wil ik graag de leden van mijn kleine commissie bedanken voor alle ingestoken tijd en moeite voor het beoordelen van dit proefschrift; Prof. dr. V. (Vera) van Noort, Prof. dr. J.B.J. (Joyce) van Meurs & Prof. dr. H.R. (Ruud) Delwel.

Ook wil ik nogmaals mijn dank betuigen aan mijn promotoren en co-promotoren. Ronald en Guido, ontzettend bedankt voor de vele adviezen en inzichten in de verscheidenheid aan disciplines. Een (klein) inkijkje in zowel het klinische en experimentele deel van de wetenschap hebben vele leuke en innovatieve resultaten opgeleverd.

Harmen, ik kan mij nog goed herinneren hoe wij samen op mijn allereerste werkdag naar een symposium in Leuven zijn geweest; met daarna een bezoekje aan de oude brouwzaal van Stella Artois. Daarna zijn er nog vele gezellige dagen geweest en hebben we soms flinke (maar altijd leuke) discussies gehad over van alles. Ik had het niet beter kunnen treffen met jou als co-promoter; er is een stroom aan geweldige publicaties uitgekomen door de jaren heen. Ik wens je alle geluk en plezier in je nieuwe rol bij het CBBI!

Martijn, wij hebben elkaar leren kennen tijdens de CPCT-projecten en ik heb sindsdien veel van je geleerd en je was altijd wel met een sportwijsheid paraat. Er is door de jaren heen van alles voorbijgekomen, van analyses op solide en vloeibare biopten tot het opschrijven van de CABA-V7 studie. Er zijn vele mooie publicaties uit voortgekomen en er komen er binnenkort zelfs nog een aantal bij. Bedankt dat je de gok heb genomen om samen met een bio-informaticus al deze projecten op te pakken.

Dr. Youri Hoogstrate en Krystany Isebia, met jullie als paranimfen en vertegenwoordigers van zowel de bio-informatica en kliniek moet het tijdens de verdediging zeker lukken.

Als laatste wil ik ook mijn familie en vrienden bedanken, met in het bijzonder Sonja en onze Reus.

Propositions

The Era of Next-Generation Sequencing in Clinical Oncology

Job van Riet

1. Open-source software easing the processing, standardization and interpretation of large quantities of biological data are crucial in daily practice and research.
   This Thesis

2. Next-generation sequencing may enable the detailed and extensive discovery of genetic factors driving a multitude of malignancies.
   This Thesis

3. Whole-genome sequencing allows clinically-relevant stratification of malignancies based on large-scale genomic features.
   This Thesis

4. Uncovering the genome of malignancies could lead to the discovery of supplemental or novel treatment-targets for personalized medicine.
   This Thesis

5. Secondary and retrospective analysis of publicly-available next-generation sequenced cohorts allows for complimentary yet original research aims.
   This Thesis

6. We are more efficient in discovering erroneous biological mechanisms than their well-meaning counterparts.

7. In the last decade, we have witnessed tremendous growth in sequencing capability, accompanied by growing sophistication in computational tools. As DNA sequencing is now a commodity, the amount of data generated by the cancer research community will continue to increase.
   I. Cortés-Ciriano et al., "Computational analysis of cancer genome sequencing data.", Nature Reviews Genetics, 1-17 (2021)

8. We must learn to embrace uncertainty.
   V. Amrhein, S. Greenland & B. McShane, "Scientists rise up against statistical significance", Nature 567, 305-307 (2019)

9. We continue to foresee cancer research as an increasingly logical science, in which myriad phenotypic complexities are manifestations of a small set of underlying organizing principles.
   D. Hanahan & R.A. Weinberg, "Hallmarks of Cancer: The Next Generation.", Cell 144, 646-674 (2011)

10. Health is a core human concern, even if it is not consciously considered, or is valued only for instrumental reasons.
    A.D. Napier et al., "Culture and health." The Lancet 384.9954, 1607-1639 (2014)

11. If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat.
    Douglas Adams