Genetic Epidemiology of Cancer in Romania

Paul Iordache

Doctor of Philosophy September 2018

Applied Biostatistics

Ph.D. Dissertation

Genetic Epidemiology of Cancer in Romania

Dissertation of 180 ECTS credits submitted to the School of Science and Engineering at Reykjavík University in partial fulfillment of the requirements for the degree of **Doctor of Philosophy (Ph.D.)** in **Applied Biostatistics**

September 2018

Supervisor:

Bjarni V. Halldórsson, Supervisor

Associate Professor, Reykjavik University, Iceland

Thesis Committee

Andrei Manolescu, Committee Member

Professor, Reykjavík University, Iceland

Thorunn Rafnar, Committee Member

Ph.D., Head of Division of Oncology, DeCODE Iceland

Examiner:

Kathleen C. Barnes, Examiner

PhD, Director, Colorado Center for Personalized Medicine

Head, Division of Biomedical Informatics & Personalized Medicine

University of Colorado, Department of Medicine

Genetic Epidemiology of Cancer in Romania

Paul Iordache

September 2018

Abstract

In Romania, there is a lack of quantitative data associating cancer risk with genetic characteristics. The aim of this thesis is to evaluate genetic risk factors associated with two major cancer types, prostate cancer and colorectal cancer. The first result is the determination of the profile of common prostate cancer risk variants in the Romanian population. The second result is the identification of high-risk mutations in colorectal cancer genes associated with Lynch Syndrome.

Early diagnosis and treatment are key factors in determining the clinical development and survival of cancer patients. Genetic variants that can determine which early stage tumors will progress to an aggressive form of the disease can change decision making for clinicians, patients, and their families. If genetic findings are to be translated into clinical utility, it is critical to understand the particularities of genetic epidemiology of cancer in Romania.

Genetic Epidemiology of Cancer in Romania

Paul Iordache

September 2018

Útdráttur

In Romania, there is a lack of quantitative data associating the cancer risk with genetic characteristics. The aim of the present thesis is to evaluate the genetic risk factors associated with two major cancer types (prostate cancer and colorectal cancer). The first outcome is to determine the profile of common prostate cancer risk variants in the Romanian population. The second outcome is the identification of high-risk mutations in colorectal cancer genes associated with Lynch Syndrome.

Early diagnosis and treatment are key factors in determining the evolution and survival of cancer patients. Genetic variants that can determine which early stage tumors will progress to an aggressive form of the disease can change decision making for clinicians, patients, and their families. If genetic findings are to be translated into clinical utility, it is vital to understand the particularities of genetic epidemiology of cancer in Romania.

The undersigned hereby certify that they recommend to the School of Science and Engineering Reykjavík University for acceptance this Dissertation entitled **Genetic Epidemiology of Cancer in Romania** submitted by **Paul Iordache** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy (Ph.D.) in Applied Biostatistics**

Date 20.09.2018

Bjarni V. Halldórsson, Supervisor

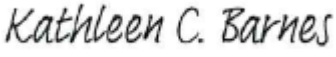
Associate Professor, Reykjavik University, Iceland

Andrei Manolescu, Committee Member

Professor, Reykjavík University, Iceland

Þórunn Rafnar, Committee Member

PhD. Head of Oncology, DeCODE Genetics



Kathleen C. Barnes, Examiner

PhD, Director, Colorado Center for Personalized Medicine

Head, Division of Biomedical Informatics & Personalized Medicine

University of Colorado, Department of Medicine

The undersigned hereby grants permission to the Reykjavík University Library to reproduce single copies of this Dissertation entitled **Genetic Epidemiology of Cancer in Romania** and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the Dissertation, and except as herein before provided, neither the Dissertation nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

20.09,2018

Date

Paul Iordache

Doctor of Philosophy

I dedicate this thesis to Viorel Jinga, Dana Mates and Daniela Pitiogi.

Acknowledgements

This study was funded in part by the European Union FP7 Program (ProMark project 202059) and by the EEA grant (ROMCAN project RO14-0017; EEAJRP-RO-NO-20131-10191).

List of Figures

List of Tables

Table of Contents

Chapter 1: Introduction

The primary objective of these studies is to identify novel cancer predisposition genetic variants through large-scale case-control studies of selected cancers and to investigate how genetic variants may aid in the early detection of disease in the Romanian population. The secondary goal is to integrate the Romanian case-control and cohort studies into international meta-analyses and develop collaborations with large consortia.

To date, genetic epidemiology represents a relatively new field for the Romanian scientific community. In Romania, there is a lack of quantitative data associating cancer risk with lifestyle factors and even less information regarding the genetic characteristics of cancer. Using biological samples and information from the ROMCAN ("Genetic Epidemiology of Cancer in Romania") project, we evaluated genetic risk factors related to prostate and colorectal cancers. The ROMCAN Project aimed to define genetic profiles for breast cancer (BrCa), colorectal cancer (CRC), prostate cancer (PCA) and lung cancer (LuCa) in the Romanian population. An assessment of genetic markers associating with prostate cancer in the Romanian population is compared with previously reported results in other populations. We also discovered genetic markers associated with Lynch Syndrome (LS) providing valuable information on the genetic makeup of familial colorectal cancer in the Romanian population. These studies can be expected to lead to a better understanding of genetic cancer susceptibility of the cancers in Romania and take the first steps towards screening for high-risk families.

1.1 Background

1.1.1 Historical perspective

Genetics is the study of heredity and the variation of inherited elements known as genes. Historically, it took the effort of three individuals: Charles Darwin, Gregor Mendel and Alfred Wallace for genetics to become a unified field.

Charles Darwin developed his theory of natural selection, publishing his observations in the book "On the Origin of Species" in 1859(Darwin, 1859). The theory of natural selection was illustrated for the first time by the research of the Austrian monk Gregor Mendel in 1866(Mendel, 1866). Despite Mendel publishing "Experiments on Plant Hybridization" in 1866, the scientific community overlooked his research until 1900. The rediscovery of Mendel's work sparked the progress in genetic studies of the twentieth century. Alfred Russel Wallace had a critical contribution to the development of the theory of evolution(Lloyd, Wimpenny, & Venables, 2010).

Genetics has an important impact on medicine, primarily on inherited disorders, but increasingly also in wider areas of medical research and practice(Murgatroyd, 2015). Relatively isolated experimental observations in biology in the nineteenth century made its entrance into medicine and exploded in the past century(Radick, 2001).

Wilhelm Weinberg is one of the first pioneers of human and medical genetics; his name is commonly associated with the English mathematician G. H. Hardy in the Hardy–Weinberg equation formulated in 1908(Hardy, 1908; Weinberg, 1908). Wilhelm Weinberg's contribution to segregation analysis was to verify that Mendel's segregation law still held in the setting of human heredity(Weinberg, 1912). He proved that the

proportion of recessive offspring genotypes aa in human parental crossings Aa × Aa (the segregation ratio for such a setting) was indeed p = ¼ (Stark & Seneta, 2013). This discovery provided the basis of the Hardy-Weinberg equilibrium and is one of the fundamental principles used in genetic epidemiology, in particular in genetic association studies. Around the same time as Wilhelm Weinberg and G. H. Hardy made their discoveries, Wilhelm Johannsen formulated the genotype theory defining the terms genotype and phenotype. In 1909(Johannsen, 1909), he made the distinction between the genotypes represented by hereditary dispositions of organisms and phenotypes represented by the ways in which those dispositions manifest themselves in the physical characteristics of those organisms(B. R. Erick Peirson, 2012). The distinction between genotypes and phenotypes represents the cornerstone of genetic epidemiology and the starting point for the development of genome-wide association studies.

The last significant development of this era, coined the “pre DNA era“ was represented by the publication of “The Mechanism of Mendelian Heredity“ in 1915(Morgan, 1915). This study was published by four prominent Drosophila geneticists: Thomas H. Morgan, Alfred H. Sturtevant, Hermann J. Muller, and Calvin B. Bridges. Their results stated that genes form linkage groups on chromosomes inherited in a Mendelian fashion and laid the genetic foundation that promoted Drosophila as a model organism(Bellen & Yamamoto, 2015).

The “DNA era“ was pioneered by James Watson and Francis Crick in their groundbreaking conclusion of 1953: that the DNA molecule exists in the form of a three-dimensional double helix(Watson & Crick, 1953). This discovery was awarded The Nobel Prize in Physiology or Medicine in 1962 "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material." Another Nobel Prize in Physiology or Medicine was awarded in 1959 to Severo Ochoa and

Arthur Kornberg "for their discovery of the mechanisms in the biological synthesis of ribonucleic acid and deoxyribonucleic acid."

Continued discoveries in the field of molecular genetics have improved knowledge of cell cycles and cell functions. Drawing from increasingly detailed and accurate information about the physiological behavior of human body cells, essential aspects of the development of neoplastic processes have been elucidated. Each cell has a control center - the nucleus - which contains the information necessary for the cell for the processes of growth, division and fulfillment of specific functions. Cells use the genes selectively, i.e. some of them are activated or inactivated at certain times of development and differentiation. The processes by which cells establish their patterns of gene expression without affecting the primary DNA sequence information are called epigenetic processes(Felsenfeld, 2014). These represent a whole level of coding that goes further than the primary nucleotide sequence and is the reason why there is such a vast spectrum of ways in which genes can be expressed during development and differentiation.

1.1.2 Genetic Variation

The last two decades have seen extensive efforts to catalog human genetic variation and correlate it with phenotypic differences. The first complete human genome sequence, 17 years ago(Lander et al., 2001; Venter et al., 2001), opened the possibility of investigating the various forms of human genetic variation. Based on the frequency of the minor allele (MAF) in the human population, human genetic variants are classified as common (MAF>5%), low frequency (MAF 1-5%) or rare (MAF<1%).

In terms of nucleotide composition, variants in the human genome can be separated into two different classes: single nucleotide variants and structural variants(Human

Genome Structural Variation Working Group et al., 2007).

Single nucleotide variants or SNPs are the most frequent class of genetic variation observed in the human genome. Sequencing results have estimated that the human genome contains at least 20 million SNPs and 1.5 million insertions-deletions(Gudbjartsson et al., 2015). Also, in each human generation, there are a large number of rare and novel or *de novo* single nucleotide variants, in some cases present in only a family or a single individual(Frazer, Murray, Schork, & Topol, 2009).

Structural variants are characterized as multiple base pairs that differ between individuals and that are not single nucleotide variants. The most common structural variants are insertion-deletions, inversions of DNA sequences and copy number differences. Structural variants represent around 1% of the human genome and influence genome organization, contributing to human disease(Collins et al., 2017; Frazer et al., 2009). Despite the observed genetic variation, the DNA sequences of any two individuals are 99.9 percent identical. The variations, however, may substantially affect an individual's disease risk.

Large numbers of single nucleotide variants on the same chromosome are inherited in blocks; these blocks define haplotypes. Blocks may contain a large number of SNPs, but a few SNPs are enough to identify the haplotypes in a block. The HapMap project was the first effort to define a map of these haplotype blocks and the specific SNPs that identify the haplotypes. The HapMap project provided valuable information by reducing the number of SNPs required to examine a large part of the genome for association with a phenotype from 10 million SNPs to roughly 500,000 tag SNPs. The HapMap project helped genome scan approaches to find regions with genes that affect diseases in a much more efficient and comprehensive way than was possible before.

Launched in 2008 and defined as an international research consortium aiming to

sequence the genomes of at least 1000 individuals, the 1000 Genomes Project aimed to create a deep catalog of human genetic variation(Auton et al., 2015). This project sequenced genomes from more than 1,000 volunteers worldwide ensuring representation of African, Asian and European populations. The 1000 Genomes Project characterized over 95% of variants that are in genomic regions accessible to high throughput sequencing technologies and that have an allele frequency of 1% or greater in one of five major population groups(Abecasis et al., 2010).

The development of next-generation sequencing (NGS) technologies since 2005 has brought revolutionary benefits to medical genetics studies by reducing costs and increasing yield by several orders of magnitude(Metzker, 2010). A comprehensive collection of published germline mutations in nuclear genes that underlie, or are closely associated with, human inherited disease are collected by The Human Gene Mutation Database. At the time of writing (May 2018), the database contained more than 224,000 different gene lesions identified in over 8,000 genes manually curated from over 2,600 journals. The Human Gene Mutation Database represents the central unified gene/disease-oriented repository of heritable mutations causing human genetic disease. It is used worldwide by researchers, clinicians, diagnostic laboratories and genetic counselors, and is an essential tool for the annotation of next-generation sequencing data(Stenson et al., 2017).

Despite all these improvements, understanding the relationship between genotype and phenotype is still one of the central goals in medical genetics and genetic epidemiology. Genome-wide association studies (GWAS) have evolved over the last fifteen years into a tool for investigating the genetic risk factors for human disease. They have identified new genetic risk factors for many common human diseases and have forced the genetics community to think on a genome-wide scale(Bush & Moore, 2012).

Integrating these type of studies in the Romanian medical genetics landscape is crucial for future genetic studies of the population. The data from HapMap allows for the comparison between the LD blocks structure observed in the Romanian population and other European populations. The dataset provided by the full 1000 Genomes Project allowed more accurate imputation of variants in the Romanian GWAS and thus more accurate localization of disease-associated variants. Combining all these databases was pivotal for completing large-scale GWAS´s using Romanian data.

1.2 Genetics of Cancer

Cancer is characterized by a diversity of genetic and epigenetic alterations occurring in both the germline and somatic genomes(Feigelson et al., 2014). Reflecting these two types of genetic alterations, there are two types of approaches to the genetics of cancer. The first approach is examining the inherited risk of cancer defined as susceptibility or predisposition and the second one, the somatic approach, refers to the actual carcinogenic processes on a genetic level. There is an expanding interest in identifying germline genomic variants associated with different types of cancer, the past decade has seen a dramatic increase in the identification of germline variants that associate with the disease.

1.2.1 Germline variants - inherited risk of cancer

The advancement of genome-wide association studies and genome sequencing techniques has led to improvements in the processes of estimating risk of germline mutations in cancer susceptibility genes and assessing risks of cancer based on personal

and family histories. Family history has been examined extensively as a risk factor for cancer. Neoplastic disease serves as a useful model for studying heritability since the familial contribution to this kind of disease risk is usually high in the general population(Pomerantz & Freedman, 2011). One of the most comprehensive studies on familial risk and heritability of cancer was performed using 80,309 monozygotic and 123,382 same-sex dizygotic twin individuals within the population-based registers of the Nordic countries(Mucci et al., 2016). The study reported significant excess familial risk for cancer overall and for specific types of cancer, including prostate, melanoma, breast, ovary, and uterus.

Familial clustering of cancer has proven to be relatively common and is likely to be due to a combination of environmental factors, rare gene mutations with high penetrance, and more common lower penetrant gene variants are acting together to alter disease susceptibility. Only a small proportion of cancers are due to highly penetrant inherited mutations in genes(Hodgson, 2008). As the field of genetics of cancers has matured, alternate methods to assess familial risk have been developed. There are good reasons to expect that common genetic variants explain a large fraction of the inherited risk of the common cancers(Bodmer & Tomlinson, 2010).

1.2.2 Somatic variants

Cancers appear in individual cells that have accumulated one or more mutations in the DNA sequence, mutations that lead to malignant transformations. Most of these mutations affect the genes involved in signaling for cell proliferation, cell cycle control, apoptosis, and DNA repair. Mutations that activate protein function in signal transduction, advance cell cycle progression or inhibit apoptosis are found in dominant oncogenes that

affect cellular phenotype despite the contralateral presence of the normal allele(Abreu Velez & Howard, 2015). Suppressor proteins are affected by the loss or disruption of genes involved in cell cycle control, apoptosis, or DNA repair. Genes encoding cell membrane surface molecules involved in adhesion or growth inhibition may act as tumor suppressors. In some cases, the genes may be haploinsufficient, and the loss or inactivation of a single allele is sufficient to influence the pathogenesis of cancer(Sherr, 2004).

1.2.3 Oncogenes

Oncogenes are variants of normal (proto-oncogenic) genes that have undergone mutations. Proto-oncogenes control the cell cycle by transmitting to the cell information related to the time and frequency of cell divisions. A gain-of-function mutation in a proto-oncogene that has a dominant cellular effect can be enough to initiate oncogenesis (Lodish et al., 2000).

The first information on the influence of viruses on the occurrence of cancers was obtained from studies by Peyton Rous in 1911. In the 60's and 70's, the molecular mechanisms of virus action were identified, so it was demonstrated that Rous sarcoma virus (RSV) is a retrovirus whose RNA genome is reverted into DNA that will be incorporated into the host cell genome(Weiss & Vogt, 2011).

In 1977, Michael Bishop and Harold Varmus showed that normal human cell DNA contains sequences similar to those of retroviruses. These genes have been called proto-oncogenes. Inappropriate activation of these may cause the process of tumorigenesis to occur. About 100 oncogenes have been identified so far(Varmus, 2017). They encode various cellular proteins with essential roles in cell structure and cell function. Activation of oncogenes is involved in various stages of progression of cancer, together with the loss

of tumor suppressor gene function.

Activation mechanisms of oncogenes:

1. Oncogenic point mutations (SNP) - these changes often occur in functionally important parts of genes, resulting in the continuous and uncontrolled activity of proteins involved in intracellular signaling pathways. The RAS gene is an example, but mutations in this gene occur in ~ 15% of human cancers. The mutations lead to a protein that is constitutively active driving cell growth.

2. Gene amplification. This mechanism results in overexpression of the protein encoded by the gene which then leads to overactivation of the respective biological pathways. A good example is the c-Myc proto-oncogene which is amplified in a large fraction of cancers.

3. Chromosomal translocation - acting through two mechanisms:

a) Translocation brings a gene that controls cell growth near a strong gene promoter, resulting in over-expression of the gene. The oncogenic effect of this mechanism is due to a normal protein that has an increased level of expression.

b) Translocation to obtain a hybrid protein with new properties.

4. Insertion of retroviruses into the proto-oncogene sequence (e.g., HTLV1 virus, HPV).

1.2.4 Tumor Suppressor Genes

Tumor suppressor genes control cell division, slowing down this process. Additionally, proteins that are encoded by these genes play a role in repairing DNA damage or control apoptosis. Mutations occurring in tumor suppressor genes lead to loss of protein function and usually have a recessive character in the cell phenotype.

1.3 Medical Genetics of Cancer

1.3.1 The Multistage Model of Cancer Development

Both SNPs and structural variation present in oncogenes or tumor suppressor genes can initiate a carcinogenic process. Carcinogenesis is a long-lasting multistage process involving the following mandatory stages(Hanahan & Weinberg, 2000):

- The initiation of tumorigenesis begins with the carcinogenic insult on the genetic material causing a lesion that can be repaired or not. The time factor is essential in repairing DNA. The impossibility of repairing DNA damage determines the stage of initiation of tumorigenesis.

- Immortalization of tumor cells is another mandatory event of multistage cancer development. Each cell in the human body undergoes a limited number of replications due to the shortening of the chromosomal ends (telomeres) with each replication. Cancer cells acquire unlimited proliferation properties by re-activating telomerase that rebuilds the telomeres. There is a small percentage of cancers in which the unlimited cell proliferative properties are due to other phenomena than telomerase activation.

- Epigenetic changes also have an essential role in tumorigenesis, intervening in maintaining tissue-specific patterns of gene expression.

- Angiogenesis - Tumor cells synthesize numerous proteins that play a role in angiogenesis.

- Resistance to programmed cell death mechanisms. Cells that fail to repair the lesions of the genetic material initiate the action of the p53 gene triggering programmed cell death (apoptosis). Most human cancers have mutations in the p53 gene.

- Resistance to the action of immunological mechanisms of antitumor defense. The immune system recognizes altered cellular proteins on the surface of tumor cells. However, tumors evolve various mechanisms for evading the immune surveillance.

- Metastasis of cancer cells**.**

1.3.2 The Genetic Epidemiology of Prostate Cancer

The familial aggregation of prostate cancer was first reported by Morganti and coauthors in 1956 (Morganti, Gianferrari, Cresseri, Arrigoni, & Lovati, 1956), who observed that patients with prostate cancer reported a higher frequency of the disease among relatives than did hospitalized controls. After this initial observation, several reports have suggested that ∼10% of all prostate cancers are hereditary, with an autosomal dominant inheritance(Carter, Beaty, Steinberg, Childs, & Walsh, 1992). The effect was strongest among first-degree relatives, where the relative risk estimates were in the range of 1.7–3.7(Fincham, Hill, Hanson, & Wijayasinghe, 1990). Also, younger age at diagnosis and multiple relatives with prostate cancer were both associated with even higher relative risks. As described by National Cancer Institute (NCI), factors suggestive of a genetic contribution to prostate cancer include multiple affected first-degree relatives, early-onset prostate cancer and prostate cancer with a family history of other cancers (breast, ovarian, pancreatic).

Linkage studies have been successfully applied for the discovery of highly penetrant cancer genes. For example, multiple linkage signals were identified for breast cancer (BRCA1)(Miki et al., 1994), colon cancer (HNPCC)(Froggatt et al., 1995) and in renal carcinoma (VHL, MET)(Yap et al., 2015). Linkage analysis is based on the observation that genes that reside physically close on a chromosome remain linked during

meiosis(Pulst, 1999). However, prostate cancer is one of the cancer types where highly-penetrant alleles have not been identified using linkage analysis. The general consensus is that this disease is heterogeneous, with both highly and weakly penetrating genetic variants contributing to the phenotype. Through continuous investigations, the discovery of a critical gene for susceptibility to prostate cancer, leading to early diagnosis and deciphering specific etiology, is being attempted.

Through linkage analysis, numerous prostate cancer susceptibility loci have been identified, including: HPC1 (1q24-25), PCaP (1q42.2-43), HPCX (Xq27-28), CAPB (1p36), HPC2 (17p12) and HPC20 (20q13)(Heise & Haus, 2014). These loci have been analyzed from the perspective of their role in increasing the risk of prostate cancer. Common variants of the HPC (hereditary prostate cancer) gene may increase the risk of men with family history but also in case of sporadic occurrence of the disease(Lynch et al., 2016).

The genetic heritability of prostate cancer is largely due to commonly occurring variants conferring lower risks. The number of identified variants has increased dramatically in the last ten years with the development of the genome-wide association study (GWAS) and the collaboration of international consortia that have led to the sharing of large-scale genotype data(Benafif, Kote-Jarai, & Eeles, 2018).

GWAS' have been remarkably successful in identifying common sequence variants affecting the risk of PCA. More than 200 SNPs have been identified at 70 loci, explaining 30% of the familial risk of this disease. These observations agree with a high familial prostate cancer risk proposed by epidemiological studies that estimate a two-to-five-fold increased relative risk of prostate cancer for men with a familial history of PCA(Bratt, Drevin, Akre, Garmo, & Stattin, 2016).

Replication studies for prostate cancer risk variants have become increasingly

important to validate associations in diverse ethnic populations and to help develop models that predict individual disease risks (Ishak & Giri, 2011). Not all reported SNPs have been replicated, and the strength of associations for those that have is often variable across populations(Virlogeux, Graff, Hoffmann, & Witte, 2015). Also, risk allele frequencies (RAF) of prostate cancer-associated SNPs show substantial variation across populations(Eeles et al., 2013).

Most GWAS' on prostate cancer have been conducted in populations with high rates of PSA screening and include indolent disease with undetermined clinical significance. Prostate-specific antigen (PSA) levels have been used for the detection and surveillance of prostate cancer. Not surprisingly, some of the prostate cancer variants reported have subsequently been shown to associate with PSA levels rather than prostate cancer. In our work on prostate cancer, we tried to find sequence variants affecting prostate cancer susceptibility in an unscreened Romanian population using a GWAS.

1.3.3 The Genetic Epidemiology of Colorectal Cancer

Hereditary cancer syndromes are classically characterized by markedly increased lifetime risks of multiple cancers, typically at young ages. Identifying individuals with inherited predispositions to cancer thus greatly impacts risk counseling for affected patients and their families, including the type and timing of cancer surveillance and potential recommendations for prophylactic surgery(Yurgelun et al., 2015).

Lynch syndrome is an autosomal dominant genetic condition that confers a high risk of colon cancer as well as other cancers including endometrial cancer and cancers of the ovary, stomach, small intestine, hepatobiliary tract, upper urinary tract, brain, and skin (Jasperson, Tuohy, Neklason, & Burt, 2010). Lynch syndrome is among the most common

hereditary cancer syndromes, and estimates suggest that as many as 1 in every 300 people may be carriers of an alteration in a gene associated with Lynch syndrome and it may account for as much as 3% of all colon and endometrial cancers(Cohen & Leininger, 2014). Although 30% of individuals diagnosed with CRC report a family history of the disease, only a small fraction carry germline mutations in genes associated with known hereditary cancer syndromes(Stoffel & Kastrinos, 2014).

Lynch syndrome is caused by germline mutations in one of the DNA mismatch repair (MMR) genes (MSH2, MLH1, MSH6, PMS2, EPCAM) and is transmitted in an autosomal dominant fashion(Carethers & Stoffel, 2015; Peltomäki, 2005). The two most commonly mutated genes, MSH2 and MLH1, account for approximately 90% of mutations found in Lynch cases in most populations tested (Carethers & Stoffel, 2015). Mutation of MSH6 and PMS2 are identified in < 10% of Lynch cases, with EPCAM accounting for the rest.

Next-generation sequencing assays are rapidly being incorporated into clinical laboratory practices and have diagnostic applications for hereditary cancer syndromes(Tafe, 2015). Whole exome sequencing (WES) uses next-generation sequencing technology to provide information on nearly all functional, protein-coding regions in an individual's genome(Hitch et al., 2014). Sequencing can provide valuable information regarding pathogenic germline variants in individuals with suspected Lynch syndrome.

In our work, we start assessing the impact of LS variants in CRC in Romania and provide a framework for screening in high-risk families. Our present study was designed to identify high-risk mutations in six CRC genes using whole-genome sequencing (WGS).

1.4 Genetic Epidemiology of cancer in Romania.

The ROMCAN project is an EEA-funded research program that aims to strengthen the basis of genetic epidemiology of cancer in Romania. In the project, a large number of biological samples and clinical data on cancer cases (prostate, lung, breast and lung cancers) and controls were collected from multiple hospitals in Bucharest. All DNA samples are then subjected to whole-genome genotyping and selected samples underwent whole-genome sequencing (WGS) at deCODE genetics.

My Ph.D. thesis, being part of the ROMCAN Project, was focused on a systematic evaluation of genetic risk factors associating with prostate and colorectal (CRC) cancers and providing GWAS results for lung cancer and breast cancer to other ROMCAN researchers for collaborative work. These four types of cancer represent almost half of the overall burden of cancer in the Romanian population. Also, I was aiming to define high-risk groups for whom specific preventive measures can be implemented.

With the ROMCAN Project, the Romanian scientists have been able to start building the necessary infrastructures for transferring and analyzing genetic data. In addition to the grant from the EEA, deCODE funded a large part of the cost of genotyping and sequencing done in the project and had significant involvement in the processing of biological samples, genotyping and analysis of genetic data. As a Ph.D. student at Reykjavik University, I also contributed to setting up the training program for the Romanian partners in addition to the genetic epidemiology studies completed as the main body of his thesis.

Alongside deCODE, we decided to conduct two sizeable genetic epidemiology studies using the ROMCAN data. The first study was a GWAS study investigating the genetic profile of common prostate cancer risk variants in the Romanian population. Because PSA screening is still rare in Romania and most prostate cancer cases have a clinically significant disease at diagnosis, we believe this study population has less

confounders than studies in heavily screened populations. This study provides evidence that a substantial fraction of previously validated prostate cancer variants associate with risk in this unscreened Romanian population with a high proportion of clinically significant disease.

The objective of the second study was to start assessing the impact of LS variants in CRC in Romania and provide a framework for screening in high-risk families. The study was designed to identify high-risk mutations in six CRC genes using WGS. The frequencies of all candidate variants were then assessed in the entire ROMCAN cohort of 688 CRC cases and 4,567 cancer cases (other than CRC) and controls. In addition to the five MMR genes, we focused the analysis on mutations in two other CRC-associated genes, APC and MUTYH. The papers reporting the results of these studies are presented in Chapters 2 and 3 of the results.

To further define the contribution of this work to the field of genetic epidemiology, I was involved in multiple scientific studies both with Icelandic and Romanian partners. A complete list of the projects and short reviews on each of them can be found in Chapter 4 Collaborative Work Results.

Chapter 2 Results: Profile of common prostate cancer risk variants in an unscreened Romanian population.

Paul D. Iordache, Dana Mates, Bjarni Gunnarsson, Hannes P. Eggertsson, Patrick Sulem, Júlíus Guðmundsson, Stefania Benonisdottir, Irma Eva Csiki, Stefan Rascu, Daniel Radavoi, Radu Ursu, Catalin Staicu, Violeta Calota, Angelica Voinoiu, Mariana Jinga, Gabriel Rosoga, Razvan Danau, Sorin Cristian Sima, Daniel Badescu, Nicoleta Suciu, Viorica Radoi, Andrei Manolescu, Thorunn Rafnar, Bjarni V. Halldórsson, Viorel Jinga, Kári Stefánsson.

Abstract:

Background: Genome-wide association studies have yielded a large number of common sequence variants that associate with prostate cancer (PCA) susceptibility. PSA screening is still rare in Romania, and most PCA cases have a clinically significant disease at diagnosis. Here we report the results of a GWAS of PCA in Romania.

Methods: The study population included 990 unrelated pathologically confirmed PCA cases and 1,034 male controls. DNA was genotyped using Illumina SNP arrays, and 24.295.558 variants were imputed using the 1000 Genomes dataset. An association test was performed between the imputed markers and PCA. A systematic literature review for variants associated with PCA risk identified 115 unique variants that were tested in the Romanian sample set.

Results: None of the variants tested in the Romanian GWAS reached genome wide significance (p-value < $5 \cdot 10^{-8}$) but 807 markers had p-values < $1 \cdot 10^{-4}$. Thirty of the previously-reported SNPs replicated (p-value < 0.05), with the strongest associations observed at: 8q24.21, 11q13.3, 6q25.3, 5p15.33, 22q13.2, 17q12 and 3q13.2. The replicated variants showing the most significant association in Romania are rs1016343 at 8q24.21 (P = $2.2 \cdot 10^{-4}$), rs7929962 at 11q13.3 (P = $2.7 \cdot 10^{-4}$) and rs9364554 at 6q25.2 (P = $4.7 \cdot 10^{-4}$).

Conclusion: Here we report the results of the first GWAS of PCA performed in a Romanian population.

Impact: Our study provides evidence that a substantial fraction of previously validated PCA variants associates with risk in this unscreened Romanian population with a high proportion of clinically significant disease.

Introduction:

PCA is the fourth most common cancer and the second most common cancer in men worldwide(Jacques Ferlay et al., 2015). Prostate cancer is the third most commonly diagnosed cancer in Europe and has emerged as the most frequent cancer in men, reaching an age-standardized rate of 96 per 100.000 men in 2012(J Ferlay, Steliarova-Foucher, et al., 2013). Incidence has been increasing rapidly over the past two decades in most European countries, particularly in the wealthiest countries in Northern and Western Europe(Bray, Lortet-Tieulent, Ferlay, Forman, & Auvinen, 2010; J Ferlay, Steliarova-Foucher, et al., 2013). More than 1.1 million new cases of prostate cancer were diagnosed in 2012 worldwide, accounting for approximately 8% of all new cancer cases. The incidence is expected to grow to 1.7 million new cases and 500,000 deaths by 2030 worldwide, mainly due to the growth and aging of the global population(Jacques Ferlay et al., 2010a).

The incidence of prostate cancer differs between countries, in part due to differences in the prevalence of prostate-specific antigen (PSA) screening. PSA screening has a much greater effect on incidence than on mortality; hence, there is less variation in mortality rates worldwide (10-fold) than is observed for incidence (25-fold). In 2012, the age-standardized mortality rate in Europe was 19 per 100.000 men, and the mortality rate was almost the same in developed and developing regions of Europe(J Ferlay, Steliarova-Foucher, et al., 2013; Jacques Ferlay et al., 2010a). Prostate cancer screening with PSA has been shown to decrease prostate cancer mortality in the European Randomized Study of Screening for Prostate Cancer (ERSPC)(Schröder et al., 2014). However, the possibility of negative effects of screening on over-diagnosis and over-treatment cannot be ignored(Gomella et al., 2011). Screen-detected prostate cancer typically runs an indolent

course, less than 13% of those diagnosed will succumb to the disease (Schröder et al., 2014). In order to improve the outcome of screening, it is essential to find prognostic biomarkers that can distinguish between indolent and aggressive disease(Attard et al., 2016). Sequence variants that associate with aggressive PCA could be useful for this purpose.

GWAS‘ have been remarkably successful in identifying common sequence variants affecting the risk of PCA (Mucci et al., 2016). More than 200 SNPs have been identified at 70 loci, explaining 30% of the familial risk of this disease(Eeles et al., 2013). Most GWAS‘ have been conducted in populations with high rates of PSA screening and include indolent disease with undetermined clinical significance. Not surprisingly, some of the PCA variants reported have subsequently been shown to associate with PSA levels rather than PCA(Gudmundsson et al., 2010).

In Romania, the estimated age-standardized incidence of PCA was 37.9 per 100,000 men in 2012 and the estimated age-standardized mortality rate for PCA was 16.9 per 100,000 men(J Ferlay, Steliarova-Foucher, et al., 2013). Due to the poor health status of the Romanian population and difficulties in health care accessibility(Bara, van den Heuvel, & Maarse, 2002), PCA might be an underdiagnosed condition. PSA screening is not common in Romania (Jinga et al., 2016) and consequently, more than 95% of patients have advanced disease at the time of diagnosis(Waidelich et al., 2011). Here we report the first GWAS on PCA in Romania and profile the known PCA risk variants in this population of patients with the clinically significant disease.

Materials and Methods:

Study population

Subjects included in this study were male patients admitted between 2008 and 2012 to two clinics in Bucharest (Urology Clinic "Th. Burghele" and General Surgery Clinic "St. Mary") for various medical conditions. The study consists of 2,024 hospital patients; 990 unrelated histopathologically-confirmed PCA cases, most of which had abnormal PSA levels, and 1,034 controls, consisting of patients admitted for urological and surgical conditions other than cancer. Blood samples were collected for the measurement of biomarkers and genotyping. PSA levels in plasma were measured for all subjects at hospital admission but were not used as exclusion criteria. All subjects gave written informed consent prior to enrolment and accepted the use of personal and clinical data and biological samples for genetic research. The Bioethical Committee of the Romanian College of Physicians approved the study, and the study protocols were approved by the National Ethical Board of the Romanian Medical Doctors Association in Romania. Trained interviewers performed face-to-face interviews, using standardized questionnaires, to collect personal data (ethnicity, marital status, education, height, and weight), lifestyle data (occupation, smoking, coffee and tea consumption) and medical history (personal and familial). All subjects were of self-reported European descent. No significant difference was observed between the average age of the cases (66,9) and controls (64,3). No significant differences were observed in other epidemiological features; BMI, smoking or alcohol consumptions (Table 1).

The UICC–TNM staging system was used(Mohler et al., 2010). For the T stage, more than 75 percent of the cases were graded as T3 or T4. The N and M stages were distributed similarly, a vast majority were staged as Mx or Nx. For the Gleason score, the majority of cases were graded as Gleason 7 or 8 (45.1% and 20.3% respectively). A complete description of the clinical characteristics of the cohort can be found in Table 1.

Genotyping and analysis of SNP data

DNA was extracted from whole blood at deCODE Genetics (Reykjavik, Iceland) and genotyped using Infinium OmniExpress-24 bead chips (Illumina). 716,503 SNPs were genotyped for each individual included in the study. The genotype data were filtered using Plink! v1.07 (O'Connell et al., 2014). Approximately 10% of the SNPs genotyped were removed using a Hardy-Weinberg equilibrium significance threshold of $5 \cdot 10^{-6}$ and by excluding markers with a minor allele frequency lower than 1%. Prior to the imputation, each chromosome was phased in a single run using SHAPEIT(Delaneau, Howie, Cox, Zagury, & Marchini, 2013). Markers from Phase 3 October 2014 of the 1000 genomes(Auton et al., 2015) were imputed into the 2,024 chip-typed individuals using the IMPUTE2 software(Howie et al., 2009) with a posterior probability of 0.9 as a threshold to call genotypes. The set of genotypes were tested for population heterogeneity using principal component analysis in the ADMIXTURE software(Alexander, Novembre, & Lange, 2009) and the results were consistent with a homogeneous population.
A total of 24,295,558 markers were generated by imputation for each individual in the study.

Quality control for the imputation results was performed by removing markers with minor allele frequency less than 1%, call rate of 0.95 and info of 0.8. In total, 8,506,022 markers met the filtering criteria. An association test was performed between the 8.5 million imputed markers and a phenotype represented by positive biopsy for prostate cancer. The association test was calculated using SNPTEST(Marchini & Howie, 2010), using a single binary variable as a response, all reported p-values are two-sided.

Selection of SNPs for replication of previous findings

A systematic literature review of variants associated with prostate cancer from previous GWAS' was completed on October 4th 2016 using the NHGRI catalog of published genome-wide association studies(Welter et al., 2014) as a starting point. A

search query with "prostate cancer "as a keyword was performed and the inclusion criteria for selection were: p-values < $5 \cdot 10^{-8}$ and a minor allele frequency above 5%. For each study, the following variables were collected: country and ethnicity of the participants, genotyping method, the source of controls and source of replication cohort, and a number of cases and controls in both discovery and replication study.

A total of 37 articles were obtained initially from the GWAS catalog based on the keyword search. Twelve of the studies reported results only tangentially related to prostate cancer, while the remaining 25 studies reported associations with prostate cancer risk. After removing duplicate markers, we obtained 173 unique markers. Out of the 173, 58 markers either did not report ORs and corresponding 95% CI or the tested allele. These markers were excluded from the study, resulting in a final set of 115 unique variants used in our replication.

Results:

In order to search for new susceptibility loci for prostate cancer, we tested a total of 8.5 million variants of frequency above 1%. No variants tested in the Romanian GWAS reached genome-wide significance (p-value lower than $5 \cdot 10^{-8}$), while 635 markers showed association p-values < $1 \cdot 10^{-4}$ (Supplementary Table 1) and 41 markers, at 16 genetic loci, showed association p-values < $1 \cdot 10^{-5}$. Figure 1 shows a Manhattan plot of the results. The 16 markers with the lowest p-values at each locus are shown in Table 2. We observe no excess signal in the Q-Q plot when testing all marker (Figure 2-A); the observed p-values (blue line) show a comparable trend to the expected p-values (the red line).

Next, we tested the effect of 115 previously reported PCA variants in the Romanian population. Thirty SNPs from 13 loci replicated in the Romanian cohort (p-value < 0.05)

(Table 3). Eighty-nine (77%) of the markers selected in the systematic literature review show effects consistent with reported studies although the p values were not < 0.05. We observe an excess of the signal in the Q-Q plot when restricted to this set of previously reported variants (Figure 2-B); the observed p-values (blue line) show a steeper slope than the expected p-values (the red line).

Replication, or lack thereof, allows us to refine association signals and rule out associations due to differences in phenotype definitions between cohorts. Compared to the original studies, replication studies may use cohorts with slightly different ethnic and pathologic characteristic. Differences in ethnic characteristics lead to differences in LD structure, and consequently, markers that were previously found to be correlated with a risk variant may not show an association in a population of different ethnicity. We determined whether the effects of the 115 reported SNPs are similar in the Romanian population as in the discovery cohorts, by conducting a weighted linear regression, modeling the relationship between the log-odds ratio of each of the 115 SNP (Figure 3). We observed a highly significant correlation of R=0.66 (p-value= $5 \cdot 10^{-16}$) for the 115 markers represented by the grey (non-replicating) and orange (replicating) dots. Most markers are near the diagonal, indicating that the effect in the Romanian population is similar to that previously reported.

The locus showing the strongest replication in the Romanian GWAS is 8q24 represented by 12 variants with p-values ranging from $2 \cdot 10^{-4}$ to $4 \cdot 10^{-2}$. These 12 SNPs are in high LD (average R^2=0.81) clustering in a 500kb region, all representing the same association signal. The closest gene to this locus is the *MYC* gene. The locus showing the second strongest replication in Romania is 11q13.3 located close to the *MYEVO* gene [20]. This locus is represented by 4 SNPs with p-values between $2.7 \cdot 10^{-4}$ and $2.1 \cdot 10^{-2}$. All 4 SNPs are in high LD (R^2 >0.93) clustering in a 10KB region and represent the same

association signal. This locus was previously reported to associate with early-onset PCA (Lange et al., 2014). We assessed the association with early-onset PCA in the Romanian cohort using the same criteria as in the original study, but could not replicate this result (P=0.41, OR=0.81), possibly due to lack of power in our set of 128 early-onset PCA cases. The locus showing the third strongest replicated association in the Romanian results is 6q25.3, represented by a pair of markers (rs7758229 p= $1.5 \cdot 10^{-3}$ and rs9364554 p=$4.7 \cdot 10^{-4}$) in strong LD ($R^2 > 0.78$). The markers are located in the proximity of *SLC22A3*, a gene that has been implicated in prostate cancer pathogenesis (Grisanzio et al., 2012). The 17q12 locus was replicated by a pair of markers in high LD (rs8064454 p =$3.1 \cdot 10^{-3}$ and rs4430796 p=$1.5 \cdot 10^{-2}$, $R^2 > 0.96$) clustering in a 5KB region next to the *HNF1B* gene, representing the same association signal.

Discussion:

Genetic epidemiology straddles between statistically driven research and research inspired by clinical needs. Genome-wide association studies have successfully yielded loci associated with PCA risk. However none of the variants at these loci conclusively separate aggressive from indolent disease. Most previous GWAS' investigating PCA are based on cohorts including indolent cancer forms, including cases with low stage and grade. In an attempt to search for loci of clinical importance, the present study focused on refining associations in men with clinical presentations and not those identified solely by an elevated PSA. More than 70% of the cases included in our study presented with a Gleason score equal to or greater than 7 and a majority were staged at T3 and T4. This is a clear indication of aggressiveness of the tumors. Therefore, the replicated variants are likely to represent associations with clinically significant disease although they may also associate

with the indolent form of the disease.

At least two studies of similar size have been performed including clinically advanced cases (Schumacher et al., 2011; Sun et al., 2009). In both studies, the patients had less advanced clinical characteristics than the Romanian cohort. In both studies, fewer than 50% of cases presented with stage T3 and T4 or Gleason score equal to or higher than 7(Sun et al., 2009; Tao et al., 2012). Despite the clinically distinct population, no variants tested in the Romanian GWAS reached genome-wide significance (p-value lower than $5 \cdot 10^{-8}$). The GWAS Q-Q plot (Figure 2-A) and the lack of novel genome-wide significant results suggest that our dataset is underpowered to detect genome-wide significant associations on its own.

Although only 30 of the 115 previously-reported markers showed p-values < 0.05, the effects of additional 59 markers were consistent with the reported results. The "winner's curse," the observation that effect sizes are often larger in the populations in which they are discovered, may be one reason why some SNPs failed to replicate, and why ORs were generally smaller in our cohort than previously found(Hoffmann et al., 2015). Previous studies have shown the utility of including functional evaluation, in an attempt to identify candidate risk loci below currently accepted statistical levels of genome-wide significance(Stegeman et al., 2015). Functional characterization of the variants described here remains to be done. However, the GTEx database(GTEx Consortium, 2013), suggests that some of the markers may influence gene expression.

It is interesting to note that many of the variants showing the strongest replication in the Romanian population reside at loci that have been associated with several cancer types, so-called cancer hubs. The locus showing the strongest replication p-value ($2 \cdot 10^{-4}$) in the Romanian GWAS is 8q24, one of the first hotspots for cancer risk alleles reported. In addition to PCA, the locus was previously reported to associate with breast cancer (Easton

et al., 2007), colorectal cancer (Tomlinson et al., 2007; Zanke et al., 2007), ovarian cancer (Goode et al., 2010), pancreatic cancer (Wolpin et al., 2014), renal cell carcinoma(Gudmundsson et al., 2013), urinary bladder cancer (Kiemeney et al., 2008) and Hodgkin's lymphoma(Enciso-Mora et al., 2010). The closest gene to this locus is the *MYC* gene.

A similar situation is found in the case of 11q13.3, a locus associated with breast cancer (Michailidou et al., 2013; Turnbull, Ahmed, et al., 2010) and early onset breast cancer(Ahsan et al., 2014), renal cell carcinoma (Purdue et al., 2011) and multiple myeloma(Weinhold et al., 2013), in addition to PCA(Eeles et al., 2008; Thomas et al., 2008) and early onset PCA(Lange et al., 2014).

Yet another locus replicating in our study that is associated with several types of cancer is the TERT locus at 5p15.33. Variants at this locus have been associated with risk of lung cancer(Landi et al., 2009), pancreatic cancer (Petersen et al., 2010), breast cancer(Haiman et al., 2011), testicular cancer(Turnbull, Rapley, et al., 2010) and bladder cancer(Rafnar et al., 2009). The two markers replicated in this region, rs2242652 and rs7725218, are both located in the intron region of the *TERT* gene, a gene known to be involved in the activation of oncogenic pathways.

Conclusion:

Our study provides evidence that a large fraction of previously validated prostate cancer SNPs associates with risk in the unscreened Romanian population. These variants are likely to have clinical importance and can be considered for inclusion in future risk models of potential clinical utility.

Table 2.1 - Desciption of the Romanian case-control population

Age	% cases(n=990)	% controls(n=1034)
under 50	0.3%	15.5%
50-60	1%	20.1%
60-70	35.5%	26.8%
70-80	44%	31.2%
80-90	15%	6.1%
Over 90	0.2%	0.1%
T Staging	% cases(n=990)	% controls(n=1034)
1A	2.3%	-
1B	1.1%	-
1C	15.6%	-
2A	1.7%	-
2B	1.9%	-
2C	4.0%	-
3A	43.8%	-
3B	6.06%	-
4	23.3%	-
Gleason Score	% cases(n=990)	% controls(n=1034)
2	0.2%	-
3	0.3%	-
4	1%	-
5	3.3%	-
6	13.2%	-
7	45.1%	-
8	20.3%	-
N Staging	% cases(n=990)	% controls(n=1034)
N0	21.5%	-
N1	3.2%	-
Nx	75.3%	-
M Staging	% cases (n=990)	% controls(n=1034)
M0	22%	-
M1	10%	-
Mx	68%	-
PSA levels	% cases(n=990)	% controls(n=1034)
<4	42%	-
4.0-9.99	18%	-
9.99-19.99	12%	-
19.99-49.99	9.5%	-
49.99-99.99	6.7%	-
>100	9.3%	-
NA	1.5%	-
Alcohol Consumptions	%cases (n=979)	%controls (n=1030)
No	43.4%	45.1%
Yes	56.6%	54.9%
Smoking	%cases (n=975)	%controls (n=1022)
No	88.1%	81.9%
Yes	11.9%	19.1%
BMI	%cases (n=979)	%controls (n=1023)
underweight	1.5%	0.5%
normal weight	37.5%	31.5%
overweight	45.6%	47.7%
obese	15.1%	20.3%

Note. Reprinted from "Profile of common prostate cancer risk variants in an unscreened Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine 22(3) · December 2017.

Table 2.2: The variants in the Romanian GWAS with lowest p-values for each locus.

RS ID	Chr	Position	Reference Allele	Tested Allele	Info	MAF* (%)	OR	P-Value
rs55960139	13	95288608	T	C	0.93	22.3	1.45	$1.83 \cdot 10^{-7}$
rs146493482	16	8002169	C	T	0.93	2.4	2.86	$8.25 \cdot 10^{-7}$
rs17467679	2	16133863	A	G	1	37.8	0.74	$9.48 \cdot 10^{-7}$
rs35890542	4	177243229	A	G	1	6.6	0.54	$1.67 \cdot 10^{-6}$
rs187936586	11	21614186	T	C	0.88	2.4	0.38	$3.50 \cdot 10^{-6}$
rs13111983	4	710801	T	G	0.91	27.1	0.74	$4.08 \cdot 10^{-6}$
rs1383	14	73129765	T	A	0.83	23.4	1.36	$4.18 \cdot 10^{-6}$
rs6834053	4	127918594	C	A	0.93	3.9	2.14	$4.70 \cdot 10^{-6}$
rs35544574	13	37172379	CAA	C	0.95	9.2	1.6	$4.72 \cdot 10^{-6}$
rs74437803	22	17089228	G	A	0.81	8.4	0.63	$5.60 \cdot 10^{-6}$
rs71751677	16	11314438	GTGTTT	G	0.86	48.7	0.78	$6.16 \cdot 10^{-6}$
rs201872456	2	115129517	C	G	0.8	17.3	1.38	$6.59 \cdot 10^{-6}$
rs183478269	1	161032417	G	C	0.82	1.5	3.08	$6.97 \cdot 10^{-6}$
rs13253942	8	126154649	G	A	1	9.5	1.63	$7.93 \cdot 10^{-6}$
rs148921321	8	76468497	C	T	0.93	2.1	0,31	$8.67 \cdot 10^{-6}$
rs133917	22	44524314	C	T	0.81	47.3	1.27	$8.89 \cdot 10^{-6}$

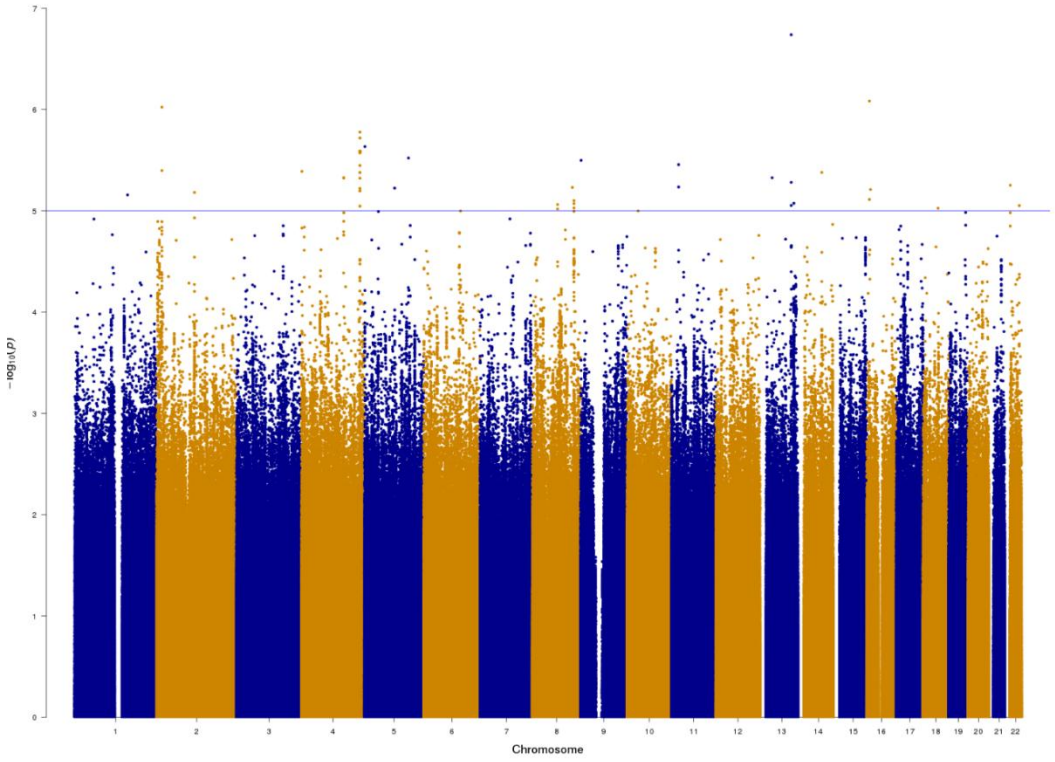
Note. Reprinted from "Profile of common prostate cancer risk variants in an unscreened Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine 22(3) · December 2017.

Table 2.3: Previously reported PCA risk markers that associate with PCA risk in the Romanian population with a p value < 0.05

Rs Number	Chr	Position	OR (95%CI)	P-value	Tested Allele	Mapped Gene
rs636291	1	10496040	1.19 (1.05, 1.35)	$8 \cdot 10^{-3}$	A	PEX14
rs1218582	1	154861707	1.13 (1.01, 1.26)	$4 \cdot 10^{-2}$	G	KCNN3
rs7611694	3	113556777	1.14 (1.01, 1.29)	$3 \cdot 10^{-2}$	A	SIDT1
rs7679673	4	105140377	1.15 (1.02, 1.29)	$2 \cdot 10^{-2}$	C	TET2
rs2242652	5	1279913	1.23 (1.06, 1.42)	$5 \cdot 10^{-3}$	C	TERT
rs7725218	5	1282299	1.21 (1.09, 1.36)	$7 \cdot 10^{-4}$	G	TERT
rs9364554	6	160412632	1.28 (1.11, 1.47)	$4 \cdot 10^{-4}$	T	SLC22A3
rs7758229	6	160419220	1.23 (1.08, 1.40)	$1 \cdot 10^{-3}$	T	SLC22A3
rs1016343	8	127081052	1.30 (1.13, 1.50)	$2 \cdot 10^{-4}$	T	PCAT2, PRNCR1
rs13254738	8	127092098	1.13 (1.01, 1.28)	$4 \cdot 10^{-2}$	C	PCAT2, PRNCR2
rs12682344	8	127094539	1.58 (1.13, 2.20)	$7 \cdot 10^{-3}$	G	PCAT2, PRNCR3
rs6983561	8	127094635	1.58 (1.13, 2.21)	$7 \cdot 10^{-3}$	C	PCAT2, PRNCR4
rs16901979	8	127112671	1.58 (1.13, 2.21)	$7 \cdot 10^{-3}$	A	PCAT2, PRNCR5
rs10505483	8	127112950	1.58 (1.13, 2.21)	$7 \cdot 10^{-3}$	T	PCAT2, PRNCR6
rs445114	8	127310936	1.23 (1.09, 1.39)	$1 \cdot 10^{-3}$	T	PCAT2, PRNCR7
rs6983267	8	127401060	1.16 (1.03, 1.31)	$1 \cdot 10^{-2}$	G	PCAT2, PRNCR8
rs1447295	8	127472793	1.35 (1.11, 1.64)	$3 \cdot 10^{-3}$	A	PCAT2, PRNCR9
rs4242382	8	127505328	1.33 (1.10, 1.62)	$3 \cdot 10^{-3}$	A	PCAT2, PRNCR10
rs4242384	8	127506309	1.33 (1.10, 1.62)	$3 \cdot 10^{-3}$	T	PCAT2, PRNCR11
rs10090154	8	127519892	1.32 (1.09, 1.60)	$4 \cdot 10^{-3}$	T	PCAT2, PRNCR12
rs11228565	11	69211113	1.19 (1.03, 1.38)	$2 \cdot 10^{-2}$	A	MMP7, MMP20
rs7929962	11	69218116	1.25 (1.11, 1.40)	$3 \cdot 10^{-4}$	T	MMP7, MMP20
rs7931342	11	69227030	1.23 (1.09, 1.39)	$6 \cdot 10^{-4}$	G	MMP7, MMP20
rs10896449	11	69227200	1.24 (1.10, 1.40)	$3 \cdot 10^{-4}$	G	MMP7, MMP20
rs11568818	11	102530930	1.19 (1.05, 1.34)	$5 \cdot 10^{-3}$	A	MMP7, MMP20
rs10875943	12	49282227	1.16 (1.02, 1.32)	$3 \cdot 10^{-2}$	C	TUBA1C
rs4430796	17	37738049	1.16 (1.03, 1.31)	$2 \cdot 10^{-2}$	A	HNF1B
rs8064454	17	37741595	1.20 (1.06, 1.35)	$3 \cdot 10^{-3}$	C	HNF1B
rs2735839	19	50861367	1.20 (1.01, 1.42)	$3 \cdot 10^{-2}$	G	KLK3
rs5759167	22	43104206	1.21 (1.08, 1.36)	$1 \cdot 10^{-3}$	G	BIK

Note. Reprinted from "Profile of common prostate cancer risk variants in an unscreened Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine 22(3) · December 2017.

Figure 2.1: Manhattan plot of GWAS findings in the Romanian sample.



Y-axis shows –log10 P-values and x-axis shows the chromosomal position

Note. Reprinted from "Profile of common prostate cancer risk variants in an unscreened Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine 22(3) · December 2017.

Figure 2.2: Q-Q plot of the association's results.

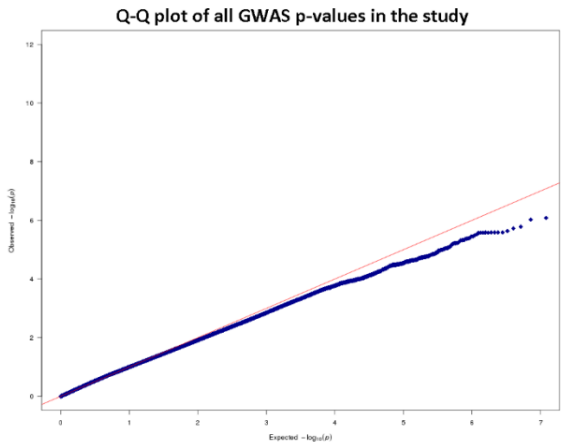


Figure 2 - A

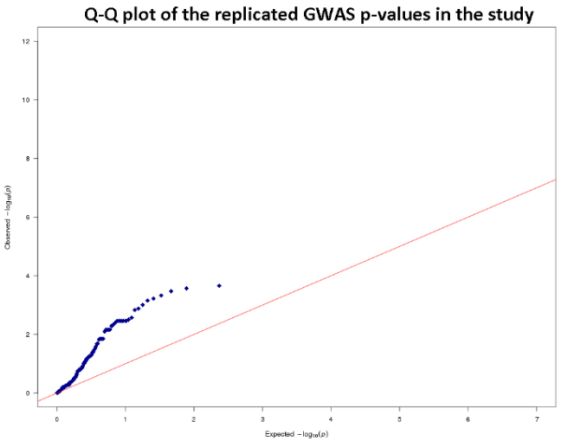
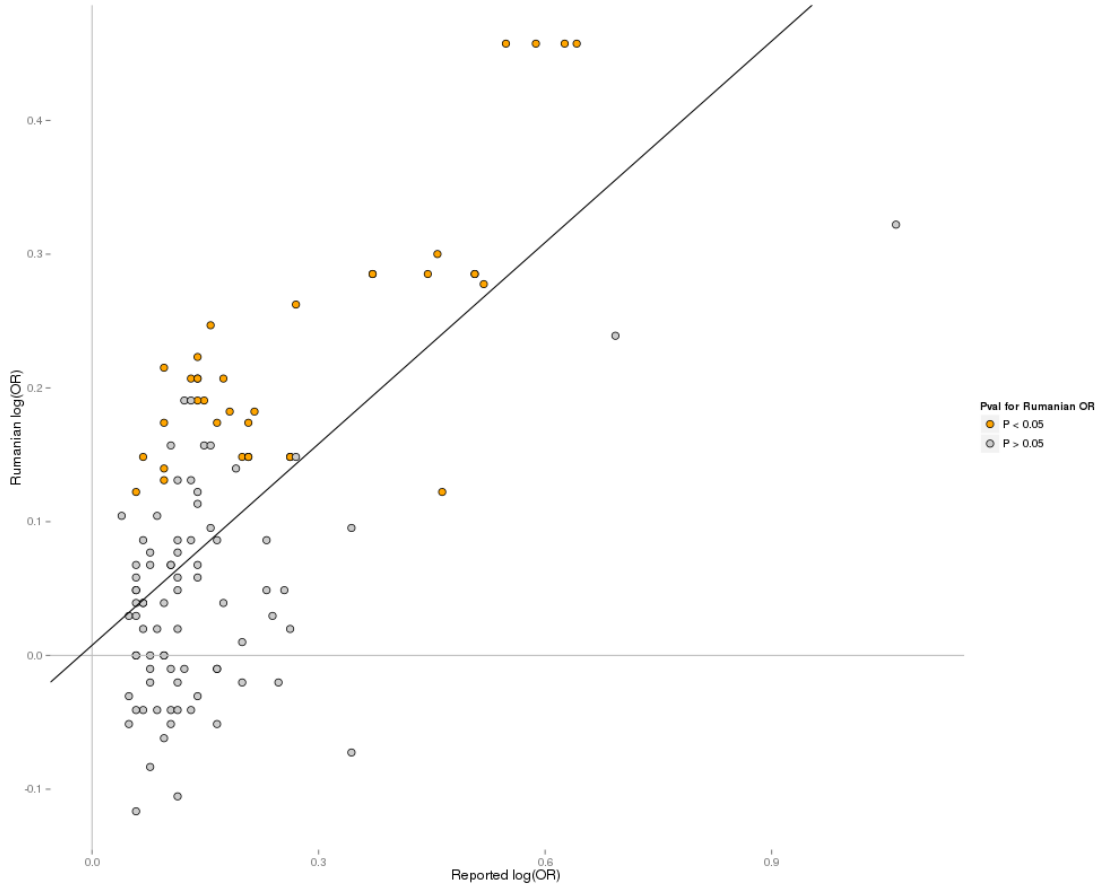


Figure 2 - B

Blue dots show observed p-values, and the red line shows expected p-values. Figure 2-A shows results from genome-wide analysis; Figure 2-B shows results when restricted to GWAS catalog markers.

Note. Reprinted from "Profile of common prostate cancer risk variants in an unscreened Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine 22(3) · December 2017.

Figure 2.3: Scatter plot showing the association of the 115 previously-reported SNPs with PCA (log(OR)) in the Romanian dataset (x-axis) and in reported articles (y-axis).



Note. Reprinted from "Profile of common prostate cancer risk variants in an unscreened Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine 22(3) · December 2017.

Chapter 3 Results: Identification of Lynch Syndrome risk variants in the Romanian population.

Paul D. Iordache, Dana Mates, Bjarni Gunnarsson, Hannes P. Eggertsson, Patrick Sulem, Stefania Benonisdottir, Irma Eva Csiki, Stefan Rascu, Daniel Radavoi, Radu Ursu, Catalin Staicu, Violeta Calota, Angelica Voinoiu, Mariana Jinga, Gabriel Rosoga, Razvan Danau, Sorin Cristian Sima, Daniel Badescu, Nicoleta Suciu, Viorica Radoi, Ioan Nicolae Mates, Mihai Dobra, Camelia Nicolae, Sigrun Kristjansdottir, Jon G. Jonasson, Andrei Manolescu, Gudny Arnadottir, Brynjar Jensson, Aslaug Jonasdottir, Asgeir Sigurdsson, Louise le Roux, Hrefna Johannsdottir, Thorunn Rafnar, Bjarni V. Halldorsson, Viorel Jinga and Kari Stefansson.

Abstract:

Two familial forms of colorectal cancer (CRC), Lynch syndrome (LS) and Familial Adenomatous Polyposis (FAP), are caused by rare mutations in DNA mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*) and the genes *APC* and *MUTYH*, respectively. No information is available on the presence of high-risk CRC mutations in the Romanian population. We performed whole-genome sequencing of 61 Romanian CRC cases with a family history of cancer and/or early onset of disease, focusing the analysis on candidate variants in the LS and FAP genes. The frequencies of all candidate variants were assessed in a cohort of 688 CRC cases and 4,567 controls. Immunohistochemical (IHC) staining for *MLH1*, *MSH2*, *MSH6*, and *PMS2* was performed on tumor tissue. We identified 11 candidate variants in 11 cases; six variants in *MLH1*, one in *MSH6*, one in *PMS2* and three in *APC*. Combining information on the predicted impact of the variants on the proteins, IHC results and previous reports, we found three novel variants likely to be pathogenic (*MSH6*:p.Arg1068Ter, *MLH1*:p.Ala586CysfsTer7, *PMS2*:p.Arg211ThrfsTer38), and two novel variants that is unlikely to be pathogenic. Also, we confirmed three previously published pathogenic LS variants and suggested to reclassify a previously reported variant of uncertain significance (VUS) to pathogenic (*MLH1*:c.1559-1G>C).

KEYWORDS: Lynch Syndrome; Romania; Colorectal Cancer; *MLH1*, *PMS2*, *MSH2*, *APC*, *MSH6*, *MUTYH*.

Introduction:

Colorectal cancer (CRC) is the third most common cancer worldwide and the fourth most common cause of death from cancer (Jacques Ferlay et al., 2010b), causing an estimated 8% of all cancer deaths. Lynch syndrome (LS) or hereditary nonpolyposis colorectal cancer is an autosomal dominant syndrome that accounts for about 1 to 3% of all CRC cases(Aaltonen et al., 1998; Samowitz et al., 2001). LS is the most common inherited cause of CRC and is caused by pathogenic germline mutations in one of four DNA mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*)(Peltomäki, 2005). Carriers of LS mutations have an estimated 25% to 75% lifetime risk of colorectal cancer as well as increased risk of several other cancer types, including endometrial and ovarian cancer(Weissman et al., 2012). Due to the low frequency of LS and heterogeneity in phenotypic expression, it has proven difficult to establish population prevalence accurately and to assess the penetrance of LS mutations(Haraldsdottir et al., 2017).

About 15% of CRC cases are somatically hypermutated as a consequence of MMR deficiency. Of this 1-3 % are due to LS while most of the remaining MMR deficient tumors have somatic inactivation of *MLH1* via hypermethylation of the gene promoter(Haraldsdottir et al., 2016). In MMR deficient tumors, both copies of the same MMR gene have been inactivated, resulting in no production of the respective protein product. MMR deficient tumors exhibit several clinical characteristics that have implication for therapy, in particular with regard to the use of immune system modulators(Gatalica, Vranic, Xiu, Swensen, & Reddy, 2016). Therefore, universal screening of CRC tumors, using microsatellite testing or immunohistochemistry for the MMR proteins, has been recommended by several groups in Europe and the US(Richman, 2015).

To date, nothing is known about the prevalence of LS in the Romanian population,

an d no mutations have been reported. The ROMCAN Project ("Genetic Epidemiology of Cancer in Romania") started in 2012 with the major aim of characterizing genetic risk factors for CRC, breast cancer (BRC), prostate cancer (PCA) and lung cancer (LuCa) in the Romanian population. A second goal of the ROMCAN Project is to define high-risk groups for whom specific preventive measures can be implemented.

The objective of this study is to start assessing the impact of LS variants in CRC in Romania and provide a framework for screening in high-risk families. Our present study was designed to identify high-risk mutations in six CRC genes using whole-genome sequencing (WGS). The frequencies of all candidate variants were then assessed in the entire ROMCAN cohort of 688 CRC cases and 4,567 controls. In addition to the four MMR genes we focused the analysis on mutations in two other CRC-associated genes, *APC* and *MUTYH* (Joint Test and Technology Transfer Committee Working Group, American College of Medical Genetics, 9650 Rockville Pike, Bethesda, MD 20814-3998, United States., Murphy, Petersen, Thibodeau, & Fishel, 2000).

Materials and methods:

Selection of CRC cases for WGS: We selected 61 CRC cases from the ROMCAN and ProMark projects sample collection, a hospital-based sample set of 4,567 cancer cases and controls recruited from 5 major hospitals in Bucharest between 2008 and 2017. The 61 CRC cases were selected using the following criteria: age at diagnosis lower than 40 years or family history of CRC, endometrial or gastrointestinal tumors. Two of the selected cases are of Roma origin.

All subjects gave written informed consent prior to enrolment and accepted the use of personal and clinical data and biological samples for genetic research. The Bioethical Committee of the Medical School "Carol Davila "approved the study protocols. Trained

interviewers performed face-to-face interviews, using standardized questionnaires, to collect personal data (ethnicity, marital status, education, height, and weight), lifestyle data (occupation, smoking history, coffee and tea consumption) and medical history (personal and familial). A description of relevant epidemiological and clinical information can be found in Table 1.

Whole genome sequencing and variant calling: DNA isolated from buccal samples from the 61 individuals was subjected to WGS to an average targeted depth of 30x. The samples were prepared following the TruSeq Nano sample preparation method and sequenced on Illumina HiSeq X machines. Sequencing reads were aligned to build 38 of the human reference sequence (GRCh38) using the Burrows-Wheeler Aligner (BWA)(Li & Durbin, 2009). Alignments were merged into a single BAM file and marked for duplicates using Picard. Only non-duplicate reads were used for the downstream analyses. Variants were called using version 3.8-0 of the Genome Analysis Toolkit (GATK) (McKenna et al., 2010), using a multi-sample configuration.

Variant annotation and filtering: Variants were annotated using release 8.0 of the Variant Effect Predictor (VEP-Ensembl) (McLaren et al., 2016). To filter out variants over a certain frequency threshold we used a reference set of 38,000 non-Romanian individuals whole-genome sequenced at deCODE genetics, an extension of a previously described set of 15,220 WGS Icelanders (Jónsson et al., 2017). Additional frequency filtering was performed using alleles from publicly available datasets of the Exome Aggregation Consortium(Lek, 2017).

Genetic analysis: Only rare (below 1% allelic frequency) coding and splice region variants were considered, including variants with predicted high (stop, frameshift, and splice essential) and moderate (missense, inframe and splice region) impact on protein function. We focused on single-nucleotide polymorphisms (SNPs) and small indels (< 20

base pairs). We analyzed pathogenic and expected pathogenic mutations in 6 genes, defined by the American College of Medical Genetics to be high-risk genes in colorectal cancer: *MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC* and *MUTYH*(Kalia et al., 2017).

Frequency assessment in the Romanian population: All 11 coding variants found in the study were genotyped in the entire ROMCAN sample collection of 688 colorectal cases, 254 breast cancer cases, 1,457 prostate cancer cases, 1,317 lung cancer cases and 1,409 cancer-free controls. The variants were genotyped using one of two assays: Centaurus [12] or KASP [13]. Sanger sequencing was used for one variant that failed in both assays. The primer sequences for the assays are listed in Supplementary Table 1.

Immunohistochemistry (IHC): Paraffin blocks with tumor samples from all 11 carriers of variants in t he LS genes were collected and sections from them stained for *MLH1*, *MSH6*, *PMS2* to assess if the protein was present. Immunohistochemistry was performed on 3 μm sections. Following deparaffinization in xylene, samples were rehydrated in ethanol and subjected to heat-induced epitope retrieval (HIER) (Tris/EDTA buffer, pH 9) in a 98.2°C water bath. Endogenous peroxidase activity was blocked with 3% hydrogen peroxide (DAKO). After incubation with the respective primary antibodies for 30 min. at 20°C, EnVision FLEX Kit (DAKO ref:K8000) was used for detection. Antibodies used: Anti-Human MutL Protein Homolog 1, Clone ES05, (DAKO ref:IR079), Anti-Human Postmeiotic Segregation, Clone EP51 (DAKO ref:IR087, Anti-Human MutS Protein Homolog 6, Clone EP49 (DAKO ref:IR086)(Peltomäki, 2003).

Results:

Sequencing of the 61 CRC patients revealed 11 rare coding variants in CRC genes: 6 variants in *MLH1*, 1 variant in *MSH6*, 1 variant in *PMS2* and 3 variants in *APC*. By examining literature and clinical trial submission data collected by ClinVar with respect to

pathogenicity,6 out of the 11 variants have been previously reported by ClinVar. All 6 previously reported variants have a frequency lower than 1% in the ExAC database(Lek, 2017).

For each of the eleven variants, we assessed whether the variants had a high or moderate impact on protein function, based on the location of the mutation within its gene and its predicted molecular consequences (Table 2); frameshift, splice donor or acceptor and stop-codon gain variants are predicted to have a high impact while in-frame, missense and splice region variants are predicted to have a modarate impact. To assess the frequencies of the variants in the Romanian population, we genotyped all 11 variants in the ROMCAN cohort: 688 colorectal cases and 4,567 cases with cancers other than CRC and controls (254 breast cancer cases, 1457 prostate cancer cases, 1317 lung cancer cases and 1409 cancer-free controls) (Table 3). To test for loss of protein product, we stained MLH1, MSH6 and PMS2 in tumor tissue from paraffin blocks, collected from the carriers of the coding variants.

We divide our results into 2 categories; novel variants and previously documented variants. We summarized all reports regarding the pathogenicity of the previously-reported variants in Supplementary Table 2, using the output from the ClinVar database. An overview of personal and familial history of cancer for the eleven carriers is listed in Table 4.

Novel variants:

MLH1: MLH1: c.251_255delAACTG is a frameshift variant with predicted amino acid change Lys84ThrfsTer4, and consequently assessed as a high impact variant. IHC staining of the tumor of the carrier of this mutation revealed a loss of *MLH1* protein, indicating pathogenicity. We classify this variant as a pathogenic for Lynch syndrome based on the impact of the frameshift mutation and IHC results.

MLH1:c.1755dupT is a frameshift variant with predicted amino acid change Ala586CysfsTer7, and consequently annotated as a high impact variant. IHC staining of the patient‘s tumor revealed a loss of *MLH1* protein. Based on the impact of the frameshift mutation and IHC result, we classify this variant as a pathogenic for Lynch syndrome.

MLH1:c.2104-6T>C is a splice region variant, and consequently annotated as having moderate impact. IHC staining of the patient‘s tumor did not reveal a loss of *MLH1* protein expression. Our results do not support that his variant should be classified as pathogenic.

PMS2: The frameshift variant *PMS2*:c.630dupA results in the predicted protein change Arg211ThrfsTer38, and consequently annotated as having high impact. IHC staining showed a loss of *PMS2* protein expression in the carrier of this variant. The patient was diagnosed at age 44 and had extensive family history of gastrointestinal tract cancers. Based on the impact of the frameshift mutation, and IHC, we classify this variant as pathogenic for Lynch syndrome.

APC: The missense variant *APC*:c.5116T>A results in a protein change of Ser1706Thr, and is consequently annotated as having moderate impact. It has not been previously reported either by ClinVar or other studies. According the ACMG guidelines for classification of sequence variants we consider this variant to be likely benign. The carrier of this mutation also had a previously-reported *APC* variant of uncertain significance, *APC*:c.2780C>G (Ala927Gly) described below.

Previously documented variants:

MSH6: MSH6:c.3202C>T (RS63749843), is a stop-gained variant, assessed as a high impact variant with a protein change of Arg1068Ter. This variant was previously reported as pathogenic in ClinVar by 10 different submitters involved in clinical testing

and research. Tumor sample from the carrier of this variant had loss of *MSH6* protein expression, further supporting its pathogenicity.

MLH1: MLH1:c.1148T>C (RS141344760) is a missense variant, and consequently of moderate predicted impact, resulting in a protein change of Met383Thr. IHC staining of the carrier‘s tumor showed positive protein expression for *MLH1*. The variant was previously reported as being of uncertain significance in ClinVar by 5 different submitters from clinical testing and research and our results do not support pathogenicity.

MLH1:c.1559-1G>C is a splice acceptor variant, annotated as high impact. This was the only candidate mutation found in 2 CRC cases. The tumor samples from both carriers had lost MLH1 protein and both had documented family history or CRC. The variant has been reported previously by 2 different submitters as likely pathogenic for Lynch syndrome and our results indicate that the variant is pathogenic.

MLH1:c.2041G>A (rs63750217) is a missense with a predicted protein change of Ala681Thr, and consequently of moderate impact. The tumor of the carrier had loss of *MLH1* protein staining. This variant was previously reported as pathogenic in ClinVar by 9 different submitters from clinical testing and research. In addition, OMIM has classified the variant as pathogenic for Lynch syndrome II.

APC: APC:c.2780C>G (rs587781500) results in the amino acid change Ala927Gly and consequently of moderate impact. It was previously reported as a variant of uncertain significance in ClinVar by 4 different submitters from clinical testing. As mentioned above, the carrier also had another likley benign variant in *APC*, *APC*:c.5116T>A (Ser1706Thr).

APC:c.3682C>T is a stop gained variant resulting in the protein change Gln1228Ter, and consequently of high impact. It was reported in ClinVar by a single submitter as pathogenic for familial multiple polyposis syndrome and was found in a recent

study investigating somatic *APC* mutations and loss of heterozygosity (LOH) status for 630 patients with sporadic CRC[14].

Discussions:

This study is the first assessment of rare variants underlying LS in colorectal cancer patients in Romanians. We identify new variants specific to the Romanian population and show that some variants previously reported to be pathogenic in other populations also occur in Romania.

We identified three novel pathogenic variants, two novel variants that is unlikely to be pathogenic. Also, we confirmed three previously published pathogenic variants and suggested to reclassify a variant previously classified as VUS as pathogenic. Due to study limitations, we were not able to classify the three *APC* variants identified in the Romanian population. We note that out of the two rare missense variants in APC identified in the same individual, we classify one as a likely benign variant based on ACMG's guidelines for classification of sequence variants(Gussow, Petrovski, Wang, Allen, & Goldstein, 2016). The other variant, p.Ala927Gly, has been reported previously as a VUS, but we note that it is located within a critical domain, intolerant to mutations. Our present study is the first one, to our knowledge, to examine rare sequence variants associated with colorectal cancer in the Romanian population.

In order to determine the prevalence of these variants in Romania, we assessed the frequencies of the 11 variants in the full ROMCAN cohort. As described in Table 3, none of the mutations were found in more than 1 CRC patients except for *MLH1*:c.1559-1G>C. Our results do not suggest any strong association between the eleven variants identified here and breast, lung or prostate cancer.

Identification of LS variants in the Romanian population is important in order to reduce the incidence and mortality of this multicancer disorder. Our present study is the largest effort, to our knowledge, to examine the genetic profile of this pathology in Eastern Europe. This study is the first step towards improving our understanding of the genetic particularities of this pathology in Romania and provides new insights for the scientific community studying the genetic epidemiology of LS.

Table 3.1. Patient and tumour characteristics of the 61 CRC cases selected for whole-genome sequencing

	N	%
TNM- T 1	11	18,03
TNM- T 2	35	57,38
TNM- T 3	8	13,11
TNM- T 4	2	3,28
TNM- T X	5	8,2
TNM-N X	17	27,87
TNM-N 0	2	3,28
TNM-N 1	8	13,11
TNM-N 2	8	13,11
TNM-N 3	21	34,43
TNM-N NA	5	8,2
TNM-N total	61	100
TNM-M 0	48	78,69
TNM-M 1	2	3,28
TNM-M X	1	1,64
TNM-M NA	10	16,39
Gender M	40	65,57
Gender F	21	34,43
ICD-O code C18.0	3	4,92
ICD-O code C18.2	5	8,2
ICD-O code C18.3	2	3,28
ICD-O code C18.4	5	8,2
ICD-O code C18.5	2	3,28
ICD-O code C18.6	6	9,84
ICD-O code C18.7	12	19,67
ICD-O code C18.8	1	1,64
ICD-O code C19.9	8	13,11
ICD-O code C20.9	16	26,23
ICD-O code NA	1	1,64
Age 30-39	9	14,75
Age 40-49	16	26,23
Age 50-59	14	22,95
Age 60-69	15	24,59
Age 70-79	7	11,48
Morphology M8480/3	1	1,64
Morphology M8144/3	1	1,64

Morphology M8140/3	59	96,72

Note. Reprinted from "Identification of Lynch Syndrome risk variants in the Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine

Table 3.2: Description of the 11 variants in CRC-associated genes observed in the Romanian population.

Position	Ref	Alt	Consequence	ClinVar	Impact	Gene	IHC	Nucleotide change	Predicted protein effect	Exon	APRP	Comment
chr2:47803449	C	T	stop gained	P	HIGH	MSH6	-	NM_000179.2:c.3202C>T	NP_000170.1:p.Arg1068Ter	5/10	P	
chr3:37000997	AAACTG	A	frameshift	NA	HIGH	MLH1	-	NM_000249.3:c.251_255delAACTG	NP_000240.1:p.Lys84ThrfsTer4	3/19	P	Novel
chr3:37025746	T	C	missense	LP	MODERATE	MLH1	+	NM_000249.3:c.1148T>C	NP_000240.1:p.Met383Thr	12/19	VUS	
chr3:37040185	G	C	splice acceptor	VUS	HIGH	MLH1	-	NM_000249.3:c.1559-1G>C	.	intron	P	Suggest to classify as P
chr3:37047540	C	CT	frameshift	NA	HIGH	MLH1	-	NM_000249.3:c.1755dupT	NP_000240.1:p.Ala586CysfsTer7	16/19	P	Novel
chr3:37048955	G	A	missense	P	MODERATE	MLH1	-	NM_000249.3:c.2041G>A	NP_000240.1:p.Ala681Thr	18/19	P	
chr3:37050480	T	C	splice region	NA	MODERATE	MLH1	+	NM_000249.3:c.2104-6T>C	.	intron	NP	Novel
chr5:112838374	C	G	missense	VUS	MODERATE	APC	NA	NM_000038.5:c.2780C>G	NP_000029.2:p.Ala927Gly	16/16	NA	
chr5:112839276	C	T	stop gained	P	HIGH	APC	NA	NM_000038.5:c.3682C>T	NP_000029.2:p.Gln1228Ter	16/16	NA	
chr5:112840710	T	A	missense	NA	MODERATE	APC	NA	NM_000038.5:c.5116T>A	NP_000029.2:p.Ser1706Thr	16/16	NA	Novel
chr7:5999182	G	GT	frameshift	NA	HIGH	PMS2	-	NM_000535.5:c.630dupA	NP_000526.1:p.Arg211ThrfsTer38	6/15	P	Novel

Position – position of the variant in build38, Ref- Reference allele, Alt- Alternative allele, ClinVar classification - P for pathogenic, VUS for variant of uncertain significance, LP for likely pathogenic, NP for not pathogenic and NA for not listed in ClinVar, IHC – Results of protein staining of the individuals tumor, Exon – location of mutation/ total number of exons, APRP – Assessment of pathogenicity in the Romanian population P for pathogenic, VUS for variant of uncertain significance and NA for not available, Comment – this indicates the

status of the variant compared to ClinVar reports.

Note. Reprinted from “Identification of Lynch Syndrome risk variants in the Romanian population.” by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine

Table 3.3: Frequencies of the 11 variants in CRC-associate genes observed in the Romanian population

Position	Reference allele - Alternative allele	N carriers/N controls genotyped	N carriers/ N lung cancer cases genotyped	N carriers/ N breast cancer cases genotyped	N carriers/ N prostate cancer cases genotyped	N carriers/ N colorectal cancer cases genotyped	N alleles in EXAC/ Total alleles in EXAC
chr2:47803449	C/T	0/1388	0/1148	0/248	0/1446	1/655	18/121286
chr3:37000997	AAACTG/A	0/1392	0/1143	0/239	0/1436	1/634	NA
chr3:37025746	T/C	0/1357	1/1151	0/242	0/1439	1/616	NA
chr3:37040185	G/C	0/1396	0/1151	0/246	0/1447	2/654	NA
chr3:37047540	C/CT	0/1395	0/1140	0/239	0/1440	1/645	NA
chr3:37048955	G/A	0/1396	0/1149	0/243	0/1445	1/658	NA
chr3:37050480	T/C	2/1385	0/1138	0/241	0/1434	1/634	NA
chr5:112838374	C/G	0/1437	0/1148	0/243	0/1422	1/654	4/121082
chr5:112839276	C/T	0/1385	0/1147	0/243	0/1443	1/642	NA
chr5:112840710	T/A	0/1231	1/1082	0/244	0/1269	1/608	NA
chr7:5999182	G/GT	0/1373	0/1104	0/237	0/1434	1/640	1/121410

Note. Reprinted from "Identification of Lynch Syndrome risk variants in the Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine

Table 3.4: Description of clinical information for the 11 patients

Patient	Variant	Sex	Age at diagnostic	ICD-10 code	SNOMED code	Cancer Grade	TNM - T	TNM - N	TNM - M	Relative 1	Age at dignostic relative 1	ICD10-CM code for relative 1	Relative 2	Age at dignostic relative 2	ICD10-CM code for relative 2
1	MSH6:p.Arg1068Ter	male	44	C18.3	M8140/3	2	T3	N2	M1	grandfrather	54	C18	uncle	40	C50
2	MLH1:p.Lys84ThrfsTer4	female	45	C18.4	M8140/3	2	T3	N1	M0	sister	42	C18			
3	MLH1:p.Met383Thr	female	60	C18.7	M8140/3	2	T3	N1	M0	father	59	C41	brother	63	C18
4	MLH1:c.1559-1G>C	female	41	C18.6	M8140/3	2	T3	N1	M0	father	49	C18	brother	42	C18
5	MLH1:c.1559-1G>C	male	43	C18.2	M8140/3	2	T3	N0	M0	sister	NA	C18			
6	MLH1:p.Ala586CysfsTer7	male	65	C18.6	M8480/3	2	T3	N1	M0	brother	37	C18			
7	MLH1:p.Ala681Thr	male	47	C18.6	M8140/3	2	T3	N0	M0	father	65	c18			
8	MLH1:c.2104-6T>C	female	50	C18.7	M8140/3	2	T3	N0	M0	brother	55	C18			
9	APC:p.Ser1706Thr	female	66	C18.0	M8140/3	2	T3	N0	M0	father	64	C18	mother	79	C18
9	APC:p.Ala927Gly	female	66	C18.0	M8140/3	2	T3	N0	M0	father	64	C18	mother	79	C18
10	APC:p.Gln1228Ter	male	58	C18.8	M8140/3	1	Tis	N0	M0	father	52	C18			
11	PMS2:p.Arg211ThrfsTer38	male	44	C18.7	M8140/3	1	T4	N0	M0	father	64	C16	brother	36	C16

Patient - a number used in the paper for this individual, Relative 1 – the first relative with a neoplasic pathology reported by the patient, Relative 2 – the second relative with a neoplasic pathology reported by the patient

Note. Reprinted from "Identification of Lynch Syndrome risk variants in the Romanian population." by Paul D. Iordache et al. in Journal of Cellular and Molecular Medicine

Chapter 4 Collaborative Work Results

Besides the two primary studies described in the objective section of this thesis, I participated in a small-scale case-control study on prostate cancer with the Romanian partners. Also, I expanded the areas of research of my thesis by collaborating on two GWAS studies on non-cancer phenotypes and providing one Icelandic study with a replication cohort for a small number of PCA variants.

4.1 Replication study of 34 common SNPs associated with prostate cancer in the Romanian population

The first project, published in January 2016 in the Journal of Cellular and Molecular Medicine, was a collaboration with the Romanian partners from the ROMCAN Project(Jinga et al., 2016) . This study provided the first assessment of how previously reported prostate cancer SNPs associate with risk in the Romanian population.

My contributions to this study were: verifying the quality of the genotypes, performing the association tests, assisting with manuscript preparation and interpreting the statistical results.

Abstract

Prostate cancer is the third-most common form of cancer in men in Romania. The unscreened Romanian population represents a good sample to study common genetic risk variants. However, a comprehensive analysis has not been conducted yet. Here, we report our replication efforts in a Romanian population of 979 cases and 1027 controls, for the potential association of 34 literature-reported single nucleotide polymorphisms (SNPs)

with prostate cancer. We also examined whether any SNP was differentially associated with tumor grade or stage at diagnosis, with disease aggressiveness, and with the levels of PSA (prostate-specific antigen). In the allelic analysis, we replicated the previously reported risk for 19 loci on 4q24, 6q25.3, 7p15.2, 8q24.21, 10q11.23, 10q26.13, 11p15.5, 11q13.2, 11q13.3. Statistically significant associations were replicated for other six SNPs only with a particular disease phenotype: a low-grade tumor and low PSA levels (rs1512268), high PSA levels (rs401681 and rs11649743), less aggressive cancers (rs1465618, rs721048, rs17021918). The strongest association of our tested SNP's with PSA in controls was for rs2735839, with 29% increase for each copy of the major allele G, consistent with previous results. Our results suggest that rs4962416, previously associated only with prostate cancer, is also associated with PSA levels, with 12% increase for each copy of the minor allele C. The study enabled the replication of the effect for the majority of previously reported genetic variants in a set of clinically relevant prostate cancers. This is the first replication study on these loci, known to associate with prostate cancer, in a Romanian population(Jinga et al., 2016).

4.2 Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer.

The second goal of the ROMCAN project was to integrate the Romanian case-control and cohort studies in international meta-analyses and develop collaborations with international consortia. This paper represents the first international collaboration that used Romanian data from the PCA GWAS in a replication study including cohorts from USA, Spain, Netherlands and the UK. My contribution to this study was to provide the replication results from the Romanian cohort for this particular locus, process the

Romanian data and assist with interpreting the result.

Abstract

Transcriptional and splicing anomalies have been observed in intron 8 of the CASP8 gene (encoding procaspase-8) in association with cutaneous basal-cell carcinoma (BCC) and linked to a germline SNP rs700635. Here, we show that the rs700635-C allele, which is associated with increased risk of BCC and breast cancer, is protective against prostate cancer (odds ratio (OR) = 0.91, $P = 1.0 \times 10^{-6}$). rs700635-C is also associated with failures to correctly splice out CASP8 intron 8 in breast and prostate tumors and in corresponding normal tissues. Investigation of rs700635[C] carriers revealed that they have a human-specific short interspersed element-variable number of tandem repeat-Alu (SINE-VNTR-Alu), subfamily-E retrotransposon (SVA-E) inserted into CASP8 intron 8. The SVA-E shows evidence of prior activity because it has transduced some CASP8 sequences during subsequent retrotransposition events. Whole-genome sequence (WGS) data were used to tag the SVA-E with a surrogate SNP rs1035142[T] (r2 = 0.999), which showed associations with both the splicing anomalies ($P = 6.5 \times 10^{-32}$) and with protection against prostate cancer (OR = 0.91, $P = 3.8 \times 10^{-7}$)(Stacey et al., 2016).

4.3 Epigenetic and genetic components of height regulation.

This study was based on adult height measurements and birth length measurements that were corrected for the year of birth and standardized separately for each of the sexes to have a standard normal distribution. These measurements were tested in a GWAS, observing 13 novel height associations loci. My contribution to this study was to assist with interpreting the result from a biological perspective and provide a medical context to the results.

Abstract

Adult height is a highly heritable trait. Here we identified 31.6 million sequence variants by whole-genome sequencing of 8,453 Icelanders and tested them for association with adult height by imputing them into 88,835 Icelanders. Here we discovered 13 novel height associations by testing four different models including parent-of-origin (|β|=0.4-10.6 cm). The minor alleles of three parent-of-origin signals associate with less height only when inherited from the father and are located within imprinted regions (IGF2-H19 and DLK1-MEG3). We also examined the association of these sequence variants in a set of 12,645 Icelanders with birth length measurements. Two of the novel variants, (IGF2-H19 and TET1), show a significant association with both adult height and birth length, indicating a role in early growth regulation. Among the parent-of-origin signals, we observed opposing parental effects raising questions about underlying mechanisms. These findings demonstrate that common variations affect human growth by parental imprinting(Benonisdottir et al., 2016).

4.4 A sequence variant associating with educational attainment also affects childhood cognition.

This study computed a polygenic score for educational attainment from Social Sciences Genetics Association Consortium (SSGAC) study and predicted the educational attainment in an independent Icelandic sample and correlated the two results. My contribution to this study was to assist with interpreting the genetic pathways involved in educational attainment and childhood cognition and provide a medical and biological context for the results.

Abstract

Only a few common variants in the sequence of the genome have been shown to impact cognitive traits. Here we demonstrate that polygenic scores of educational attainment predict specific aspects of childhood cognition, as measured with IQ. Recently, three sequence variants were shown to associate with educational attainment, a confluence phenotype of genetic and environmental factors contributing to academic success. We show that one of these variants associating with educational attainment, rs4851266-T, also associates with Verbal IQ in dyslexic children ($P = 4.3 \times 10^{-4}$, $\beta = 0.16$ s.d.). The effect of 0.16 s.d. corresponds to 1.4 IQ points for heterozygotes and 2.8 IQ points for homozygotes. We verified this association in independent samples consisting of adults ($P = 8.3 \times 10^{-5}$, $\beta = 0.12$ s.d., combined $P = 2.2 \times 10^{-7}$, $\beta = 0.14$ s.d.). Childhood cognition is unlikely to be affected by education attained later in life, and the variant explains a greater fraction of the variance in verbal IQ than in educational attainment (0.7% vs. 0.12%, $P = 1.0 \times 10^{-5}$)(Gunnarsson et al., 2016).

4.5 Unpublished work: Profile of common colorectal cancer risk variants in the Romanian population

In the present study, we investigated for the first time the profile of common colorectal cancer risk variants in a Romanian population. My contribution to this study was to complete the association test using the Romanian infrastructure developed during the ROMCAN Project and perform quality control checks on the results.

Abstract:

The study population consisted of 576 unrelated histo-pathologically confirmed colorectal cancer (CRC) cases and 1069 controls. DNA was extracted from buccal swabs at deCODE Genetics (Reykjavik, Iceland) and genotyped using Illumina SNP arrays,

24.295.558 variants were imputed using the 1000 Genomes dataset. A systematic literature review for variants associated with CRC in previous GWAS' was done using the NHGRI catalog, identifying 85 unique variants.

A total of 24.295.558 markers were analyzed. Of all the variants tested in the Romanian samples, 2 previously-published markers show genome-wide significance (p<5e-8). Figure 4.1 presents a Manhattan plot of the results. Our present study is the first GWAS on CRC performed in a Romanian population.

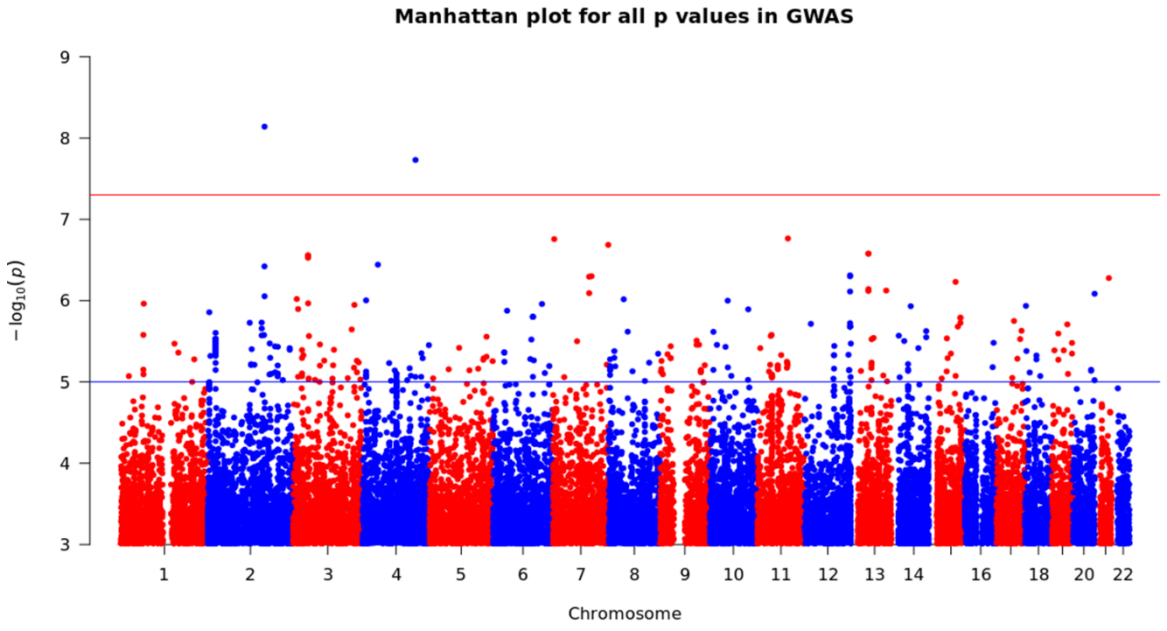


Figure 4.1 Manhattan plot of GWAS findings in the Romanian colorectal cancer sample

4.6 Unpublished work: Genetic Risk Factors Associated with Breast Cancer in Romania

In this study, I worked with Romanian scientists to investigate for the first time the profile of common breast cancer risk variants in a Romanian population. My contribution to this study was to complete the association test using imputed genotype data from Romanian breast cancer cases and controls collected under the ROMCAN Project and perform quality control checks on the results.

Abstract:

In this study, we investigated for the first time the profile of common breast cancer risk variants in the Romanian population. The study population consisted of 198 unrelated pathologically confirmed breast cancer cases and 1073 controls. DNA was genotyped using Illumina SNP arrays, 24.295.558 variants were imputed using the 1000 Genomes dataset. A systematic literature review of variants associated with breast cancer risk identified 198 unique variants that were tested in the Romanian sample set. Using the GWAS, 19 million markers passed quality criteria and were analyzed. A strong signal was observed on chromosome 10 (rs2912779, rs2981579, rs2912780). We replicated the great majority of markers previously published for breast cancer. The analysis of both p-value and OR suggests that breast cancer risk profile of the Romanian population is comparable to other European populations. Figure 4.2 presents a Manhattan plot of the results.

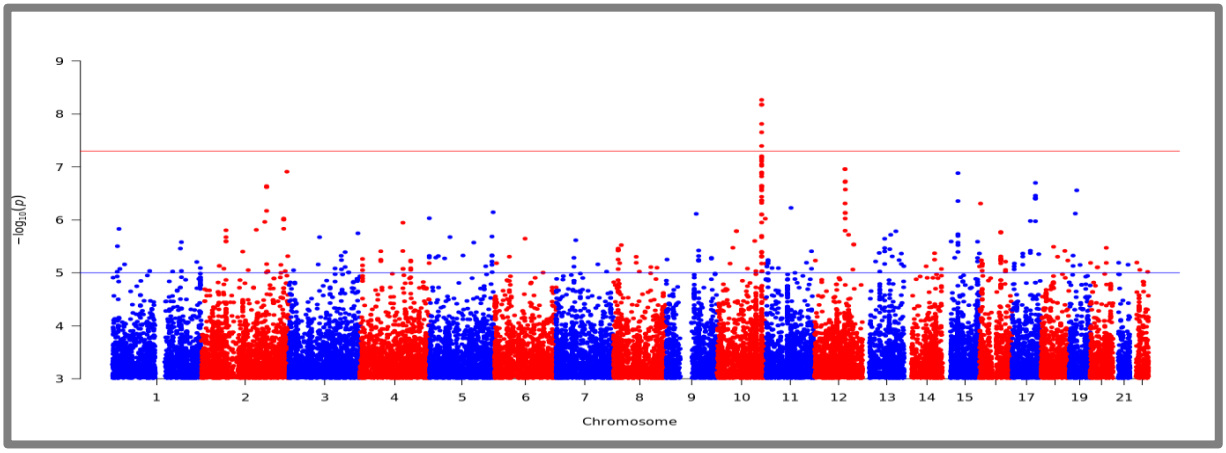


Figure 4.2 Manhattan plot of GWAS findings in the Romanian breast cancer sample

4.7 Unpublished work: Obesity, diabetes, and hypertension - genetic risk factors

In the present study, we investigated for the first time the profile of common obesity, diabetes and hypertension risk variants in a Romanian population. My contribution to this study was to complete the association test using the Icelandic (deCODE) infrastructure and perform quality control checks on the results.

Abstract:

Background. The current research wishes to determine high-risk polymorphisms (Single-Nucleotide-Polymorphisms) associated with obesity, diabetes (DM) and hypertension (HTA) in the Romanian population. It is the first large obesity-related genome-wide-association study (GWAS) in Romania and the follow-up of previous research on 2 polymorphisms correlated with obesity, DM and HTA (FTO-rs9939609, ADRB3-rs4994)

Materials and Methods. A GWAS was performed on 2024 Romanian men (1302 obese/overweight, 250 diabetics, 694 with hypertension. The research included 716503 SNPs and was performed at deCode Genetics Reykjavik. Statistical analysis was performed

using Plink! v1.07, IBM SPSS Statistics and R. Softwares. Genotyping results were analyzed in correlation with obesity (weight, BMI), DM and HTA.

Results. The results revealed genetic markers correlated with obesity, DM, and HTA both separately and in common. SNPs in the PIK3C2G (12p12.3), PCDH17 (13q21.1), ADCY9 (16p13.3), BVES (6q21), SLC7A7 (14q11.2), MBOAT (6p22.3), NID1 (1q24.3) and TBX20 (7p14.3) genes showed strong statistical correlations with DM ($p=10^{-5}$-10^{-6}). Polymorphisms on chromosomes 2, 9, 4, 12 and 13 (BANK1, PPP3CA, ATB2A, FRY genes) revealed associations with HTA, without obesity and DM mediation. Signals near the TMEM108 (3q23) and IREB2 (15q25.1) genes were in common of risk for all 3 pathologies. Obesity and DM were correlated with variants in the SPATA3, NQO2, WIP11, NRG1, FTO genes. SNPs in the TBX20 and ANK2 genes revealed correlations both with HTA and DM.

Conclusions. The results revealed several gene clusters correlated with obesity and DM. The implications of these variants for the Romanian population need to be further analyzed, replication studies having to be undertaken for confirmation.

Chapter 5 Discussion

5.1 The impact of this thesis on public health policies and medical practice in Romania

Prostate cancer is the second most common cancer in men worldwide, with the lifetime risk of being diagnosed with prostate cancer being 1 in 8(J Ferlay, Soerjomataram, et al., 2013). World Health Organization (WHO) statistics suggest that prostate cancer incidence varies more than 25-fold worldwide, 1.1 million cases were diagnosed worldwide with prostate cancer in 2012, accounting for 15% of the cancers diagnosed in men. (J Ferlay, Soerjomataram, et al., 2013).

According to the American Cancer Society, 89% of men diagnosed with prostate cancer survive at least five years after diagnosis, and 63% survive ten years. If the cancer is discovered while still localized, the five-year relative survival rate is nearly 100%(Brawley, 2012). The prostate-specific antigen (PSA) test is currently the most frequently used tool for detecting prostate cancer in its earliest and most curable stages. The test measures levels of prostate-specific antigen in the blood; however, PSA levels are not specific for prostate cancer but are also influenced by other conditions such as inflammation. Physicians and patients often overestimate the capabilities of PSA testing, paving the way for both overtreatment and undertreatment of many patients and subsequently suboptimal care(M. E. V. Caram, Skolarus, & Cooney, 2016). Despite the extensive use of PSA testing, there is no precise way to determine, whether the cancers detected would have ever caused symptoms or harm during a man's lifetime. One study estimated over detection to rise with age, from 27% at age 55 to 56% by age 75(A

Stangelberger, M Waldert,).

Although PSA may be helpful in managing and treating patients with PCA, using PSA as a sole measure is not sufficient to guide treatment decision making(M. E. V Caram, Skolarus, & Cooney, 2016). Developing additional genetic screening solutions for PCA could improve treatment decisions, impacting the outcome of the disease. Also, the finding of a positive result through genetic testing in patients with prostate cancer can lead to a recommendation that family members be tested.

With the PCA GWAS(Iordache et al., 2018b), I performed the first evaluation of genetic factors in prostate cancer (PCA) in the Romanian population using the ROMCAN Project data. As the Romanian medical community learn more about the potential role that genetic testing may in the future play in prostate cancer screening, the next challenge needs to be in determining the best way to put this knowledge to practice.

One of the first challenges is to decide who should be tested based on the evaluation of the populations most at risk and who is going to benefit the most from testing and evaluation the costs versus benefit of testing. A better understanding of a large number of genetic variants involved in prostate cancer in the Romanian population can help researchers overcome this challenge. With our study, we highlighted the population at risk based on the correlation between the clinical evaluation of the cohort and the genetic profile of each individual.

To date a large number of common genetic variants are associated with small PCA risk. Current evidence supports the hypothesis that excess familial risk of prostate cancer could be due to the inheritance of multiple moderate-risk genetic variants(Kommu, Edwards, & Eeles, 2004). Identifying genetic variants that confer genetic risk can change decision making for clinicians, patients, and their families. One important achievement of this thesis was the replication of a large number of genetic variants in the Romanian

population and validating them as possible starting points for future screening programs (Iordache et al., 2018b). Preliminary studies (Cucchiara et al., 2018) suggest that this will reduce unnecessary biopsies, discriminate clinically insignificant disease from aggressive ones, and help choose the best therapy in the metastatic patient. One of the outcomes of this part of the thesis is new knowledge on the similarities between predisposing genetic factors that are involved in carcinogenesis of the prostate in the Romania population compared to other populations. Incorporating all the prostate cancer risk SNPs into a screening panel can lead to improvements in clinical practice in Romania.

Large-scale studies are required to develop and validate guidelines for using genetic variants in clinical practice. As a first step towards this goal, we performed a GWAS on prostate cancer and looked up the association between previously reported PrC risk variants and prostate cancer in Romania. 30 of the 115 previously reported markers showed P-values < 0.05, and the direction of effects of an additional 59 markers were consistent with the reported results.

Our prostate cancer GWAS allows us to refine association signals and rule out associations due to differences in phenotype definitions between cohorts. Compared to the original studies, replication studies may use cohorts with slightly different ethnic and pathologic characteristic. Differences in ethnic characteristics lead to differences in LD structure, and consequently, markers that were previously found to be correlated with a risk variant may not show an association in a population of different ethnicity.

It is interesting to note that many of the variants showing the most robust replication in the Romanian population reside at loci that have been associated with several cancer types, so-called cancer hubs. The locus showing the strongest replication P-value (2×10^{-4}) in the Romanian GWAS is 8q24, one of the first hotspots for cancer risk alleles reported. The gene closest to this locus is the MYC gene, an oncogene known to contribute

to the genesis of many human cancers.

Identifying individuals with high genetic risk of cancer represents only the first step in the management of the disease and is of little use unless the corresponding changes in public health policies are implemented that prescribe the measures that should be taken. One of the most relevant hereditary pathologies in Romania that would benefit from genetic counseling and DNA testing is Lynch syndrome (LS). Lynch syndrome is the most common hereditary colon cancer syndrome, accounting for 3-5% of colorectal cancer (CRC) cases, and it is associated with the development of other cancers. Early detection of individuals with LS is relevant since patients can take advantage of particular medical care and health surveillance(Møller et al., 2017). Despite being a dominantly inherited cancer syndrome, the incidence and mortality show wide geographical variation across the world. LS is present in families in an autosomal dominant inheritance pattern, with a 50 percent chance that will be passed on to next generation. The genes affected in Lynch syndrome are known as mismatch repair genes (MMR) and are responsible for correcting changes in the genetic code. Identification of LS variants in the Romanian population is essential in order to reduce the incidence and mortality of this multicancer disorder. The work done on the LS cohort represents the most extensive effort, to our knowledge, to examine the genetic profile of this pathology in Eastern Europe. We identify a new set of variants specific to the Romanian population and further confirmed that some variants previously reported to be pathogenic in other populations also occur in Romania.

The results of our study identified three novel pathogenic variants and two novel variants that are unlikely to be pathogenic. Also, we confirmed three previously published pathogenic variants and suggested to reclassify a variant previously classified as VUS as pathogenic. All seven pathogenic variants can be included in future LS screening panels for the Romanian population. It is essential for Romanian clinicians to be able to recognize

individuals and families who are at risk of LS in order to provide adequate treatment and medical care. Members of families with suspected LS should receive genetic testing for detecting mutations in all LS-associated genes. This practice will provide predictive testing of at-risk relatives and provide a genetic context for the variants observed in the initial carriers. This study is the first step towards improving our understanding of the genetic particularities of this pathology in Romania. Correlating these with patients' genetic profiles is likely to play a critical role in the development of a screening protocol for the Romanian population. Combining our results with other similar studies will increase the knowledge of the genetic profile of LS worldwide. Integrating the results of this study in the worldwide context of genetic screening for LS will make it easier to diagnose and treat LS patients and their families.

Despite being a relatively new medical field in Romania, genetic epidemiology represents an essential area of research bringing together epidemiology, genetics, and public health. Genetic epidemiology could eventually influence both public health practice and clinical medical practice in Romania. Based on the ROMCAN Project results, we hope to start the second wave of Romanian genetic epidemiological studies, complementing the results from this present Ph.D. thesis.

5.2 The contribution of this thesis to genetic epidemiology in Romania.

Large-scale, population-based studies of genetic epidemiology are underway or planned in multiple countries, and the results of these studies will define the landscape of genetic epidemiology for the near future. Also, next-generation sequencing technologies are being established as platforms of choice for routine screening of tumor samples in a clinical setting. With routine whole exome or even whole genome sequencing of patients

becoming affordable for many hospitals and academic institutions, we believe that genetic screening can be a financially viable solution for medical practice. Both approaches show great potential in integrating of genetic epidemiology elements into the medical practice(Fallin, Duggal, & Beaty, 2016).

The two primary studies presented here provide good examples of genetic epidemiology studies of cancer. Both studies are designed to catalog genetic variation in the Romanian population, one focusing on common variants and the second on genome sequencing for high impact variants.

GWAS' have many limitations, such as their inability to fully explain the genetic risk of common diseases. Despite these limitations, variants obtained from these studies could be used to stratify the population by level of risk for particular diseases. Another important role of GWAS studies can be in identifying disease subtypes that have different causes or responses to treatment.

Genome sequencing will facilitate the reach and power of traditional genetic approaches to discovering disease genes, involved in the etiology of rare disease. This new way of analyzing the genetic profile of an individual will lead to more accurate diagnosis, predictions, and better treatment. These two directions were represented by the 2 main studies of this thesis. Besides providing a state-of-the-art analysis for genetic variants, we also integrated the Romanian variants into genetic variants databases for future reference. All genetic variants identified were analyzed immunohistological for confirmation of pathogenicity, and the results were registered to the InSIGHT- LOVD database.

The replicated results from the PCA GWAS are available in the GWAS catalog(Welter et al., 2014) for future replications and comparisons. Also, GWAS results from the Lung, Colorectal, and Breast cohorts will be included in the same database mentioned above, after publication. The other three cancer GWAS' mentioned above

represent the main body of future Romanian work in the genetic epidemiology of cancer and use the template generated by the Romanian PCA GWAS. Generating a complete genetic profile for the four main cancer types in the Romanian population will add significant information on the genetic epidemiology of cancer in Eastern Europe and be useful in further collaborative efforts in the field.

5.3 The integration of this thesis into future genetic epidemiology projects.

Genetic epidemiology of cancer is a discipline with the final aim of identifying and characterizing population-level factors that contribute to particular types of cancer. Genetic epidemiologists often pursue this aim through the design and implementation of studies that simultaneously invoke principles in population genetics, epidemiology, molecular biology, and biostatistics. Mastering all the disciplines mentioned above by one individual represents an impossible task, consequently multi-disciplinary projects need to be developed in the near future by the Romanian scientific community.

The new era of mass genetic sequencing enables the expansion of the screening panels for all the highly heritable cancer types. These fast and affordable methods for analyzing the DNA sequence can be utilized in targeted diagnostics to increase substantially clinical testing of genetic predisposition variants to cancer. Targeted sequencing of a limited set of clinically essential genes has been the most practical approach for clinical applications. These panels present multiple advantages, such as a small panel size, low DNA input requirement, and a relatively low cost. Personalized genetic testing offers opportunities to improve the quality of care provided to patients with cancer or other diseases. The primary limitation of these panels is that their reliability is directly correlated to known pathogenic variants particular to the tested population. The

study(Iordache et al., 2018) provided the first insights into rare variants associated with LS in Romania. The inclusion of these variants in online databases such as InSIGHT(Peltomäki, 2005) and ClinVar(Landrum et al., 2018) will help future researchers and medical doctors to screen and provide adequate treatment to patients carrying these particular variants. Also, suggesting to reclassify one of the variants previously known as a "conflict of interpretation" as "pathogenic" will benefit the scientific community, elucidating the clinical interpretation of this variant.

Future projects can build upon the work done in this Ph.D. thesis and in the ROMCAN Project. In order to make use of the results, there is a need to focus on developing strong Romanian biostatistics and bioinformatic component to complement the genetic data available now in Romania. Another essential direction will be represented by the continuous collection of DNA samples and phenotypic information for new patients in the hope of continually updating and expanding our results. The last, and probably the most critical direction, is represented by developing a cohesive dissemination strategy of scientific results through scientific papers, presentations at international conferences and public information campaigns. The integration of genetic testing within mainstream medical services will require education of physicians, remodeling of existing medical protocols and adding additional infrastructure, including specialized laboratories with bioinformatics and clinical interpretive capabilities.

Bibliography

Aaltonen, L. A., Salovaara, R., Kristo, P., Canzian, F., Hemminki, A., Peltomäki, P., … Valkamo, E. (1998). Incidence of Hereditary Nonpolyposis Colorectal Cancer and the Feasibility of Molecular Screening for the Disease. *New England Journal of Medicine*, *338*(21), 1481–1487. https://doi.org/10.1056/NEJM199805213382101

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Abreu Velez, A. M., & Howard, M. S. (2015). Tumor-suppressor Genes, Cell Cycle Regulatory Checkpoints, and the Skin. *North American Journal of Medical Sciences*, *7*(5), 176–188. https://doi.org/10.4103/1947-2714.157476

Ahsan, H., Halpern, J., Kibriya, M. G., Pierce, B. L., Tong, L., Gamazon, E., … Whittemore, A. S. (2014). A Genome-wide Association Study of Early-Onset Breast Cancer Identifies PFKM as a Novel Breast Cancer Gene and Supports a Common Genetic Spectrum for Breast Cancer at Any Age. *Cancer Epidemiology Biomarkers & Prevention*, *23*(4), 658–669. https://doi.org/10.1158/1055-9965.EPI-13-0340

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Attard, G., Parker, C., Eeles, R. A., Schröder, F., Tomlins, S. A., Tannock, I., … al., et. (2016). Prostate cancer. *The Lancet*, *387*(10013), 70–82. https://doi.org/10.1016/S0140-6736(14)61947-4

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., …

Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

B. R. Erick Peirson. (2012). Wilhelm Johannsen's Genotype-Phenotype Distinction. Retrieved from https://pdfs.semanticscholar.org/d6c7/e12a9aa20ac1488a75cfacb8062238a55cd8.pdf

Bara, A.-C., van den Heuvel, W. J. A., & Maarse, J. A. M. (2002). Reforms of health care system in Romania. *Croatian Medical Journal*, *43*(4), 446–452. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12187523

Bellen, H. J., & Yamamoto, S. (2015). Morgan's legacy: fruit flies and the functional annotation of conserved genes. *Cell*, *163*(1), 12–14. https://doi.org/10.1016/j.cell.2015.09.009

Benafif, S., Kote-Jarai, Z., & Eeles, R. A. (2018). A review of prostate cancer genome wide association studies (GWAS). *Cancer Epidemiology Biomarkers & Prevention*, cebp.1046.2017. https://doi.org/10.1158/1055-9965.EPI-16-1046

Benonisdottir, S., Oddsson, A., Helgason, A., Kristjansson, R. P., Sveinbjornsson, G., Oskarsdottir, A., … Stefansson, K. (2016). Epigenetic and genetic components of height regulation. *Nature Communications*, *7*, 13490. https://doi.org/10.1038/ncomms13490

Bodmer, W., & Tomlinson, I. (2010). Rare genetic variants and the risk of cancer. *Current Opinion in Genetics & Development*, *20*(3), 262–267. https://doi.org/10.1016/j.gde.2010.04.016

Bratt, O., Drevin, L., Akre, O., Garmo, H., & Stattin, P. (2016). Family History and Probability of Prostate Cancer, Differentiated by Risk Category: A Nationwide Population-Based Study. *Journal of the National Cancer Institute*, *108*(10), djw110. https://doi.org/10.1093/jnci/djw110

Brawley, O. W. (2012). Trends in prostate cancer in the United States. *Journal of the National Cancer Institute. Monographs*, *2012*(45), 152–156.

https://doi.org/10.1093/jncimonographs/lgs035

Bray, F., Lortet-Tieulent, J., Ferlay, J., Forman, D., & Auvinen, A. (2010). Prostate cancer incidence and mortality trends in 37 European countries: an overview. *European Journal of Cancer (Oxford, England : 1990)*, *46*(17), 3040–3052. https://doi.org/10.1016/j.ejca.2010.09.013

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Caram, M. E. V., Skolarus, T. A., & Cooney, K. A. (2016). Limitations of Prostate-specific Antigen Testing After a Prostate Cancer Diagnosis. *European Urology*, *70*(2), 209–210. https://doi.org/10.1016/j.eururo.2015.12.045

Caram, M. E. V, Skolarus, T. A., & Cooney, K. A. (2016). Platinum Opinion Limitations of Prostate-specific Antigen Testing After a Prostate Cancer Diagnosis. https://doi.org/10.1016/j.eururo.2015.12.045

Carethers, J. M., & Stoffel, E. M. (2015). Lynch syndrome and Lynch syndrome mimics: The growing complex landscape of hereditary colon cancer. *World Journal of Gastroenterology*, *21*(31), 9253–9261. https://doi.org/10.3748/wjg.v21.i31.9253

Carter, B. S., Beaty, T. H., Steinberg, G. D., Childs, B., & Walsh, P. C. (1992). Mendelian inheritance of familial prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(8), 3367–3371. https://doi.org/10.1073/PNAS.89.8.3367

Cohen, S. A., & Leininger, A. (2014). The genetic basis of Lynch syndrome and its implications for clinical practice and risk management. *The Application of Clinical Genetics*, *7*, 147–158. https://doi.org/10.2147/TACG.S51483

Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., … Talkowski, M. E. (2017). Defining the diverse spectrum of inversions, complex structural variation,

and chromothripsis in the morbid human genome. *Genome Biology*, *18*(1), 36. https://doi.org/10.1186/s13059-017-1158-6

Darwin, C. R. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray. [1st edition]*.

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., & Marchini, J. (2013). Haplotype Estimation Using Sequencing Reads. *The American Journal of Human Genetics*, *93*(4), 687–696. https://doi.org/10.1016/j.ajhg.2013.09.002

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., … Ponder, B. A. J. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, *447*(7148), 1087–1093. https://doi.org/10.1038/nature05887

Eeles, R. A., Kote-jarai, Z., Giles, G. G., Amin, A., Olama, A., Guy, M., … Easton, D. F. (2008). Multiple newly identified loci associated with prostate cancer susceptibility, *40*(3), 316–321. https://doi.org/10.1038/ng.90

Eeles, R. A., Olama, A. A. Al, Benlloch, S., Saunders, E. J., Leongamornlert, D. A., Tymrakiewicz, M., … Easton, D. F. (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nature Genetics*, *45*(4), 385–391, 391e1-2. https://doi.org/10.1038/ng.2560

Enciso-Mora, V., Broderick, P., Ma, Y., Jarrett, R. F., Hjalgrim, H., Hemminki, K., … Houlston, R. S. (2010). A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). *Nature Genetics*, *42*(12), 1126–1130. https://doi.org/10.1038/ng.696

Fallin, M. D., Duggal, P., & Beaty, T. H. (2016). Genetic Epidemiology and Public Health: The Evolution From Theory to Technology. *American Journal of Epidemiology*, *183*(5), 387–

393. https://doi.org/10.1093/aje/kww001

Feigelson, H. S., Goddard, K. A. B., Hollombe, C., Tingle, S. R., Gillanders, E. M., Mechanic, L. E., & Nelson, S. A. (2014). Approaches to integrating germline and tumor genomic data in cancer research. *Carcinogenesis*, *35*(10), 2157–2163. https://doi.org/10.1093/carcin/bgu165

Felsenfeld, G. (2014). A brief history of epigenetics. *Cold Spring Harbor Perspectives in Biology*, *6*(1). https://doi.org/10.1101/cshperspect.a018200

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010a). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer. Journal International Du Cancer*, *127*(12), 2893–2917. https://doi.org/10.1002/ijc.25516

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010b). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, *127*(12), 2893–2917. https://doi.org/10.1002/ijc.25516

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., … Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, *136*(5), E359–E386. https://doi.org/10.1002/ijc.29210

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., … Bray, F. (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Retrieved January 1, 2016, from http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W. W., Comber, H., … Bray, F. (2013). Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *European Journal of Cancer (Oxford, England : 1990)*, *49*(6), 1374–

1403. https://doi.org/10.1016/j.ejca.2012.12.027

Fincham, S. M., Hill, G. B., Hanson, J., & Wijayasinghe, C. (1990). Epidemiology of prostatic cancer: a case-control study. *The Prostate*, *17*(3), 189–206. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2235728

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, *10*(4), 241–251. https://doi.org/10.1038/nrg2554

Froggatt, N. J., Koch, J., Davies, R., Evans, D. G., Clamp, A., Quarrell, O. W., … al., et. (1995). Genetic linkage analysis in hereditary non-polyposis colon cancer syndrome. *Journal of Medical Genetics*, *32*(5), 352–357. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7616541

Gatalica, Z., Vranic, S., Xiu, J., Swensen, J., & Reddy, S. (2016). High microsatellite instability (MSI-H) colorectal carcinoma: a brief review of predictive biomarkers in the era of personalized medicine. *Familial Cancer*, *15*(3), 405–412. https://doi.org/10.1007/s10689-016-9884-6

Gomella, L. G., Liu, X. S., Trabulsi, E. J., Kelly, W. K., Myers, R., Showalter, T., … Wender, R. (2011). Screening for prostate cancer: the current evidence and guidelines controversy. *The Canadian Journal of Urology*, *18*(5), 5875–5883.

Goode, E. L., Chenevix-Trench, G., Song, H., Ramus, S. J., Notaridou, M., Lawrenson, K., … Pharoah, P. D. P. (2010). A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nature Genetics*, *42*(10), 874–879. https://doi.org/10.1038/ng.668

Grisanzio, C., Werner, L., Takeda, D., Awoyemi, B. C., Pomerantz, M. M., Yamada, H., … Freedman, M. L. (2012). Genetic and functional analyses implicate the NUDT11, HNF1B,

and SLC22A3 genes in prostate cancer pathogenesis. *Proceedings of the National Academy of Sciences*, *109*(28), 11252–11257. https://doi.org/10.1073/pnas.1200853109

GTEx Consortium, T. Gte. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., … Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444. https://doi.org/10.1038/ng.3247

Gudmundsson, J., Besenbacher, S., Sulem, P., Gudbjartsson, D. F., Olafsson, I., Arinbjarnarson, S., … Stefansson, K. (2010). Genetic Correction of PSA Values Using Sequence Variants Associated with PSA Levels. *Science Translational Medicine*, *2*(62).

Gudmundsson, J., Sulem, P., Gudbjartsson, D. F., Masson, G., Petursdottir, V., Hardarson, S., … Stefansson, K. (2013). A common variant at 8q24.21 is associated with renal cell cancer. *Nature Communications*, *4*. https://doi.org/10.1038/ncomms3776

Gunnarsson, B., Jónsdóttir, G. A., Björnsdóttir, G., Konte, B., Sulem, P., Kristmundsdóttir, S., … Stefánsson, K. (2016). A sequence variant associating with educational attainment also affects childhood cognition. *Scientific Reports*, *6*(1), 36189. https://doi.org/10.1038/srep36189

Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S., & Goldstein, D. B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology*, *17*(1), 9. https://doi.org/10.1186/s13059-016-0869-4

Haiman, C. A., Chen, G. K., Vachon, C. M., Canzian, F., Dunning, A., Millikan, R. C., … Couch, F. J. (2011). A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor–negative breast cancer. *Nature Genetics*, *43*(12), 1210–1214.

https://doi.org/10.1038/ng.985

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10647931

Haraldsdottir, S., Hampel, H., Wu, C., Weng, D. Y., Shields, P. G., Frankel, W. L., … Bekaii-Saab, T. (2016). Patients with colorectal cancer associated with Lynch syndrome and MLH1 promoter hypermethylation have similar prognoses. *Genetics in Medicine*, *18*(9), 863–868. https://doi.org/10.1038/gim.2015.184

Haraldsdottir, S., Rafnar, T., Frankel, W. L., Einarsdottir, S., Sigurdsson, A., Hampel, H., … Stefansson, K. (2017). Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2. *Nature Communications*, *8*, 14755. https://doi.org/10.1038/ncomms14755

Hardy, G. H. (1908). Mendelian Proportions in a Mixed Population. *Science*, *28*(706), 49–50. https://doi.org/10.1126/science.28.706.49

Heise, M., & Haus, O. (2014). Hereditary prostate cancer. *Postępy Higieny i Medycyny Doświadczalnej*, *68*, 653–665. https://doi.org/10.5604/17322693.1104682

Hitch, K., Joseph, G., Guiltinan, J., Kianmahd, J., Youngblom, J., & Blanco, A. (2014). Lynch syndrome patients' views of and preferences for return of results following whole exome sequencing. *Journal of Genetic Counseling*, *23*(4), 539–551. https://doi.org/10.1007/s10897-014-9687-6

Hodgson, S. (2008). Mechanisms of inherited cancer susceptibility. *Journal of Zhejiang University. Science. B*, *9*(1), 1–4. https://doi.org/10.1631/jzus.B073001

Hoffmann, T. J., Eeden, S. K. Van Den, Sakoda, L. C., Jorgenson, E., Habel, L. A., Graff, R. E., … Witte, J. S. (2015). A large multi-ethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. *Cancer Discovery*,

5(8), 878. https://doi.org/10.1158/2159-8290.cd-15-0315

Howie, B. N., Donnelly, P., Marchini, J., Rioux, J., Xavier, R., Taylor, K., … Abecasis, G. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Human Genome Structural Variation Working Group, Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., … Waterston, R. H. (2007). Completing the map of human genetic variation. *Nature*, *447*(7141), 161–165. https://doi.org/10.1038/447161a

Iordache, P. D., Mates, D., Gunnarsson, B., Eggertsson, H. P., Sulem, P., Guðmundsson, J., … Stefánsson, K. (2018a). Identification of Lynch Syndrome risk variants in the Romanian population. *Journal of Cellular and Molecular Medicine*.

Iordache, P. D., Mates, D., Gunnarsson, B., Eggertsson, H. P., Sulem, P., Guðmundsson, J., … Stefánsson, K. (2018b). Profile of common prostate cancer risk variants in an unscreened Romanian population. *Journal of Cellular and Molecular Medicine*, *22*(3), 1574–1582. https://doi.org/10.1111/jcmm.13433

Ishak, M. B., & Giri, V. N. (2011). A systematic review of replication studies of prostate cancer susceptibility genetic variants in high-risk men originally identified from genome-wide association studies. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *20*(8), 1599–1610. https://doi.org/10.1158/1055-9965.EPI-11-0312

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., & Burt, R. W. (2010). Hereditary and familial colon cancer. *Gastroenterology*, *138*(6), 2044–2058. https://doi.org/10.1053/j.gastro.2010.01.054

Jinga, V., Csiki, I. E., Manolescu, A., Iordache, P., Mates, I. N., Radavoi, D., … Mates, D. (2016). Replication study of 34 common SNPs associated with prostate cancer in the Romanian population. *Journal of Cellular and Molecular Medicine*, *20*(4), 594–600. https://doi.org/10.1111/jcmm.12729

Johannsen, W. L. (1909). Elemente der exakten Erblichkeitslehre. Retrieved from http://caliban.mpipz.mpg.de/johannsen/elemente/index.html

Joint Test and Technology Transfer Committee Working Group, American College of Medical Genetics, 9650 Rockville Pike, Bethesda, MD 20814-3998, United States., J. T. and T. T. C. W., Murphy, P., Petersen, G., Thibodeau, S., & Fishel, R. (2000). Genetic testing for colon cancer: joint statement of the American College of Medical Genetics and American Society of Human Genetics. Joint Test and Technology Transfer Committee Working Group. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *2*(6), 362–366. https://doi.org/10.109700125817-200011000-00011

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., … Stefansson, K. (2017). Whole genome characterization of sequence diversity of 15,220 Icelanders. *Scientific Data*, *4*, 170115. https://doi.org/10.1038/sdata.2017.115

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., … Miller, D. T. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine*, *19*(2), 249–255. https://doi.org/10.1038/gim.2016.190

Kiemeney, L. A., Thorlacius, S., Sulem, P., Geller, F., Aben, K. K. H., Stacey, S. N., … Stefansson, K. (2008). Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nature Genetics*, *40*(11), 1307–1312. https://doi.org/10.1038/ng.229

Kommu, S., Edwards, S., & Eeles, R. (2004). The Clinical Genetics of Prostate Cancer. *Hereditary Cancer in Clinical Practice*, *2*(3), 111. https://doi.org/10.1186/1897-4287-2-3-111

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., … Caporaso, N. E. (2009). A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma. *The American Journal of Human Genetics*, *85*(5), 679–691. https://doi.org/10.1016/j.ajhg.2009.09.012

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lange, E. M., Johnson, A. M., Wang, Y., Zuhlke, K. A., Lu, Y., Ribado, J. V, … Cooney, K. A. (2014). Genome-Wide Association Scan for Variants Associated with Early-Onset Prostate Cancer, *9*(4). https://doi.org/10.1371/journal.pone.0093436

Lek, M. et al. (2017). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *106*(3), 201–207. https://doi.org/10.1002/bdra.23483.Autoantibodies

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lloyd, D., Wimpenny, J., & Venables, A. (2010). Alfred Russel Wallace deserves better. *Journal of Biosciences*, *35*(3), 339–349. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/20826943

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). Proto-Oncogenes and Tumor-Suppressor Genes. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK21662/

Lynch, H. T., Kosoko-Lasaki, O., Leslie, S. W., Rendell, M., Shaw, T., Snyder, C., … Powell, I. (2016). Screening for familial and hereditary prostate cancer. *Int. J. Cancer UICC International Journal of Cancer*, *138*, 2579–2591. https://doi.org/10.1002/ijc.29949

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*, 122. https://doi.org/10.1186/s13059-016-0974-4

Mendel, G. (1866). *Versuche über Pflanzen-Hybriden*. Retrieved from http://www.bshs.org.uk/bshs-translations/mendel/2016?page=1

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., … Easton, D. F. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics*, *45*(4), 353–361. https://doi.org/10.1038/ng.2563

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., …

Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science (New York, N.Y.), 266*(5182), 66–71. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7545954

Mohler, J., Bahnson, R. R., Boston, B., Busby, J. E., D'Amico, A., Eastham, J. A., … Walsh, P. C. (2010). NCCN clinical practice guidelines in oncology: prostate cancer. *Journal of the National Comprehensive Cancer Network : JNCCN*, *8*(2), 162–200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20141676

Møller, P., Seppälä, T., Bernstein, I., Holinski-Feder, E., Sala, P., Evans, D. G., … Mallorca Group (http://mallorca-group.eu), in collaboration with T. M. G. (2017). Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut*, *66*(3), 464–472. https://doi.org/10.1136/gutjnl-2015-309675

Morgan, T. H. (1915). *The Mechanism of Mendelian Heredity*. Retrieved from https://archive.org/details/mechanismofmende00morgiala

Morganti, G., Gianferrari, L., Cresseri, A., Arrigoni, G., & Lovati, G. (1956). RECHERCHES CLINICO-STATISTIQUES ET GÉNÉTIQUES SUR LES NÉOPLASIES DE LA PROSTATE. *Human Heredity*, *6*(2), 304–305. https://doi.org/10.1159/000150844

Mucci, L. A., Hjelmborg, J. B., Harris, J. R., Czene, K., Havelick, D. J., Scheike, T., … ES, L. (2016). Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*, *315*(1), 68. https://doi.org/10.1001/jama.2015.17703

Murgatroyd, C. (2015). Editorial to the Special Issue Historical Medical Genetics II. *Gene*, *555*(1), 1. https://doi.org/10.1016/j.gene.2014.11.023

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., … Marchini, J. (2014). A general approach for haplotype phasing across the full spectrum of relatedness.

PLoS Genetics, *10*(4), e1004234. https://doi.org/10.1371/journal.pgen.1004234

Peltomäki, P. (2003). Role of DNA Mismatch Repair Defects in the Pathogenesis of Human Cancer. *Journal of Clinical Oncology*, *21*(6), 1174–1179. https://doi.org/10.1200/JCO.2003.04.060

Peltomäki, P. (2005). Lynch Syndrome Genes. *Familial Cancer*, *4*(3), 227–232. https://doi.org/10.1007/s10689-004-7993-0

Petersen, G. M., Amundadottir, L., Fuchs, C. S., Kraft, P., Stolzenberg-Solomon, R. Z., Jacobs, K. B., … Chanock, S. J. (2010). A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics*, *42*(3), 224–228. https://doi.org/10.1038/ng.522

Pomerantz, M. M., & Freedman, M. L. (2011). The genetics of cancer risk. *Cancer Journal (Sudbury, Mass.)*, *17*(6), 416–422. https://doi.org/10.1097/PPO.0b013e31823e5387

Pulst, S. M. (1999). Genetic linkage analysis. *Archives of Neurology*, *56*(6), 667–672. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10369304

Purdue, M. P., Johansson, M., Zelenika, D., Toro, J. R., Scelo, G., Moore, L. E., … Brennan, P. (2011). Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. *Nature Genetics*, *43*(1), 60–65. https://doi.org/10.1038/ng.723

Radick, G. (2001). The Century of the Gene. Evelyn Fox Keller. Harvard University Press, Cambridge, MA. 2000. pp. 186. Price £15.95, hardback. ISBN 0 674 00372 1. *Heredity*, *86*(5), 639–640. https://doi.org/10.1046/j.1365-2540.2001.0946c.x

Rafnar, T., Sulem, P., Stacey, S. N., Geller, F., Gudmundsson, J., Sigurdsson, A., … Stefansson, K. (2009). Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nature Genetics*, *41*(2), 221–227. https://doi.org/10.1038/ng.296

Richman, S. (2015). Deficient mismatch repair: Read all about it (Review). *International Journal of Oncology*, *47*(4), 1189–1202. https://doi.org/10.3892/ijo.2015.3119

Samowitz, W. S., Curtin, K., Lin, H. H., Robertson, M. A., Schaffer, D., Nichols, M., … Slattery, M. L. (2001). The colon cancer burden of genetically defined hereditary nonpolyposis colon cancer. *Gastroenterology*, *121*(4), 830–838. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11606497

Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L. J., Zappa, M., Nelen, V., … ERSPC Investigators. (2014). Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *The Lancet*, *384*(9959), 2027–2035. https://doi.org/10.1016/S0140-6736(14)60525-0

Schumacher, F. R., Berndt, S. I., Siddiq, A., Jacobs, K. B., Wang, Z., Lindstrom, S., … Kraft, P. (2011). Genome-wide association study identifies new prostate cancer susceptibility loci. *Human Molecular Genetics*, *20*(19), 3867–3875. https://doi.org/10.1093/hmg/ddr295

Sherr, C. J. (2004). Principles of Tumor Suppression. *Cell*, *116*(2), 235–246. https://doi.org/10.1016/S0092-8674(03)01075-4

Stacey, S. N., Kehr, B., Gudmundsson, J., Zink, F., Jonasdottir, A., Gudjonsson, S. A., … Stefansson, K. (2016). Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. *Human Molecular Genetics*, *25*(5), 1008–1018. https://doi.org/10.1093/hmg/ddv622

Stark, A., & Seneta, E. (2013). Wilhelm Weinberg's early contribution to segregation analysis. *Genetics*, *195*(1), 1–6. https://doi.org/10.1534/genetics.113.152975

Stegeman, S., Amankwah, E., Klein, K., O'Mara, T. A., Kim, D., Lin, H.-Y., … Batra, J. (2015). A Large-Scale Analysis of Genetic Variants within Putative miRNA Binding Sites

in Prostate Cancer. *Cancer Discovery*, *5*(4). Retrieved from http://cancerdiscovery.aacrjournals.org/content/5/4/368?iss=4

Stenson, P. D., Mort, M., Ball, E. V, Evans, K., Hayden, M., Heywood, S., … Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, *136*(6), 665–677. https://doi.org/10.1007/s00439-017-1779-6

Stoffel, E. M., & Kastrinos, F. (2014). Familial Colorectal Cancer, Beyond Lynch Syndrome. *Clinical Gastroenterology and Hepatology*, *12*(7), 1059–1068. https://doi.org/10.1016/j.cgh.2013.08.015

Sun, J., Zheng, S. L., Wiklund, F., Isaacs, S. D., Li, G., Wiley, K. E., … Chang, B. (2009). Sequence Variants at 22q13 Are Associated with Prostate Cancer Risk, (1), 10–16. https://doi.org/10.1158/0008-5472.CAN-08-3464

Tafe, L. J. (2015). Targeted Next-Generation Sequencing for Hereditary Cancer Syndromes: A Focus on Lynch Syndrome and Associated Endometrial Cancer. *The Journal of Molecular Diagnostics*, *17*(5), 472–482. https://doi.org/10.1016/J.JMOLDX.2015.06.001

Tao, S., Wang, Z., Feng, J., Hsu, F.-C., Jin, G., Kim, S.-T., … Sun, J. (2012). A genome-wide search for loci interacting with known prostate cancer risk-associated genetic variants. *Carcinogenesis*, *33*(3), 598–603. https://doi.org/10.1093/carcin/bgr316

Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., … Chanock, S. J. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, *40*(3), 310–315. https://doi.org/10.1038/ng.91

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., … Houlston, R. (2007). A genome-wide association scan of tag SNPs identifies a

susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics*, *39*(8), 984–988. https://doi.org/10.1038/ng2085

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., … Easton, D. F. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics*, *42*(6), 504–507. https://doi.org/10.1038/ng.586

Turnbull, C., Rapley, E. A., Seal, S., Pernet, D., Renwick, A., Hughes, D., … UK Testicular Cancer Collaboration. (2010). Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nature Genetics*, *42*(7), 604–607. https://doi.org/10.1038/ng.607

Varmus, H. (2017). How Tumor Virology Evolved into Cancer Biology and Transformed Oncology. *Annual Review of Cancer Biology*, *1*(1), 1–18. https://doi.org/10.1146/annurev-cancerbio-050216-034315

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Virlogeux, V., Graff, R. E., Hoffmann, T. J., & Witte, J. S. (2015). Replication and heritability of prostate cancer risk variants: impact of population-specific factors. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *24*(6), 938–943. https://doi.org/10.1158/1055-9965.EPI-14-1372

Waidelich, R., Bumbu, G., Raica, M., Toma, M., Maghiar, T., & Hofstetter, A. (2011). Screening for prostate cancer in Romania. *International Urology and Nephrology*, *34*(4), 503–505. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14577492

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for

Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte Vereins Vaterländische Natur, Württember*, 369–382. Retrieved from https://archive.org/details/cbarchive_35716_berdennachweisdervererbungbeim1845

Weinberg, W. (1912). Weitere Beiträge zur Theorie der Vererbung. *Rassen Gesellschafts-Biol*, (9), 165–174.

Weinhold, N., Johnson, D. C., Chubb, D., Chen, B., Försti, A., Hosking, F. J., … Hemminki, K. (2013). The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nature Genetics*, *45*(5), 522–525. https://doi.org/10.1038/ng.2583

Weiss, R. A., & Vogt, P. K. (2011). 100 years of Rous sarcoma virus. *The Journal of Experimental Medicine*, *208*(12), 2351–2355. https://doi.org/10.1084/jem.20112160

Weissman, S. M., Burt, R., Church, J., Erdman, S., Hampel, H., Holter, S., … Senter, L. (2012). Identification of Individuals at Risk for Lynch Syndrome Using Targeted Evaluations and Genetic Testing: National Society of Genetic Counselors and the Collaborative Group of the Americas on Inherited Colorectal Cancer Joint Practice Guideline. *Journal of Genetic Counseling*, *21*(4), 484–493. https://doi.org/10.1007/s10897-011-9465-7

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001-6. https://doi.org/10.1093/nar/gkt1229

Wolpin, B. M., Rizzato, C., Kraft, P., Kooperberg, C., Petersen, G. M., Wang, Z., … Amundadottir, L. T. (2014). Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nature Genetics*, *46*(9), 994–1000. https://doi.org/10.1038/ng.3052

Yap, N. Y., Rajandram, R., Ng, K. L., Pailoor, J., Fadzli, A., & Gobe, G. C. (2015). Genetic and

Chromosomal Aberrations and Their Clinical Significance in Renal Neoplasms. *BioMed Research International*, *2015*, 476508. https://doi.org/10.1155/2015/476508

Yurgelun, M. B., Allen, B., Kaldate, R. R., Bowles, K. R., Judkins, T., Kaushik, P., … Syngal, S. (2015). Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome. *Gastroenterology*, *149*(3), 604–13.e20. https://doi.org/10.1053/j.gastro.2015.05.006

Zanke, B. W., Greenwood, C. M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M., … Dunlop, M. G. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genetics*, *39*(8), 989–994. https://doi.org/10.1038/ng2089

.

School of Science and Engineering Reykjavík University

Menntavegur 1

101 Reykjavík, Iceland

Tel. +354 599 6200

Fax +354 599 6201

www.ru.is