Ressortforschungsberichte zum Strahlenschutz

Genomweite Analyse genetisch bedingter Strahlenempfindlichkeit in Wismut-Bergarbeitern – Datenauswertung und Bewertung der Assoziationsanalysen - Vorhaben 3615S32253

Auftragnehmer:
Universitätsmedizin Göttingen (UMG)

Prof. Dr. H. Bickeböller
Dr. A. Rosenberger

Das Vorhaben wurde mit Mitteln des Bundesministeriums für Umwelt, Naturschutz, Bau und Reaktorsicherheit (BMUB) und im Auftrag des Bundesamtes für Strahlenschutz (BfS) durchgeführt.

Bundesamt für Strahlenschutz

BfS-RESFOR-131/18

Bitte beziehen Sie sich beim Zitieren dieses Dokumentes immer auf folgende URN:
urn:nbn:de:0221-2018050314833

Salzgitter, Mai 2018

Forschungsvorhaben:	Genomweite Analyse genetisch bedingter Strahlenempfindlichkeit in Wismut-Bergarbeitern (Vertragsnr. 3615 S 32253)
Auftraggeberin:	Die Bundesrepublik Deutschland, vertreten durch das Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit, dieses vertreten durch den Präsidenten des Bundesamtes für Strahlenschutz
Auftragnehmer:	Universitätsmedizin Göttingen (UMG)
Projektleitung:	Prof. Dr. Heike Bickeböller UMG, Institut für Genetische Epidemiologie Wissenschaftlicher Mitarbeiter: Dr. Albert Rosenberger
Laufzeit	1.3.2016-28.2.2017
Fachbetreuung BfS:	Dr. Maria Gomolka

Schussbericht

Genomweite Analyse genetisch bedingter Strahlenempfindlichkeit in Wismut-Bergarbeitern – Datenauswertung und Bewertung der Assoziationsanalysen

Der Bericht gibt die Auffassung und Meinung des Auftragnehmers wieder und muss nicht mit der Meinung der Auftraggeberin übereinstimmen.

Göttingen, den 19.6.2017

Inhalt

TABELLENVERZEICHNIS

ABBILDUNGSVERZEICHNIS

1 Kurzfassung

1.1 Hintergrund

Lungenkrebs ist weltweit ein großes Thema des Gesundheitswesens. Im Laufe ihres Lebens werden einer von 14 Männern und eine von 17 Frauen einen invasiven Lungen- oder Bronchialtumor entwickeln.[1] Außerdem überleben nur ein bis zwei von 5 Patienten die ersten 5 Jahre nach der Diagnosestellung.[2] Lungenkrebs kann durch das Inhalieren von Tabakrauch oder Feinstaub, aber auch durch ionisierende Strahlung infolge der Inhalation von Radon und Radonfolgeprodukten ausgelöst werden. Ein erhöhtes, durch Strahlung induziertes Lungenkrebsrisiko konnte durch mehrere Studien zur Radon-Exposition in Wohnräumen (z. B. Darby, et al., 2005 [3]) unter Uranbergarbeitern (z. B. Grosche, et al., 2006 [4]; BEIR IV Report 1999 [5]) belegt werden.

Das *International Lung Cancer Consortium* (ILCCO), aus dem das Konsortium *Transdisciplinary Research in Cancer of the Lung* (TRICL) hervorging, wurde 2004 mit dem Ziel gegründet, vergleichbare Daten von laufenden Studien zu Lungenkrebs-Erkrankung zusammen zu bringen.[6] Die teilnehmenden Studien stammen aus verschiedenen geografischen Regionen und umfassen mehrere Ethnien. Auf der Basis aller genomischen Daten des ILCCO/TRICL war es möglich, die Existenz von genomischen Risikofaktoren für Lungenkrebs in europäisch stämmigen Populationen auf den Chromosomen 5p15.33, 6p21-22 und 15q25.3-10 zu identifizieren und zu verifizieren.[7-14] Ebenso wurde durch das Konsortium die Untersuchung von genetischen Risikofaktoren der Tabaksucht in Abgrenzung zur Tabakrauch als bedeutendster Risikofaktor für Lungenkrebs selbst möglich.[15,16] In den Jahren 2013-2016 wurde von ILCCO/TRICL eine großangelegte Genotypisierung mit dem OncoArray durchführt.

Beginnend 1946 waren Bergbaubeschäftigte der Wismut−AG über lange Jahre hinweg Strahlung durch Radon und Radonfolgeprodukte während des Abbaus von Uranerz ausgesetzt. Von diesen Personen liegen dem BfS gute Abschätzungen deren berufsbedingter Exposition gegenüber Strahlung vor. Beschäftigte, deren kumulierte Exposition 50 WLM übersteigt, werden von der dt. gesetzlichen Unfallversicherung (DGUV) jährlich, Beschäftigte mit weniger als 50 „Working Level Months „ (WLM) alle drei Jahre zu Vorsorgeuntersuchungen einbestellt. Die Gesundheitsdaten sowie andere epidemiologische Merkmale (z.B. das Rauchverhalten) dieser Personen sind gut erfasst. Die Datensammlung dieses Kollektivs aus Bergarbeitern ist in seinem Umfang weltweit einmalig. In den Jahren 2009/2011 wurde eine Bio- und Datenbank von gesunden, ehemaligen Wismut-Bergarbeitern aufgebaut, die auch Blutproben bzw. DNA-Proben zur Analyse des Genoms enthält. Dies Bioprobenbank umfasst Proben und Daten von 292 hochexponierten Probanden (>750 WLM, 66%) und 150 gering exponierten Probanden (<50 WLM, 34%). Die Probanden waren zum Zeitpunkt der Probenahme zwischen 70 und 90 Jahren alt. 63% waren ehemalige Raucher, 5% rauchten noch zum Zeitpunkt der Untersuchung und 32% hatten nie geraucht.[17] Die meisten der in diese Untersuchten eingehenden Lungenkrebsfälle der Wismut-Bergarbeiter wurden im Rahmen einer Studie zu Innenraumbelastung durch Radon zwischen 1990 und 1997 rekrutiert.[18] Sie waren zwischen 49 und 81 Jahren alt und mit einer Ausnahme Raucher zum Zeitpunkt der Diagnosestellung.

1.2 Projektziel

Das Ziel des hier berichteten Forschungsvorhabens bestand in der Suche nach genetischen Varianten, die im Zusammenhang mit Strahlung zu einem erhöhten Lungenkrebsrisiko beitragen können. Im Rahmen zweier Vorgängerprojekten des BfS (3614S10013 und 3614S10014) wurden 516 Proben von Wismut-Bergarbeitern ebenfalls mit dem OncoArray typisiert. Diese Daten bilden zusammen mit denen der LUCY-Study, der German Lung Cancer Study (GLC) sowie anderen Lungenkrebsstudien (soweit vom *Transdisciplinary Research In Cancer of the Lung / International Lung Cancer Consortium* TRICL/ILCCO zur Verfügung gestellt) die Grundlage der hier berichteten genomweite Asso-

ziationsuntersuchung (GWAS) der Interaktion zwischen genomischen Markern und der Strahlenexposition hinsichtlich des Lungenkrebsrisikos.

1.3 Ergebnisse

Die untersuchte Stichprobe bestand aus insgesamt 28.599 kaukasischen Studienteilnehmern - 463 Studienteilnehmer waren Wismut-Bergarbeiter, 946 Fälle und Kontrollen stammen von der *German Lung Cancer Study* und 27.187 stammen vom OncoArray-Konsortium ILCCO/TRICL. Davon waren 49 der 15.077 (0,3%) Fälle und 259 der 13.522 Kontrollen (1,9%) gegenüber Strahlung aus natürlichen Radionukliden exponiert (WLM>50).

Die Datenauswertung umfasste sowohl Einzel- als auch Multimarker-Assoziationsmodelle, wobei in zweiterem alle Marker eines LD-Blocks gemeinsam berücksichtigt wurden. Alle Modellschätzungen sind hinsichtlich Alter, Geschlecht und Rauchverhalten sowie bezüglich genomischer Populationsstrukturen adjustiert. Aus methodischen Gründen sollten die als Oddsratio (OR) geschätzten Assoziationsstärken nicht als unverzerrte Schätzwerte eines relativen Lungenkrebsrisikos betrachtet werden, sondern als studien-interne Größe angesehen werden.

Darüber hinaus wurde die gemeinsam Assoziation von Genen, bzw. den zugeordneten Markern, die einem von 119 vordefinierten Gen-Sets angehören, im Rahmen einer *"gene-set enrichment analysis"* untersucht. Bei dieser Gene-Set-Analyse wurde die Häufigkeit von interaktionstragenden LD-Blöcken eines betrachteten Gene-Sets mit der unter allen verbleibenden Genen verglichen. Gesucht wurden also Gene-Sets mit einer relativen „Anreicherung" an auffälligen GxE-Interaktionen, bzw. jene Gene die zu dieser Anreicherung wesentlich beitragen.

Fasst man die Ergebnisse der Einzel- und Multimaker-Assoziationsanalysen und der Gen-Set-Analyse zusammen, so kann für einige Gene bzw. Gene-Sets eine Rolle in der Tumorgenese des Lungenkrebses, in Interaktion mit einer berufsbedingten Radon/Strahlen-Belastung vermutet werden. Die bedeutsamsten Gene und Gene-Sets mit signifikanter GxE Interaktion sind:

Für eine genomische Region – **Chromosom 1p31.3** - konnte eine genomweit signifikante Interaktion mit einer berufsbedingten Radon/Strahlen-Exposition beobachtet werden ($p=5x10^{-7}$). Diese Region überdeckt das Gene **UBE2U**. Der Intron-Marker rs2029868 markiert dabei mit einer geschätzten OR=22,6 (95%-CI: 4,7- 109; p=0,0001) eine positive Interaktion für das jeweils seltene Allel unter Radon-exponierten Bergarbeitern.

Zwei weitere Gene fielen durch die Gene-Set-Analyse hinsichtlich einer nominal-signifikanten „Anreicherung" an auffälligen GxE-Interaktionen auf.

Dem Gen **FTO**/**ALKBH9** – auf Chromosom 16q12.2 – kommt hinsichtlich des molekularen Signalwegs der ***„DNA dealkylation involved in DNA repair"*** *(*Gen-Set *GO:0006307)* eine zentrale Rolle zu. Das Gen-Set selbst erzielte in der Gene-set-Analyse mit p=0,0139 den niedrigsten p-Wert. Ein Intron-Marker des Gens FTO an Position 53.995.500 markiert dabei mit einer geschätzten OR=7,9 (95%-CI: 4,7- 109; p=0,0001) eine positive Assoziation für das jeweils seltene Allel unabhängig von einer Radon-Exposition. Gleichzeitig wird für diesen Marker aber eine negative Interaktion für das jeweils seltene Allel unter Radon-exponierten Bergarbeitern (OR=0,03; p=0,0130) geschätzt.

Das Gen **CUBN** – auf Chromosom 10p13 - war hinsichtlich des weitgefassten GO-Begriffs GO:0016020 ***„Membrane (cellular-component)"***, der eine Signifikanz nur knapp verfehlte (p=0,0558), das am stärksten assoziierte Gen. Ein Multimarker-Modell ergab ebenso für den innerhalb des Genes liegenden LD-Block Nr. 58899 eine suggestive Signifikanz (p=0,000013). Der Intron-Marker rs4748341 markiert dabei mit einer geschätzten OR>36 die stärkste positive Interaktion für das jeweils seltene Allel unter Radon-exponierten Bergarbeitern ($p=5,6x10^{-6}$).

Darüber hinaus zeigten die Genfamilie der ***„microRNAs" (*HGNC:476)** sowie der **„acyl-CoA metabolic process"** (GO:0006637) eine auffällige „Anreicherung" an auffälligen GxE-Interaktionen. Es konn-

te jedoch keinem einzelnen Gen eine zentralen Rolle hinsichtlich der untersuchten GxE-Interaktion in diesen Gen-Sets zugeschrieben werden.

Durch die Einzel- oder Multimarker-Modellierung konnten 10 weitere genomische Regionen mit suggestiver Signifikanz identifiziert werden, die möglicherweise weitere Gen x Radon-Interaktionen beherbergen.

Zum ersten Mal wurden in einer genomweiten Untersuchung nach genomischen Prädispositionen gesucht, die nach einer lebenslangen, berufsbedingten Strahlenbelastung das Risiko einer Lungenkrebserkrankung erhöhen oder davor schützen. Die ausreichende Fallzahl für diese Untersuchung wurde erst durch das Zusammenführen von Studiendaten des BfS und aller in TRILC/ILCCO vereinigten, internationalen Studien erreicht. Die Konformität der genotypischen Daten wurde durch identische Typisierungsarrays und übereinstimmenden Typisierungsprotokolle gewährleistet. Die Wirkmechanismen der durch diese Untersuchung durch statistische Modelle identifizierten Gene und molekular-biologischen Signalwege müssen jedoch weiter untersucht und validiert werden. Die Ergebnisse sind aber dennoch schon jetzt für den Strahlenschutz von Bedeutung. Es konnte gezeigt werden, dass das Risiko für eine strahlungsinduzierte Lungenkrebserkrankung durch eine individuell genetische Prädisposition mit determiniert ist.

1.4 Zukünftige Arbeiten und Auswahl zu validierender Kandidatengene (AP 3.3)

Die Assoziation dieser Kandidatengene mit Lungenkrebs sollte durch DNA-Proben der derzeit rund 250 Lungenkrebsfälle des *Wismut-Pathologiearchivs der BfS* zu validiert werden. Dabei handelt es sich um folgende Gene (Tabelle 1):

Tabelle 1 Auswahl der Kandidatengene für eine Validierung

Gene	Chromosom	
Genomweite Signifikanz (Multimarker-Modelle)		
UBE2U	1p31.3	Multimarker-Analyse
LOC107985187, LOC105372156, RP11-325K19.1	18q21.32	Multimarker-Analyse
Durch Gen-Set-Analyse identifiziert		
FTO/ALKBH9	16q12.2	Gene-Set-Analyse
CUBN	10p13	Gene-Set-Analyse
SOX5, MIR920	12p12.1	Gene-Set-Analyse
Suggestive Signifikanz (Einzel- oder Multimarker-Modelle)		
CSNK1G3, LINCO1170	5q23.2	Einzelmarker-Analyse
NAV2, NAV2AS4, NAV2AS5	11p15.1	Multimarker-Analyse
CD163L1, LOC101927882, ACSM4, PEX5	12p13.31	Multimarker-Analyse
C15orf32, ST8SIA2, snoU109, RP11763K1511, RP11763K1521	15q26.1	Multimarker-Analyse

Sofern noch ausreichend Mittel zur Verfügung stehen, wäre es ebenso sinnvoll, die Assoziation von durch andere Studien endeckten Genen zu validieren. Diese sind (Tabelle 2):

Tabelle 2 Kandidatengene für eine Validierung mit externer Evidenz

Gene	Chromosom	Referenz
GSTM1	1p13.3	[19]
EPHX1	1q42.12	[19]
IL6	7p15.3	[20]
CDKN2A	9p21.3	[21]
SIRT1	10q21.3	[22]
MGMT	10q26.3	[21]
P53	17p13.1	[23-25]
GSTT1/GSTT2	22q11.23	[19]

2 Summary

2.1 Background

Lung cancer is a major topic of health care worldwide. During their lifetime, one in 14 men and one in 17 women will develop an invasive lung or bronchial tumor. [1] In addition, only one to two out of five patients survive the first 5 years after diagnosis. [2] Lung cancer can be caused by the inhalation of tobacco smoke or fine dust but also be triggered by ionizing radiation due to the inhalation of radon and radon derived products. Increased radiation-induced lung cancer risk has been demonstrated by several studies of radon exposure in dwellings (e.g. Darby, et al., 2005 [3]) or for uranium miners (e.g. Grosche, et al., 2006 [4]; BEIR IV Report 1999 [5]).

The International Lung Cancer Consortium (ILCCO), from which the Transdisciplinary Research consortium in Cancer of the Lung (TRICL) emerged, was founded in 2004 with the aim of sharing comparable data from ongoing studies on lung cancer disease.[6] The participating studies were carried out in different geographical regions and include several ethnic groups. Based on all genomic data from ILCCO/TRICL, it was possible to identify and verify the existence of genomic risk factors for lung cancer in European populations on the chromosomes 5p15.33, 6p21-22 and 15q25.3-10. [7-14] The collaboration in the consortium has also enabled to study of genetic risk factors of tobacco addiction as distinct from tobacco smoke, the major risk factor for lung cancer. [15,16] In the years 2013-2016, within ILCCO / TRICL a large-scale genotyping of blood samples with OncoArray was carried out.

Beginning in 1946, mining employees of the Wismut-AG were exposed for many years to radiation by radon and radon products during the digging of uranium. Of these miners, the BfS holds good estimates of their occupational exposure to radiation. Miners whose cumulative exposure exceeds 50 “Working Level Months“ (WLM) are examined annually by German statutory accident insurance (DGUV), employees with less than 50 WLM every three years for check-ups. Health data as well as other epidemiological characteristics (e.g., smoking behavior) of these individuals are well recorded. The data collection of this collective of miners is unique in its scope worldwide. In 2009/2011, a bio and database of healthy, former Wismut miners was set up, which also contains blood or DNA samples for genetic investigations. This Biobank includes samples and data from 292 highly exposed volunteers (>750 WLM, 66%) and 150 low-exposure subjects (<50 WLM, 34%). The subjects were between 70 and 90 years of age at the time of sampling. 63% were former smokers, 5% still smoked at the time of examination and 32% had never smoked.[17] Most of the lung cancer cases of the Wismut miners entering this investigation were recruited from 1990 to 1997 for an indoor radon study.[18] They were between 49 and 81 years of age and with one exception smokers at the time of diagnosis.

2.2 Project Goal

The aim of the reported research project was to search for genetic variants that could contribute to increased lung cancer risk associated with radiation. In the context of two previous projects of the BfS (3614S10013 and 3614S10014), 516 samples of Wismut miners were typed with the OncoArray. These data together with those of the LUCY study, the German Lung Cancer Study (GLC) and other lung cancer studies (as provided by the Transdisciplinary Research In Cancer of the Lung / International Lung Cancer Consortium TRICL / ILCCO) form the basis of the genome-wide association study (GWAS) of the interaction between genomic markers and radiation exposure in respect to the lung cancer risk reported here.

2.3 Results

The study sample consisted in total of 28,599 Caucasian study participants - 463 study participants were Wismut miners, 946 cases and controls were from the *German Lung Cancer Study* and 27,187 were from the OncoArray consortium ILCCO/TRICL. 49 of the 15,077 (0.3%) cases and 259 of the 13,522 controls (1.9%) were exposed to radiation from natural radionuclides (WLM> 50).

The data analysis included both single and multi-marker association models, while in the second all markers of an LD block were considered together. All model estimates are adjusted for age, gender and smoking behavior, as well as genomic population structures. For methodological reasons, the association strengths estimated as odds ratio (OR) should not be considered as unbiased estimates of relative lung cancer risk, but should be considered as study-internal variables.

In addition, the joint association of genes or the associated markers, which belong to one of 119 predefined gene sets, was investigated in a *gene-set enrichment analysis*. In this gene-set analysis, the frequency of interacting LD blocks of a considered gene set was compared with that within all remaining genes. We were looking for gene sets with a relative "*accumulation*" of conspicuous GxE interactions, or genes that contribute significantly to this accumulation.

Summarizing the results of single and multi-maker association analyzes and gene set analysis, some genes or sets of genes may be implicated in tumorigenesis of lung cancer interacting with occupational radon/radiation exposure. The most prominent genes and gene sets with significant GxE interaction are:

For a genomic region - **chromosome 1p31.3** - a genome-wide interaction with occupational radon/radiation exposure could be observed ($p = 5x10^{-7}$). This region covers the Gene **UBE2U**. The intron marker rs2029868, with an estimated OR = 22.6 (95% CI: 4.7-109; p = 0.0001), indicates a positive interaction for the rare allele among radon-exposed miners.

Two other genes were detected by gene set analysis for nominal-significant "accumulation" of prominent GxE interactions.

The gene **FTO/ALKBH9** - on chromosome 16q12.2 - plays a central role in the molecular signal pathway of the ***"DNA dealkylation involved in DNA repair"*** (gene set GO: 0006307). The gene set achieved the lowest p-value in the gene-set analysis with p = 0.0139. An intron marker of the FTO gene at position 53,995,500, with an estimated OR=7.9 (95% CI: 4.7-109; p=0.0001), marks a positive association for the respective rare allele independently of a radon exposure. At the same time, however, a negative interaction for the respective rare allele among radon-exposed miners (OR=0.03, p=0.0130) is estimated for this marker.

The gene **CUBN** - on chromosome 10p13 - was the most strongly associated gene of the broadly defined GO term: 0016020 ***"membrane (cellular-component)"***, which only barely missed significance (p=0.0558). A multi-marker model also revealed suggestive significance for the LD block no. 58899 located within this gene (p=0.000013). The strongest positive interaction for the rare allele among radon-exposed miners was estimated for then intron marker rs4748341 (OR>36; p=5.6x10-6).

In addition, the gene family of ***"microRNAs"*** **(HGNC: 476)** and the ***"acyl-CoA metabolic process"*** (GO: 0006637) showed a remarkable "*accumulation*" of prominent GxE interactions. However, no single gene could be attributed a central role in the investigated GxE interaction in these gene sets.

Single or multi-marker modeling identified 10 additional genomic regions of suggestive significance that may harbor additional gene x radon interactions.

For the first time, the genetic predisposition that influences the risk of lung cancer given a lifelong occupational radiation exposure was investigated in a genome-wide study. The sufficient number of cases for this study was only achieved by merging study data of the BfS and the collection of international studies of TRILC/ILCCO. The conformity of the genotypic data was ensured by identical typing arrays and consistent typing protocols. However, the mechanisms of action of the genes and molecular biological pathways identified by this study through statistical models need further investigation and validation. Nevertheless, the results are already important for radiation protection. It has been shown that the risk of radiation-induced lung cancer is also determined by an individual genetic predisposition.

2.4 Additional project task

The association of these candidate genes with lung cancer need to be validated by DNA samples from the at present around 250 lung cancer cases of the BfS Wismut pathology archive. These are the following genes (Tabelle 3):

Tabelle 3 Selection of candidate genes for validation

Gen	Chromosome	
Genome-wide significant (multi-marker-model)		
UBE2U	1p31.3	Multi-marker analysis
LOC107985187, LOC105372156, RP11-325K19.1	18q21.32	Multi-marker analysis
Identified by the gene-set-analysis		
FTO/ALKBH9	16q12.2	Gene-set analysis
CUBN	10p13	Gene-set analysis
SOX5, MIR920	12p12.1	Gene-set analysis
Suggestive significant (single- or multi-marker-model)		
CSNK1G3, LINCO1170	5q23.2	Single-marker analysis
NAV2, NAV2AS4, NAV2AS5	11p15.1	Multi-marker analysis
CD163L1, LOC101927882, ACSM4, PEX5	12p13.31	Multi-marker analysis
C15orf32, ST8SIA2, snoU109, RP11763K1511, RP11763K1521	15q26.1	Multi-marker analysis

If there are still sufficient resources available, it would be also useful to validate the association of genes revealed by other studies. These are (Tabelle 4):

Tabelle 4 Candidate genes for validation with external evidence

Gene	Chromosome	Reference
GSTM1	1p13.3	[19]
EPHX1	1q42.12	[19]
IL6	7p15.3	[20]
CDKN2A	9p21.3	[21]
SIRT1	10q21.3	[22]
MGMT	10q26.3	[21]
P53	17p13.1	[23-25]
GSTT1/GSTT2	22q11.23	[19]

3 Kurzdarstellung des FE-Vorhabens

3.1 Aufgabenstellung

Ziel des Vorhabens ist abzuklären, welche genetischen Varianten im Zusammenhang mit Strahlung zu einem erhöhten Lungenkrebsrisiko beitragen.

Die im Rahmen der Vorgängerprojekte des BfS (3614S10013 und 3614S10014) gewonnenen genomischen Daten der Wismut-Bergarbeiter der Bioprobenbank (3608S04532), sowie die Daten der LUCY-Studie, der Deutschen Lungenkrebsstudie (*German Lung Cancer Study,* GLC) und anderen Lungenkrebsstudien (soweit diese vom *Transdisciplinary Research In Cancer of the Lung / International Lung Cancer Consortium* TRICL/ILCCO zur Verfügung gestellt werden) bilden die Grundlage für eine genomweite Assoziationsuntersuchung (GWAS) der Interaktion zwischen genomischen Markern und der Strahlenexposition hinsichtlich des Lungenkrebsrisikos. In die Analyse fließen die Expositionsdaten, medizinische und epidemiologische Daten ein. Tabelle 64 im Anhang enthält eine detaillierte Studienbeschreibung der Originalstudien.

Die Qualität der Genotypisierung der Proben von Wismut-Bergarbeitern und Proben anderer Studien wurde im Auftrag des BfS am Institut für Genetische Epidemiologie der Universitätsmedizin Göttingen bereits geprüft. Von Bergarbeitern, die bei der Wismut-AG beschäftigt waren, liegen gute Abschätzungen über ihre Exposition gegenüber Strahlung, als auch zum Teil durch andere Noxen, wie Quarzstaub und Arsen, vor. Beschäftigte, deren kumulierte Lebenszeitexposition 50 „Working Level Months“ (WLM) übersteigt, werden von der dt. gesetzlichen Unfallversicherung (DGUV) jährlich, Beschäftigte mit weniger als 50 WLM alle drei Jahre zu Vorsorgeuntersuchungen einbestellt. Die Gesundheitsdaten sowie andere epidemiologische Merkmale (z.B. das Rauchverhalten) dieser Personen sind gut erfasst.

3.2 Voraussetzungen, unter denen das FE-Vorhaben durchgeführt wurde

Das Forschungsvorhaben FKZ 615S32253 schließt direkt an die Vorgängerprojekte des BfS (3614S10013 und 3614S10014) an. Im Projekt FKZ 3614S10013 wurde das Gesamtprojekt geplant und die Kooperation mit ILCCO/TRICL etabliert. Im Projekt FKZ 3614S10014 wurden die Qualität der an der HGMU gewonnenen Genotypen geprüft und mit der der Proben von ILCCO/TRICL verglichen.

3.3 Planung und Ablauf des Vorhabens

Die vertraglich vereinbarte Leistung des Instituts für Genetische Epidemiologie umfasst folgende Arbeitspakete:

AP1. **Arbeitspaket 1:** Entwicklung eines geeigneten Modells zur Untersuchung von Gen x Umwelt (Gen x Strahlung) Interaktionen

AP2: **Arbeitspaket 2:** Genomweite Assoziationsanalyse zum Lungenkrebsrisiko in Abhängigkeit der Strahlenexposition unter Einbeziehung der Daten aus verschiedenen nationalen (Wismut, LUCY, KORA) und internationalen Kollektiven (z.B. ILCCO, TRICL, GAME-ON).

- **AP 2.1a Zusammenführen und Harmonisieren phänotypischer Daten**
- **AP 2.1b Zusammenführen und Harmonisieren genomischer Daten**
- **AP 2.1c Durchführen der GWAS**
- **AP 2.2 Erstellung eines Zwischenberichts**

AP3: Arbeitspaket 3: Identifizierung krankheitsrelevanter genomischer Regionen, „Gene-sets“ und/oder regulatorischer Netzwerke und andere biologischer Signalwege hinsichtlich des strahleninduzierten Lungenkrebses. Auswahl geeigneter SNPs, die in einem weiteren Bergarbeiterkollektiv (Lungenkrebsfälle/gesunde exponierte Kontrollen) validiert werden.

AP 3.1 Zusammenstellen der zu untersuchenden „Gene-sets“
AP 3.2 Durchführung einer GSA
AP 3.3 Auswahl zu validierender Kandidaten-SNPs
AP 3.4 Erstellung eines Abschlussberichts

Vertraglich wurde ferner folgender zeitliche Projektablauf vereinbart:

Übersicht: Zeitlicher Ablauf

Jahr	Kalender-monat	Quartal	Projektmo-nat	Arbeitspaket		Dauer	*von*	*bis*
2016	März	4	1	**AP 1**	**AP 1** *Modellentwicklung*	2 Monate	*1.3.*	
	Apr	1	2					*30.4*
	Mai		3	**AP 2** ***GWA***	**AP 2.1a** Harmonisieren der Phänotypen	2 Monate	*1.5.*	
	Jun		4					*30.6.*
	Jul	2	5		**AP 2.1b** Harmonisieren der Genotypen	2 Monate	*1.7.*	
	Aug		6					*31.8.*
	Sep		7		**AP 2.1c** GWA durchführen	1½ Monate	*1.9.*	
	Okt	3	8					*15.10.*
					AP 2.2^{B}	½ Monat	*16.10.*	*31.10.*
	Nov		9	**AP 3** ***GSA***	**AP 3.1** Gene-Sets definieren	1 Monat	*1.11.*	*30.11.*
	Dez		10		**AP 3.2** GSA durchführen	1½ Monate	*1.12.*	
2017	Jan	4	11					*15.1.*
					AP 3.3 SNPs auswählen	1 Monat	*16.1.*	*15.2.*
	Feb		12		**AP 3.4**^{B,PT}	½ Monat	*16.2.*	*28.2.*

AP… Arbeitspaket, B… Bericht verfassen, PT…Projekttreffen

Dieser Bericht enthält die Ergebnisse der Arbeitspakete AP1 bis AP3.
Die Auswahl des für die Analyse geeigneten Basismodells erfolgte in Abhängigkeit von den zur Verfügung stehenden Daten. Daher ist dieser Bericht im Weiteren nicht nach Arbeitspaketen strukturiert, sondern folgt einer inhaltlich auf sich aufbauenden Gliederung.

3.4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

In den vergangenen 10 Jahren wurden mehrere genomweite Assoziationsanalysen (GWAS) durchgeführt, um genetischen Risikofaktoren einer Lungenkrebserkrankung zu identifizieren. 45 Loci wurden dabei mit unterschiedlicher Assoziationsstärke für die histologischen Subtypen des Lungenkrebses gefunden. Das TRICL/ILCCO-Konsortiums wurde 2004 mit dem Ziel gegründet, Daten vergleichbarer Fall-Kontroll- und Kohortenstudien zu Lungenkrebs auszutauschen und zusammenzubringen um die Fallzahlen für statistische Analysen zu vergrößern.[6] Im Rahmen von TRICL/ILCCO wurden koordiniert GWAS durchgeführt. Eine Übersicht über mit Lungenkrebs assoziierte Gene, auch aus Studien die Kandidatengene untersucht haben, wurde im Rahmendes Forschungsvorhaben FKZ 3614S10013 zusammengestellt. Ein Übersichtsarbeit zu den Fortschritten in der Erforschung genetischen Prädisposition für Lungenkrebs in GWAS innerhalb der vergangenen 10 Jahre wurde von Bosse and Amos, 2017 [26] veröffentlicht.

Ein erhöhtes Risiko für Lungenkrebs, verursacht durch Inhalation von Radon, wurde in mehreren Studien zur Innenraumexposition in Wohnhäusern als auch für Uranbergarbeiter nachgewiesen. Es wurde geschätzt, dass ionisierende Strahlung durch Radon 3% bis 15% der LC-Fälle verursacht. [27]

Bisher wurde die individuelle (genetisch bedingte) Strahlenempfindlichkeit vor allem hinsichtlich Akut– bzw. Spätfolgen einer Strahlentherapie unter Tumorpatienten untersucht. Als Untersuchungsmethoden für die Strahlenempfindlichkeit (bzw. DNA-Reparaturfähigkeit) in der Normalbevölkerung standen bisher nur in vitro Versuchssysteme zur Verfügung. Das individuelle strahleninduzierte Tumorrisiko konnte bislang in der Normalbevölkerung nicht abgeschätzt werden.

Seit 2009 hat das BfS eine Biobank mit Proben von Bergarbeitern des Uranerzabbaus (ehemals Wismut AG) aufgebaut. Biologisches Material steht sowohl von gesunden, als auch von an Lungenkrebs verstorbenen Bergarbeitern zur Verfügung. Dieses bildet zusammen mit den Expositions–, medizinischen und epidemiologischen Daten die Grundlage für diese genomweite Assoziationsuntersuchung der Interaktion zwischen genomischen Markern und einer Strahlenexposition hinsichtlich des Lungenkrebsrisikos.

3.5 Zusammenarbeit mit anderen Stellen

Das Projekt forderte eine enge Zusammenarbeit mit dem Bundesamt für Strahlenschutz (BfS), das die Rahmenbedingung des Gesamtprojektes festlegt und über die notwendigen Bioproben von Wismut-Bergarbeitern verfügt.

Das Institut für Genetische Epidemiologie der UMG betreute das Projekt statistisch, sowohl in der Projektplanung, als auch während der Genotypisierung und führte die Qualitätskontrolle und -sicherung der Genotypisierung durch.

Die Abteilung für Molekulare Epidemiologie, Genome Analysis Center des Helmholtz-Zentrums München (HGMU) führte die Genotypisierung im Rahmen des Forschungsvorhabens FKZ 3614S10013 durch.

4 Eingehende Darstellung AP1-AP3

4.1 Datentransfer und -harmonisierung

Die Proben der in dieses Projekt eingehenden Wismut-Bergarbeiter wurden der BfS-Bioprobenbank (BfS-Projekt 3608S04532) entnommen. Die Bioprobenbank enthält Proben von freiwillig teilnehmenden, männlich Bergarbeitern, die entweder einer hohen kumulativen Exposition (>750 WLM) oder einer niedrigen kumulativen Exposition (<50 WLM) ausgesetzt waren und an keiner chronische Erkrankung (insbesondere Krebs, dekompensierte Nieren- oder Leberinsuffizienz) litten.[17]

Die genotypischen Daten der am Helmholtz-Zentrum München *Deutsches Forschungszentrum für Gesundheit und Umwelt* (HGMU) typisierten Personen (der Vorgängerprojekte 3614S10013 und 3614S10014) waren zu Projektbeginn bereits am *Institut für Genetische Epidemiologie* (GenEpi) vorhanden.

Die genotypischen und die phänotypischen Daten des TRICL/ILCCO-Konsortiums können aus rechtlichen Gründen derzeit nur auf dem Computer-Cluster des *Dartmouth College* in Hanover, New Hampshire 03755, USA genutzt werden. Ein Zugang zum Computer-Cluster, und auch zu OncoArray-Daten besteht für das *Institut für Genetische Epidemiologie* seit dem 6. April 2016.

4.1.1 Zusammenführen und Harmonisieren phänotypischer Daten (AP 2.1a)

4.1.1.1 Phänotyp-Daten Wismut-Bergarbeiter

Einige Expositionsdaten, im Wesentlichen *Working Level Months* (WLM), der Wismut-Bergarbeiter wurden vom BfS bereits für die Projektplanung im Jahr 2014 an die *GenEpi* übermittelt. Die Phänotyp-Daten der Indoor-Radon-Studie (Wismut-Fälle) wurden am 3. März 2016 an die *GenEpi* übergeben. Durch Rückfragen an das BfS, bzw. in Kooperation mit der *HGMU* konnten fehlende Daten identifiziert und vom BfS nachgeliefert werden. Die Phänotypen der Wismut-Bergarbeiter waren am 16 August 2016 vollständig und bereinigt. Für 6 Wismut-Bergarbeiter waren in der Datenbank dem BfS jedoch keine Phänotypen gespeichert. Nach Rückfrage an Frau Dr. Brüske (*HGMU*) wurde eine fehlende Angabe der Strahlenexposition bei Teilnehmern der INDOOR-Radon-Studien als „nicht exponiert“ (WLM=0) gewertet.

4.1.1.2 Phänotyp-Daten – OncoArray-Konsortium

Zugang zu den Phänotyp-Daten des *TRICL/ILCCO-Konsortium* wurde uns am 6. April 2016 gewährt. Die Prüfung der Daten auf Vollständigkeiten und ob sich Genotypen zuordnen lassen ergab einige Inkonsistenzen. Diese wurden innerhalb des Konsortiums abgeklärt. Bereinigte Phänotyp-Daten liegen seit dem 31. Juli 2016 vor. Bei einer Person des OncoArray-Konsortiums waren die Angaben zum Erkrankungsstatus widersprüchlich.

4.1.1.3 Vollständigkeit aller Phänotyp-Daten

Von den insgesamt 28.606 genotypisierten Personen liegen vollständige Angaben zu *Alter* bei Diagnose bzw. Interview, *Geschlecht* und dem *Raucherstatus* vor. Hingegen fehlen Angaben zu *Packungsjahren* (*pack years)* bei 2.595 (9%) Personen, zu *Zigaretten pro Tag* bei 3.424 Personen (12%) und zur *Rauchdauer* von 3.249 Personen (11%). In den statistischen Modellen kann daher nur der *Raucherstatus* zum Adjustieren für das rauchbedingte Lungenkrebsrisiko verwendet werden. Nach Ausschluss der 6+1 Personen mit fehlenden WLM bzw. Alter können die Informationen von **28.599 Personen** in der Datenauswertung berücksichtigt werden.

4.1.2 Deskriptives: Phänotypen

Das Kollektiv besteht aus insgesamt 28.599 kaukasischen Studienteilnehmern (7 genotypisierte Personen mussten wegen fehlender oder widersprüchlicher Phänotypen ausgeschlossen werden).

Insgesamt waren 49 der 15.077 (0,3%) Fälle und 259 der 13.522 Kontrollen (1,9%) gegenüber Strahlung aus natürlichen Radionukliden exponiert (WLM>50). Für 10 der 15.077 Fälle wurde eine Strahlenexposition WLM<50 dokumentiert; diese werden in der Datenauswertung als nicht-exponiert betrachtet. 27.187 Studienteilnehmer sind nicht exponierte Fälle oder Kontrollen des OncoArray-Konsortiums; 949 nicht exponierte Fälle und Kontrollen der *German Lung Cancer Study* (GLC) wurden mit dem Illumina Human-550K Chip typisiert, und 463 Fälle und Kontrollen waren ehemalige Wismut-Bergarbeiter.

Von 20 Studienteilnehmern kommen 9 aus Nordamerika, 9 weitere aus Europa und zwei aus Israel oder Russland. Etwa 1/3 der Studienteilnehmer sind Frauen, nur jeder 5-te hat nie geraucht. Das mediane Alter bei Diagnose bzw. Interview lag bei 63 Jahren, war jedoch unter den Wismut-Bergarbeitern mit medianen 76 Jahren deutlich höher und unter den Studienteilnehmern der GLC mit medianen 46 Jahren deutlich niedriger (gemäß des Einschlusskriteriums Alter≤50 Jahre der GLC). Die Strahlenexposition unter den 308 Exponierten streute von 51 bis 1.479 WLM (im Mittel 966 WLM) (siehe Tabelle 5 und Tabelle 6; Details je Originalstudie siehe Tabelle 64, 63 und 64).

Tabelle 5 Studienteilnehmer: Strahlenexposition

			nicht-exponiert			*exponiert*	*nicht-exponiert*	*exponiert*
	gesamt		***WLM=0***	***niedrig***	***moderat***	***hoch***	***≤50 WLM***	***>50 WLM***
gesamt	**28.599**	100%	28.138	97	56	308	99%	1%
Lungenkrebs								
Fälle	**15.077**	53%	15.018	4	6	49	100%	0%
Kontrollen	**13.522**	47%	13.120	93	50	259	98%	2%
Originalstudie								
German Lung Cancer Study	949	3%	949				100%	
OncoArray-Konsortium	27.187	95%	27.187				100%	.
Wismut-Bergarbeiter	463	2%	2	97	56	308	33%	67%
Geschlecht								
männlich	18.059	63%	17.598	97	56	308	98%	2%
weiblich	10.540	37%	10.540				100%	.
Rauchverhalten								
Nie-Raucher	5.676	20%	5.542	38	17	79	99%	1%
Ex-Raucher	9.518	33%	9.496	5	4	13	100%	0%
Raucher	12.039	42%	11.793	49	28	169	99%	1%
Jemals-Raucher	1.366	5%	1.307	5	7	47	97%	3%

Strahlenexposition: niedrig WLM >0 bis <20, moderat WLM 20 bis <50, hoch WLM>50;
Klassifikation für die Datenanalyse: nicht-exponiert WLM≤50, exponiert WLM>50
Sieben genotypisierte Studienteilnehmer werden auf Grund fehlender Phänotypen von der Datenauswertung ausgeschlossen.

Tabelle 6 Studienteilnehmer: Alter bei Diagnose / Interview

	gesamt				***Fälle***				***Kontrollen***			
Alter	***N***	***Min***	***Max***	*Median*	***n***	***Min***	***Max***	*Median*	***n***	***Min***	***Max***	*Median*
gesamt	28.599	15	96	**63**	13.522	15	96	**62**	15.077	21	95	**64**
Originalstudie												
GLC	949	27	51	**46**	478	29	51	**47**	471	27	50	**46**
OncoArray-Konsortium	27.187	15	96	**63**	12.642	15	96	**62**	14.545	21	95	**64**
Wismut	463	28	90	**76**	402	28	90	**77**	61	49	81	**67**

4.1.3 Zusammenführen und Harmonisieren genomischer Daten (AP 2.1b)

4.1.3.1 Imputation nicht typisierter Genotypen

Die *Genotypen* von 186 Lungenkrebsfällen der Heidelberg Lungenkrebsstudie, von 295 Lungenkrebsfällen der *LUCY*-Studie sowie von 478 krebsfreien Kontrollen der *KORA*-Studie (alle Teil der *German Lung Cancer Study* -GLC, einer Studie des TRICL/ILCCO-Konsortiums) wurden im März 2008 mit dem Illumina Human550K SNP-Array bestimmt. Dieses konnten in die hier beschriebene Untersuchung einbezogen werden, nachdem fehlende Genotypen (gegenüber einer Typisierung mit dem OncoArray) imputiert wurden.

Die Imputation wurde im Dezember 2012 durch das TRICL/ILCCO-Konsortium durchgeführt (Zhaoming Wang, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD 20892, USA: Imputation von >10 Million SNPs, mit den Daten des *1000 Genomes Projects* als Referenz. Verwendetet Programme: IMPUTE2, MACH, Minimac). Details der Imputation wurden an anderer Stelle publiziert.[13]

Aus den *dosage-files* der Imputation wurde der jeweils wahrscheinlichste Genotyp pro Person und Marker verwendet. Genotypen, bei dem die Wahrscheinlichkeit des wahrscheinlichsten Genotyps weniger als 1,25-fache der des zweit-wahrscheinlichsten Genotyps war, wurden nicht weiter verwendet

(Bsp (a): W(AA; AB; BB)=(0,90; 0,09; 0,01) → AA imputiert da 0,90/0,09=10;
Bsp (b): W(AA; AB; BB)=(0,50; 0,40; 0,10) → AA imputiert da 0,50/0,40=1,25;
Bsp (c): W(AA; AB; BB)=(0,40; 0,30; 0,30) → AA imputiert da 0,40/0,30=1,33;

Personen und Marker mit einer Rate fehlender Genotypen von jeweils >10% wurde ausgeschlossen.

4.1.3.2 Vereinen der GLC-500K-Daten mit den OncoArray-Daten

411.984 SNPs von 949 Studienteilnehmern der *GLC wurden schließlich mit* 456.699 SNPs und 27.651 mit dem OncoArray typisierten Personen vereint. Der initiale Genotyp-Datensatz besteht somit **aus 456.699 genomischen Markern** von **27.599 Individuen**.

Das Kollektiv dieses Datensatzes wurde auf mögliche genomische Substrukturen untersucht, um die Modellschätzungen geeignet adjustieren zu können. Hierbei wurde die Anzahl notwendiger „genomic principal components“ (PC) bestimmt (Hauptkomponenten, die diese Substrukturen wiederspiegeln) und auf genomische Unabhängigkeit der Studienteilnehmer geprüft.

4.1.3.3 Prüfen der genetischen Abstammung aller Personen

Vor jeglicher Modellschätzung ist zu prüfen, ob die genotypisierten Studienteilnehmer als Kaukasier gelten können, bzw. ob einzelne Personen von einer genomischen Mischpopulation („admixed source-population“) abstammen. Dazu wurde mit dem Programm ADMIXTURE[28] für jede Person die Wahrscheinlichkeit der Zugehörigkeit zu den Populationen Kaukasier, Asiaten oder Afrikaner geschätzt. Als Referenz dienten die Genotypen von nicht miteinander verwandte Personen, CEU-Kaukasier, YRI-Afrikaner und CHB-Asiaten, aus HapMap.

„Although millions of SNPs have been identified, only a small subset needs to be genotyped in order to accurately predict ancestry with a minimal error rate“.[29] Solche abstammungs-informative Marker (AIMs: *„Ancestry Informative Markers“*) sind genomische Varianten mit hohem Informationsgehalt zur Unterscheidung regionaler Herkunft auf kontinentaler und globaler Ebene (z. B. gemäß des “*informativeness for assignment measures*” $I_{n,}$ das mit der Divergenz der Häufigkeiten des selteneren Alleles (*minor allele frequencies*, MAFs) von AIMs zwischen Populationen steigt; oder Marker mit extremen Fixierungsindex $F_{ST,}$ einem Maß der Wahrscheinlichkeit, Kopien desselben Alleles nach Herkunft – *identity by descent* – in verschiedenen Personen zu beobachtet).[30]

Für die Klassifikation in die genomischen Ethnizitäten *Kaukasier, Asiaten* und *Afrikaner* wurden 178 abstammungs-informative Marker (25 AIMs und 25 PCAIMS - *PCA-Informative Markers* - zur Identifikation inner-europäischer Feinstrukturen[31], 128+26 AIMs zur Identifikation genomischer Ethnizitäten[32,33]) der Fachliteratur entnommen (siehe Tabelle 7). Von diesen wurden 91 Marker mit dem OncoArray typisiert. Von 68 nicht-typisierten Markern konnte je ein typisierter tagSNP (r^2>0,8 und Δ_{MAF}<0,1) bestimmt werden. Von diesen zusammen 159 AIMs wurden 107 Marker in allen drei HapMap-Populationen typisiert, und stehen somit für die ADMIXTURE-Analyse zur Verfügung.

Zusätzlich zur Kontrolle des Clusterings wurde eine Zufallsauswahl an 7.899 SNPs getroffen (MAF>5%, keine fehlenden Werte, p_{HWE}>0,05, kein nennenswertes „linkage disequilibrium LD" – untereinander, wurde durch gezielte Markerauswahl - „*pruning*" - sichergestellt), die dann gemeinsam mit den AIMs in die ADMIXTURE-Analyse eingehen.

Tabelle 7 Abstammungs-informative Marker (AIMs)

	Gesamt	auf OncoArray			als AIM verwendbar		HapMap ok*	Angereichtert durch eine Zufallsauswahl an SNPs*
		nein	ja	tagAIM	nein	ja		
AMIS genomischer Ethnizitäten [32]	**128**	**7**	**78**	**43**	**7**	**121**	**85**	**121**
AMIS genomischer Ethnizitäten [33]	**26**	**14**	**4**	**8**	**14**	**12**	**8**	**12**
AMIS inner-europäische [31]	**25**	**12**	**3**	**10**	**12**	**13**	**7**	**13**
PCAMIS inner-europäische [31]	**25**	**12**	**6**	**7**	**12**	**13**	**7**	**13**
Zufallsauswahl an SNPs								**7899**
gesamt	**204**	**45**	**91**	**68**	**45**	**159**	**107**	**8084**

* Genotypen für alle 3 HapMap-Samples (CEU, AFR, ASI) vorhanden

4.1.3.4 Prüfen der genetischen Substrukturen

Die Prüfung nach genetischen Subpopulationen erfolgte mit dem Clusteralgorithmus des Programms ADMIXTURE [28]. Alle typisierten Personen werden k „*ancestral populations clusters*" (ap-Cluster) zugeordnet. Diese ap-Cluster müssen dabei nicht notwendiger Weise mit den HapMap-Populationen (CEU, ARF, ASI) übereinstimmen. Die optimale Anzahl *k* an ap-Clustern wird durch die Bestimmung des Cross-Validierungsfehlers für k=1 bis 8 bestimmt. Die ap-Cluster stellen eine Art Dimensionsreduktion dar. Für jede Person werden die Wahrscheinlichkeiten p_i, einem ap-Cluster zuzugehören ausgegeben („ancestry fractions"). Aus diesen p_i können nun Profile für die 3 HapMap-Populationen erstellt werden. Mittels *k-means FastClustering* (PROC FASTCLUS) kann nun a) die Zugehörigkeit jeder Person zu den genomischen Populationen CEU, ARF, ASI bestimmt werden und b) eventuell weitere genomische Subcluster (gsub-Cluster) innerhalb der CEU und damit auch Inhomogenität zwischen den Studien untersucht werden. Die optimale Anzahl an genomischen Subclustern wird durch Maximierung der pseudo-F Statistik bestimmt.[34]

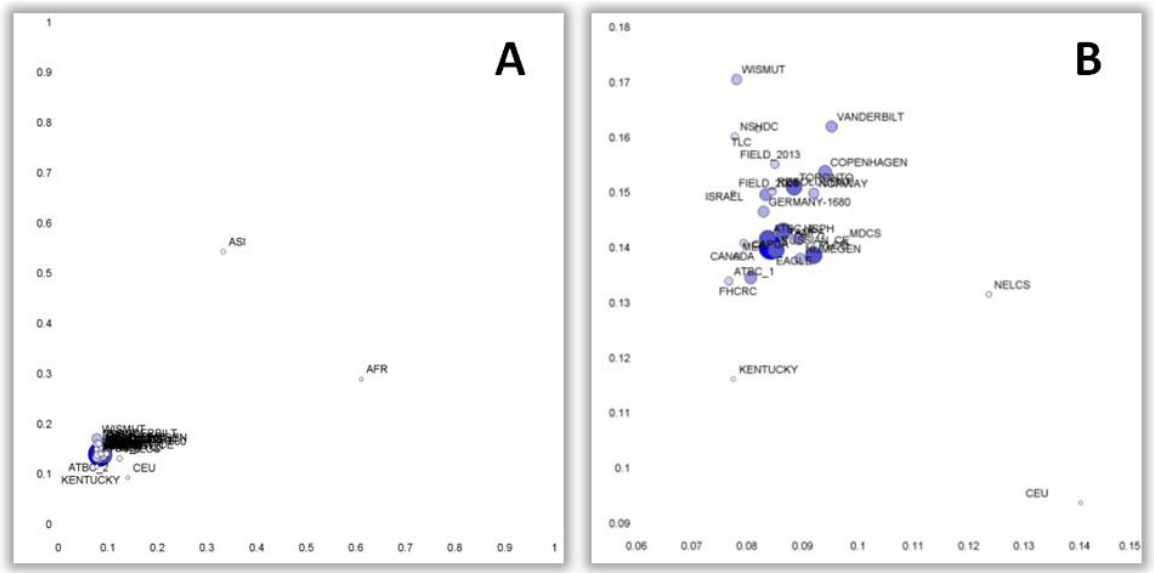
4.1.3.5 Ergebnis für die abstammungs-informativen Marker

Das am besten passende ADMIXTURE-Modell enthielt k=4 *ancestral populations.* Mit diesen lassen sich die Stichproben der Originalstudien klar als „kaukasisch" identifizieren und von HapMap-Afrikanern und HapMap-Asiaten unterscheiden (siehe Abbildung 1). Ein Unterschied in der Häufigkeit der ap-Cluster konnte weder hinsichtlich des Geschlechtes (HapMAp-Populationen, p=0,5773, χ^2-Test) noch hinsichtlich des Erkrankungsstatus (OncoArray-Stichprobe, p=0,2748, χ^2-Test) beobachtet werden. Es zeigt sich aber auch eine gewisse (zu erwartende) Inhomogenität innerhalb der kaukasischen Stichproben (siehe Abbildung 1). Am auffälligsten dabei ist die klare Abgrenzung der OncoArray-Stichproben gegenüber dem CEU-Sample von HapMap. Je nachdem welche Häufigkeit einer *ancestral population* (1-4) man betrachtet, zeigen die Wismut-Studien, KENTUCKY, TAMPA oder NELCS auffällige Besonderheiten.

Anhand der 4 ap-Cluster konnte nur eine einzige Person (ID=7855; Studie: HSPH) als „genomischer Ausreißer“ klassifiziert werden.

Anhand dieser 4 *„ancestral population“* wurden zunächst die Mittelpunkte von ethnischen ap-Clustern (HapMap-Populationen) bestimmt. Danach wurde jede einzelne Person einer diese k-means Cluster zugeordnet. Es zeigt sich, dass 6 genomische Subcluster (gsub-Cluster) gebildet werden konnten (siehe Tabelle 8). Die gsub-Cluster 2 und 5 enthielten alle HapMap-Afrikanern und HapMap-Asiaten (grafisch Dargestellt in Abbildung 1), interessanter Weise aber auch 5 der 60 HapMap-Kaukasier. Die anderen 4 genomische Subcluster (gsub-Cluster 1, 3, 4 und 6) dienen offensichtlich zur Differenzierung der Kaukasier (grafisch Dargestellt in Abbildung 1).

Abbildung 1 Genetischen Substrukturen (ausschließlich anhand abstammungs-informativer Marker AIMs)



A: Zuordnung der Stichproben als „kaukasisch“; gsub-Clustern 2 (horizontal) und 5 (vertikal)
B: Inhomogenität innerhalb der kaukasischen Stichproben; gsub-Clustern 1 (horizontal) und 3 (vertikal)
Auf den Achsen sind die mittleren „ancestry fractions“ ausgewählter gsub-Clustern je Originalstudien aufgetragen. Der Kreisdurchmesser ist proportional zur Stichprobengröße. Die gsub-Clustern wurden ausschließlich anhand abstammungs-informativer Marker AIMs gebildet.

Es konnte nur vereinzelt besondere Anhäufung einzelner gsub-Cluster in der einen oder anderen Originalstudie identifiziert werden. Die meisten Studien zeigen einen Anteil von 11% Personen die dem gsub-Cluster 2 und rund 2% Personen die dem gsub-Cluster 5 zugeordnet werden. Auffällig war vor allem die NELCS-Studie, die einen höheren Anteil Personen im gsub-Cluster 5 aber auch im gsub-Cluster 6 aufweist. Keine Auffälligkeiten wurden für die Proben der Wismut-Studie beobachtet (siehe Anhang Tabelle 68 und Abbildung 41).

Tabelle 8 Verteilung der genomischen Subcluster innerhalb der HapMap-Individuen

		1		2		3		4		5		6	
	N	n	%	n	%	n	%	n	%	n	%	n	%
Alle	210	17	8%	91	43%	11	5%	7	3%	64	30%	20	9%
AFR	60	.	.	22	36%	.	.	.	.	38	63%	.	.
ASI	90	.	.	68	75%	.	.	.	.	22	24%	.	.
CEU	60	17	28%	1	1%	11	18%	7	11%	4	6%	20	33%

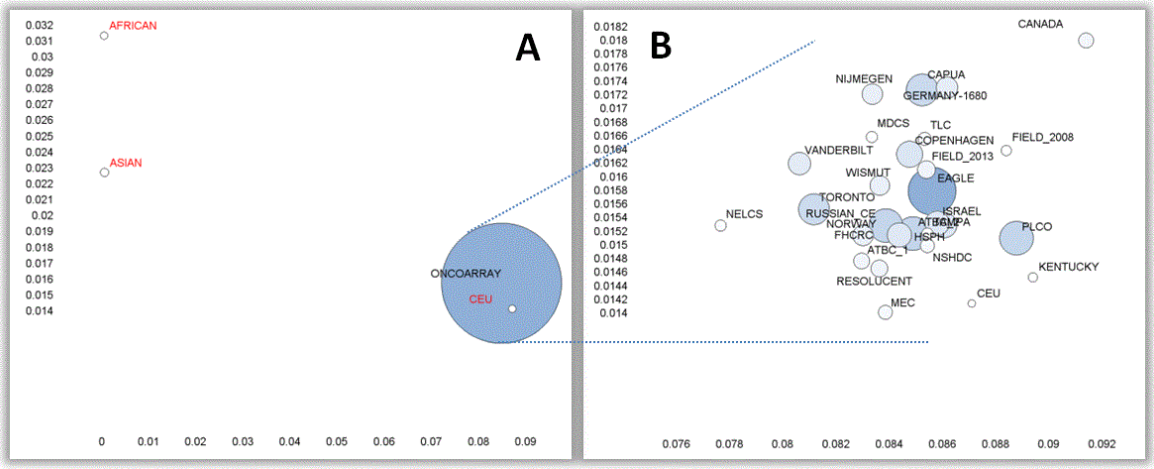
4.1.3.6 Ergebnis für die abstammungs-informativen Marker plus einer Zufallsauswahl an SNPs

Werden neben der AIMs (n=159) auch eine große Zahl (n=7899) an zufällig ausgewählten SNPs zum Bestimmen von genomischen Strukturen verwendet, erhält das am besten passende ADMIXTURE-Modell k=25 *„ancestral populations“.* Durch dieses Modell konten jedoch nur n=162 Studienteilnehmer (=0,92%) mit über 90%igen Wahrscheinlichkeit einem einzigen der ap-Cluster zugeordnet werden.
Anhand der Zugehörigkeit zu ap-Clustern lassen sich die Stichproben (je Studie) klar als Kaukasier identifizieren und von HapMap-Afrikanern und HapMap-Asiaten unterschieden (siehe Abbildung 2).

Für nur einen dieser 25 ap-Clustern konnte eine, auf dem 5% Niveau signifikant unterschiedliche Häufung zwischen Fällen und Kontrollen erkannt werden (ap-Cluster Nr. 17, p=0,0377, logistischen Regression; ebenso robuste lineare Regression, für etwa 15% der Studienteilnehmer ist dieser das wahrscheinlichste ap-Cluster). Angesichts des multiplen Testens, kann dies aber als Zufallsbefund gewertet werden.

Ein Unterschied in der Häufigkeit der ap-Cluster konnte weder hinsichtlich des Geschlechtes (HapMap-Populationen, p=0,6823, χ^2-Test) noch hinsichtlich des Erkrankungsstatus (OncoArray-Stichprobe, p=0,1862, χ^2-Test) beobachtet werden. Es zeigt sich aber auch eine gewisse (zu erwartende) Inhomogenität innerhalb der kaukasischen Stichproben (siehe Abbildung 2).

Abbildung 2 Genetischen Substrukturen (abstammungs-informativer Marker und Zufallsauswahl an SNPs)



A: Zuordnung der Stichproben als „kaukasisch“; gsub-Clustern 2 (horizontal) und 3 (vertikal)
B: Inhomogenität innerhalb der kaukasischen Stichproben; gsub-Clustern 2 (horizontal) und 3 (vertikal)
Auf den Achsen sind die mittleren „ancestry fractions“ ausgewählter gsub-Clustern je Originalstudien aufgetragen. Der Kreisdurchmesser ist proportional zur Stichprobengröße. Die gsub-Clustern wurden anhand abstammungs-informativer Marker und Zufallsauswahl an SNPs gebildet.

Anhand dieser 25 *„ancestral populations“* wurden nun zunächst die Mittelpunkte von ethnischen ap-Clustern bestimmt. Danach wurde jede einzelne Person einem dieser k-means Cluster zugeordnet. Es zeigt sich, dass 4 gsub-Cluster gebildet werden konnten (siehe Tabelle 9). Die gsub-Cluster 2 und 4 enthielten alle HapMap-Afrikanern und HapMap-Asiaten. Interessanter Weise aber auch 5 der 60 HapMap-Kaukasier. Der gsub-Cluster Nr. 3 dienen offensichtlich zur Identifizierung von Kaukasier, wobei je ein HapMap-Afrikaner und HapMap-Asiate diesem gsub-Cluster zugeordnet wurde (siehe Tabelle 9).

Es konnte keine besondere Anhäufung einzelner gsub-Cluster in der einen oder anderen Studie der OncoArray-Studienteilnehmer identifiziert werden (p=0,1183, χ^2-Test). Auffällig ist nur, dass etwa 9% der „Kaukasier“ dem gsub-Cluster 2 zugeordnet werden (siehe Anhang Tabelle 69).

Tabelle 9 Verteilung der „ancestral populations“-Cluster unter HapMap-Kontrollen (AIMS + Zufallsauswahl an SNPs)

	gesamt	1		2		3		4	
	n	n	%	n	%	n	%	n	%
Alle	210	1	0%	89	42%	57	27%	63	30%
AFR	60	.	.	21	35%	1	1%	38	63%
ASI	90	1	1%	67	74%	1	1%	21	23%
CEU	60	.	.	1	1%	55	91%	4	6%

4.1.3.7 Fazit

Die mit dem OncoArray typisierte Studienpopulation kann als „kaukasisch“ angesehen werden. Eine Untermischung von afrikanischer oder asiatischer DNA ist, wenn vorhanden, dann limitiert und über alle Originalstudien gleichmäßig verteilt. Die Wismut-Stichprobe zeigt keine besonderen Auffälligkeiten. Unterschiede zwischen Fällen und Kontrollen hinsichtlich der gesamt-genomischen Ethnizität konnten keine gefunden werden. Innerhalb der kaukasischen Originalstudien kann aber ein begrenztes Maß an Heterogenität nicht ausgeschlossen werden.

4.1.4 Deskriptives: Genotypen

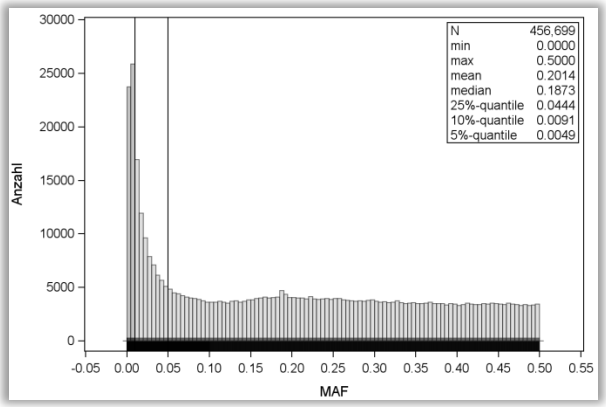
Von den 456.699 für die Analyse verwertbaren Markern sind nur 792 (0,1%) monomorph (siehe Tabelle 10). Etwa 10% (n=48.790) haben aber eine MAF<1%, sind also seltene Varianten. Die Teststärke einen Haupteffekt G und/oder eine Interaktion GxE statistisch nachzuweisen ist für diese Marker sehr gering. Dieser Anteil seltener Varianten schwankt nur gering zwischen den Autosomen (Chromosom 1-22: 8,8% bis 12,0%), ist jedoch am X-Chromosom mit 21% deutlich höher. Die gezielte Anreicherung des OncoArrays mit seltenen Varianten kann aus Abbildung 3 erkannt werden.

Tabelle 10 **Allelhäufigkeit unter genotypisierten Kontrollen (MAF: *minor allele frequency*)**

	Alle	MAF					
		monomorph		selten		häufig	
Chromosom	N	N	%	N	%	N	%
Alle	456.699	792	0,1%	48.790	10,6%	407.117	89,1%
1	32.473	54	0,1%	3.444	10,6%	28.975	89,2%
2	38.696	50	0,1%	3.887	10,0%	34.759	89,8%
3	29.334	37	0,1%	3.342	11,3%	25.955	88,4%
4	25.012	39	0,1%	2.665	10,6%	22.308	89,1%
5	28.909	57	0,1%	3.394	11,7%	25.458	88,0%
6	36,410	53	0,1%	3.864	10,6%	32.493	89,2%
7	*23.095*	*34*	*0,1%*	*2.263*	*9,7%*	*20.798*	*90,0%*
8	*24.513*	*37*	*0,1%*	*2.335*	*9,5%*	*22.141*	*90,3%*
9	*19.815*	*25*	*0,1%*	*1.794*	*9,0%*	*17.996*	*90,8%*
10	*23.299*	*25*	*0,1%*	*2.151*	*9,2%*	*21.123*	*90,6%*
11	*21.766*	*29*	*0,1%*	*2.465*	*11,3%*	*19.272*	*88,5%*
12	*23.353*	*29*	*0,1%*	*2.244*	*9,6%*	*21.080*	*90,2%*
13	12.696	35	0,2%	1.709	13,4%	10.952	86,2%
14	14.547	22	0,1%	1.479	10,1%	13.046	89,6%
15	13.293	11	<0,1%	1.607	12,0%	11.675	87,8%
16	13.494	21	0,1%	1.409	10,4%	12.064	89,4%
17	14.908	21	0,1%	1.635	10,9%	13.252	88,8%
18	12.641	20	0,1%	1.240	9,8%	11.381	90,0%
19	11,826	15	0,1%	1,253	10,5%	10,558	89.2%
20	12,706	17	0,1%	1,121	8,8%	11,568	91.0%
21	5,567	7	0,1%	501	8,9%	5,059	90.8%
22	8,271	10	0,1%	879	10,6%	7,382	89.2%
X	10,075	144	1,4%	2,109	20,9%	7,822	77.6%

monomorph: MAF=0%; selten: MAF<1% („rare variant“); häufig: MAF≥1% („polymorphism“)

Abbildung 3 **Allelhäufigkeit unter genotypisierten Kontrollen (MAF: minor allele frequency)**



4.2 Basismodell ohne Genotypen

4.2.1 Bestimmung der Gewichtung der Studienteilnehmer aus Nicht-Wismut-Studien

In der zur Verfügung stehenden Stichprobe der Wismut-Bergarbeiter (*BfS-Bioprobenbank* plus Fälle der *Indoor-Radon-Studie*) kann das rohe Odds Ratio (OR) für Lungenkrebs unter Strahlenexposition (>50 WLM, mittlere WLM=966) mit

Odds Ratio-Schätzer (95%-CI): 2,25 (1,16-4,38)

geschätzt werden (siehe Tabelle 11*)*. Aufgrund der besonderen Zusammensetzung der Fälle und Kontrollen ist dieses OR eine stichproben-interne Größe, die zwar die Fallwahrscheinlichkeit in der analysierten Stichprobe quantifiziert, jedoch nicht also unverzerrter Schätzwert des relativen Lungenkrebsrisikos in der Grundgesamtheit angesehen werden kann. Siehe dazu Kapitel 4.7.

Adjustiert für Alter und Rauchen steigt der Schätzwert auf OR=5,28 an. Dieser Wert ist vergleichbar mit dem in einer anderen Untersuchung geschätztem relativen Risiko RR von 5,74 (95%-CI: 8,85-8,58) für Uranbergarbeiter, bei einer Exposition ≥400 WLM und ebenfalls adjustiert für Rauchen.[35]

Würde man nun alle nicht-exponierten Fälle und Kontrollen der Nicht-Wismut-Studien hinzufügen, erhielte man ein geschätztes rohes OR von nur 0,17, da das Verhältnis der genotypisierten Fälle und Kontrollen nicht mit dem der Wismut-Studie korrespondiert. Daher ist es zwingend notwendig, die Fälle und Kontrollen der Nicht-Wismut-Studien zu gewichten, um die Schätzung des ORs der Strahlenexposition und deren Interaktionen nicht zu verzerren.

Eine geeignete Gewichtung ergibt sich dabei aus der Gleichheitsbedingung $OR_{Wismut} = OR_{alle}$ der geschätzten ORs:

$$OR_{Wismut} = \frac{143 \times 49}{259 \times 12} \quad \text{mit} \quad OR_{alle} = \frac{(143 + 13.121) \times 49}{259 \times (12 + 15.016 \cdot k)}$$

Daraus ergibt sich $$k = \frac{12}{143} \times \frac{13.121}{15.016} = 0{,}073326 = 1{:}\,13{,}6.$$

Definiert man das Verhältnis *Fälle : Kontrollen* als $\kappa_{Wismut} = 12{:}\,143$, bzw. $\kappa_{nicht-Wismut} = 15.016{:}\,13.121$, ergbit sich der Gewichtungsfaktor für nicht-exponierte Nicht-Wismut-Kontrollen ebenso als $k = \kappa_{Wismut}/\kappa_{nicht-Wismut}$ (siehe Tabelle 12, Tabelle 13 und Tabelle 14).

Wenn alle genotypisierten Fälle und Kontrollen einbezogen werden, beträgt das mit dieser Gewichtung geschätzte rohe OR schließlich

Odds Ratio-Schätzer (95%-CI): 2,25 (1,65- 3,07).

Als Alternative zur Gewichtung von nicht-exponierten Nicht-Wismut-Kontrollen kann auch (nur) eine Zufallsauswahl von $15016 \times k = 1101$ Nicht-Wismut-Kontrollen berücksichtigt werden. Das ist dann sinnvoll, wenn die verwendeten Programme eine Gewichtung einzelner Personen nicht zulassen (z.B. PLINK).

Tabelle 11 Odds Ratio für Lungenkrebs nach Strahlenexposition: nur Wismut-Bergarbeiter

Strahlenexposition (WLM>50)	*Lungenkrebs*		
	Kontrollen	*Fälle*	*gesamt*
nicht-exponiert	**143**	**12**	155
exponiert	**259**	**49**	308
gesamt	402	61	463

Odds Ratio-Schätzer (95%-CI): 2,25 (1,16-4,38)
Odds Ratio-Schätzer (95%-CI): 5,78 (2,37-14,1) adjustiert für Alter
Odds Ratio-Schätzer (95%-CI): 5,28 (2,05-13,6) adjustiert für Alter und Rauchen

Tabelle 12 Odds Ratio für Lungenkrebs nach Strahlenexposition: alle Studienteilnehmer (ungewichtet)

Strahlenexposition (WLM>50)	*Lungenkrebs*		
	Kontrollen	*Fälle*	*gesamt*
nicht-exponiert	**143**+13.121	**12**+15.016	
exponiert	**259**	**49**	308
gesamt	13.523	15.077	28.600

Odds Ratio-Schätzer (95%-CI) : 0,17 (0,12-0,23)

Tabelle 13 Odds Ratio für Lungenkrebs nach Strahlenexposition: alle Studienteilnehmer (gewichtet)

Strahlenexposition (WLM>50)	*Lungenkrebs*		
	Kontrollen	*Fälle*	*gesamt*
nicht-exponiert	**143**+13.121	**12**+15.016 x ***k*≈ 1.113**	14.377
exponiert	**259**	**49**	308
gesamt	13.523	1.162	14.685

Odds Ratio-Schätzer (95%-CI): 2,25 (1,65- 3,07).

Tabelle 14 Gewichtung für Nicht-Radonexponierte Nicht-Wismut-Bergarbeiter

Fälle:Kontrollen (Wismut)	*Fälle: Kontrollen (Nicht-Wismut)*	*Gewichtung der Nicht-Wismut-Fälle**	*OR (nur Wismut)*	*OR (Wismut+ gewichtet Nicht-Wismut)*
143: 12	13.121: 15.016	**0,073326**	2,2545	2,2545

* Lösung der Gleichheitsbedingung der Fälle: Kontrollen zwischen nur-Wismut- und einer kombinierten Stichprobe: 143:12=(143+13.121):(12+15.016k)

4.2.2 Auswahl des besten mehrerer alternativer Basismodelle

Da das Risiko an Lungenkrebs zu erkranken auch wesentlich von anderen Faktoren, wie Alter, Geschlecht und dem Rauchverhalten abhängt und sich die exponierten Fälle (Wismut-Bergarbeiter) durch diese Faktoren von allen anderen Personen unterscheiden, ist es notwendig im statistischen Modell für diese Faktoren bestmöglich zu adjustieren. Es wurde daher ein geeignetes Basismodell, d.h. ohne genetische Komponenten, gesucht, das alle relevanten Störfaktoren auf adäquate Weise berücksichtigt. Als Störfaktoren wurden in Betracht gezogen:

- Alter, Geschlecht, Raucherstatus
  (als einzig rauchrelevante Variable, für die suffizient Information von alle Personen vorliegt),
- das Design der Originalstudien
  (im Sinne des Rekrutierungsmechanismus; da nur die Studie des Moffitt Cancer Center im *case only*-Design durchgeführt wurde, wurde diese zu *hosp. CC* hinzugezählt)

- der Kontinent der Originalstudie
  (Russland und Israel wurden ob ihrer Bevölkerungsstrukturen gesondert berücksichtigt)

Der Einfluss des Raucherstatus bzw. des Alters wurde ggf. geschlechtsspezifisch modelliert. Die Wahl des die Daten am besten erklärenden Modells wurde anhand des AIC (Akaike's Informationskriterium) getroffen. Grundlage hierfür bildet die Wahrscheinlichkeit der der beobachteten Daten für eine Modell mit erfolgter Parameterschätzung nach dem Maximum-Likelihood-Prinzip, die sogenannte maximale „Likelihood L" (üblicherweise als -2ln(L) angegeben). Das AIC ist definiert als $AIC = -2L + 2k$, mit k der Anzahl der zu schätzende Parameter im Modell. Dies trägt also dem Umstand Rechnung, dass die Modelanpassung (L) steigt wenn Variablen einem bestehendem Modell hinzugefügt werden (also k ebenso steigt).

Ebenso wurde der Anteil konkordanter Personenpaare (es wird der tatsächliche mit dem durch das Modell geschätzten Erkrankungsstatus verglichen) und Somers' D (als Maß der Korrelation zweier ordinal-skalierter Variablen) bestimmt.

Die Anpassungsgüte der einzelnen Modelle an die beobachteten Daten ist in Tabelle 15 aufgelistet. **Modell Nr. 7 zeigt das kleinste AIC und ist daher mit 70% konkordanten Personenpaaren das Basismodell der Wahl.** Die Anpassungsgüte (AIC) des Modell 5 unterscheidet sich jedoch nur geringfügig, die des Modells 6 nur äußerst geringfügig von der des Modells 7, wenn auch ob der großen Fallzahl beide Male signifikant. Das adjustierte OR einer Strahlenexposition (>50 WLM) wird durch das Modell 7 mit 3,18 (95%-CI: 2,16-4,49) geschätzt und weicht nur geringfügig von dem rohen OR ab, das unter ausschließlich den Wismut-Bergarbeiter geschätzt wurde (siehe Tabelle 16 und Tabelle 13).[36] Das Adjustieren nach Design und Kontinent der Originalstudie erscheint essentiell, da sich der AIC zwischen Modell 4 und 5 sprunghaft verändert. Gleichzeitig gleicht sich das OR einer Strahlenexposition wieder den rohen OR unter nur Wismut-Bergarbeitern an.

Wird die dichotome Zielvariable Strahlenexposition (≤50 bzw. >50 WLM) durch die kontinuierliche Größe WLM ersetzt (eine lineare Assoziation zum logit(OR) wird postuliert), so verschlechtert sich die Modellpassung (Modelle Nr. 8-10) wieder geringfügig, wobei unter diesen dreien das Modell Nr. 10 als Pendant zu Modell 7 die beste Anpassungsgüte aufweist.

Es existieren hinreichende wissenschaftlichen Belege für eine lineare Assoziation zwischen dem WLM und dem „*Excess Relative Risk*" ERR=RR-1 (bzw. dem RR).[3,5,35,37,38] Dieses Modell wird üblich mit $RR = e^{(\beta_0+\beta_1 x)}(1 + \beta_2 WLM)$ (ERR-Modell) definiert.[5]

Deshalb wurden zusätzlich nicht-lineare, logistische Regressionsmodelle der Form $OR = e^{(\beta_0+\beta_1 x)}(1 + \beta_2 WLM)$ (ERR-Modell) an die Daten angepasst.[39] (Das von Richardson und Kaufmann 2009 publizierte SAS-Makro zur Bestimmung der logLikelihood-basierten 95%-Konfidenzintervalle wurde wie folgt adaptiert: Das logistische Sub-Modell wurde auf mehrere Kovariablen erweitert. Die aus dem OR abgeleitete Erkrankungswahrscheinlichkeit wurde auf den Bereich 10^{-100} bis $1\text{-}10^{-100}$ beschränkt, um die Berechnung der logLikelihood aus der Bernoulli-Verteilung zu gewährleisten. Die Berechnung der logLikelihood-basierten Konfidenzgrenzen wurde für alle Modellparameter automatisiert.)

Alle drei entsprechenden Modelle (Nr. 11-13) haben eine erkennbar schlechtere Anpassungsgüte gegenüber Modell Nr.7, wobei unter diesen dreien das Modell Nr. 12 als Pendant zu Modell Nr.6 die beste Anpassungsgüte aufweist. Die Steigerung des ORs wird im Modell Nr. 13 mit 0,000129 pro ein WLM, in den Modellen Nr. 11 und 12 sogar als negativ geschätzt. In keinem Modell ist das OR_{WLM} (pro ein WLM) signifikant. Der Schätzwert des Modells Nr. 13 fällt auch im Vergleich zu publizierten Schätzung um bis zu einer 10er-Potenz niedrig aus (ERR_{WLM}=0,0021 [40]; =0,008 [35]; =0,013 [37]; =0,0012 [3,37], =0,0021 [4], =0,0053 [5,38]). Es kann vermutet werden, dass sowohl die Annahme: „WLM=0 (keinerlei Strahlenbelastung)" für Nicht-Wismut-Bergarbeiter als auch die Zusam-

mensetzung des Studienkollektive der Wismut-Bergarbeiter (DNA zur Genotypisierung musste vorhanden sein; Auswahl der Fälle aus der *Indoor-Radon Studie* sowie der Kontrollen aus der *BfS-Bioprobenbank*) zu diesen deutlichen Verzerrung der Schätzwerte für ERR/WLM führen. Damit disqualifiziert sich das von andere favorisierte ERR-Modell zusätzlich zur geringeren Anpassungsgüte für die hier berichtete Datenauswertung.

Tabelle 15 Anpassungsgüte verschiedener Basismodelle

					Log-Likelihood-Test																
Modell	***AIC***	***Δ_{AIC}***	***%cp***	***Somers' D***	***-2 ln(L)***	***df***	***p-Wert***	*Expo*	*WLM*	*PC1*	*PC2*	*PC3*	*PC4*	*Alter*	*Geschlecht*	*Rauchverhalten*	*Design*	*Kontinent*	*Rauchen(Geschl.)*	*Alter(Geschl.)*
1	8107,1	531	0,3%	-0,02	8103,2	1	[a]3x10^{-101}	x												
2	8112,8	537	20%	0,00	8100,9	5	[a]1x10^{-105}	X		X	X	X	X							
3	8097,4	522	51%	0,07	8081,4	7	[a]5x10^{-104}	X		X	X	X	X	X	X					
4	7709,5	134	66%	0,34	7687,5	10	[a]2x10^{-24}	X		X	X	X	X	X	X	X				
5	7598,9	23	70%	0,41	7448,9	17	[a]1x10^{-4}	X		X	X	X	X	X	X	X	X	X		
6	7579,6	4	70%	0,41	7424,5	20	[a]0,0483	X		X	X	X	X	X	X	X	X	X	X	
7	7575,7	0	**70%**	**0,41**	**7419,2**	**21**	- -	**X**		**X**	**X**	**X**	**X**			**X**	**X**	**X**	**X**	**X**
8 (÷5)	7624,3	49	70%	0,41	7592,3	15	[b]0,0003		X	X	X	X	X	X	X	X	X	X		
9 (÷6)	7604,5	29	70%	0,41	7566,5	18	[b]0,2943		X	X	X	X	X	X	X	X	X	X	X	
10 (÷7)	7603,4	28	70%	0,41	7563,4	19	- -		X	X	X	X	X			X	X	X	X	X
11 (÷5)	7975,7	400	67%	0,35	7943,7	17	[c]2x10^{-5}		X	X	X	X	X	X	X	X	X	X		
12 (÷6)	7958,0	382	67%	0,35	7920,0	20	[c]3x10^{-11}		X	X	X	X	X	X	X	X	X	X	X	
13 (÷7)	8002,5	427	64%	0,32	7964,5	21	- -		X	X	X	X	X			X	X	X	X	X

X = Einflussvariable im Modell

Δ_{AIC}: Differenz von AIC to AIC_{min} (Modell 7); %cp = % konkordante Paare,

[a] im Vergleich zu Modell 7, [b] im Vergleich zu Modell 10 (÷7), [c] im Vergleich zu Modell 13 (÷7); ÷ identische Kovariablen wie Modell #;

Modell 1-12: logistische Regression; Modell 11-13.: $OR = e^{(\beta_0+\beta_1 x)}(1+\beta_3 WLM)$

Rauchen(Geschl.): das Rauchverhalten wurde geschlechtsspezifisch modelliert; Alter(Geschl.): das Alter wurde geschlechtsspezifisch modelliert;

Das Basismodell der Wahl wurde durch Fettdruck hervorgehoben.

Tabelle 16 Odds Ratios verschiedener Basismodelle für Lungenkrebs nach Strahlenexposition

Variable	***Modell***	***OR***	***(95%-CI)***	***p-Wert***	
Strahlenexposition exponiert (WLM>50) vs. nicht-exponiert	1	2,25	(1,65- 3,07)	3,1x10^{-7}	
	2	2,20	(1,60- 3,00)	7,9x10^{-7}	
	3	1,79	(1,29- 2,49)	4,0x10^{-4}	
	4	1,38	(0,98- 1,93)	0,061	
	5	2,91	(2,00- 4,25)	2,4x10^{-8}	
	6	2,90	(1,99- 4,24)	3,1x10^{-8}	
	7	**3,18**	**(2,16- 4,69)**	**3,7x10^{-9}**	
pro 1 WLM	8 (÷5)	0,000304	(-0,000064- 0,000672)	0,1071	s.e. 1,9x10^{-4}
	9 (÷6)	0,000310	(-0,000060- 0,000680)	0,1005	s.e. 1,9x10^{-4}
	10 (÷7)	0,000365	(-0,000094- 0,000739)	0,0561	s.e. 1,9x10^{-4}
ERR pro 1 WLM	11 (÷5)	-0,000120	(-0,000294- 0,000043)	0,4364	s.e. 1,5x10^{-4}
	12 (÷6)	-0,000120	(-0,000282- 0,000037)	0,4411	s.e. 1,5x10^{-4}
	13 (÷7)	0,000129	(-0,000097- 0,000355)	0,5673	s.e. 2,6x10^{-4}

Modell 1-12: logistische Regression; Modell 11-13: $OR = e^{(\beta_0+\beta_1 x)}(1+\beta_3 WLM)$;

ERR Excess Relative Risk; s.e. Standardfehler (standard error)

Tabelle 17 Odds Ratios und p-Werte verschiedener Basismodelle

*Strahlenexposition (WLM>50)**			*OR*							*p-Wert*						
			1	*2*	*3*	*4*	*5*	*6*	*7*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
			2,25	2,20	1,79	1,38	2,91	2,90	3,18	$3{,}1\times10^{-7}$	$7{,}9\times10^{-7}$	$4{,}0\times10^{-4}$	0,061	$2{,}4\times10^{-8}$	$3{,}1\times10^{-8}$	$3{,}7\times10^{-9}$
*Alter bei Diagnose/ Interview***	*unter Frauen*								1,02							$6{,}0\times10^{-8}$
	unter Männern				1,01	1,01	1,01	1,01	1,01			$7{,}2\times10^{-4}$	$4{,}8\times10^{-9}$	$3{,}6\times10^{-8}$	$2{,}2\times10^{-8}$	0,00275
Studiendesign (… vs. nested CC)	*hosp, cc+*						2,40	2,42	2,45					$<1\times10^{-25}$	$<1\times10^{-25}$	$<1\times10^{-25}$
	pop, cc						1,35	1,35	1,34					0,020	0,022	0,024
Kontinent des Studien-orts (vs. Europa)	*Amerika*						1,65	1,66	1,65					$2{,}9\times10^{-10}$	$2{,}4\times10^{-10}$	$4{,}0\times10^{-10}$
	Israel						1,90	1,91	1,92					$5{,}2\times10^{-4}$	$4{,}7\times10^{-4}$	$4{,}1\times10^{-4}$
	Russland						0,82	0,82	0,80					0,177	0,174	0,144
Rauchverhalten (… vs. Nie-Raucher)	*Raucher*	*unter Frauen*						7,58	7,89				$<1\times10^{-25}$	$<1\times10^{-25}$	$<1\times10^{-25}$	$<1\times10^{-25}$
		unter Männern				6,40	8,33	10,5	10,4						$<1\times10^{-25}$	$<1\times10^{-25}$
	Jemals-Raucher	*unter Frauen*						1,64	1,63						0,082	0,084
		unter Männern				5,49	4,52	7,93	7,82				$<1\times10^{-25}$	$<1\times10^{-25}$	$<1\times10^{-25}$	$<1\times10^{-25}$
	Ex-Raucher	*unter Frauen*						3,25	3,25						$2{,}6\times10^{-14}$	$3{,}0\times10^{-14}$
		unter Männern				3,34	3,96	5,25	5,33				$<1\times10^{-25}$	$<1\times10^{-25}$	$<1\times10^{-25}$	$<1\times10^{-25}$
Geschlecht (weiblich vs. männlich)					0,84	1,10	1,12	1,10	--			0,00863	0,134	0,089	0,133	--

* exponiert vs. nicht-exponiert; ** pro Jahr; hosp. cc+ = Krankenhausbasierte Fall-Kontroll-Studie (hospital based case-control study) + Case-only Studie; pop. cc = populationsbasierte Fall-Kontroll-Studie (population based case-control study)

4.2.3 Propensity-Score zur Zusammenfassung der nicht-genomischen Störgrößen

In der Datenauswertung von Beobachtungsstudien sind Störgrößen besonders zu beachten, da deren Einfluss auf die Zielgröße, anders als in einem Experiment, nicht durch Randomisierung kontrolliert werden kann. Das Fehlen der Randomisierung in Beobachtungsstudien kann dazu führen, dass sich Fälle und Kontrollen nicht nur durch die zu untersuchende Exposition unterscheiden, sondern durch eine Vielzahl anderer (messbarer oder unbekannter) gegebenenfalls miteinander interagierender Störgrößen, die die Wahrscheinlichkeit zu erkranken jedes Studienteilnehmers mit bestimmen. Wenn diese Störfaktoren wiederum mit der zu untersuchenden Zielgröße korrelieren (Confounding) können in den Studiendaten Scheineffekte beobachtet werden. Die *Propensity-Score (PS) Methode* ist eine statistische Technik mit deren Hilfe versucht werden kann, diese Art der Verzerrung durch mehrere interagierenden Störgrößen zu reduzieren, im besten Fall gänzlich zu eliminieren.[41,42] Mit dem PS sollen alle Fälle und Kontrollen hinsichtlich der Störgrößen „ausbalanciert" werden. Die Ergebnisse von Beobachtungsstudien, die mit Hilfe der PS-Methode ausgewertet wurden, sind in der Regel – bei klinischen Studien – den Ergebnissen von randomisierten Experimenten vergleichbar. Im Kontext solcher Studien konnte kein systematisches Über- oder Unterschätzen von „Behandlungseffekten" belegt werden.[43]

Der PS wird in der Fachliteratur definiert als die Wahrscheinlichkeit eines Studienteilnehmers als Fall für eine Beobachtungsstudie rekrutiert geworden zu sein, bedingt einer Anzahl dafür relevanter Störfaktoren $\boldsymbol{X}$ (Kovariablen) $PS_{prob} \stackrel{\text{def}}{=} P(case|\boldsymbol{X})$.

Vorteile der Propensity-Score-Methode sind:

- Mit Hilfe des PS können Fälle und Kontrollen hinsichtlich deren Hintergrund-Fallwahrscheinlichkeit (hinsichtlich gemessener Störgrößen) in Balance gebracht werden.
- Wenn die gemessenen Störgrößen ausreichend Information beinhalten, können Kontrollen mit nur geringer Hintergrund-Fallwahrscheinlichkeit, bzw. Fälle mit sehr hoher Hintergrund-Fallwahrscheinlichkeit anhand des PS identifiziert und gegebenenfalls ausgeschlossen werden („*subject pruning*"). Dies kann transparenter dargestellt werden.[42]
- Komplexe Verflechtungen der Störgrößen können best-möglichst modelliert werden (z.B. nested effects, CART-Analysen, etc.). Für die Analyse der Zielgröße selbst kann dann ein einfacheres, für die Fragestellung besser angepasstes Modell verwendet werden.
- Auch wenn eine Vielzahl an Störgrößen zur Bestimmung des PS herangezogen wird, bindet der PS im finalen Analysemodell der Zielgröße nur einen Freiheitsgrad.
- In GWAS wird der PS nur einmal bestimmt, aber in der Modellierung jedes einzelnen Markers angewandt. Die dem PS zugrundeliegende innere Struktur der Störgrößen untereinander wird nicht verändert.

Als Nachteile der Propensity-Score-Methode gelten:

- Der *PS* kann nur auf Basis gemessener (bekannter) Störgrößen geschätzt werden. Verzerrungen der Studienergebnisse durch „*unmeasured confounding*" kann nicht vermieden werden.[41-43]
- Besteht nur ein geringer Überlapp zwischen den Verteilungen des PS in Fällen und Kontrollen, ist das Risiko verzerrter Studienergebnisse hoch.[42,44]
- Die Schätzung des PS muss auf hinreichend suffizienten Daten (ausreichend großen Stichproben) basieren, da die Unsicherheit des PS nicht weiter berücksichtigt wird.
- Es konnte kein Vorteil im Sinne von Power oder Signifikanz des Adjustierens anhand des PS gegenüber dem Adjustieren direkt anhand der Störgrößen gezeigt werden.
- Die PS-Methode kennt mehrere Optionen, Adjustieren, Matchen, Stratifizieren oder Gewichten, um den PS im finalen Modell zu berücksichtigen. Die Studienergebnisse können

stark von der angewandten Option abhängen. Die Wahl der Option muss daher vor der Analyse getroffen werden.

In der Anwendung der PS-Methode müssen gemäß Arbogast and Ray, 2011 [42] drei wesentliche Schritte eingehalten werden:

1. „Building the Propensity Score“:
   Die sorgfältige Wahl eines Modells zur Berechnung des PS
2. „Propensity Score Use - Restriction“:
   Der Vergleich der Verteilungen des PS im Fällen und Kontrollen, ggf. der Ausschluss von Studienteilnehmer
3. „Propensity-Score Use in the Analysis“
   Die sorgfältige Wahl der Art und Weise, wie der PS im finalen Analysemodell berücksichtigt werden soll, d.h. Adjustieren, Matchen, Stratifizieren oder Gewichten

4.2.3.1 Wahl eines Modells zur Berechnung des Propensity-Scores

Der PS wurde für die weitere Analyse als der lineare Prädiktor $\beta \boldsymbol{x}$ aus dem Modell Nr. 7 (logistisches Modell $\ln\left(\frac{p_{case}}{p_{control}}\right) = \beta \boldsymbol{x}$, dem Basismodell der Wahl, bestimmt. Der vom Modell geschätzte PS wurde um die Effekte der Strahlenexposition (≤50 bzw. >50 WLM), sowie der vier PCs für genetische Stratifikation bereinigt. Damit vereinigt der PS folgende Störgrößen: Alter, Geschlecht, Rauchverhalten, Studiendesign und Kontinent.

Aus dem PS kann die Hintergrund-Fallwahrscheinlichkeit $PS_{prob} = \frac{e^{PS}}{1+e^{PS}}$ abgeleitet werden.

4.2.3.2 Vergleich der Verteilungen des Propensity-Scores in Fällen und Kontrollen

Der PS reicht unter allen Studienteilnehmern von -5,40 bis -0,28. Das entspricht einer Hintergrund-Fallwahrscheinlichkeit von 0,5% bis 43%. Der mediane PS ist unter exponierten Fällen mit -2,85 ($\widehat{=}$ PS_{prob}=5,5%) am niedrigsten, unter nicht-exponierten Fällen mit -2,09 ($\widehat{=}$ PS_{prob} =11%) am höchsten (siehe Tabelle 18). Im Allgemeinen ist der PS unter Kontrollen erwartungsgemäß geringer als unter Fällen. Die Wismut–Bergarbeiter (÷Exponierte) zeigen jedoch keine systematische höheren oder niedrigeren PS (siehe Abbildung 4).

Nur 18 der Fälle weisen einen PS auf, der größer ist als der Maximalwert unter den Kontrollen. Diese Fälle hätten damit per se eine höhere Chance zu erkranken als jede andere Kontrollperson. Ebenso weisen 3 Kontrollen einen PS kleiner als das Minimum unter den Fällen auf. Diese Kontrollen hätten also per se eine kleinere Chance zu erkranken als jeder andere Fall. 10 der Fälle stammen aus der MSH-PMH Studie, 7 aus der HSPH- Studie. Die Restlichen stammen aus je einer anderen Studie.

Da der Überlapp der Verteilungen des PS in Fällen und Kontrollen über 99,9% (28.578/28.599) beträgt und keiner der 21 (18+3) auffälligen Studienpersonen gravierend hervorsticht, muss kein Studienteilnehmer von der Datenanalyse ausgeschlossen werden.

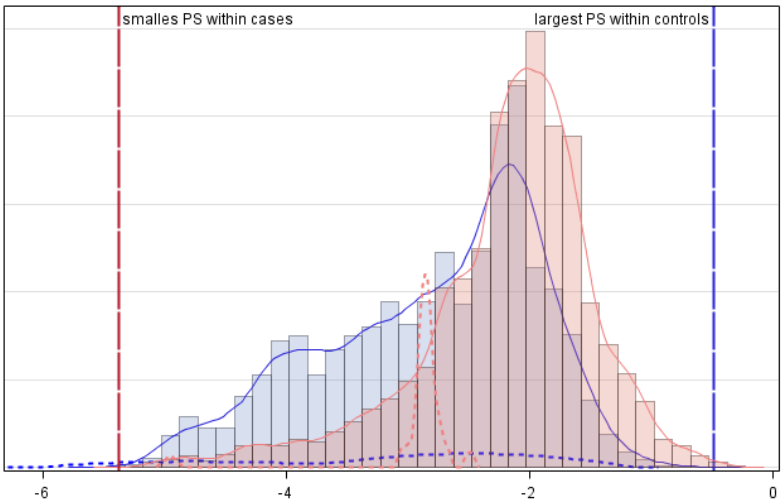
Tabelle 18 Verteilung des Propensity-Scores zwischen Fällen und Kontrollen

Propensity Score	gesamt	Kontrollen		Fälle	
		nicht-exponiert	exponiert	nicht-exponiert	exponiert

Propensity Score	gesamt	Kontrollen				Fälle	
		nicht-exponiert		exponiert		nicht-exponiert	exponiert
	N=28.599	N=13.263		N=259		N=15.028	N=49
Minimum	-5,40	-5,40		-4,96		-5,38	-4,99
Mittelwert	-2,49	-2,81		-3,16		-2,19	-2,92
Median	-2,26	-2,64		-2,46		-2,09	-2,85
Maximum	-0,28	-0,49		-2,27		-0,28	-2,49
Propensity-Score (gerundet)							
-5	822	581	4%	77	30%	162 1%	2 4%
-4	3,295	2.572	19%	--	--	723 5%	-- --
-3	7,087	4.038	30%	24	9%	2.980 20%	45 92%
-2	15,286	5.738	43%	158	61%	9.388 62%	2 4%
-1	2,090	333	3%	--	--	1.757 12%	-- --
-0	19	1	<1%	--	--	18 <1%	-- --
gesamt	28,599	13.263	100	259	100	15.028 100	49 100
		GLC	**OncoA.**	**Wismut**		**GLC**	**OncoA.** **Wismut**
PS-größer als unter Kontrollen	18	--	--	--		--	18 --
PS-Überlapp	28.578	478	12.640	401		471	14.527 61
PS-kleiner als unter Fällen	3	--	2	1		--	-- --

OncoA.. … OncoArray-Konsortium (TRICL/ILCCO)

Abbildung 4 Verteilung des Propensity-Scores zwischen Fällen und Kontrollen



Die Abbildung zeigt Histogramme und Kernel-Dichten des Propensity-Scores;
in Blau = Kontrollen; in Rot = Fälle;
gepunktete Linien = Dichte der strahlenexponierten Fälle oder Kontrollen (nur Wismut-Bergarbeiter)

4.2.3.3 Optionen für die Berücksichtigung des Propensity-Scores im finalen Analysemodell

4.2.3.3.1 Paarweises Zuordnen (Matchen)

Paarweises Zuordnen (Matching) anhand des PS besteht in der Auswahl einer (1:1 Matching) oder mehrerer (m:1 Matching) Kontrollen je Fall, mit identischem oder vergleichbaren PS. „*Matching on the propensity score as a single variable has the effect of matching on all of the components of the propensity score, without the drawback of matching on numerous individual variables, which leads to greater and greater difficulty in finding appropriate matches due to the expansion in the number of potential matching categories*".[42] Matching führt zu einem direkten Vergleich von Fällen und Kontrollen, ohne den Effekt von Störgrößen schätzen zu müssen. Matching kann zu einer höheren Effizienz der Datenanalyse führen, d.h. zu höherer Power statischer Tests und kürzeren Konfidenzintervallen.[45]

In einer GWAS mit Zielerkrankung "depressive Störung" wurde 1:1 Matching innerhalb von „*ancestry-informative*" Strata (definiert über PCs zum Adjustieren für Populations-Stratifikation) um unter anderem „stressful live events" (SLE) adäquat zu berücksichtigen. Die Analyse zeigte, dass wenn die zugrundeliegende Heritabilität der Zielerkrankung (Depression) mit der Heritabilität einer wesentlichen Komponenten des PSs (hier SLE) überlappt, ungünstiger Weise eine genetische Ähnlichkeit („*genetic similarity*") von Fällen und Kontrollen erzeugt wird. Matching führt in einem solchen Fall zu Powerverlusten in Folge von Überkorrektur.[46]

Matching führt ferner zu dem Problem, dass bei geringem Überlapp der PS-Verteilungen für einige Fälle, respektive Kontrollen, kein Matching-Partner gefunden werden kann (implizites *„subject pruning")* [42,46]. Bei sehr geringem Überlapp kann der Powervorteil des Matching verloren gehen. Es gibt auch Überlegungen, nach denen bei kombiniertem Matching und *subject pruning* die PS-Methode zu verzerrten Ergebnissen führen kann (PSM-Paradox).[44]

Da 45 der 49 exponierten Fälle (95%) einen PS in dem engen Bereich zwischen -3,5 und -2,5 ($\div PS_{prob}$ =3%-8%) aufweisen, dies aber nur für 24 der 249 exponierten Kontrollen (9%) der Fall ist, ist ein suffizientes, alle Fälle und Kontrollen einbeziehendes Matching ohne Gefahr des PSM-Paradoxes kaum möglich.

4.2.3.3.2 Stratifizieren

Anhand des PS können Strata (homogene Schichten im Sinne von ähnlichen Personen) gebildet werden. Die Schätzung des Einflusses der Zielgröße im finalen Modell erfolgt dann bedingt auf die Strata-Zugehörigkeit der Studienteilnehmer. Wie für das Matching gilt, dass die meisten exponierten Fälle in einem Stratum (PS zwischen -3,5 und -2,5) liegen. Somit würde Stratifizieren faktisch dem Ausschluss aller anderen Strata gleichkommen (*subject pruning).* Die Stichprobe würde unnötig verkleinert werden. Powerverlust wäre die Folge.

4.2.3.3.3 Gewichten

Es kann versucht werden, durch Gewichten die Verteilung des PS der Kontrollen und der Fälle anzugleichen. Dabei bestimmt sich das Gewicht nach der „*inverse probability of treatment weight*" (IPTW). Entscheidend ist hier vor allem die Frage nach der Referenz-Verteilung: die der Fälle, die der Kontrollen oder die aller Studienteilnehmer. Diese Gewichte sind vergleichbar mit „*sampling weights*", die dazu dienen die Repräsentativität einer Stichprobe herzustellen. Zur Bestimmung der Gewichte können folgende Funktionen verwendet werden:

Referenz Fälle und Kontrollen[41]:	$\omega_i = \frac{Z_i}{e_i} + \frac{1-Z_i}{1-e_i}$	
Referenz Fälle[47]:	Fälle: $\omega_i = 1$	Kontrollen: $\omega_i = \frac{e_i}{1-e_i}$
Referenz Kontrollen[47]:	Fälle: $\omega_i = \frac{1-e_i}{e_i}$	Kontrollen: $\omega_i = 1$

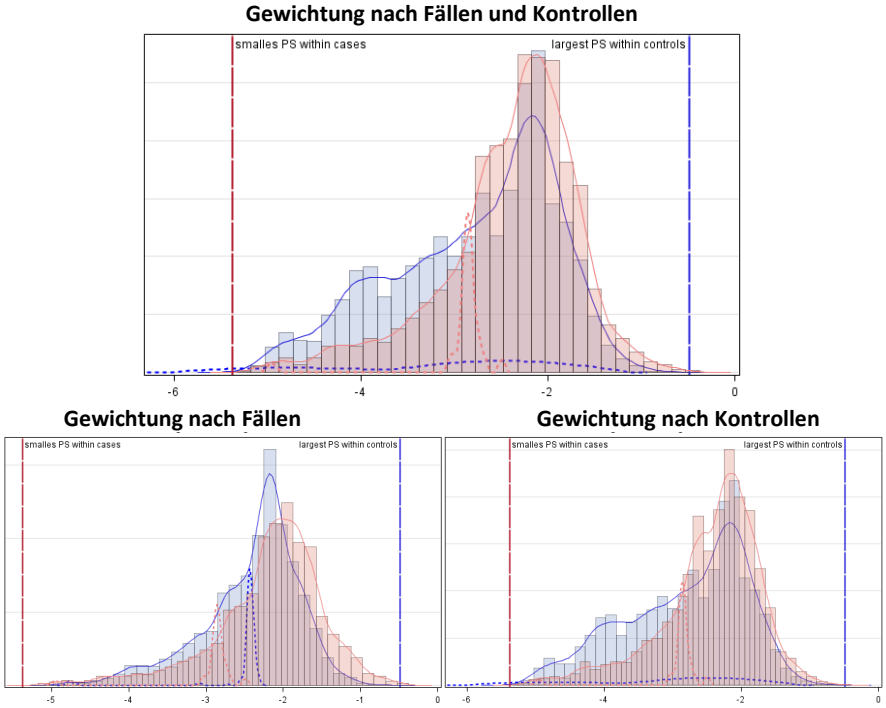
Mit $Z_i = 1$ für Fälle und $Z_i = 0$ für Kontrollen, sowie der Fallwahrscheinlichkeit $e_i = \frac{e^{PS}}{1+e^{Ps}}$.

Unter den beteiligten Originalstudien ist die Rekrutierung der Kontrollen uneinheitlich. Die in diese Analyse eingehenden Wismut-Bergarbeiter stammen aus der BfS-Bioprobenbank (Wismut-Kontrollen) und der Indoor-Radon-Studie (Wismut-Fälle). Folglich würde die gesamte Kohorte (hier Fälle und Kontrolle) als Referenz am sinnvollsten sein. Bei den Fall-Kontrollstudien wiederum können die Fälle als repräsentativ für Erkrankte gelten, die Kontrollen wurden in der Regel nach Alter und Geschlecht gematched rekrutiert. Folglich würden die Fälle als Referenz am sinnvollsten sein. Durch eine Simulationsuntersuchung von Mansson, et al., 2007 [48] bezüglich möglicher Verzerrungen, die durch die PS-Methode in Fall-Kontroll-Studien erst erzeugt werden, konnte gezeigt werden, dass *„[l]ittle to no effect modification by propensity score was induced by estimating the pro-*

pensity score using the unweighted case-control or the modeled control methods. We expect that this will only be true … if the model for exposure probability used does not include interactions between the covariates and case-control status".

Kontrollen mit einem niedrigem PS würden ein Gewicht von weit über dem 15-fachen erhalten (bei Referenz Fall und Kontrollen, als auch bei Referenz Fälle) (siehe Abbildung 5). Auf der andern Seite würden Fälle mit einem sehr hohem PS ein Gewicht von bis zum 8-fachen erhalten (Referenz Kontrolle). Damit würde diesen einzelnen, gemäß PS extremen Beobachtungen ein sehr hoher Einfluss auf die Modellschätzung gewährt. Ungeachtet welche Gewichtung vorgenommen wird verbleibt ein höherer Anteil mit niedrigerem PS unter Kontrollen als unter Fällen. Ferner kann Gewichten nur als PS-Option angewandt werden, wenn die verwendeten Programme der GWAS-Analyse eine Gewichtung zulassen.

Abbildung 5 Verteilung des gewichten Propensity-Scores zwischen Fällen und Kontrollen

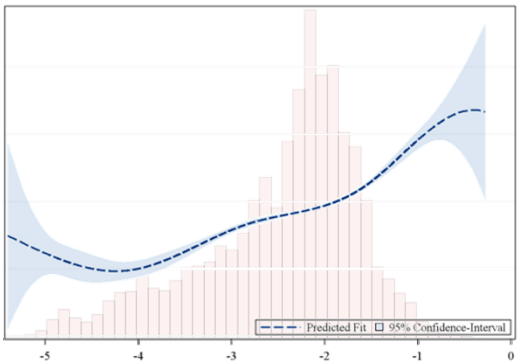


Die Abbildung zeigt Histogramme und Kernel-Dichten des Propensity-Scores;
in Blau = Kontrollen; in Rot = Fälle;
gepunktete Linien = Dichte der strahlenexponierten Fälle oder Kontrollen (nur Wismut-Bergarbeiter)

4.2.3.3.4 Adjustieren

Schließlich verbleibt als PS-Option das Einschließen des PS als Kovariable im finalen Analysemodell (Adjustieren).[41] Dies ist die am wenigsten problematische Option und kommt einer üblichen multivariablen-adjustierten Modellschätzung am nächsten. In der Simulationsstudie von Månsson et al. konnte auch gezeigt werden, dass eine als Artefakt durch die PS-Methode hervorgerufene Effektmodifikation nicht durch Adjustieren mit dem PS zu erwarten ist.[48] Wie in Abbildung 6 erkennbar, steht der PS im Allgemeinen in einer linearen Korrelation zum Logit der Fallwahrscheinlichkeit. Lediglich in den sehr seltenen Fällen besonders niedrigem oder besonders hohem PS wird die Linearitäts-Annahme verletzt. Eine weitere Transformation des PS für die hier beschriebene Datenanalyse ist daher nicht notwendig.[42]

Abbildung 6 Spline-Linearität des Propensity-Scores in einem logistischen Regressionsmodell



Die Abbildung zeigt den linearen Prädiktor $\hat{y} = \ln(p_{case}/(1 - p_{case}))$ (y-Achse) einer nicht-parametrischen logistischen Regression einer Spline-Funktion des Propensity-Scores (x-Achse) als erklärende Variable vor dem Hintergrund des Histogramms des Propensity-Scores aller Fälle und Kontrollen.

4.2.3.4 Fazit

Die Störgrößen Alter, Geschlecht, Rauchverhalten, Studiendesign und Kontinent können adäquat in einem PS abgebildet werden. Kein Studienteilnehmer muss auf Grund eines markant hohen oder markant niedrigen PS ausgeschlossen werden, da sich die Verteilungen des PS der Kontrollen und der Fälle fast vollständig überschneiden.

Matchen und Stratifizieren anhand des PS sind als Analyseoption ungeeignet, da dies zu impliziten „*subject pruning*“ führen würde. Gewichten anhand des PS ist als Analyseoption ungeeignet, da einzelne, extreme Beobachtungen die Analyseergebnisse sehr stark beeinflussen würden. Daher wird der PS als Kovariable in das finale Analysemodell aufgenommen.

4.2.4 Basismodell mit Propensity-Score

Schlussendlich wurde überprüft, inwieweit sich die Güte der Modellanpassung des gewählten Basismodells (Modell Nr. 7) ändert, wenn die einbezogenen Kovariablen durch den PS ersetzt werden. Nicht alle der verwendeten Programme lassen eine Gewichtung einzelner Personen zu. Alternativ zur Gewichtung von nicht-exponierten Nicht-Wismut-Kontrollen kann auch nur eine Zufallsauswahl von $15.016 \times k = 1.101$ Nicht-Wismut-Kontrollen berücksichtigt werden. Deshalb wurde ebenso überprüft, inwieweit sich die Güte der Modellanpassung des gewählten Basismodells unter Einbeziehung nur einer Zufallsauswahl an nicht-exponierten Nicht-Wismut-Kontrollen ändert. Zehn solche Zufallsauswahlen wurden getroffen.

Das Basismodell mit PS zeigt im Vergleich zu einem multiplen Basismodell eine höhere Modellanpassung, bemessen am AIC (kleinere AIC → höher Anpassungsgüte). Dies ist auf die geringere Anzahl zu schätzender Parameter zurück zu führen, da die Likelihood der Daten im Basismodell mit PS höher ist als im multiplen Basismodell (kleinere -2lnL → höher Anpassungsgüte) (siehe Tabelle 19 – Spalte -2lnL). Die Punktschätzung des ORs bezüglich der Strahlungsexposition ist in beiden Modellen mit 3,18 identisch; das 95%-Konfidenzintervall im Modell mit PS ist kürzer, die Signifikanz höher ($p=3{,}7\times10^{-9}$ gegenüber $p=2{,}1\times10^{-12}$).

Fazit: Das Modellieren mit PS ist gegenüber einem multiplen Modell vorzuziehen. Verzerrende Effekte sind nicht zu erwarten.

Das Ersetzen der notwendigen Gewichtung von nicht-exponierten Nicht-Wismut-Kontrollen durch eine entsprechende Zufallsauswahl führt zu einer geringfügigen Streuung der Punktschätzung der OR bezüglich der Strahlungsexposition (OR zwischen 3,07 und 3,29) und der Signifikanz (p-Werte zwischen $6{,}3\times10^{-13}$ und $1{,}5\times10^{-12}$) (siehe Tabelle 20). Der Vorteil gegenüber dem multiplen Modell bleibt hingegen bestehen. Das OR pro Einheit PS wird über alle Modell mit Zufallsauswahl nahezu

unverändert dem gewichteten Modell mit PS geschätzt. Die Güte der Modellanpassung in den Modellen mit Zufallsauswahl ist bemessen am AIC vernachlässigbar geringer.

Fazit: Die Alternativen „gewichtetes Modell mit PS“ und „ungewichtetes Modell mit Zufallsauswahl und PS“ sind in der Schätzung des stichproben-internen ORs der Strahlenexposition und somit der stichproben-internen Hintergrund-Fallwahrscheinlichkeit ebenbürtig.

Tabelle 19 Modellanpassung des Basismodells mit Propensity-Score

Modell	AIC	Δ_{AIC}	-2lnL	% cp	Somers D	Fälle	Kontrollen	gesamt
Bestes multiples Modell Nr.7 / gewichtet	7.575,7	24	**7.419,2**	70%	0,41	15.077	13.522	28.599
Propensity-Score / gewichtet	**7.549,6**	0	7.535,6	70%	0,41	15.077	13.522	28.599
Propensity-Score / Zufallsauswahl 1	7.545,5	4	7.531,5	71%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 2	7.548,5	1	7.534,5	70%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 3	7.540,5	9	7.526,5	71%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 4	7.552,4	3	7.538,4	70%	0,41	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 5	7.544,9	5	7.530,9	70%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 6	7.548,1	1	7.534,1	70%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 7	7.552,3	3	7.538,3	70%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 8	7.545,4	4	7.531,4	70%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 9	7.546,3	3	7.532,3	71%	0,42	1.162	13.522	14.684
Propensity-Score / Zufallsauswahl 10	7.549,0	1	7.535,0	70%	0,42	1.162	13.522	14.684

Δ_{AIC}: Differenz zum AIC des Modells “Propensity-Score / gewichtet”; %cp … % konkordante Paare

Tabelle 20 Schätzung des Odds Ratios für Strahlenexposition im Modell mit Propensity-Score

Modell	OR (95%-CI)	p-Wert
Strahlenexposition		
Bestes multiples Modell Nr.7 / gewichtet	3,18 (2,16- 4,69)	**$3{,}7\times10^{-9}$**
Propensity-Score / gewichtet	3,18 (2,30- 4,40)	$2{,}1\times10^{-12}$
Propensity-Score / Zufallsauswahl 1	3,19 (2,31- 4,42)	$1{,}8\times10^{-12}$
Propensity-Score / Zufallsauswahl 2	3,21 (2,32- 4,43)	$1{,}5\times10^{-12}$
Propensity-Score / Zufallsauswahl 3	3,07 (2,22- 4,24)	$9{,}5\times10^{-12}$
Propensity-Score / Zufallsauswahl 4	3,29 (2,38- 4,54)	$5{,}3\times10^{-13}$
Propensity-Score / Zufallsauswahl 5	3,15 (2,28- 4,35)	$3{,}4\times10^{-12}$
Propensity-Score / Zufallsauswahl 6	3,16 (2,29- 4,37)	$2{,}7\times10^{-12}$
Propensity-Score / Zufallsauswahl 7	3,27 (2,37- 4,53)	$6{,}3\times10^{-13}$
Propensity-Score / Zufallsauswahl 8	3,18 (2,30- 4,39)	$2{,}2\times10^{-12}$
Propensity-Score / Zufallsauswahl 9	3,17 (2,29- 4,38)	$2{,}5\times10^{-12}$
Propensity-Score / Zufallsauswahl 10	3,14 (2,27- 4,34)	$3{,}9\times10^{-12}$
Propensity-Score		
Propensity-Score / gewichtet	2,71 (2,47- 2,98)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 1	2,71 (2,47- 2,98)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 2	2,71 (2,47- 2,98)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 3	2,70 (2,46- 2,97)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 4	2,71 (2,47- 2,98)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 5	2,71 (2,46- 2,97)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 6	2,71 (2,46- 2,97)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 7	2,71 (2,46- 2,97)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 8	2,72 (2,47- 2,99)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 9	2,71 (2,47- 2,98)	$<1\times10^{-25}$
Propensity-Score / Zufallsauswahl 10	2,70 (2,46- 2,97)	$<1\times10^{-25}$

4.3 Genomische Blockstruktur

Das menschliche Genom besteht aus einer Vielzahl an Regionen mit stark aneinander gekoppelten Markern (den LD- oder Haplotyp-Blöcken), die durch Rekombinations-„*hot spots*„ voneinander getrennt sind. Die Kopplung genomischer Marker innerhalb eines LD-Blocks bezieht sich auf die Tatsache, dass bestimmte Allele an nahe gelegenen Stellen auf demselben Haplotyp häufiger auftreten können als dies durch Zufall zu erwartet ist.[49,50] LD ist von grundlegender Bedeutung für die Assoziationsstudien, da Assoziation zu einem Phänotyp nicht nur am ursächlichen Locus, sondern auch an benachbarten Markern beobachtet werden kann. LD-Muster variieren aber zwischen Ethnizitäten.[51] Ebenso können Markerpaare, die mehrere zehn Kilobasen voneinander entfernt sind, in vollständigem LD zueinander sein, während naheliegende Paare aus der gleichen Region in schwachen LD zueinander stehen. Darüber hinaus variiert die Länge der LD-Blöcke von einem genomischen Bereich zum nächsten.

Die beobachteten Assoziationen mehrerer Marker eines LD-Blocks zu einem Phänotyp müssen daher als miteinander korreliert angesehen werden. Es ist daher von Bedeutung, die LD-Struktur der typisierten Marker im Studienkollektiv zu ermitteln, um die „Anzahl unabhängiger Tests“ auf Assoziation abschätzen zu können. Dies ist notwendig, um in der Bestimmung der Signifikanz für multiples Testen adäquat adjustieren zu können. Ebenso kann eine Modellschätzung für alle Marker eines LD-Blocks gemeinsam erfolgen.

Definitionen solcher Blöcke basieren entweder auf der Schätzung von Haplotypen oder auf der Schätzung des LDs zwischen Markern.[50,52] Für diese Untersuchung wurden Haplotyp-Blöcke nach der in PLINK implementierten Routine bestimmt (= gemäß Handbuch identisch mit der Voreinstellung von Haploview: LD wird nur für Markerpaare mit weniger als 500 kb Distanz zueinander, min-

destens 50% verfügbaren Genotypen und einer minimalen MAF>5% geschätzt. Ein Block enthält mindestens 95% Markerpaare mit "*strong LD*" gemäß den 95%-Konfidenzintervallen für D'). [53] LD-Blöcke werden ferner aus allen benachbarten Markerpaaren mit einem $r^2 \geq 0{,}025$ definiert. Marker-Blöcke ergeben sich aus der Kombination beider Definitionen.

Berisa and Pickrell, 2016 [54] haben auf Basis der 1000-Genom-Projekt-Phase-I Daten (mapping GRCh37/hg19) 1.083 sich nicht überlappende LD-Blöcke u.a. für Kaukasier definiert. Diese Blöcke sind im Mittel 2,531 kb lang und umfassen 1 bis 4.300 häufige und 1 bis 765 selten Varianten des OncoArrays (im Mittel 412 Marker).

Im Vergleich dazu werden 456.699 Blöcke (74.660 Haplotyp-Blöcke und 54.491 hot spots) durch die in PLINK implementierte Routine identifiziert. Diese sind im Mittel nur 14 kb lang und umfassen durchschnittlich 3,5 Marker. Fasst man alle, in LD zueinander stehende Marker ($r^2 \geq 0{,}025$) zusätzlich zusammen, können 103.983 Blöcke (67.161 LD/Haplotyp-Blöcke und 36.822 hot spots) definiert werden. Diese sind im Mittel nur 20 kb lang und umfassen durchschnittlich 4,4 Marker. 5% dieser Blöcke umfassen mehr als 13 Marker (siehe Tabelle 22 und Tabelle 23).

Die Blockstruktur nach Berisa und Pickrell erscheint als für eine Analyse zu grob. Immerhin wird das kürzeste aller Chromosomen (Nr. 22) in nur 3 Blöcke aufgeteilt. Bei der kombinierten Blockdefinition (Haplotyp + LD) sind das immerhin 2.584 Blöcke.

Fazit: Die Analyse aller häufigen Varianten erfolgt in 103.983 Blöcken. Davon umfassen 10.497 Blöcke (~10%) nur selten Varianten, 70.289 Blöcke (~68%) nur häufige. Die verbleibenden 23.197 Blöcke (22%) beinhalten beides (siehe Tabelle 21).

Da die Blöcke als unabhängig angesehen werden können, lässt sich das genomweite Signifikanzniveau gemäß Bonferroni auf $\alpha' = 0{,}05/103.983 \sim 0{,}5 \times 10^{-7}$ festlegen. Als suggestives Signifikanzniveau wir ein $\alpha' = 1 \times 10^{-5}$ betrachtet. (Seltene Varianten blieben bei der Datenauswertung jedoch unberücksichtigt.)

Tabelle 21 Anzahl LD-Blöcke mit seltenen und häufigen Varianten

Anzahl seltener Varianten (MAF≤1%1)	*Anzahl häufiger Varianten (MAF>1%)*		
	keine	*1 oder mehr*	*gesamt*
keine	--	70.289	**70.289**
1 oder mehr	10.497	23.197	**33.694**
gesamt	**10.497**	**93.486**	**103.983**

Block-Definition: Haplotyp +LD

Tabelle 22 LD-Blöcke je Chromosom

LD-block	Blocktyp			Anzahl SNPs		Anzahl häufiger Varianten		Anzahl seltener Varianten		Anzahl monomorpher Varianten		Länge
	gesamt	cold spot	hot spot									
	N	N	N	Mittel	Median	Min.	Max.	Min.	Max.	Min.	Max.	Mittel
gesamt	103.983	67.161	36.822	4.4	3	1	1.230	1	76	1	6	19.658
1	7.461	4.818	2.643	4.4	3	1	280	1	76	1	3	21.553
2	8.664	5.550	3.114	4.5	3	1	142	1	25	1	2	19.785
3	6.705	4.330	2.375	4.4	3	1	206	1	38	1	2	20.656
4	5.954	3.894	2.060	4.2	3	1	100	1	23	1	1	22.393
5	6.366	4.107	2.259	4.5	3	1	156	1	61	1	2	20.185
6	7.510	4.865	2.645	4.8	3	1	205	1	22	1	1	16.473
7	5.278	3.439	1.839	4.4	3	1	250	1	21	1	3	21.067
8	5.410	3.525	1.885	4.5	3	1	101	1	13	1	2	19.124
9	4.581	2.956	1.625	4.3	3	1	73	1	28	1	2	17.477
10	5.215	3.403	1.812	4.5	3	1	85	1	23	1	2	18.323
11	4.905	3.159	1.746	4.4	3	1	85	1	62	1	3	19.492
12	5.142	3.350	1.792	4.5	3	1	349	1	18	1	1	18.407
13	3.232	2.047	1.185	3.9	3	1	107	1	56	1	4	20.394
14	3.414	2.205	1.209	4.3	3	1	193	1	12	1	1	18.664
15	3.180	2.057	1.123	4.2	3	1	51	1	38	1	2	18.518
16	3.264	2.103	1.161	4.1	3	1	82	1	12	1	1	20.054
17	3.185	2.055	1.130	4.7	3	1	1.230	1	48	1	2	18.006
18	3.007	1.943	1.064	4.2	3	1	126	1	16	1	1	18.684
19	2.765	1.788	977	4.3	3	1	180	1	14	1	1	14.138
20	2.910	1.875	1.035	4.4	3	1	159	1	8	1	1	15.811
21	1.357	886	471	4.1	3	1	42	1	14	1	1	19.829
22	1.894	1.214	680	4.4	3	1	221	1	8	1	1	12.875
23	2.584	1.592	992	3.9	3	1	295	1	34	1	6	40.000

cold spot: Marker in LD; *hot spot*: genomische Regionen unkorrelierter Marker; seltene Varianten: MAF≤1%

Tabelle 23 Anzahl Marker pro LD-Block

Anzahl Blöcke	SNPs je Block						
	1	2	3	4	5-10	11-50	>50
103.983	21.596	26.005	15.231	10.092	22.990	7.916	153
	20,7%	25,0%	14,6%	9,7%	22,1%	7,6%	0,1%

4.4 Genomische Stratifikation

Um für vorhandene Populationsstratifikation möglichst effizient adjustieren zu können, wurde eine geeignete Hauptkomponentenanalyse (PCA, *„principal component analysis“*) auf einer Zufallsauswahl typisierter Marker durchgeführt.[55,56] Die Beschränkung der Anzahl berücksichtigter Marker ist durch den enormen Rechenaufwand der PCA bei großen Stichproben notwendig. Ebenso wäre eine PCA durch die Multikolinearität von Markern im LD zueinander nicht mehr valide berechenbar. Im finalen Schätzmodell werden dann eine ausreichend Anzahl an PCs (Hauptkomponenten) berücksichtigt um hinreichend für Populationsstratifikation zu adjustieren, wobei sich die Frage stellt, wie viele PCs „ausreichen“.

Zwei Ansätze hierfür wurden verfolgt:

a) Alle PCs mit signifikanter Varianzzerlegung (Tracy-Widom Statistik)
b) Alle PCs mit signifikanter Änderung im Schätzer des Dezentralitätsparameters λ der χ^2-Verteilung der Teststatistiken aller Marker.

Die Berechnung der PCs erfolgte mit den Programmen SMARTPCA und EIGENSTRAT.[56] Die Schätzung erfolgte auf Basis aller verfügbaren Genotypen des OncoArray-Konsortium, der Wismut und KORA-Studien ungeachtet der Verfügbarkeit und Validität von Phänotypen. Für die Schätzung wurde ein Set an zufällig ausgewählten, nicht-korrelierten Markern (m=26.600; *„pruned“*) verwendet, wobei gemäß Price, et al., 2008 [57] Marker in „long-range LD“-Regionen ausgeschlossen wurden ebenso wie Regionen um bekannte „susceptibility genes“ für Lungenkrebs (5p15.33 *TERT* 1.200-1.400kb; 15q25.1 *CHRNA3* 78.800-79.000kb; 15q25.1 *HYKK* 78.700-78.900kb; 6p21.33 *BAG6/BAT3* 31.600-31.700kb) ergänzt durch jene 22 disponierende Gene (*„novel identified susceptibility loci“*) aus dem OncoArray-Projekt.[58] Da einige SNPs nahe ausgeschlossener Regionen Einfluss auf signifikant zwischen Fällen und Kontrollen differenzierende PCs nehmen, wurden die Regionen nachträglich erweitert. Ebenso wurden alle Marker ausgeschlossen, die eine signifikante „crude“ Assoziation zum Fall/Kontroll-Status aufweisen (p<0,05 oder FDR<0,05). Diese 99 Regionen umfassen insgesamt 6.907 der m=26.600 Zufallsmarker (siehe Anhang Tabelle 75).

Für alle Berechnungen wurden nur Marker mit einer Allelhäufigkeit MAF>1%, einer Call-Rate>90% und Unterschieden zwischen erwarteter und beobachteter Heterozygosität von Δ_{het}<8%-Punkte* sowie Personen mit einer Call-Rate>90% verwendet. (*Ob der großen Fallzahl wurden schon kleinste absolute Abweichungen vom HWE mit $p_{HWE}<1x10^{-100}$ – χ^2-Test mit einem Freiheitsgrad - als höchst-signifikante ausgewiesen. p_{HWE} ist deshalb kein sinnvoll verwendbares Ausschlusskriterium ungeeigneter Marker)

Die Berechnung der PCA erfolgte in zwei Schritten:

1. Prüfschritt – zur Bereinigung des Markersatzes und Suche nach genomischen *Einzelfällen*
    a. Bei PCs mit signifikanten Unterschieden zwischen Fällen und Kontrollen (cc-signifikant gemäß ANOVA) wurden jene Marker, die mit dem betroffenen PC am stärksten korrelieren (*„eigenbestsnps“*) aus dem Markersatz ausgeschlossen, um ein Überkorrigieren durch für den Fall/Kontroll-Status relevante PCs zu verhindern. Ebenso wurden alle mit einer betroffenen PC signifikant korrelierten Marker ausgeschlossen (p_{GC}<0,05, p-Wert korrigiert gemäß *genomic control*) oder die mindestens 1% der Variabilität einer betroffenen PC erklären.
    b. Ebenso wurde nach Personen gesucht, die als genomische Einzelfälle (Ausreißer) anzusehen sind. Der Ausreißer-Grenzwert wurde mit σ=45 (Programmparameter *outlier sigma thresh*) festgelegt. Die Anzahl möglicher Ausreißer wurde mit 10 begrenzt. Alle auffälligen Personen wurden mit jenen der ADMIXTURE-Analyse abgeglichen und bei Übereinstimmung aus den Kollektiven als *„genomische Einzelfälle“* ausgeschlossen.

Dieser 1. Prüfschritt wurde maximal 10x wiederholt, oder bis keine PC einen signifikanten Unterschied zwischen Fällen und Kontrollen aufwies.

2. Nach Bereinigung inadäquater Marker und auffälliger Personen wurde die PCA ohne Prüfung auf Ausreißer durchgeführt.

Die maximale Anzahl zur Korrektur von genomischen Substrukturen notwendiger PCs wurde durch die Zahl signifikanter PCs gemäß der Tracy-Widom-Statistik, wie in SMARTPCA implementiert, bestimmt. Die Anzahl notwendiger PCs orientiert sich an der Schätzung des „genome wide inflation factors λ „gemäß *genomic control*".[56,59] Für die Analyse wird jene Anzahl PCs mit dem geringstem λ herangezogen, wenn signifikant mit dem CC-Status assoziierte PCs unberücksichtigt bleiben.

4.4.1 Ergebnis Markersatz I (m=26.600 Zufallsmarker)

Aus den 472.998 zur Analyse zur Verfügung stehenden Markern, wurden zufällig 26.600 über alle Chromosomen gleichmäßig verteilte Marker für die Bestimmung der PCs ausgewählt. Davon ausgeschlossen wurden Indels (gemeinsame Bezeichnung für die Mutationsformen Insertion=Einschub und Deletion=Auslassung), Marker aus "long-range LD-regions" und Marker mit auffälligen Abweichungen vom HWE. Für die Bestimmung der PCs verblieben zuletzt m=15.060 Marker. In vier Durchläufen wurden weitere 1.723 CC-signifikante Marker ausgeschlossen. Die finale PC-Schätzung beruht daher auf 13.337 Markern.

Die *Tracy-Widom-Statistik* weist für die ersten 439 der maximal 28.606 PCs einen statistisch nachweisbaren Erklärungsbeitrag für die genetischen Substrukturen in der Stichprobe aus. Über den jeweiligen Erklärungsgrad wird dabei keine Aussage getroffen. Damit kann zunächst die Obergrenze der notwendigen Anzahl PCs auf 439 beschränkt werden.

Die Chi^2-Verteilung der Teststatistiken aller m=26.600 Marker ist um den Faktor λ=1,656 (beobachteter Median/ theoretischer Median) gestreckt (=inflated). Durch Adjustieren mit 4 PCs lässt sich λ auf einen Wert von 1,139 senken, durch Adjustieren mit 5 PCs auf einen Wert von 1,108 senken. Die 5. PC korreliert jedoch signifikant (p=0,001) mit dem Erkrankungsstatus (siehe Tabelle 24 und Abbildung 7). Das Adjustieren an weiteren PCs erbringt keine weitere Senkung von λ.

Tabelle 24 Inflationsfaktor λ gemäß "genomic control": Markersatz I (m=26.600 Zufallsmarker)

PC	p-Wert Korrelation zu CC	korrigiertes λ (alle PCs)	korrigiertes λ (alle nicht CC-sig. PCs)
unkorrigiert		1,656	
1	0,069	1,436	1,436
2	0,023	1,131	--
3	0,852	1,141	1,445
4	0,732	1,139	1,455
5	0,001	1,108	--
6	0,587	1,106	1,453
7	0,741	1,109	1,457
8	0,664	1,105	1,458
9	0,415	1,110	1,470
10	0,147	1,111	1,458
11	0,417	1,106	1,457
12	0,609	1,105	1,463
13	0,435	1,107	1,459
14	0,921	1,106	1,460
15	0,139	1,097	1,463
16	0,484	1,094	1,469
17	0,756	1,097	1,469
18	0,169	1,095	1,459
19	0,464	1,099	1,460
20	0,267	1,097	1,461

PC = principal component, CC = Erkrankungsstatus (*case control status*)

4.4.2 Ergebnis Markersatz II (m=33.661 Zufallsmarker)

Die Berechnung wurden mit einer alternativen Zufallsauswahl des OncoArray-Konsortiums von m=33.661 Markern wiederholt. Die Liste diese Marker steht allen Mitgliedern von TRICL/ILCCO-Konsortium via Wiki-OncoArray zur Verfügung (http://consortia.ccge.medschl.cam.ac.uk/oncoarray/doku.php; Abschnitt: Principal Component Generator; download: PCgenerator_program.zip – darin enthalten: PCA_SNPlist.txt).

Aus den 472.998 zur Analyse zur Verfügung stehenden Markern, wurden zufällig 33.661 über alle Chromosomen gleichmäßig verteilte Marker für die Bestimmung der PCs ausgewählt. Davon ausgeschlossen wurden Indels, Marker aus "long-range LD-regions" und Marker mit auffälligen Abweichungen vom HWE. Für die Bestimmung der PCs verblieben zuletzt m=24.811 Marker. In vier Durchläufen wurden weitere 8.682 CC-signifikante Marker ausgeschlossen. Die finale PC-Schätzung beruht daher auf 16.129 Marker.

Die *Tracy-Widom-Statistik* weist für die ersten 15.753 der maximal 28.606 PCs einen statistisch nachweisbaren Erklärungsbeitrag für die genetischen Substrukturen in der Stichprobe aus. Über den jeweiligen Erklärungsgrad wird dabei keine Aussage getroffen. Damit kann zunächst die Obergrenze der notwendigen Anzahl PCs auf 15.753 beschränkt werden.

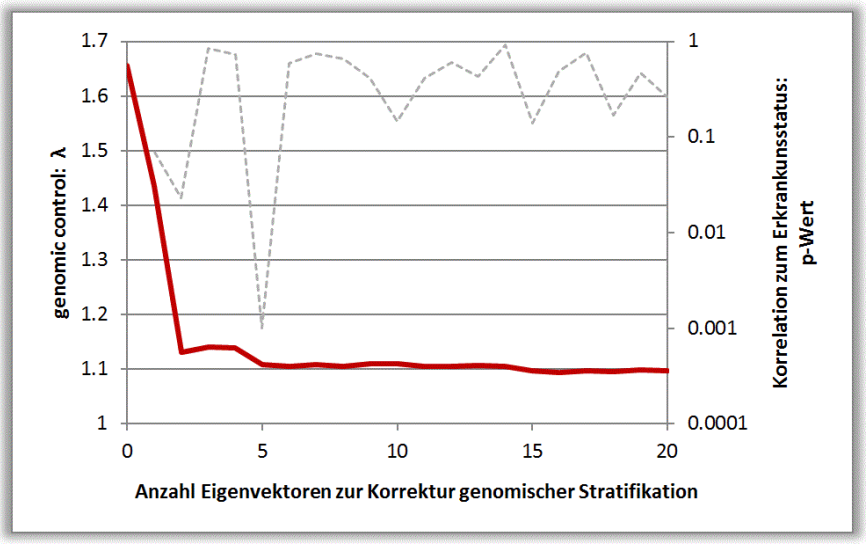
Die Chi^2-Verteilung der Teststatistiken aller m=33.661 Marker ist wie zuvor um den Faktor λ=1,656 (beobachteter Median/ theoretischer Median) gestreckt (=inflated). Durch Adjustieren mit 4 PCs lässt ich λ auf einen Wert von 1,152, durch Adjustieren mit 5 PCs aus einen Wert von 1,109 senken. Die 5. PC korreliert jedoch signifikant (p=0,0002) mit dem Erkrankungsstatus (siehe Tabelle 25 und Abbildung 8). Das Adjustieren an weiteren PCs erbringt keine weitere Senkung von λ.

Tabelle 25 Inflationsfaktor λ gemäß "genomic control": Markersatz II (m=33.661 Zufallsmarker)

PC	p-Wert Korrelation zu CC	korrigiertes λ (alle PCs)	korrigiertes λ (alle nicht CC-sig. PCs)
unkorrigiert		1,656	
1	0,170	1,617	1,617
2	0,057	1,298	1,298
3	0,120	1,158	1,158
4	0,532	1,152	1,152
5	0,0002	1,109	--
6	0,454	1,110	1,159
7	0,395	1,102	1,157
8	0,045	1,104	--
9	0,160	1,105	1,163
10	0,491	1,104	1,165
11	0,870	1,104	1,163
12	0,504	1,104	1,162
13	0,702	1,103	1,159
14	0,933	1,102	1,160
15	0,607	1,102	1,159
16	0,190	1,100	1,156
17	0,702	1,103	1,154
18	0,840	1,103	1,154
19	0,890	1,107	1,157
20	0,056	1,101	1,155

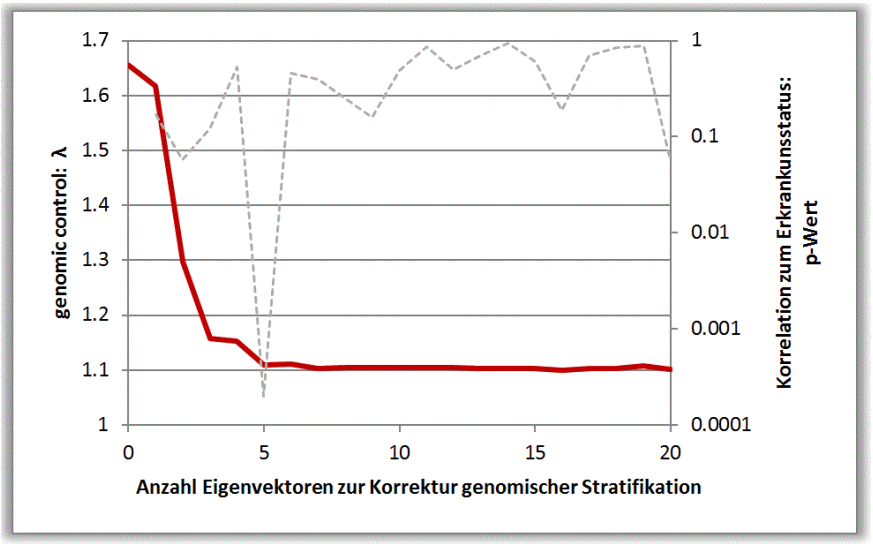
PC = principal component, CC = Erkrankungsstatus (*case control status*)

Abbildung 7 Inflationsfaktor λ gemäß "genomic control": Markersatz I (m=26.600 Zufallsmarker)



Rote durchgezogene Linie = Inflationsfaktor λ, gestrichelte Line = p-Wert

Abbildung 8 Inflationsfaktor λ gemäß "genomic control": Markersatz II (m=33.661 Zufallsmarker)



Rote durchgezogene Linie = Inflationsfaktor λ, gestrichelte Line = p-Wert

4.4.3 Fazit

Wenn man die Zufallsauswahl an Markern des OncoArray-Konsortiums (Oxford, m=33.661 Marker) zur Schätzung von λ verwendet, sind die ersten fünf PCs tendenziell stärker mit dem Fall-Kontroll-Status der Proben korreliert, als unter Verwendung der Zufallsauswahl m=26.600 Marker (siehe p-Werte). Unabhängig davon welcher Markersatz herangezogen wurde, der Parameter λ erreicht erst durch adjustieren mit jeweils 5 PCs eine Wert von ~1,1 (ursprünglich als akzeptabler Wert anerkannt [60]), ändert sich aber kaum (oder gar nicht) mehr, wenn mit weiteren PCs adjustiert wird. Jedoch ist dieser 5. PC jeweils (für beide Marker-Sätze) mit dem CC-Status signifikant korreliert und könnte daher die Power zum Aufdecken von Marker x Phänotyp (G)-Assoziation oder einer GxE-Interaktion senken.

Die Akzeptanz eines Rest-λ von ~1,1 kann durch die Untersuchung von Yang 2011 gerechtfertigt werden: "*Before large-scale GWAS being conducted, this method [genomic control method] … became a standard approach to quantify and adjust for population structure. In the first wave of GWAS, the genomic inflation factors observed in GWAS with thousands of individuals were usually <1,1, which were usually interpreted to be due to subtle population structure. Much larger inflation*

factors have been observed in GWAS with large sample size especially when pooling a number of GWAS into a meta-analysis" …. „*We have shown by theory, simulation studies and analysis of multiple data sets that a significant inflation of test statistics is to be expected under polygenic inheritance even when there is no population structure*." [60]

Es scheint daher sinnvoll mit nur 4 PC für *genomische Stratifikation* zu adjustieren. Das geschätzte λ unter Verwendung der Zufallsauswahl m=26.600 Marker ist kleiner als unter Verwendung der Zufallsauswahl m=33.661 Marker. Demzufolge wurden die diesbezüglichen Eigenwerte der PCA in den statischen Modellen zum Adjustieren als Kovariablen verwendet.

4.5 Analysemodell für eine Einzelmarker-Assoziationsanalyse

Es wurde folgendes Analysemodell zur Schätzung der GxE- Interaktion verwendet:

$$\ln(\mathrm{Odd}_D) = \ln\left(\frac{p_D}{1-p_D}\right) = \beta_o + \boldsymbol{\beta}_{1i} PC_i + \beta_{PS} PS + \beta_G G + \beta_E E + \beta_{G\times E}(G \cdot E)$$

D: Erkrankungsstatus (D=1: LK-Patienten; D=0: Kontrollen)

G: Anzahl des selteneren Allels am Marker m

E: Strahlenexposition (0: nicht-exponiert bei WLM ≤50, 1: exponiert bei WLM>50)

PS: Propensity-Score (vereinige nicht-genetischen Risikofaktoren des Lungenkrebses)

PC: Vier Eigenvektoren der PCA zum adjustieren für Populationsstratifikation

Mit diesem Modell werden nur Marker mit einer Allelhäufigkeit unter Kontrollen mit einer MAF>1% analysiert. Bei einer geringeren MAF ist die Wahrscheinlichkeit (Power) eine Assoziation aufzudecken gering. Dafür ist die Wahrscheinlichkeit hoch, in der kleinen Gruppe der exponierten Fälle kein seltenes Allel zu beobachten und damit den zentralen Modellparameter $\beta_{G\times E}$ nicht schätzen zu können.

Die **Signifikanz einer Gen x Umwelt-Interaktion** wurde mit drei Verfahren bewertet:

1. Test des Parameters $\beta_{G\times E}$ allein (H0: $\beta_{G\times E} = 0$),
   Korrektur nach Bonferroni anhand m=103.983 unabhängiger LD-Regionen.
2. Joint-Test von β_G und $\beta_{G\times E}$ nach Kraft et al. [61]
   Korrektur nach Bonferroni anhand m=103.983 unabhängiger LD-Regionen
3. Hybrid 2-Schritt (H2)-Verfahren von Murcray et al. [62]
   Aufteilung der Marker (Vorfiltern) hinsichtlich deren marginalen Assoziation zu D bzw. E – Korrektur nach Bonferroni anhand m unabhängiger LD-Regionen, wobei m vom Vorfilter abhängt.

Für Verfahren 1 und 2 werden nur alle beobachteten Daten verwendet. Bei Verfahren zwei ist zusätzlich auf die Robustheit der Schätzung von $\beta_{G\times E}$ zu achten, um Fehlschlüsse zu vermeiden.

Für das Hybrid 2-Schritt (H2)-Verfahren [62] müssen zwei weitere Modelle für den Screening-Schritt (marginale ExG- bzw. DxG-Effekte) gefittet werden. Das α-Signifikanzniveau wird im Verhältnis ρ : (1-ρ) zwischen beiden Screening-Modellen aufgeteilt, wobei ρ ein frei wählbarer Wert zwischen 0 und 1 ist. Die beiden marginalen Screening-Modelle lauten:

ExG-Modell: $$ln\left(\frac{p_E}{1-p_E}\right) = \beta'_0 + \boldsymbol{\beta'_{1i}} PC_i + \beta'_{PS} PS + \beta'_G G \qquad \alpha_{e'} = \rho\alpha$$

Das ExG-Modell musste nicht-adjustiert durch die Eigenvektoren PC_1-PC_4 geschätzt werden, da anderenfalls aufgrund der geringen Fallzahl (nur n=308 exponierte) bei den meisten Markern $\beta_{G'}$ keine Schätzergebnisse erzielt werden konnten. Für Populationsstruktur wurde daher gemäß *„genomic control"* (GC) korrigiert, der entsprechende p-Wert p_{GC} verwendet. Der von PLINK exportiere Standardfehler s.e. korrespondiert jedoch mit dem unkorrigierten p-Wert ($p \sim \frac{\ln(OR)}{s.e.}$). Daher wurde s.e. an den p_{GC} angepasst ($p_{GC} \sim \frac{\ln(OR)}{s.e._{GC}}$), während der Punktschätzer *ln(OR)* unverändert übernommen wurde.

Da mit PLINK nur eine Modellschätzung nach Zufallsauswahl nicht-exponierten Fälle möglich ist, wurden 10 solcher Zufallsauswahlen getroffen. Die Parameterschätzer wurden im

Sinne des „*modell averaging*“ zusammengefasst und ein gemeinsamer p-Wert (über mehrere Zufallsauswahlen) bestimmt.[63]

DxG-Modell: $$\ln\left(\frac{p_D}{1-p_D}\right) = \beta_0'' + \boldsymbol{\beta}_{1i}'' PC_i + \boldsymbol{\beta}_{PS}'' PS + \beta_G'' G \qquad \alpha_{d'} = (1-\rho)\alpha$$

Da die Strahlenexposition nicht in das DxG-Modell eingeht, ist weder eine Gewichtung noch eine Zufallsauswahl nicht-exponierter Fälle notwendig. Das Modell kann daher auf Basis aller zur Verfügung stehenden typisierten Personen geschätzt werden. Ebenso wurde von McKay et al. [58] eine Assoziationsanalyse je Marker des OncoArray-Konsortiums, aller typisierten Kaukasier kombiniert mit 29.863 Fällen und 55.586 Kontrollen aus bereits zuvor existierenden Lungenkrebs-Studien und nach Imputation nicht-typisierter Marker durchgeführt. So weit verfügbar, können alternativ die p-Werte dieser umfangreicheren Datenauswertung zum Screening verwendet werden, auch wenn die im Vergleich kleine Stichprobe aus Wismut-Bergarbeiter hierfür unberücksichtigt bleibt.

4.6 Analysemodell für eine Multimarker-Assoziationsanalyse

Das Analysemodell zur Schätzung der GxE- Interaktion kann durch gemeinsames Schätzen alle Marker eines LD-Blocks zu eine analogen Analysemodell für Multimarker-Interaktion erweitert werden:

$$\ln(\mathrm{Odd}_D) = \ln\left(\frac{p_D}{1-p_D}\right) = \beta_o + \boldsymbol{\beta}_{1i} PC_i + \beta_{PS} PS + \boldsymbol{\beta}_{\boldsymbol{G}} \boldsymbol{G} + \beta_E E + \boldsymbol{\beta}_{\boldsymbol{G}\times E}(\boldsymbol{G}\cdot E)$$

D: Erkrankungsstatus (D=1: LK-Patienten; D=0: Kontrollen)

G: Matrix der Anzahl seltenerer Allele aller Marker eines LD-Blocks

E: Strahlenexposition (0: nicht-exponiert bei WLM ≤50, 1: exponiert bei WLM>50)

PS: Propensity-Score (vereinigte nicht-genetischen Risikofaktoren des Lungenkrebses)

PC: Vier Eigenvektoren der PCA zum Adjustieren für Populationsstratifikation

Mit diesem Modell werden nur Marker mit einer Allelhäufigkeit unter Kontrollen mit einer MAF>1% analysiert, da die Wahrscheinlichkeit hoch ist, in der kleinen Gruppe der exponierten Fälle kein seltenes Allel zu beobachten und damit den zentralen Modellparameter $\beta_{G\times E}$ nicht schätzen zu können. Ebenso besteht die Gefahr der Ko-Linearität oder der linearen Abhängigkeit (resultierend aus „*complete separation of data points*“) zu anderen Markern.

Für den Parametervektor $\boldsymbol{\beta}_{\boldsymbol{G}\times E}$ wurde ein Test der Nullhypothese H_0: $\boldsymbol{\beta}_{\boldsymbol{G}\times E} = 0$, gleichbedeutend mit keiner Interaktion an jedem Marker, durchgeführt.

Entsprechend wurde für den Joint-Test nach Kraft et al. die Nullhypothese H_0: $\boldsymbol{\beta}_{\boldsymbol{G}} = 0$ und $\boldsymbol{\beta}_{\boldsymbol{G}\times E} = 0$ geprüft.

Sind die genomischen Marker (G) bzw. die Interaktionsvariablen (GxE) stark miteinander korreliert (Ko-Linearität) oder nur wenige Exponierte tragen einen seltenen Genotyp, dann kann die Schätzung der Modellparameter und damit verbunden das Testen der Nullhypothesen unmöglich oder nur mit geringer Präzision erfolgen. Dies tritt vor allem in LD-Blöcken auf, für die viele Marker typisiert wurden (bei extrem langen LD-Blöcken oder bei hoher Markerdichte). Um Ko-Linearität, lineare Abhängigkeiten von Modellvariablen und instabile Parameterschätzungen zu vermeiden, wurden ggf. Marker bzw. Interaktionsterme vor der Modellschätzung wie folgt eliminiert:

- Hatten alle (oder alle bis auf einen) Exponierten denselben Genotyp, wurde kein Interaktionsterm gebildet.
- Bei nahezu vollständiger Übereinstimmung zweier Marker (Korrelation r>0,99), wurde einer der beiden Marker ausgeschlossen.
- Wurden für einen Marker weniger als 2 Träger des selteneren Alleles unter den exponierten Fällen beobachtet und ebenso erwartetet (gegeben die MAF der nicht-exponierten Kontrollen), wurde kein entsprechender Interaktionsterm gebildet (analog für Kontrollen).
- Alle verbleibenden Marker wurden auf lineare Abhängigkeiten geprüft (gemäß der Routineprüfung von *proc logistic*). Abhängige Marker bzw. Interaktionsterme wurden ausgeschlossen.

- Marker bzw. Interaktionsterme wurden ausgeschlossen, denen bei der Modellschätzung von *proc logistic* eine *df=0* zugewiesen wurde (um die Likelihood maximieren zu können) oder die Teststatistik eine Wert χ^2<0,0001 angenommen hat (instabile Schätzung oder hochgradig insignifikante Effekt). Das Modell wurde erneut geschätzt.
- Konnten die Tests weiterhin nicht valide durchgeführt werden, wurden zunächst robuste Regressionsmodelle unter Maximierung von „*Firth's Penalized Likelihood*" geschätzt.[64] Danach wurden die standardisierten, geschätzten Parameter $\ddot{\beta} = \beta / s.e._{\beta}$ gebildet. Übertraf $\ddot{\beta}$ das 10-fache der Schätzung aus der vorherigen Modellanpassung, wurde der Marker bzw. die Interaktion als nur instabil schätzbar ausgeschlossen.
- Konnten die Tests weiterhin nicht valide durchgeführt werden, wurden schrittweise Marker bzw. Interaktionsterme aus den zu prüfenden Nullhypothesen ausgeschlossen. Die Reihenfolge der Elimination erfolgte gemäß dem AIC (entsprechend dem unter *„AIC-optimales Modell"* beschrieben Vorgehens, siehe nächste Seite). Die Marker und Interaktionsterme wurde jedoch nicht aus dem Modell ausgeschlossen.

Trotz allen Vorfilterns und Selektierens von Markern und Interaktionstermen, die eine zuverlässige Schätzung der Modelle stören könnten, wurden die Testergebnisse auf „überprüfenswerte" Ergebnisse (die durch nummerische Besonderheiten hervorgerufen worden sein können) untersucht. Dabei wurden folgende Heuristiken verwendet:

Der Test des Parametervektor $\boldsymbol{\beta}_{\boldsymbol{G}\times E}$ wird als „überprüfenswert" bezeichnet, wenn

a) weniger als 2 Träger des selteneren Allels unter den exponierten Kontrollen oder unter den exponierten Fällen beobachtet wurden (instabile Parameterschätzung).

Der Joint-Test der Nullhypothesen H_0: $\boldsymbol{\beta}_{\boldsymbol{G}} = 0$ und $\boldsymbol{\beta}_{\boldsymbol{G}\times E} = 0$ wird auch als „überprüfenswert" bezeichnet, wenn

b) der p-Wert für H_0: $\boldsymbol{\beta}_{\boldsymbol{G}\times E} = 0$ größer ist als 0,05
(nicht einmal lokal signifikante Interaktion)
oder
c) der p-Wert für H_0: $\boldsymbol{\beta}_{\boldsymbol{G}} = 0$ kleiner ist als der des Joint-Tests
(genomischer Haupteffekt stärker signifikant als der Haupteffekt + Interaktion).

AIC-optimales Modell

Zusätzlich wurde eine Modellselektion mit schrittweiser Elimination (*„backward selection"*) durchgeführt. Die Wahl des zu eliminierenden Markers bzw. Interaktionsterms erfolgte nach der Anpassungsgüte des Modells bemessen am AIC (*„Akaike information criterion"*). Es wurden aber nicht alle Marker bzw. Interaktionsterm eliminiert. Das schlussendlich „beste" Modell musste mindestens einen Marker bzw. einen Interaktionsterm enthalten, damit hierfür eine entsprechender Test durchgeführt und ein p-Wert berechnet werden konnte.

4.7 Einschränkungen der Interpretierbarkeit von Parameterschätzern

Es sei hier angemerkt, dass die in diese Analyse eingehenden Daten nicht einer einzelnen Erhebung mit definiertem Studiendesign entstammen. Die Daten wurden vielmehr aus drei Quellen zusammengefügt:

Die Lungenkrebsfälle der Wismut-Bergarbeiter entstammen der sogenannten *Indoor-Radon-Studie*, die zwischen 1990 und 1997 durchgeführt wurde.[65,66] Die Lungenkrebsfälle wurden dabei aus der Allgemeinbevölkerung über Studienkliniken rekrutiert. Die obere Altersgrenze betrug 75 Jahre. Die lungenkrebsfreien Kontrollen unter den Wismut-Bergarbeiter stammen aus einer Population, die an der *Gesundheitsvorsorge der Wismut AG* teilnahmen. Die Rekrutierung erfolgte ca. 2009 und somit 15-20 Jahre später als die Rekrutierung der Fälle. Es handelt sich also um Langzeitüberlebende, von denen die Hälfte über 80 Jahre alt waren. Wichtig anzumerken ist, dass die Wismut-Beschäftigten keine Zufallsauswahl darstellen, da die verwendeten Proben quotiert nach Exposition zur Genotypisierung ausgewählt wurden. Diese beiden Gruppen bilden sozusagen die Kern-Stichprobe (n=463 Personen), die durch nicht-exponierte Fälle und Kontrollen des *internationalen OncoArray-Konsortiums* ergänzt wurden (n=28.600 Personen). Diese Datenauswertung wurde daher als ein Vergleich von Fällen und Kontrollen angelegt. Um Missverständnisse, insbesondere in der Interpretation von Schätzergebnissen, vorzubeugen, wird das Design dieser Untersuchung als Fall-Kontroll-Vergleich („*case-control comparison"*) bezeichnet. Im Folgenden werden wichtige Aspekte zur Interpretation der Analyseergebnisse diskutiert:

		Krankheitsstatus krank	gesund
Exposition	ja	*a*	*b*
	nein	*c*	*d*

Das Odds-Ratio (OR) ist eine Maß zur Quantifizierung der Assoziationsstärke eine Exposition (hier Radon) mit einer dichotomen Zielgröße (hier der Krankheitsstatus Lungenkrebs). Das OR setzt dabei die Chance exponiert zu sein eines Kranken ($a:c$) ins Verhältnis zu der eines Gesunden ($b:d$). Somit ergibt sich die Schätzgleichung $\widehat{OR} = \frac{a/c}{b/d} = \frac{a \cdot d}{b \cdot c}$. (siehe Tabelle rechts). Das OR ist robust gegenüber ungleichen oder nicht repräsentativen „Stichprobenziehung" sowohl hinsichtlich der Krankheit als auch der dichotomen Exposition (wie hier <50 WLM / >50 WLM). Werden z.B. k-mal so viele Kranke wie Gesunde in die Analyse aufgenommen, so hat das keinen Einfluss auf die Schätzung der OR, da gilt: $\widehat{OR} = \frac{a \cdot kd}{kb \cdot c} = \frac{a \cdot d}{b \cdot c}$. Jedoch verbirgt sich hinter der dichotomisierten Exposition die kontinuierlicher Expositionsgröße WLM als Maß der lebenslangen, berufsbedingten Strahlenexposition. Durch die gegenüber der Grundgesamtheit (alle Wismut-Bergarbeiter) nicht-proportionale Auswahl an besonders hohen oder besonders niedrig exponierten Bergarbeiter für die BfS-Biobank [17], kann die Schätzung der ORs multiplikative verzerrt sein, wenn diese ungleich 1 sind. Aus der Ableitung des ORs einer GxE-Interaktion aus dem verwendeten Analysemodell (siehe Einschub nächste Seite) lässt sich zeigen, dass das $\mathrm{OR}_{G\times E}$ unabhängig von den ORs der Haupteffekte für G und E ist. Selbst wenn diese verzerrt geschätzt werden, ist ein Test der Null-Hypothese $\boldsymbol{H_0}: \hat{\beta}_{G\times E} \neq 0$ mit $\widehat{OR}_{G\times E} = e^{\hat{\beta}_{G\times E}}$ ein valides, statistisches Vorgehen zum Nachweis einer GxE-Interaktion, dem Ziel dieser Untersuchung.

Alle Studienteilnehmer können auf Basis ihrer DNA als Kaukasier angesehen werden und wurden in verschiedenen Regionen Europas und Nordamerikas rekrutiert. Die Ergänzungsstichprobe musste im Verhältnis *Fälle : Kontrollen* der viel kleineren Kernstichprobe der Wismut-Bergarbeiter angepasst werden, um die OR_E einer Radonexposition unter den Wismut-Bergarbeiter zu fixieren (siehe Kapitel 4.2.1). Jegliches, geschätztes OR ist damit lediglich eine stichproben-interne Größe.

Aufgrund beider Argumente wird daher verzichtet, ORs als unverzerrte Schätzer eines relativen Lungenkrebsrisikos für Kaukasier zu verallgemeinern. Ein solches Verallgemeinern z.B. der OR_E war

auch nicht Ziel dieser Untersuchung, da das durch Radon induzierte Lungenkrebsrisiko mit zahlreichen geeigneteren Studien bereits belegt wurde.[4,5,37,40,67-69]

Einschub: Aus dem vereinfachten Analysemodell (ohne den Termen: $\boldsymbol{\beta}_{1i}PC_i$ und $\beta_{PS}PS$) lässt sich folgendes ableiten:

$$\ln(\mathrm{Odd}_D) = \beta_o + \boldsymbol{\beta}_{1i}PC_i + \beta_{PS}PS + \beta_G G + \beta_E E + \beta_{G\times E}(G \cdot E)$$

$$\mathrm{OR}_{E,G=0} = \frac{\mathrm{Odd}_{E=1,G=0}}{\mathrm{Odd}_{E=1,G=0}} = \frac{e^{\beta_o}e^{\beta_E}}{e^{\beta_o}} = e^{\beta_E}$$

$$\mathrm{OR}_{E,G=1} = \frac{\mathrm{Odd}_{E=1,G=1}}{\mathrm{Odd}_{E=1,G=1}} = \frac{e^{\beta_o}e^{\beta_E}e^{\beta_{G\times E}}}{e^{\beta_o}} = \mathrm{OR}_{E,G=0}e^{\beta_{G\times E}}$$

$$\mathrm{OR}_{G\times E} = \frac{\mathrm{OR}_{E,G=1}}{\mathrm{OR}_{E,G=0}} = \frac{\mathrm{OR}_{E,G=0}\,e^{\beta_{G\times E}}}{\mathrm{OR}_{E,G=0}} = e^{\beta_{G\times E}} \quad \forall \mathrm{OR}_E$$

ebenso

$$\mathrm{OR}_{G,E=0} = \frac{\mathrm{Odd}_{E=0,G=1}}{\mathrm{Odd}_{E=0,G=0}} = \frac{e^{\beta_o}e^{\beta_G}}{e^{\beta_o}} = e^{\beta_G}$$

$$\mathrm{OR}_{G,E=1} = \frac{\mathrm{Odd}_{E=1,G=1}}{\mathrm{Odd}_{E=1,G=0}} = \frac{e^{\beta_o}e^{\beta_G}e^{\beta_{G\times E}}}{e^{\beta_o}} = \mathrm{OR}_{G=0}e^{\beta_{G\times E}}$$

$$\mathrm{OR}_{G\times E} = \frac{\mathrm{OR}_{G,E=1}}{\mathrm{OR}_{G,E=0}} = \frac{\mathrm{OR}_{G,E=0}\,e^{\beta_{G\times E}}}{\mathrm{OR}_{G,E=0}} = e^{\beta_{G\times E}} \quad \forall \mathrm{OR}_G$$

G: genomische Exposition am Marker m
E: Strahlenexposition

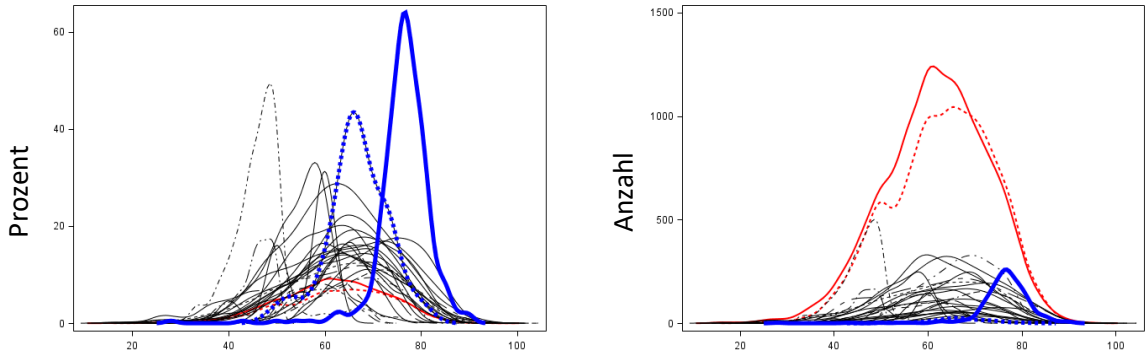
Verzerrende Effekte durch ***Beobachtungsungleichheit*** können ausgeschlossen werden, da hinsichtlich der Informationsqualität von Alter, Geschlecht, Rauchstatus und der des ärztlich diagnostizierten Primärtumors der Lunge kaum Unterschiede zwischen den Studien bestehen. Ebenso wurden die Genotypen aller Personen nach identischen Protokollen mit identischer Technologie in nur vier Zentren bestimmt.

Somit bleibt noch die Frage nach der ***Strukturgleichheit zwischen Fällen und Kontrollen***, die sich in drei Aspekte aufteilen lässt:

Strukturgleichheit hinsichtlich des Genoms: Um für genomische Substrukturen innerhalb der untersuchten Stichprobe zu adjustieren, wurde - gemäß des Standards bei GWAS - die Genotypen einer Hauptkomponenten-Analyse unterworfen und die ersten vier „principal components“ in die Analyse Modelle aufgenommen. Die Kernstichprobe aus Wismut-Bergarbeitern nimmt dabei eine zentrale Lage unter allen Teilstichproben und somit keine extreme Stellung ein (siehe Kapitel 4.1.3.4).

Strukturgleichheit hinsichtlich des Alters: In den beiden folgenden Grafiken (siehe Abbildung 9), wurde die Altersverteilung der einzelnen Teil-Stichproben (blau und schwarz) sowie aufgeteilt nach Fällen und Kontrollen (rot) dargestellt. Im linken Bild sind prozentuale Häufigkeiten innerhalb der Teil-Stichproben abgebildet. Das erkennbare höhere Alter der Wismut-Kontrollen ist jedoch nur von geringer Bedeutung, da sich deren lebenslange, berufsbedingte Radonbelastung nach Ende der Beschäftigung im Bergbau nicht mehr verändert hat. Die rechte Abbildung zeigt die Anzahl der Teilnehmer je Teil-Stichprobe und relativ zur Anzahl an Fällen bzw. Kontrollen gesamt. Die Grafik veranschaulicht, dass für eine Altersadjustierung in den statistischen Modellen auch ausreichend ältere Personen (75, 80 und älter) des OncoArray-Konsortiums zur Verfügung standen.

Abbildung 9 Altersverteilung



rot: gesamt; **blau**: Wismut-Bergarbeiter; **schwarz**: andere Teilstichproben; **punktierte Linie**; Fälle; **durchgezogene Line**: Kontrollen

Strukturgleichheit hinsichtlich des Überlebens bis zur Rekrutierung: Die Selektion der Wismut-Langzeit-Überlebenden als Kontrollen kann theoretisch zu Verzerrung durch Confounding führen, wenn einzelne Gene für eben dieses Langzeit-Überlebende verantwortlich gemacht werden können. Gene, die für sich allein einen solchen Effekt zeigen, sind nach dem Stand des derzeitigen Wissens nicht bekannt. Polygenetische Effekte, die sich aus sehr kleinen Effekten sehr vieler Gene zusammensetzen, können nicht ausgeschlossen werden. Effekte dieser Art wurden in der Datenanalyse in zweierlei Hinsicht berücksichtigt: Erstens, die Verteilung der -2·ln(p)-Werte der Einzel- als auch der Multimarkern-Analyse folgte einer Chi²-Verteilung mit einem Freiheitsgrad (df=1). Polygenetische Effekte, die Einfluss auf die Schätzung der GxE-Interaktion nehmen, konnten also nicht beobachtet werden. Zweitens, für die Gen-Set-Analyse wurde ein Verfahren gewählt, das eine sogenannte „kompetitive Null-Hypothese" prüft. Alle Gene eines betrachteten Gen-Sets wurden mit allen verbleibenden Genen (Komplementär-Set) verglichen. Polygenetische Effekte die nicht mit der Funktion des betrachteten Gen-Sets in Beziehung stehen, würden unter der Null-Hypothese (kein Unterschied zwischen den Sets) sowohl das Gen-, als auch das Komplementär-Set betreffen. Der statistische Vergleich beider Sets miteinander ist daher nicht negativ beeinflusst.

Schließlich soll hier noch darauf hingewiesen werden, dass GWAS zwar mit üblichen, epidemiologischen Studiendesigns konzipiert werden, deren Ergebnisse (statistische Assoziationen) aber nicht als Beleg eines kausal wirkenden Risikofaktors angesehen werden.[70-74] Das können statistische Assoziationen nicht leisten. Es hat sich in den letzten Jahren gezeigt, dass GWAS-Resultate auch nur einen (kleinen) Teil der Heritabilität (des vererbbaren Risikos einer Erkrankung) erklären können.[75,76] Ziel diese Projekts war es viel mehr, Hinweise auf molekularbiologischen Mechanismen zu finden, die in der Ätiologie des Lungenkrebses eine Rolle spielen, um z.B. in Zukunft Personen, die von der Natur mit Defiziten darin ausgestatten wurden und daher eine höheren Lungenkrebsrisiko bei beruflich hoher Radonexposition mit sich tragen, identifizieren zu können. Die Liste der aus dieser Untersuchung hervorgehenden Gene, Loci und Mechanismen basiert auf deren statistisch auffälligen Assoziationen (in den Daten dieser Untersuchung). Die erzielten Ergebnisse sollten jedoch repliziert und mit andern wissenschaftlichen Methoden weiter untersucht werden, *in vivo*, *in vito* oder *in silco.* Die Kausalität einer Gen-Radonexposition-Lungenkrebs-Beziehung ist damit nicht bewiesen.

4.8 Resultate der Einzelmarker-Assoziationsanalyse (AP 2.1c)

4.8.1 Signifikanz für den Interaktionseffekt GxE bzw. für den Joint-Test G/GxE

Keiner der häufigen Marker erzielte genomweite Signifikanz, weder für den Interaktionstest $\beta_{G\times E}$ noch für den Joint-Test $\beta_G \; und \; \beta_{G\times E}$. Sechs Marker erzielten einen p-Wert≤1x10^{-5} (suggestive Signifikanz), entweder im Einzel- oder im Joint-Test. Fünf dieser sechs Marker waren im jeweils anderen Test zumindest lokal signifikant (p<0,05, siehe Tabelle 26, Tabelle 27 und Abbildung 10, sowie Manhattan-Plots: Abbildung 11 und Abbildung 12).

Vier der sechs auffälligen Marker (**rs7705033**, **rs7735409**, **rs6891344** und **rs11747272**) liegen auf Chromosom 5q23.2 nahe beieinander, jedoch in 3 verschieden LD-Blöcken, und können dem Gen CSNK1G3 (casein kinase 1 gamma 3) zugeordnet werden. Für die beiden Marker **rs6891344** und **rs11747272** wurde ein Odds Ratio für Interaktion von OR_{GxE}~3,5 (hinsichtlich des selteneren Allels) mit einem p~1,5x10^{-5} geschätzt, wobei jeweils kein genetischer Haupteffekt OR_G~0,97 beobachtet wurde. Für die beiden Marker **rs7705033** und **7735409** wurde ein OR_{GxE}~0,25 (hinsichtlich des selteneren Allels, das entspricht einen OR_{GxE}~4 hinsichtlich des häufigeren Allels) mit einem p=0,8x10^{-5} bzw. 1x10^{-5} geschätzt, wobei ebenfalls jeweils kein genetischer Haupteffekt OR_G~0,94 beobachtet wurde.

CSNK1G3 kodiert ein für *Kasein-Kinase1* (CK1) relevantes Protein und wurde bislang nicht direkt mit Lungenkrebs in Verbindung gebracht; die CK1-Familie hingegen schon. *„Die Mitglieder der Casein-Kinase-1 (CK1)-Familie sind evolutionär hoch konserviert und werden in vielen Eukaryoten – von Hefen bis zum Menschen – exprimiert. Tierische CK1-Isoformen (α, β, γ, δ, ε) und deren Spleißvarianten regulieren zahlreiche zelluläre Prozesse. Zu diesen zählen unter anderem Membrantransportprozesse, tageszeitlicher Rhythmus, Zellzyklusprogression, Chromosomensegregation, Apoptose sowie Diffferenzierungsprozesse. Mutationen sowie Deregulation der Expression und Aktivität von CK1 stehen im Zusammenhang mit verschiedenen Krankheitsbildern, einschließlich neurodegenerativer Erkrankungen wie Alzheimer oder Parkinson, Schlafrhythmusstörungen sowie proliferativer Erkrankungen, insbesondere Tumorerkrankungen.“* [77]

Auffällig ist ebenso der Marker **rs10911725** auf Chromosom 1 mit einem geschätzten Odds Ratio für Interaktion von OR_{GxE}~0,21 (hinsichtlich des selteneren Allels, das entspricht einer OR_{GxE}~5 hinsichtlich des häufigeren Allels; p=5x10^{-6}). Dieser Marker kann keinem protein-kodierenden Gen zugeordnet werden, ist aber umgeben von Pseudo- und nicht-funktionalen Genen.

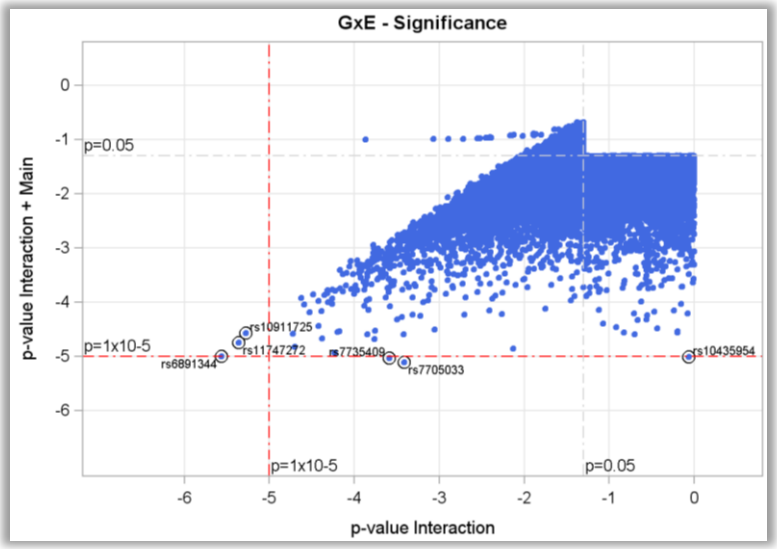
Der Marker **rs10435954** auf Chromosom 9 fällt zwar hinsichtlich eines geringen p-Werts im Joint-G/GxE-Test auf. Die Signifikanz kann aber als überprüfenswert angesehen werden, weder der genetische Haupteffekt ($p_G = 0{,}1890$) noch der Interaktionsterm ($p_{GxE} = 0{,}8691$; OR_{GxE}=0,95, 95%.CI: 0,57- 1,59) signifikant waren.

Tabelle 26 Signifikanz für den Interaktionseffekt GxE bzw. für den Joint-Test G/GxE: Übersicht

Signifikanz (significant at … level)	*Joint-G/GxE-Test*		
	nicht signifikant.	signifikant am 0.05-Level	signifikant am 1x10^{-5}-Level
Interaction GxE			
nicht signifikant.	438.857	10.261	1*
signifikant am 0.05-Level	9.677	12.807	2
signifikant am 1x10^{-5}-Level	--	3	--

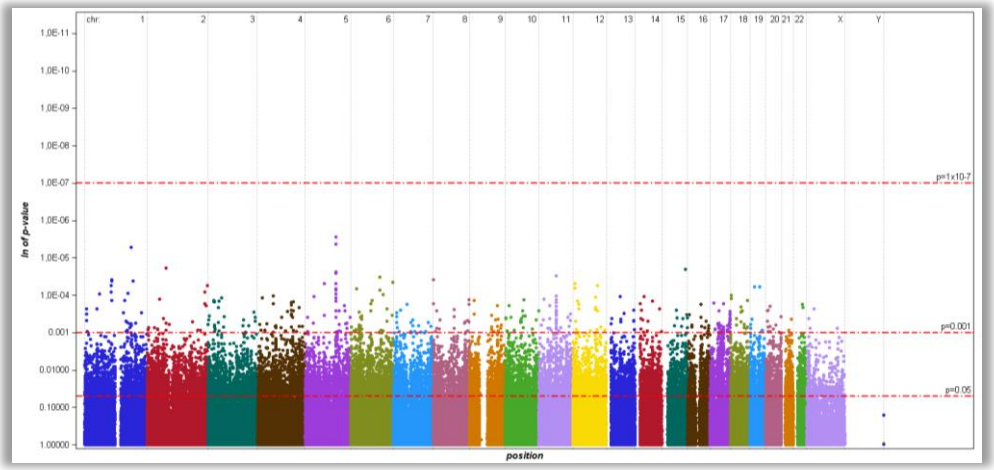
* überprüfenswert Signifikanz da $p_{G/GxE} > p_G$

Abbildung 10 Signifikanz für den Interaktionseffekt GxE bzw. für den Joint-Test G/GxE: Gegenüberstellung



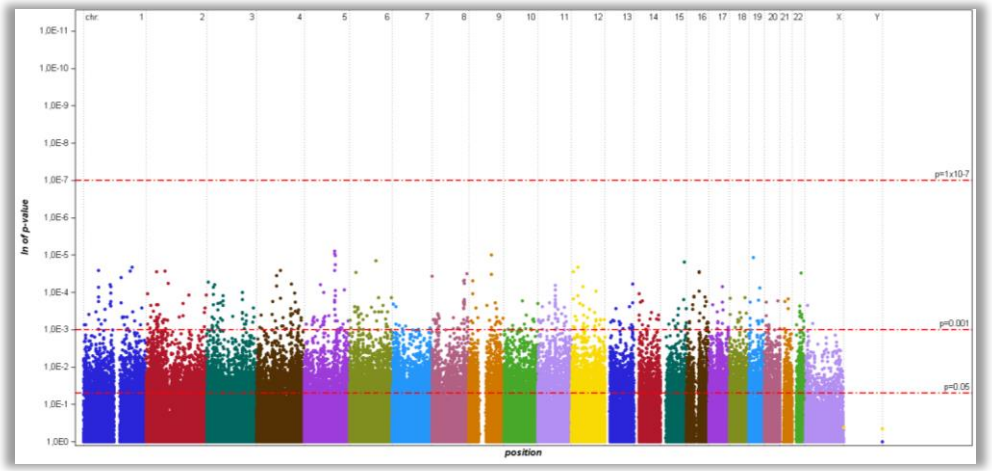
Marker mit mindestens suggestiver Signifikanz bei einem der beiden Tests wurden hervorgehoben

Abbildung 11 Manhattan-Plot: Signifikanz für den Interaktionseffekt GxE (je Marker)



Jeder Punkt entspricht einem Marker: horizontale Achse: Markerposition auf dem Genom (farblich getrennt die Chromosomen); vertikale Achse: natürlicher Logarithmus des p-Werts eines Signifikanztests eines Interaktionseffekt GxE

Abbildung 12 Manhattan-Plot: Signifikanz für den Joint-Test G/GxE (je Marker)



Jeder Punkt entspricht einem Marker: horizontale Achse: Markerposition auf dem Genom (farblich getrennt die Chromosomen); vertikale Achse: natürlicher Logarithmus des p-Werts eines Signifikanztests eines Interaktionseffekt GxE

Tabelle 27 Signifikanz für den Interaktionseffekt GxE bzw. für den Joint-Test G/GxE: ausgewählte Marker

Marker	Chr.	Gene	Position	Block	p-Wert		OR (95%-CI)		
					GxE	G/GxE	G	E	GxE
							(GxE Modell)		
rs10911725	1q25.3	--	185395182	5078	$5x10^{-6}$	$3x10^{-5}$	1,02 (0,89- 1,16)	6,70 (4,30- 10,4)	0,21 (0,11- 0,42)
rs7705033	5q23.2	--	122774785	33131	$1x10^{-4}$	$8x10^{-6}$	0,94 (0,82- 1,09)	4,98 (3,31- 7,49)	0,27 (0,13- 0,57)
rs7735409	5q23.2	--	122778705	33131	$3x10^{-4}$	$1x10^{-5}$	0,94 (0,81- 1,08)	5,03 (3,34- 7,57)	0,25 (0,11- 0,53)
rs6891344	5q23.2	CSNK1G3	123136656	33135	$3x10^{-6}$	$1x10^{-5}$	0,96 (0,84- 1,10)	1,57 (0,96- 2,55)	3,91 (2,18- 6,99)
rs11747272	5q23.2	CSNK1G3	123179990	33137	$4x10^{-6}$	$2x10^{-5}$	0,97 (0,86- 1,10)	1,23 (0,69- 2,19)	3,35 (1,98- 5,68)

chr … Chromosom; Position… Position am Chromosom [bp];
G: Haupteffekt des Genotyps; E: Haupteffekt der Strahlenexposition; GxE: Interaktion

4.8.2 Signifikanz gemäß Hybrid-2-Schritt (H2)-Verfahren nach Murcray et al.

4.8.2.1 Modell eines marginalen Haupteffekts des Genotyps (DxG-Modell, Screening)

Tabelle 28 Verteilung der p-Werte aus dem Modell eines marginalen Haupteffekts des Genotyps (DxG-Modell)

p-Wert	N	%
%gesamt	471.654	100%
<1x10^{-7}	5	0%
..x10^{-7}	2	0%
..x10^{-6}	17	0%
..x10^{-5}	71	0%
..x10^{-4}	542	0%
..x10^{-3}	5.089	1%
0,01-0,05	21.036	4%
>0,05	444.892	94%

Das Modell eines marginalen Haupteffekts des Genotyps (DxG-Modell) kann ohne Zufallsauswahl oder Gewichtung von nicht-exponierten Nicht-Wismut-Fällen geschätzt werden. Es konnten dabei 5 Marker mit einem $p<1x10^{-7}$ (genomweite Signifikanz), weitere 90 Marker mit einem $1x10^{-7}<p\leq1x10^{-5}$ (suggestive Signifikanz) gefunden werden (siehe Tabelle 28). Die genomweit signifikanten Marker liegen auf den Chromosomen 1,2,3 und 6; keine davon jedoch in einer bisher mit Lungenkrebs assoziierten genomischen Region (LK-Regionen) (siehe Tabelle 29 und Anhang Tabelle 75). Der Marker auf Chromosom 1 liegt nur knapp neben der LK-Region 1p31, die mit dem Adenokarzinom assoziiert ist. Der Marker auf Chromosom 6 liegt nur knapp neben der LK-Region 6p21_BAG6, die am stärksten mit dem Plattenepithelkarzinom (squamous cell carcinoma) assoziiert ist. Der Marker auf Chromosom 3 liegt nur knapp neben der LK-Region 3p28, die mit dem Adenokarzinom assoziiert ist.

Tabelle 29 Marker mit genomweit signifikanter Assoziation im Modell eines marginalen Haupteffekts des Genotyps (DxG-Modell)

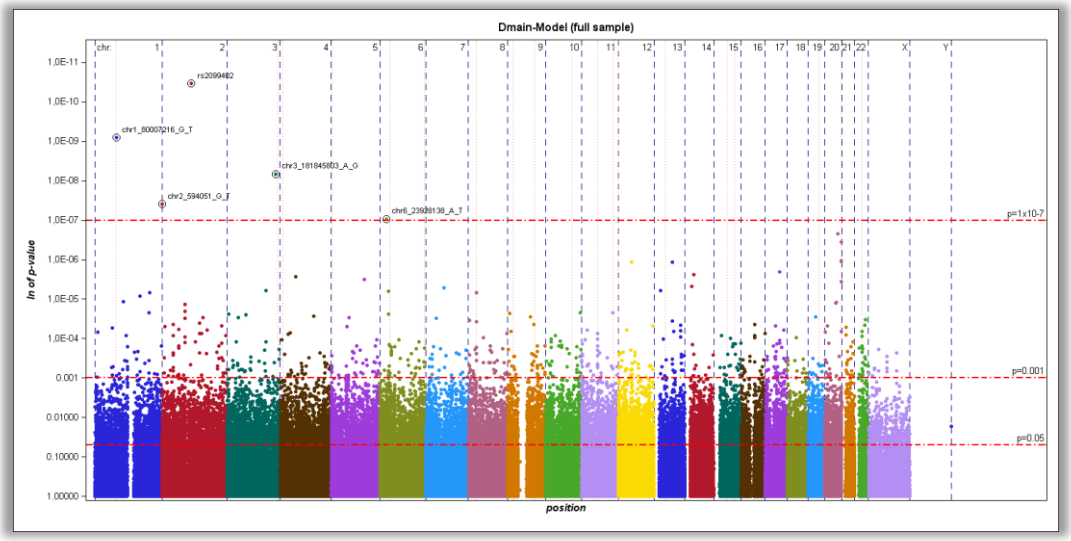
SNP	Chromosom	Position	OR (95%-CI)	p-Wert
rs2099402	2	110.442.280	0,02 (<0,01- 0,06)	$3.4x10^{-11}$
chr1_80007216_G_T	1	80.007.216	0,04 (0,01- 0,12)	$8.0x10^{-10}$
chr3_181845803_A_G	3	181.845.803	0,04 (0,01- 0,13)	$6.8x10^{-09}$
chr2_594051_G_T	2	594.051	0,10 (0,04- 0,24)	$3.9x10^{-08}$
chr6_23928138_A_T	6	23.928.138	0,13 (0,06- 0,28)	$9.3x10^{-08}$

Vergleicht man den Manhattan-Plot mit dem der Hauptanalyse des OncoArray-Konsortiums TRICL/ILCCO von McKay et al. [58], zeigen sich deutliche Unterschiede in den als mit Lungenkrebs assoziiert beobachteten Regionen und den erzielten Signifikanzen (siehe Abbildung 13 und Abbildung 14). McKay et al. konnten mit Lungenkrebs assoziierte Marker in 19 genomischen Regionen (LK-Regionen) identifizierten. Die teils sehr deutlichen Signifikanzen wurden aber erst in einer Meta-Analyse der OR-Schätzern aus der mit OncoArray typisierten Stichprobe mit den OR-Schätzern aus bereits zuvor existierenden LK-Studien (bestehend aus 29.863 Fällen und 55.586 Kontrollen, nicht-typisierte OncoArray-Marker wurden imputiert) erzielen. Sofern vergleichbar, konnten mit den Stichproben dieser Untersuchung und der von McKay et al. im Allgemeinen vergleichbare Effekte beobachtet werden. Damit lassen sich die Unterschiede zwischen beiden Untersuchungen auf die erzielte statistische Signifikanz zurückführen (nicht auf die geschätzte Richtung und Stärke der Assoziation), vor allem hervorgerufen durch unterschiedliche Fallzahlen.

Insgesamt wurden in dieser Untersuchung 29.291 Marker in diesen 19 LK-Regionen typisiert. Von 87 Markern konnten mit den zur Verfügung stehenden Stichproben keine Schätzergebnisse im DxG-Modell erzielt werden. Die meisten LK-Regionen beinhalten mindestens einen Marker mit einem p-Wert kleiner $1x10^{-3}$. Ausnahmen sind dabei die Regionen 1p31, 4p16, 11q12, 15q21 und 19q13 (siehe Tabelle 30).

Da das DxG-Modell nur dem Screening nach Regionen/Markern mit potentiell beobachtbarer GxE-Interaktion dient, können anstelle der aus den Daten ermittelten p-Werte im DxG-Modell **alternative** die von McKay et al. berechneten **p-Werte** herangezogen werden.

Abbildung 13 Manhattan-Plot: Genetischer Haupteffekt im Modell eines marginalen Haupteffekts des Genotyps (DxG-Modell)– mit OncoArray typisierter Stichprobe



Zum Vergleich …..

Abbildung 14 Manhattan-Plot: Genetischer Haupteffekt im Modell eines marginalen Haupteffekts des Genotyps (DxG-Modell) – Meta-Analyse McKay et al.

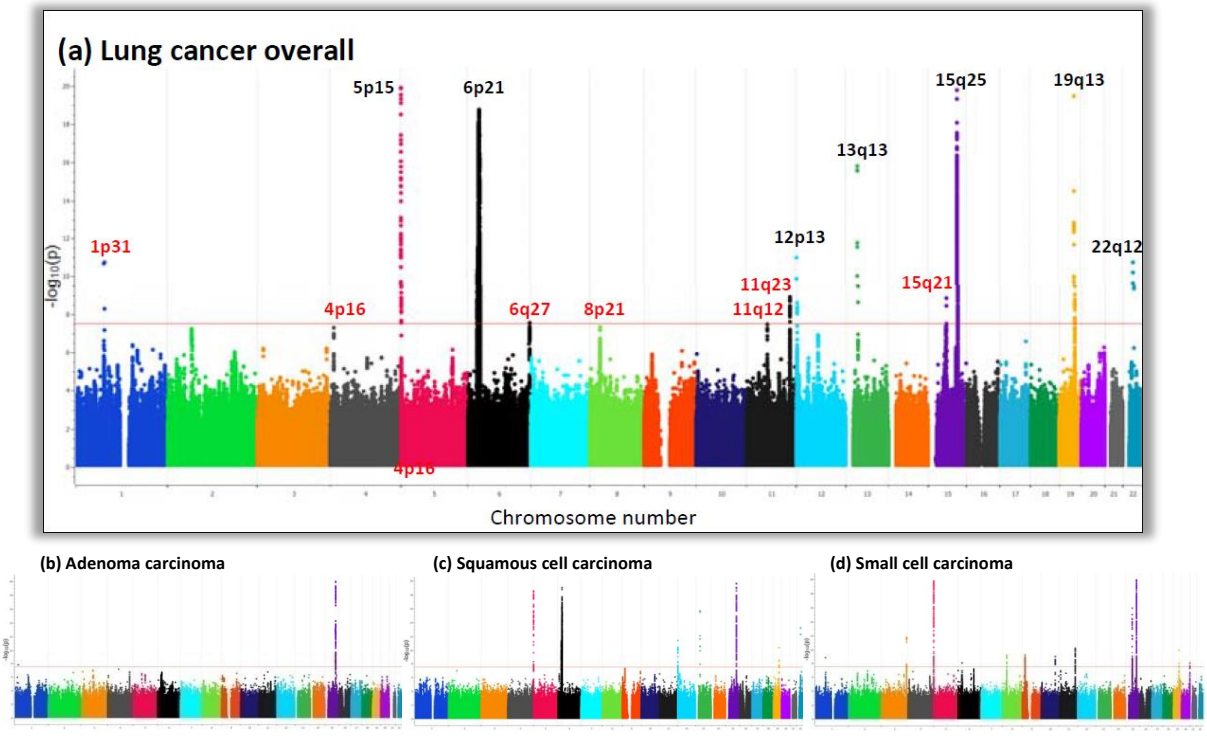


Tabelle 30 Verteilung der p-Werte aus dem Modell eines marginalen Haupteffekts des Genotyps (DxG-Modell) in genomischen Lungenkrebs-Regionen (gemäß McKay et al.)

LK-Regionen	LK Subtyp	gesamt	p-Wert <1x10^{-7}	..x10^{-7}	..x10^{-6}	..x10^{-5}	..x10^{-4}	..x10^{-3}	0,01-0,05	>0,05
1p31	LK gesamt	42								42
3q28	Adenoma-LK	51	1					2		48
4p16	LK gesamt	14								14
5p15_TERT	LK gesamt	3.583	11				5	32	130	3.405
6p21_BAG6	LK gesamt	7.333	27		1		14	93	320	6.878
6q27	LK gesamt	74						1	3	70
8p21	LK gesamt	220						2	4	214
9p21	Adenoma-LK	886	1				1	8	24	852
9q31	Plattenepithel	8.879	12			2	17	121	362	8.365
10q24	Adenoma-LK	2.396	7					26	99	2.264
11q12	LK gesamt	23						.	2	21
11q23	LK gesamt	43						2	7	34
12p13	LK gesamt	186						1	11	174
13q13	LK gesamt	2.847	21				4	21	109	2.692
15q21	LK gesamt	280							22	258
15q25_CHRNA3	LK gesamt	318	1					2	13	302
19q13	LK gesamt	75	1						3	71
20q13	Plattenepithel	1.929	3	1	3	1		15	70	1.836
22q12	LK gesamt	112	2					2	1	107
	gesamt	**29.291**	**87**							

4.8.2.2 *Korrektur für multiples Testen basierend auf der Anzahl LD-Blöcke (beobachtete p-Werte des DxG-Modells)*

Die Wahl des Signifikanz-Parameters ρ nimmt nur geringfügig Einfluss auf die Identifikation der signifikantesten Marker (siehe Tabelle 70 im Anhang). Bei ρ=0,99 wurde der geringste nach Bonferroni korrigierte p-Wert von 0,28315 eines Interaktionseffekts (GxE-Effekt) für einen Marker erzielt.

Der betreffende Marker ist **rs6891344** auf Chromosom 5q23.2, der bereits zuvor als suggestiv signifikant erkannt wurde. Der für multiples Testen (mT) korrigierte p-Wert beträgt p_{mT}=0,28315. Dieser entspricht einem rückskalierten p-Werte von $p^* = 2{,}7 \times 10^{-6}$
($p^* = 0{,}28315/103.983\ \text{LD} - \text{Blöcke} = 2{,}7 \times 10^{-6}$).

Auch durch das H2-Verfahren verfehlt dieser Marker damit genomweite Signifikanz nur knapp. Ebenso auffällig ist der benachbarte Marker **rs11747272** mit p_{mT}=0,45036. Dieser p-Wert entspricht einem rückskalierten p-Werte von $p^* = 0{,}45036\ /103.983\ \text{LD} - \text{Blöcke} = 4{,}3 \times 10^{-6}$.

Unter der Wahl von ρ=0,99 zeigen 7 weitere Marker auf den Chromosomen 1, 2, 3 (2x), 12 (12x) und 17 Hinweise hinsichtlich einer GxE-Interaktion nach dem Kriterium p_{mT}<1. Dieser Grenzwert entspricht einem unkorrigierten p-Wert von $p^* = 1\ /103.983\ \text{LD} - \text{Blöcke} \sim 1 \times 10^{-6}$ Bei einem beliebigen ρ zwischen 0,5 und 1-10^{-20} können für zwei weitere Marker von Chromosom 1 und 7 Hinweise auf GxE-Interaktionen gewonnen werden (siehe Tabelle 31 und Tabelle 32).

4.8.2.3 *Korrektur für multiples Testen basierend auf der Anzahl LD-Blöcke (p-Werte aus der Datenauswertung von McKay et al.)*

Ersetzt man die Ergebnisse des DxG-Modells durch die auf der weit umfangreicheren Analyse beruhenden p-Werte von McKay et al., nimmt die Wahl des Signifikanz-Parameters ρ deutlich Einfluss auf die Identifikation des signifikantesten Markers (siehe Tabelle 71). Bei ρ=1-1x10^{-16} (also einer nahezu kompletten Korrektur für genomweites Testen im D-Modell und kaum Korrektur im E-Modell) wird der geringste nach Bonferroni mT-korrigierte p-Wert von p_{mT}=0,03856 eines GxE-Effekts für einen Marker erzielt.

Dies betrifft den Marker **rs12440014**, für den damit ein genomweit signifikanter GxE-Effekt gezeigt werden kann. Der mT-korrigierte p-Wert von p_{mT}=0,03856 entspricht einem rückskalierter p-Werte von $p^* = 0{,}03856 / 103.983\ \text{LD} - \text{Blöcke} = 3{,}7 \times 10^{-7}$

Dieser, wie fünf weitere auffällige Marker liegen auf dem Chromosom 15q25.1 nahe oder innerhalb des Gens CHRNB4 (Cholinergic Receptor Nicotinic Beta 4 Subunit). Diese LK-Region ist als mit Lungenkrebs assoziiert bereits (bestens) bekannt, jedoch wurde die stärkste genetische Assoziation etwa 69 kb upstream für das Gen CHRNA5 beobachtet (OR_G=1,29; p=3,6x10^{-101}). Die Assoziation des Marker **rs12440014** zu LK war in der Analyse von McKay mit p=1,6x10^{-51} (OR=0,81) ebenso signifikant.

Für den Marker **rs12440014** wurde eine OR_{GxE}=0,26 (hinsichtlich des selteneren Allels, das entspricht einer OR_{GxE}~4 hinsichtlich des häufigeren Allels) mit einem p=0,00117 geschätzt, wobei kein genetischer Haupteffekt OR_G=0,99 beobachtet wurde. Die Signifikanz des GxE-Effekts wird also erst durch das massive Vorfiltern von Markern mit dem marginalen Haupteffekts des Genotyps (in dem Verfahren von Murcray et al.) sichtbar.

Für weitere 5 Marker aus anderen LD-Blöcken in dieser genomischen LK-Region konnten Hinweise auf eine GxE-Interaktion gefunden werden. Diese Marker liegen bis zu 11 kb up- bzw. 33 kb downstream von rs12440014 (siehe Tabelle 33)**.**

Auffällig (auch mit unter Vorfiltern von Markern) sind die bereits erwähnten Marker auf Chromosom 5, rs6891344 und rs11747272, sowie der Marker rs10911725 von Chromosom 1.

Tabelle 31 **Signifikante Marker gemäß Hybrid-2-Schritt (H2)-Verfahren von Murcray et al. bei ρ=0.99**

Marker	Chr.	Gen	Position	LD-Block	p-Wert[1]		OR (95%-CI)			p-Wert[2] (H2) (ρ=0.99)
					GxE	G/GxE	G	E	GxE	
							(GxE Modell)			
rs10911725	1q25.3	--	185395182	5078	$5.3x10^{-6}$	$2.7x10^{-5}$	1.02 (0.89- 1.16)	6.70 (4.30- 10.4)	0.21 (0.11- 0.42)	0.55151
rs1437730	2q37.3	--	240632788	15988	$5.5x10^{-5}$	$1.2x10^{-4}$	1.00 (0.88- 1.13)	7.98 (4.59- 13.8)	0.33 (0.19- 0.57)	0.66810
rs9825051	3p24.3	ZNF385D	21799575	17185	0.05101	$4.5x10^{-4}$	5.14 (1.87- 14.1)	5.27 (1.39- 19.9)	0.33 (0.10- 1.02)	0.25763
rs9842091	3p22.2	SLC22A13	38307510	17806	0.11504	$7.7x10^{-4}$	13.0 (2.35- 71.9)	2.22 (1.01- 4.88)	0.22 (0.03- 1.49)	0.58102
rs6891344	5q23.2	--	123136656	33135	$2.7x10^{-6}$	$1.0x10^{-5}$	0.96 (0.84- 1.10)	1.57 (0.96- 2.55)	3.91 (2.18- 6.99)	0.28315
rs11747272	5q23.2	--	123179990	33137	$4.3x10^{-6}$	$1.8x10^{-5}$	0.97 (0.86- 1.10)	1.23 (0.69- 2.19)	3.35 (1.98- 5.68)	0.45036
rs7970379	12p13.31	ACSM4	7451788	68621	$6.3x10^{-5}$	$2.8x10^{-5}$	0.91 (0.73- 1.13)	2.45 (1.70- 3.54)	10.8 (3.29- 35.8)	0.77448
rs7302538	12p13.31	CD163L1	7499479	68623	$4.8x10^{-5}$	$1.5x10^{-4}$	0.94 (0.76- 1.18)	2.45 (1.69- 3.53)	12.9 (3.66- 45.4)	0.59377
rs9895796	17q22	HLF / MMD	53412915	88292	0.14432	$1.8x10^{-4}$	4.18 (1.60- 10.9)	5.00 (1.16- 21.4)	0.46 (0.16- 1.32)	0.72889

[1] unkorrigierte p-Werte; [2] für multiples Testen korrigierte p-Werte
G: Haupteffekt des Genotyps; E: Haupteffekt der Strahlenexposition; GxE: Interaktion

Tabelle 32 **Weitere signifikante Marker gemäß Hybrid-2-Schritt (H2)-Verfahren von Murcray et al. bei beliebigem ρ**

Marker	Chr.	Gen	Position	LD-Block	p-Wert[1]		OR (95%-CI)			min. p-Wert[2] H2-methods ρ beliebig
					GxE	G/GxE	G	E	GxE	
							(GxE Model)			
rs4545364	1p21.1	--	107125001	3373	0.83118	0.83201	2.37 (0.96- 5.87)	1.85 (0.59- 5.72)	1.11 (0.40- 3.09)	0.83201
rs9692033	7p21.3	LOC105375150	10245417	43207	0.00866	0.67324	1.04 (0.35- 3.06)	2.93 (1.22- 7.00)	0.09 (0.01- 0.56)	0.67324

[1] unkorrigierte p-Werte; [2] für multiples Testen korrigierte p-Werte
G: Haupteffekt des Genotyps; E: Haupteffekt der Strahlenexposition; GxE: Interaktion

Tabelle 33 Signifikante Marker gemäß Hybrid-2-Schritt (H2)-Verfahren von Murcray et al. bei ρ=1-1x10^{-16}) / p-Werte des DxG-Modells von McKay et al.

Marker	Chr.	Gen	Position	LD-Block	p-Wert[1]		OR (95%-CI)			p-Wert[2] (H2)
					GxE	G/GxE	G	E	GxE	
							(GxE Modell)			1-ρ=1x10^{-16}
rs10911725	1q25.3	--	185395182	5078	5.3x10^{-6}	2.7x10^{-5}	1.02 (0.89- 1.16)	6.70 (4.30- 10.4)	0.21 (0.11- 0.42)	0.55151
rs4635969	5		1308552	28958	0.01147	0.01290	0.93 (0.80- 1.07)	3.98 (2.70- 5.85)	0.37 (0.17- 0.81)	0.37843
rs6891344	5q23.2	--	123136656	33135	2.7x10^{-6}	1.0x10^{-5}	0.96 (0.84- 1.10)	1.57 (0.96- 2.55)	3.91 (2.18- 6.99)	0.28315
rs11747272	5q23.2	--	123179990	33137	4.3x10-06	1.8x10^{-5}	0.97 (0.86- 1.10)	1.23 (0.69- 2.19)	3.35 (1.98- 5.68)	0.45036
rs6495309	15q25.1	CHRNA3	78915245	82002	0.00723	0.02320	0.99 (0.87- 1.13)	4.05 (2.75- 5.98)	0.35 (0.16- 0.76)	0.23866
rs28534575	15q25.1	CHRNB4	78923845	82002	0.00595	0.02075	1.00 (0.89- 1.12)	4.15 (2.80- 6.14)	0.36 (0.17- 0.75)	0.19635
rs12440014	**15q25.1**	**CHRNB4**	**78926726**	**82003**	**0.00117**	**0.00445**	**0.99 (0.88- 1.12)**	**4.43 (3.00- 6.55)**	**0.26 (0.11- 0.60)**	**0.03856**
rs1316971	15q25.1	CHRNB4	78930510	82005	0.00522	0.00593	0.97 (0.84- 1.13)	4.09 (2.78- 6.02)	0.32 (0.14- 0.72)	0.17223
rs17487514	15q25.1	CHRNB4	78953785	82008	0.00705	0.01529	1.02 (0.89- 1.17)	1.81 (1.06- 3.07)	2.01 (1.19- 3.39)	0.23249
rs6495314	15q25.1	RPL18P11	78960529	82008	0.01448	0.02987	1.02 (0.91- 1.13)	1.62 (0.86- 3.05)	1.87 (1.12- 3.12)	0.47788

[1] unkorrigierte p-Werte; [2] für multiples Testen korrigierte p-Werte
G: Haupteffekt des Genotyps; E: Haupteffekt der Strahlenexposition; GxE: Interaktion

4.9 Resultate der Multimarker-Assoziationsanalyse

Die Signifikanz einer Interaktion wurde mit vier Modelansätzen beurteilt. Es wurden entweder alle Marker eines LD-Blocks in das Analysemodell als Haupt- und Interaktionsterm aufgenommen (*taxatives Modell*), oder nur eine Auswahl an Markern gemäß dem AIC-Kriterium (*AIC-Modell*). In beiden Modellen wurden je LD-Block sowohl ein Interaktionstest $\beta_{G\times E}$ als auch der joint-Test β_G V $\beta_{G\times E}$ durchgeführt.

Die so erzielten p-Werte können mit einfacher Bonferroni-Korrektur für multiples Testen hinsichtlich statistischer Signifikanz bewertet werden. Darüber hinaus wurden die p-Werte der Interaktionstests $\beta_{G\times E}$ beider Multimarker-Modelle (taxatives und AIC-Modell) gemäß dem Hybrid-2-Schritt-Verfahren nach Murcray, et al., 2011 [62] hinsichtlich statistischer Signifikanz bewertet. Dabei wurden im Screening sowohl die p-Werte marginaler Haupteffekte aus dem DxG-Modell oder die ermittelten p-Werte von McKay et al. berücksichtigt.

Tabelle 34 Übersicht der Modelansätze und Bewertungsmethoden der Signifikanz für Multimarker-Modelle

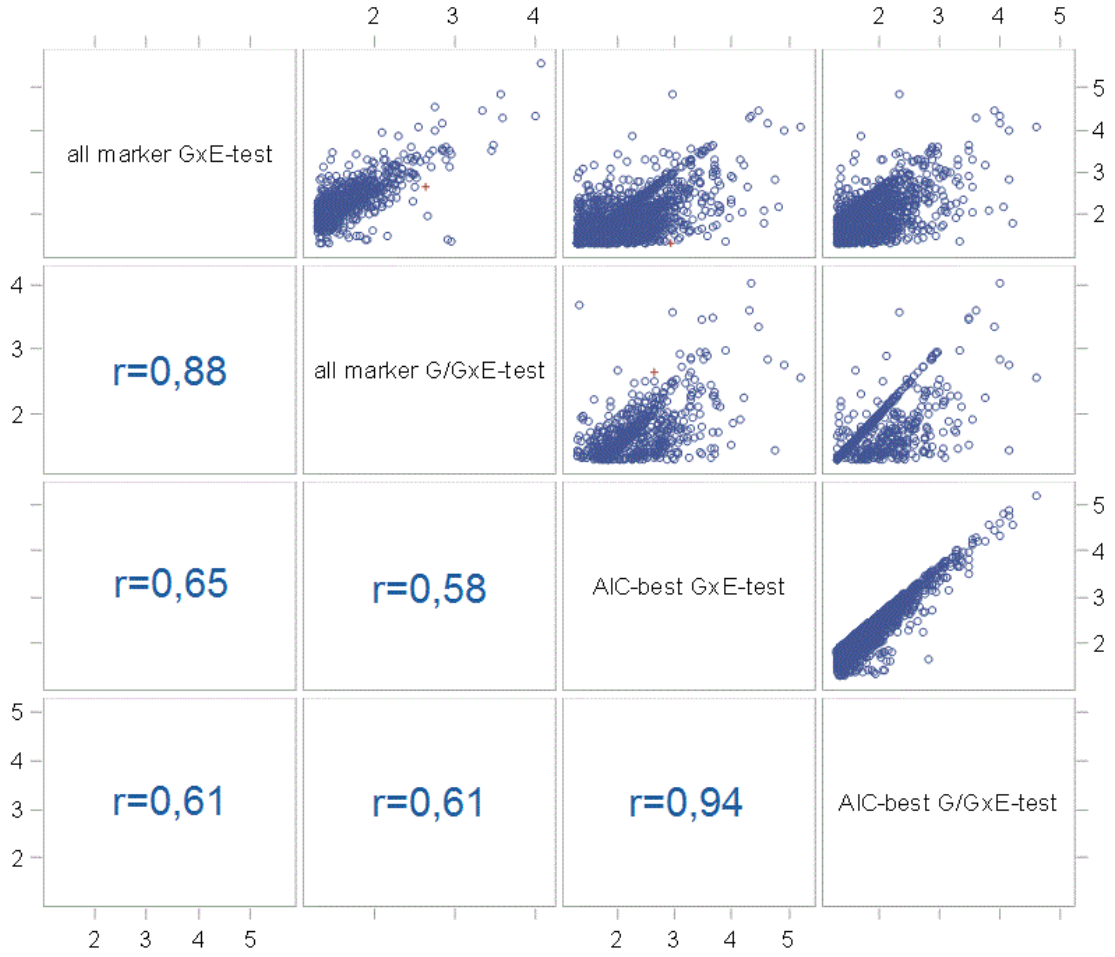
Marker je LD-Block	Test	DxG-Screening
Einfache Bonferroni-Korrektur		
alle Marker	Interaktionstest $\beta_{G\times E}$	--
Auswahl gemäß AIC	Interaktionstest $\beta_{G\times E}$	--
alle Marker	joint-Test $\beta_G \| \beta_{G\times E}$	
Auswahl gemäß AIC	joint-Test $\beta_G \| \beta_{G\times E}$	
Hybrid-2-Schritt Verfahren nach Murcray, et al., 2011 [62]		
alle Marker	Interaktionstest $\beta_{G\times E}$	DxG-Modell
alle Marker	Interaktionstest $\beta_{G\times E}$	p-Werte von McKay et al.
Auswahl gemäß AIC	Interaktionstest $\beta_{G\times E}$	DxG-Modell
Auswahl gemäß AIC	Interaktionstest $\beta_{G\times E}$	p-Werte von McKay et al.

4.9.1 Signifikanz gemäß GxE bzw. für G/GxE (joint test) mit einfacher Bonferroni-Korrektur

4.9.1.1 Korrelation der Signifikanzen der verschiedenen Modelle

P-Werte je LD-Block wurden für vier Modelansätzen (Interaktionstest bzw. Joint-Test; jeweils taxatives oder AIC-Modell) berechnet. Die Korrelation der –log(p-Werte) ist zwischen den Tests innerhalb derselben Modelle größer (r=0,88 bzw. 0,94) als zwischen den Modellen für jeweils denselben Test (r=0,65 bzw. 0,61). Die Bestimmung der Signifikanz nach Marker-Auswahl (AIC-Modelle) zeigt eine sehr hohe Übereinstimmung (r=0,94) zwischen den Test auf. Die Übereinstimmung ist etwas geringer (r=0,88), wenn alle Marker eines LD-Blocks in das Analysemodell aufgenommen werden.

Abbildung 15 Korrelation der p-Werte von vier Multimarker-Modellen/Tests



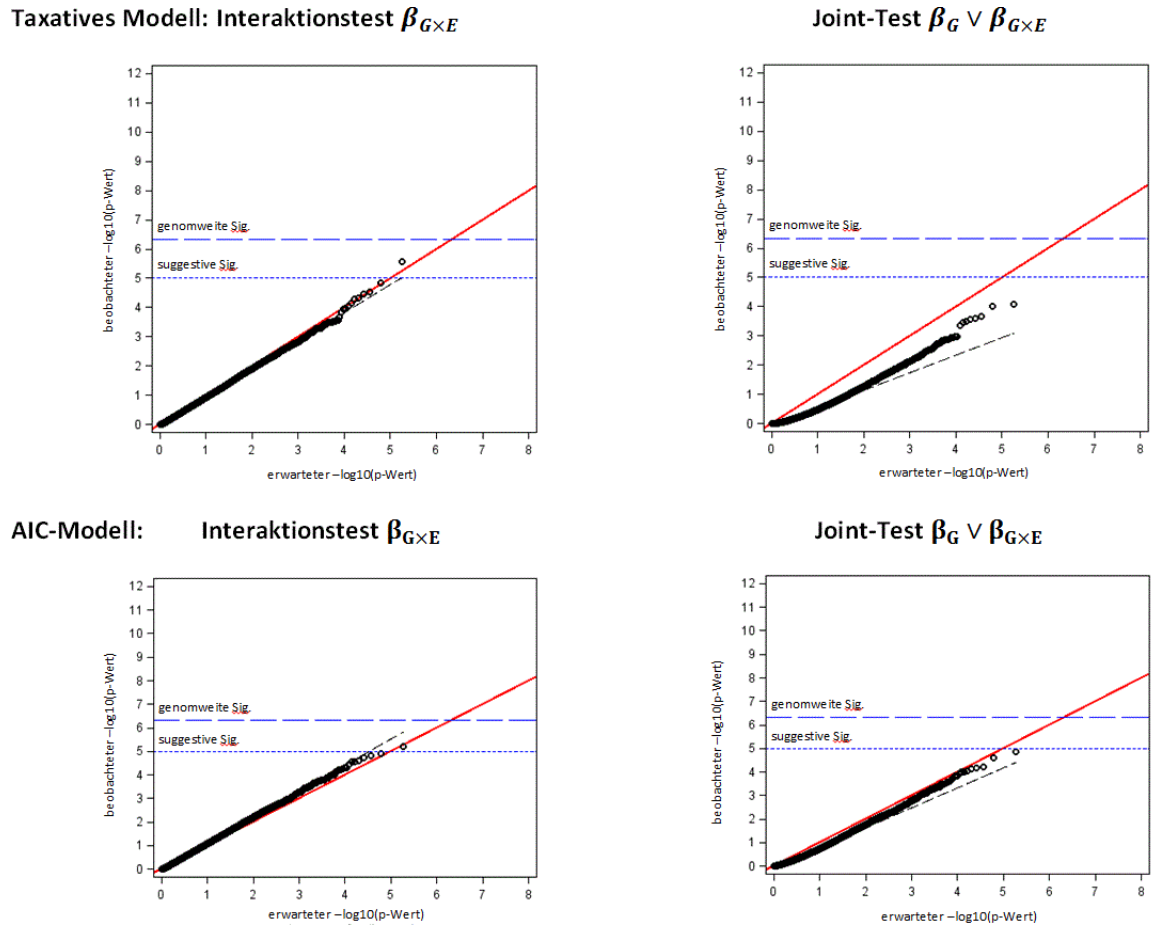
taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium;
p-Werte sind als –ln(p-Wert) dargestellt. r: Korrelationskoeffizient;
Marker mit fragwürdiger Signifikanz sind als Kreuz dargestellt.

4.9.1.2 QQ-Plot der Signifikanzen der verschiedenen Modelle

In QQ-Plots werden die erzielten Verteilungen der p-Werte der unter der Null-Hypothese (keinerlei Interaktion) theoretisch erwarteten Verteilung gegenübergestellt. Unter der Annahme, dass nur für wenige LD-Blöcke tatsächlich eine Gen-Strahlungs-Interaktion besteht, sollten sich die dargestellten Punkte auf der Diagonalen befinden. Bei konservativen Tests (Signifikanzen werden tendenziell unterschätzt) liegen die Punkte unterhalb, bei zu liberalen Test (Signifikanzen werden tendenziell zu häufig ausgewiesen) oberhalb der Diagonale.

Aus den QQ-Plots der vier Ansätze lässt sich erkennen, dass der Joint-Test $\beta_G \vee \beta_{G\times E}$ im taxativen Modell deutlich, im AIC-Modell leicht konservativ ist. Die Testergebnisse der Joint-Tests werden daher nicht weiter dargestellt. Die Verteilung der p-Werte des Interaktionstests $\beta_{G\times E}$ lassen keine Verzerrung erkennen.

Abbildung 16 QQ-Plots der p-Werte von vier Multimarker-Modellen/Tests



taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium;
p-Werte sind als –ln(p-Wert) dargestellt
gestrichelte Linie: Regressionsgerade Aller p-Werte <0.5

4.9.1.3 Ergebnisse des Interaktionstests $\beta_{G\times E}$ je LD-Block

Von 91.440 der 103.983 definierten LD-Blöcke (88%) konnten Tests durchgeführt werden. 2.046 LD-Blöcke (2%) enthielten keine valide typisierten Marker, 10.497 LD-Blöcke (10%) enthielten ausschließlich seltene Varianten (MAF<1%). Genomweit signifikante Interaktion konnte mit dem Interaktionstest $\beta_{G\times E}$ für keinen LD-Block beobachtet werden, weder im taxativen noch im AIC-Modell.

Gemäß taxativen Modell konnte für den **LD-Block Nr. 91734** auf Chromosom 18 eine suggestive Signifikanz ($p=2{,}6\times10^{-6}$) erzielt werden (siehe Abbildung 17). Dieser LD-Block beinhaltet die fünf

Marker rs1346830, rs11659206, rs7237496, rs9946324 und rs8091054 und ist umgeben von vier Genen: Zwei nicht-charakterisierte, nicht-codierende RNA Gene (LOC107985187, LOC105372156), dem rück-transkribierten Pseudo-Gen CTBP2P3 / ENSG00000267153 und dem transkribierten Pseudogen RP11-325K19.2 / ENSG00000267382.

Gemäß AIC-Modell konnte für den **LD-Block Nr. 33137** auf Chromosom 5 eine suggestive Signifikanz ($p=6{,}3x10^{-6}$) erzielt werden (siehe Abbildung 18). Dieser LD-Block beinhaltet die Marker rs3909326, rs6876166, rs12514988, rs11747272 sowie rs10052257 und liegt nahe dem Gen CSNK1G3. Eine suggestiv-signifikante Interaktion mit dem Marker rs11747272 wurde bereits im Kapitel 4.8.1 dargelegt.

Im Weitern zeigen jeweils benachbarte LD-Blöcke der Chromosomen 11 und 12 im Modell mit Marker-Auswahl eine Tendenz hin zu suggestiver Signifikanz.

Abbildung 17 Manhattan-Plot: Genetische Interaktion im Modell mit allen Markern je LD-Block

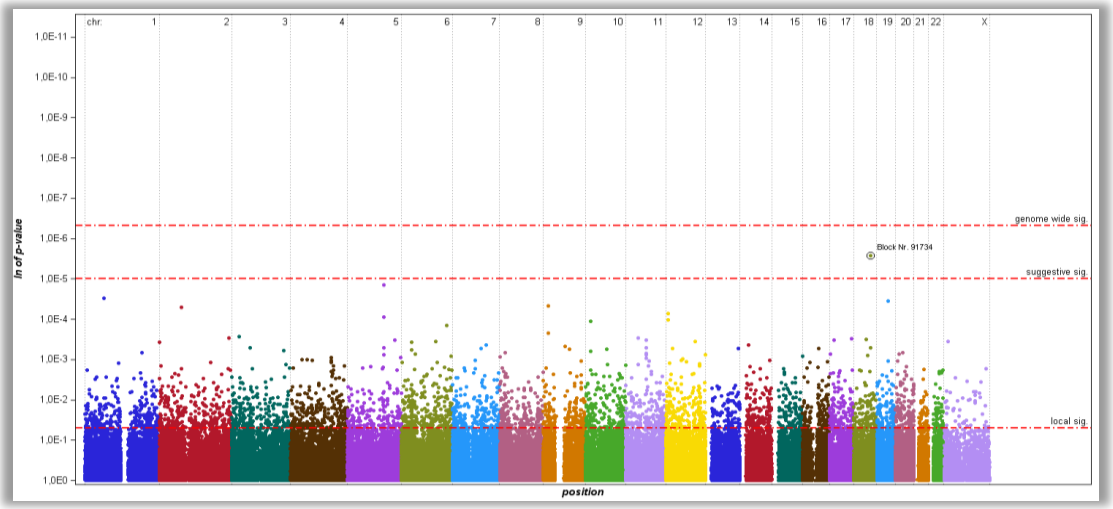
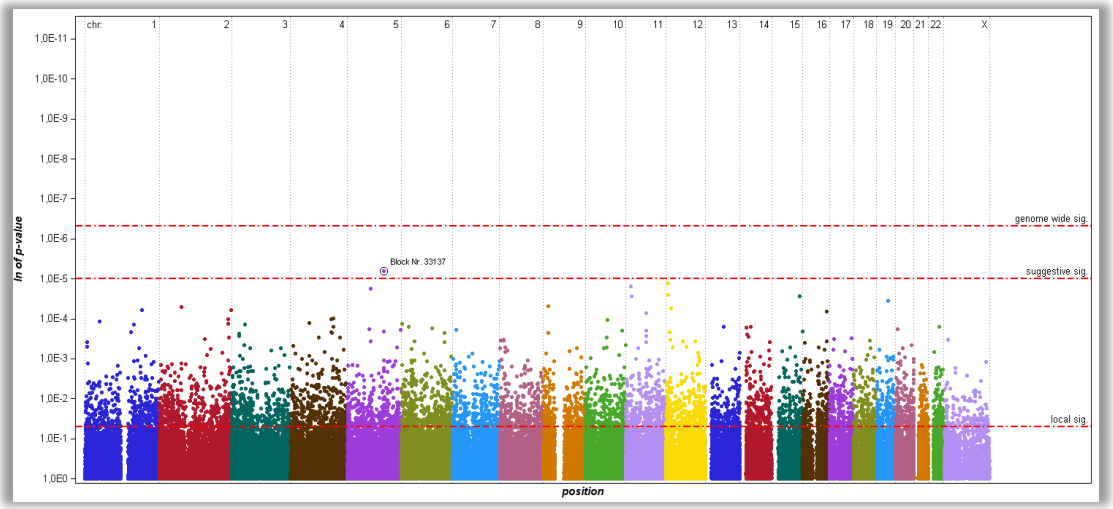


Abbildung 18 Manhattan-Plot: Genetische Interaktion im Modell mit Marker-Auswahl je LD-Block



4.9.2 Signifikanz gemäß Hybrid-2-Schritt (H2)-Verfahren im taxativen Modell

Das Hybrid-2-Schritt-Verfahren nach Murcray, et al., 2011 [62] erlaubt die Korrektur für multiples Testen der Interaktion GxE an der Signifikanzen der marginalen Haupteffekte (für G und für E). Dabei wird die Bedeutung der beiden marginalen Haupteffekte anhand eines Parameters ρ untereinander gewichtet.

4.9.2.1 *Korrektur für multiples Testen basierend auf der Anzahl LD-Blöcke (D-Screening durch das DxG-Modell)*

Die Wahl des Signifikanz-Parameters ρ nimmt nur geringfügig Einfluss auf Identifikation der signifikantesten Marker (siehe Tabelle 72 und Abbildung 42 im Anhang). Bei ρ=0,999 wurde der geringste nach Bonferroni korrigierte p-Wert eines Interaktionseffekts (p_{GxE}=0,2715) für einen LD-Block erzielt. Das entspricht suggestiver Signifikanz.

Abbildung 19 Manhattan-Plot: Signifikanz gemäß Hybrid-2-Schritt (H2)-Verfahren bei ρ=0.999 für genetische Interaktion je LD-Block im Modell mit allen Markern / D-Screening durch DxG-Modell

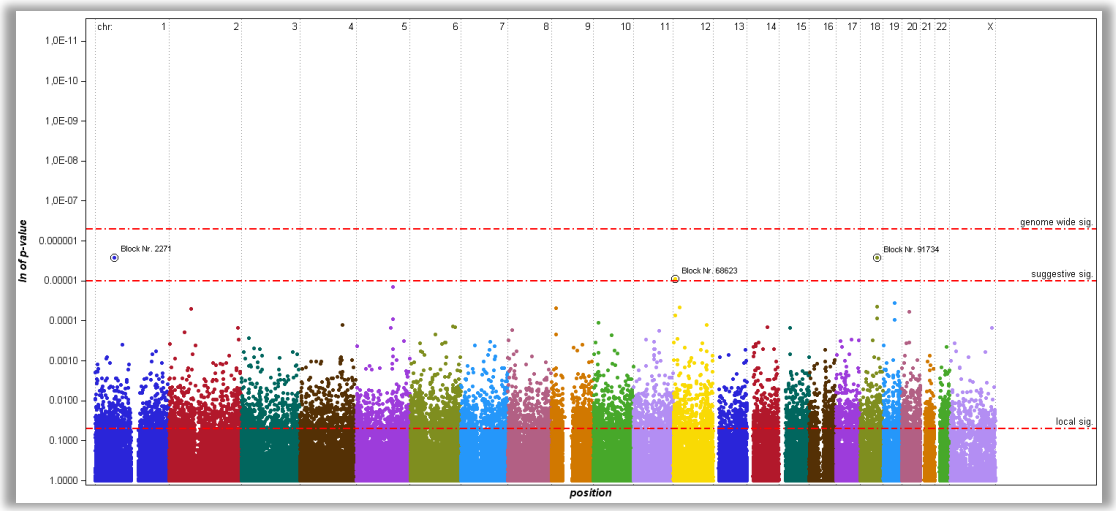


Tabelle 35 Liste suggestiv signifikanter LD-Blöcke gemäß Hybrid-2-Schritt (H2)-Verfahren bei beliebigem ρ im Modell mit allen Markern / D-Screening durch DxG-Modell

LD-Block	Chr.	Gen	ρ für min. p-Wert	p-Wert GxE	p-Wert G/GxE	min. p-Wert H2-korrigiert	rückskalierter p-Wert*
2271	1p21.3	UBE2U	0,999	$3{,}0\times10^{-5}$	0,0018	0,2715	$2{,}6\times10^{-6}$
68623	12p13.31	CD163L1 LOC101927882 ACSM4	0,999	0,0001	0,0017	0,9352	$9{,}0\times10^{-6}$
91734	18q21.32	LOC107985187 LOC105372156 RP11-325K19.1 CTBP2P3	0,500	$2{,}6\times10^{-6}$	$8{,}3\times10^{-5}$	0,2726	$2{,}6\times10^{-6}$

* nach H2-Korrektion, rückskalierter p-Werte;
GxE Gen x Radon – Interaktion; G/GxE (joint test) Haupteffekt und Interaktion

Insgesamt zeigten drei LD-Blöcke suggestive Signifikanzen über eine sehr weite Spanne für ρ von 0.5 bis $1\text{-}10^{-17}$. Dies sind die **LD-Blöcke Nr. 2271** auf Chromosom 1 (rückskalierter p-Wert: $p^* = 2{,}6 \times 10^{-6}$), **Nr. 68623** auf Chromosom 12 ($p^* = 9{,}0 \times 10^{-6}$) und **Nr. 91734** auf Chromosom 18 ($p^* = 2{,}6 \times 10^{-6}$).

Der **LD-Block Nr. 2271** enthält die Marker rs7545208, rs77199888, rs2806532, rs7526950, rs17126246, rs2029868, rs10789152, rs705540, rs11588217 und rs705551. Diese gruppieren sich um das Gene UBE2U (ubiquitin conjugating enzyme E2 U), einem Gen aus der Familie UBE2 (http://www.genenames.org/cgi-bin/genefamilies/set/102: Ubiquitin conjugating enzymes E2: „Ubiquitin-conjugating enzymes, also known as E2 enzymes and more rarely as ubiquitin-carrier enzymes, perform the second step in the ubiquitination reaction that targets a protein for degradation via the proteasome. The ubiquitination process covalently attaches ubiquitin, a short protein of 76 amino acids, to a lysine residue on the target protein."). Diese Gen-Familie ist mitentscheidend über „Leben oder Tod" eines Proteins.[78] Sie steht auch in Verbindung mit der Regulierung von

Signalwegen, die für das native Immunsystem von Bedeutung sind, sowie mit der Antigenpräsentation auf MHC Klasse I Molekülen.

Der **LD-Block Nr. 68623** enthält die Marker rs11051842, rs10844153 und rs7302538. Diese liegen innerhalb des Gens CD163L1 (CD163 molecule like 1) und neben dem nicht-kodierenden RNA-Gen LOC101927882.

4.9.2.2 Korrektur für multiples Testen basierend auf der Anzahl LD-Blöcke (D-Screening durch p-Werte des DxG-Modells von McKay et al.)

Die erzielten Signifikanzen unterscheiden sich nicht von jener, bei der das D-Screening des Hybrid-2-Schritt (H2)-Verfahren auf Basis des DxG-Modells erfolgte (sieh Kapitel 4.9.2.1).

4.9.3 Signifikanz gemäß Hybrid-2-Schritt (H2)-Verfahren im Modell mit Marker-Auswahl je LD-Block

4.9.3.1 Korrektur für multiples Testen basierend auf der Anzahl LD-Blöcke (D-Screening durch das DxG-Modell)

Die Wahl des Signifikanz-Parameters ρ nimmt nur geringfügig Einfluss auf die Identifikation der signifikantesten Marker (siehe Tabelle 73 und Abbildung 43 im Anhang). Bei ρ=0,9999 wurde der geringste nach Bonferroni korrigierte p-Wert eines Interaktionseffekts (p_{GxE}=0,0214) für einen LD-Block erzielt. Das entspricht einer genomweiten Signifikanz.

Abbildung 20 Manhattan-Plot: Signifikanz gemäß Hybrid-2-Schritt (H2)-Verfahren bei ρ=0.5 für genetische Interaktion je LD-Block im Modell mit Marker-Auswahl / D-Screening durch DxG-Modell

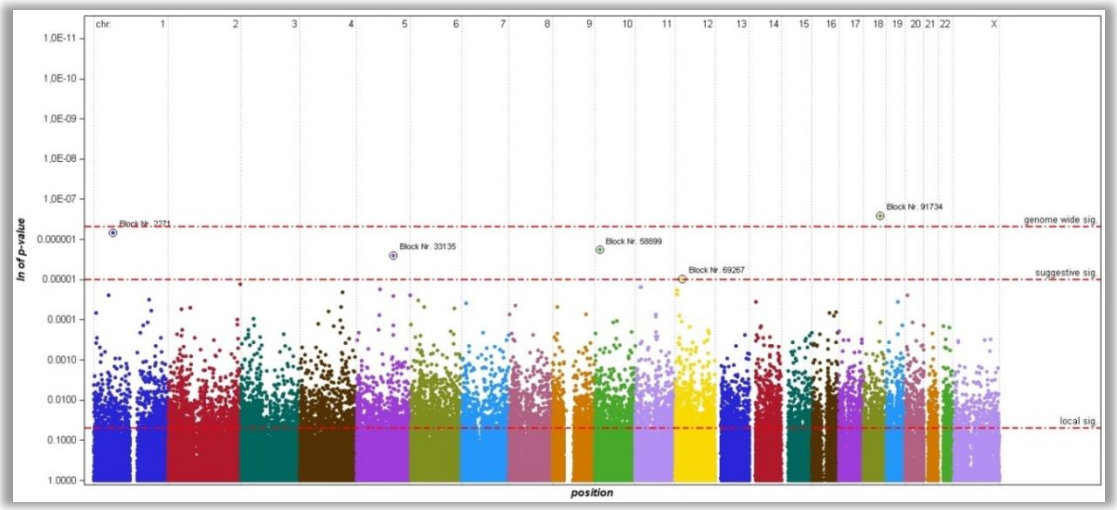


Tabelle 36 Liste suggestiv signifikanter LD-Blöcke gemäß Hybrid-2-Schritt (H2)-Verfahrens bei beliebigem ρ im Modell mit Marker-Auswahl / D-Screening durch DxG-Modell

LD-Block	Chr.	Gen	ρ für min. p-Wert	p-Wert GxE	p-Wert G/GxE	min. p-Wert H2-korrigiert	rückskalierter p-Wert*
2271	1p21.3	UBE2U	0,9999	$3{,}2\times10^{-6}$	$7{,}5\times10^{-6}$	0,0563	$5{,}4\times10^{-7}$
33135	5q23.2	CSNK1G3 LINC01170	0,5	$2{,}5\times10^{-6}$	$6{,}5\times10^{-6}$	0,2585	$2{,}5\times10^{-6}$
58899	10p13	CUBN	0,5	0,000013	0,000025	0,1878	$1{,}8\times10^{-6}$
69267	12p12.1	SOX5 MIR920	0,5	0,000071	0,000038	0,9875	$9{,}5\times10^{-6}$
91734	18q21.32	--	0,9999	$1{,}2\times10^{-6}$	$4{,}3\times10^{-6}$	**0,0214**	$2{,}1\times10^{-7}$

* nach H2-Korrektion, rückskalierter p-Werte;
GxE Gen x Radon – Interaktion; G/GxE (joint test) Haupteffekt und Interaktion

Insgesamt zeigten 5 LD-Blöcke genomweite oder suggestive Signifikanzen, drei über eine sehr weite Spanne für ρ von 0,5 bis 1-10^{-17}, zwei weitere nur bei einem ρ von 0,5.

Der **LD-Block Nr. 91734** auf Chromosom 18q21.32 (rückskalierter p-Werte: $p^* = 2{,}1 \times 10^{-7}$) erzielt dabei genomweite Signifikanz. Der **LD-Block Nr. 2271** auf Chromosom 1p21.3 (rückskalierter p-Werte: $p^* = 5{,}4 \times 10^{-7}$) verfehlt diese knapp.

Die drei LD-Blöcke mit suggestiver Signifikanz sind **LD-Block Nr. 33135** auf Chromosom 5q23.2 ($p^* = 2{,}5 \times 10^{-6}$), **LD-Block Nr. 58899** auf Chromosom 10p13 ($p^* = 1{,}8 \times 10^{-6}$) und **LD-Block Nr. 69267** auf Chromosom 12p12.1 ($p^* = 1{,}8 \times 10^{-6}$).

Der **LD-Block Nr. 91734** enthält 4 typisierte Marker zwischen rs1346830 und rs8091054. Alle Marker liegen nahe oder innerhalb der Pseudogene LOC107985187 und LOC105372156.

Der **LD-Block Nr. 2271** enthält 9 typisierte Marker zwischen rs7545208 und rs704550. Alle Marker liegen nahe oder innerhalb des Gens UBE2U. Das Gen UBE2U kodiert das Enzym *ubiquitin conjugating E2 U*. Es ist Teil der Genfamilie *Ubiquitin conjugating enzymes E2.*

4.9.3.2 Korrektur für multiples Testen basierend auf der Anzahl LD-Blöcke (D-Screening durch p-Werte des DxG-Modells von McKay et al.)

Die Wahl des Signifikanz-Parameters ρ nimmt nur geringfügig Einfluss auf die Identifikation der signifikantesten Marker (siehe Tabelle 74 und Abbildung 44 im Anhang). Bei ρ=0,9999 wurde der geringste nach Bonferroni korrigierte p-Wert eines Interaktionseffekts (p_{GxE}=0,0214) für einen LD-Block erzielt. Das entspricht suggestiver Signifikanz.

Insgesamt zeigten 4 LD-Blöcke genomweite oder suggestive Signifikanzen, drei davon über eine sehr weite Spanne für ρ von 0.5 bis 1-10^{-17}.

Der **LD-Block Nr. 91734** auf Chromosom 18q21.32 (rückskalierter p-Werte: $p^* = 2{,}1 \times 10^{-7}$) erzielt dabei genomweite Signifikanz. Der **LD-Block Nr. 2271** auf Chromosom 1p21.3 (rückskalierter p-Werte: $p^* = 5{,}4 \times 10^{-7}$) verfehlt diese knapp.

Für die beiden **LD-Block Nr. 33135 und 33135** auf Chromosom 5 (rückskalierter p-Werte: $p^* = 2{,}5 \times 10^{-6}$ bzw. $p^* = 5{,}9 \times 10^{-6}$) konnte eine suggestiv signifikante GxE-Interaktion beobachtet werden.

Abbildung 21 Manhattan-Plot: Signifikanz gemäß Hybrid-2-Schritt (H2)-Verfahrens bei ρ=0.9999 für genetische Interaktion je LD-Block im Modell mit Marker-Auswahl / D-Screening durch D-XG-Modell von McKay et al.

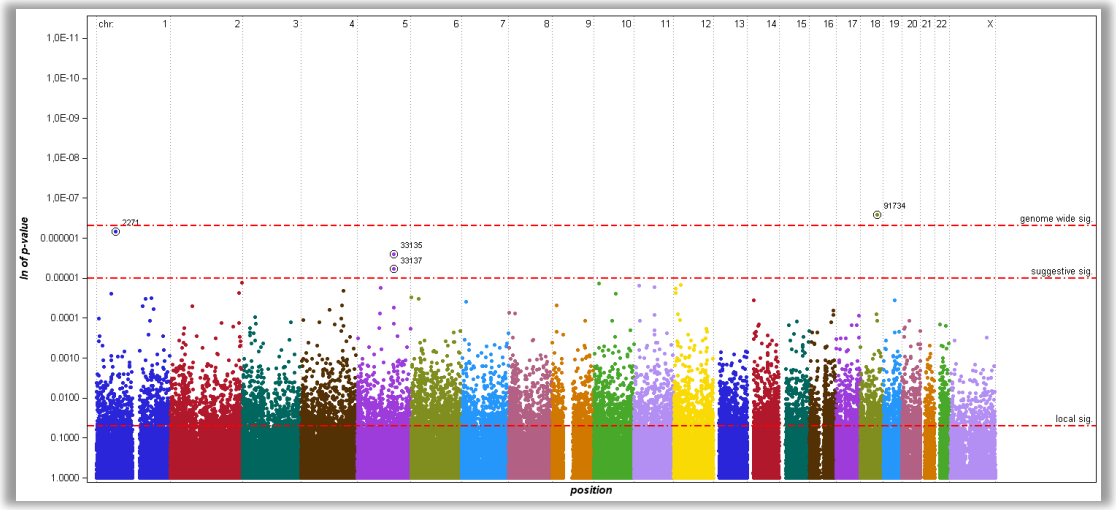


Tabelle 37 **Liste suggestiv signifikanter LD-Blöcke gemäß Hybrid-2-Schritt (H2)-Verfahrens bei beliebigen ρ im Modell mit Marker-Auswahl / D-Screening durch p-Werte des DxG-Modelles von McKay et al.**

LD-Block	Chr.		ρ für min. p-Wert	p-Wert GxE	p-Wert G/GxE	min. p-Wert H2-korrigiert	rückskalierter p-Wert*
2271	1p21.3	UBE2U	0,9999	$3{,}2\times10^{-6}$	$7{,}6\times10^{-6}$	**0,0563**	$5{,}4\times10^{-7}$
33135	5q23.2	CSNK1G3 LINC01170	0,5	$2{,}5\times10^{-6}$	$6{,}6\times10^{-6}$	0,2585	$2{,}5\times10^{-6}$
33137	5q23.2	CSNK1G3 LINC01170	0,5	0,000013	0,000025	0,6083	$5{,}9\times10^{-6}$
91734	18q21.32	--	0,9999	$1{,}2\times10^{-6}$	$4{,}3\times10^{-6}$	**0,0214**	$2{,}1\times10^{-7}$

* nach H2-Korrektion, rückskalierter p-Werte;
GxE Gen x Radon – Interaktion; G/GxE (joint test) Haupteffekt und Interaktion

4.10 Übersicht: LD-Blöcke mit mindestens suggestiver Signifikanz

Durch die verschiedenen Methoden zur Bestimmung statistischer Signifikanz konnten für insgesamt 16 LD-Blöcke in 12 genomischen Regionen eine suggestive oder genomweite signifikante GxE-Interaktion gefunden werden (siehe Tabelle 38). Eine Übersicht einer Multimarker-Assoziationsanalyse je LD-Block, bzw. je Region ist in Tabelle 39 gegeben. Die Ergebnisse von Modellschätzungen der auffälligen Regionen sind im Anschluss aufgelistet.

Die Signifikanz für LD-Blöcke nach Variablenselektion (AIC-Modell) ist stets größer als die im taxativen Modell, in dem für jeden Markern eines LD-Blocks sowohl ein Haupt- als auch ein Interaktionseffekt geschätzt wurden. Daher darf vermutet werden, dass Gen x Radon-Interaktionen auf punktuelle Varianten innerhalb der LD-Blöcke zurückzuführen sind und eher weniger durch mehrere Varianten markiert werden. Diese Vermutung wird ebenso durch die höhere Signifikanz der Modell einzelner LD-Blöcke (z.B. 33131 oder 69267) im Vergleich zu LD-Block-überspannenden Modellen (z.B.: 33131-33137 oder 69262-69267) unterstützt.

Tabelle 38 LD-Blöcke mit mindestens suggestiver Signifikanz in der Einzel- und Multimarker-Assoziationsanalyse

LD-Block	Region	Gen / Marker	n_{Marker}	p-Wert	Modell	Test	D-Screening (H2)
2271	1p21.3	UBE2U	9	5×10^{-7}	AIC-Modell	GxE (H2)	McKay et al.[a)]
				5×10^{-7}	AIC-Modell	GxE (H2)	DxG-Modell [a)]
				3×10^{-6}	AIC-Modell	GxE	
				3×10^{-6}	taxatives Modell	GxE (H2)	McKay et al. [b)]
				3×10^{-6}	taxatives Modell	GxE (H2)	DxG-Modell [b)]
5078	1q25.3	rs10911725 (--)	7	5×10^{-6}	Einzelmarker	GxE	
33131	5q23.2	rs7705033 (--)	12	8×10^{-6}	Einzelmarker	G/GxE	
		rs7735409 (--)		1×10^{-5}	Einzelmarker	G/GxE	
33135	5q23.2	rs6891344 (--)	4	1×10^{-5}	Einzelmarker	G/GxE	
		rs11747272 (--)		2×10^{-5}	Einzelmarker	GxE	
		CSNK1G3*		3×10^{-6}	AIC-Modell	GxE (H2)	DxG-Modell [a)]
		LINC01170*		3×10^{-6}	AIC-Modell	GxE (H2)	McKay et al. [a)]
33137	5q23.2	CSNK1G3*	5	6×10^{-6}	AIC-Modell	GxE (H2)	McKay et al. [c)]
		LINC01170*		6×10^{-6}	AIC-Modell	GxE	
58899	10p13	CUBN	10	2×10^{-6}	AIC-Modell	GxE (H2)	DxG-Modell [c)]
64068	11p15.1	NAV2 NAV2AS4 NAV2AS5	16	4×10^{-6}	taxatives Modell	GxE (H2)	McKay et al.[c)]
				4×10^{-6}	taxatives Modell	GxE (H2)	DxG-Modell [c)]
68621	12p13.31	CD163L1/ACSM4*	3	4×10^{-6}	taxatives Modell	GxE (H2)	McKay et al. [a)]
		PEX5*		2×10^{-6}	taxatives Modell	GxE (H2)	DxG-Modell [a)]
68623	12p13.31	CD163L1	3	2×10^{-6}	taxatives Modell	GxE (H2)	McKay et al. [b)]
		LOC101927882		9×10^{-6}	taxatives Modell	GxE (H2)	DxG-Modell [b)]
		ACSM4[$]					
69267	12p12.1	SOX5[$]	14	1×10^{-5}	AIC-Modell	GxE (H2)	DxG-Modell [c)]
		MIR920[$]					
82002	15q25.1	rs6495309 (CHRNA3)	6	2×10^{-6}	Einzelmarker	GxE (H2)	McKay et al. [d)]
		rs28534575 (CHRNB4)		2×10^{-6}	Einzelmarker	GxE (H2)	McKay et al. [d)]
82003	15q25.1	rs12440014 (CHRNB4)	4	4×10^{-7}	Einzelmarker	GxE (H2)	McKay et al. [d)]
82005	15q25.1	rs1316971 (CHRNB4)	2	2×10^{-6}	Einzelmarker	GxE (H2)	McKay et al. [d)]
82008	15q25.1	rs17487514 (RPL18P11)	14	2×10^{-6}	Einzelmarker	GxE (H2)	McKay et al. [d)]
		rs6495314 (--)		5×10^{-6}	Einzelmarker	GxE (H2)	McKay et al. [d)]
		C15orf32 ST8SIA2 snoU109	11	4×10^{-6}	AIC-Modell	GxE (H2)	DxG-Modell [c)]
82566	15q26.1	RP11763K1511					
		RP11763K1521					
91734	18q21.32	LOC107985187	4	3×10^{-6}	taxatives Modell	GxE (H2)	McKay et al. [c)]
		LOC105372156		3×10^{-6}	taxatives Modell	GxE (H2)	DxG-Modell [c)]
		RP11-325K19.1		3×10^{-6}	taxatives Modell	GxE	
		CTBP2P3		1×10^{-6}	AIC-Modell	GxE	
				2×10^{-7}	AIC-Modell	GxE (H2)	DxG-Modell [a)]
				2×10^{-6}	AIC-Modell	GxE (H2)	McKay et al. [a)]

* Intergenetischer Bereich zwischen angegebenen Genen, [$] nahe dem angegebenen Gen; n_{Marker} Anzahl Marker in LD-Block;
taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; ρ des H2-Verfahren: a) 0,9999 b) 0,999 c) 0,5 d) $1 - 1 \times 10^{-16}$

Tabelle 39 Übersicht der Signifikanz aus einer Multimarker-Assoziationsanalyse für auffällige LD-Blöcke

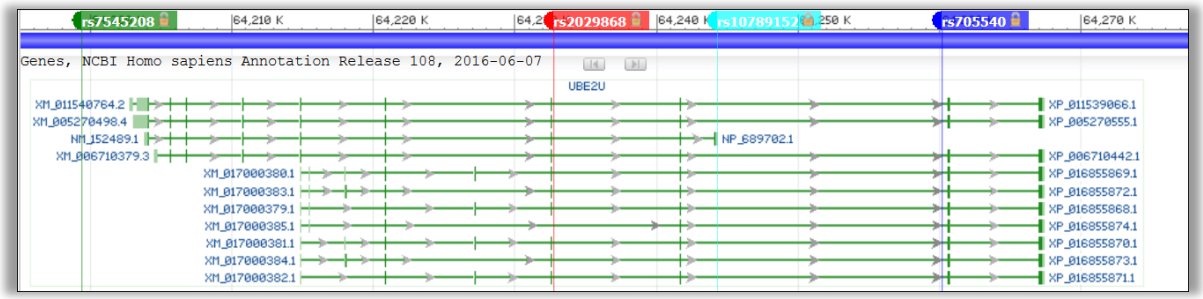
LD-Block	Region	Gene / Marker	p-Wert (GxE – Interaktion)	
			taxatives Modell	AIC-Modell
2271	1p21.3	UBE2U	$\mathbf{3 \times 10^{-5}}$	3×10^{-6}
5078	1q25.3		0,0021	1×10^{-4}
33131	5q23.2		0,0005	1×10^{-5}
33135	5q23.2	CSNK1G3* LINC01170*	1×10^{-5}	2×10^{-6}
33137	5q23.2	CSNK1G3* LINC01170*	9×10^{-5}	3×10^{-5}
33131-33137			0,0216	8×10^{-5}
58899	10p13	CUBN	0,0001	2×10^{-5}
64068	11p15.1	NAV2 NAV2AS4 NAV2AS5	0,0083	0,0001
68621-68623	12p13.31	CD163L1 LOC101927882 ACSM4$^{\$}$	0,0002	8×10^{-5}
69267	12p12.1	SOX5$^{\$}$ MIR920$^{\$}$	0,0232	7×10^{-5}
69262-69267		SOX5$^{\$}$ MIR920$^{\$}$	0,9047	0,0746
82003	15q25.1	CHRNB4	0,0297	0,0028
82002-82008	15q25.1	CHRNA3 CHRNB4	0,1965	0,0051
82566	15q26.1	C15orf32 ST8SIA2 snoU109 RP11763K1511 RP11763K1521	0,0161	0,0018
91734	18q21.32	LOC107985187 LOC105372156 RP11-325K19.1	3×10^{-6}	1×10^{-6}

* Intergenetischer Bereich zwischen angegebenen Genen, $^{\$}$ nahe dem angegebenen Gen; n_{Marker} Anzahl Marker in LD-Block;
taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium;

4.10.1 Modellschätzung LD-Block Nr. 2271 (Chr. 1p31.3; UBE2U)

Der LD-Block Nr. 2271 im Chromosomabschnitt 1p31.3 überdeckt das Gene UBE2U (Abbildung 22) Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 7 der 9 Marker verbleiben im AIC-Modell (Tabelle 40). Dabei zeigen die Marker rs705540 (OR=4,17; 95%-CI: 1,11- 15,6) und rs2029868 (OR=22,6; 95%-CI: 4,7-109) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen die Marker rs7545208 (OR=0,34; 95%-CI: 0,1-0,8) und rs10789152 (OR=0,01; 95%-CI: <0,11) eine negative Interaktion für das jeweils seltene Allel. (Zur eingeschränkten Interpretation der ORs siehe Kapitel 8.)

Abbildung 22 UBE2U mit ausgewählten Markern



Die Abbildung zeigt die Lage des Gens UBE2U (mehrere Definitionen) sowie ausgewählter Marker im Bereich 64.270K bis 64.220K des Chromosom 1p31.3 gemäß GeneDB [79]

Tabelle 40 Modellschätzung: LD-Block Nr. 2271

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,75	(2,50- 3,03)	$2{,}1 \times 10^{-95}$	2,75	(2,50- 3,03)	$1{,}6 \times 10^{-95}$
Strahlenexposition	2,23	(0,92- 5,39)	0,0729	2,37	(1,00- 5,62)	0,0482
rs705540	1,05	(0,85- 1,30)	0,6212			
rs705551	1,00	(0,85- 1,17)	0,9876			
rs2029868	0,99	(0,72- 1,37)	0,9900			
rs2806532	1,09	(0,91- 1,29)	0,3309	1,02	(0,92- 1,13)	0,6134

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
rs7526950	0,93	(0,75- 1,15)	0,5449			
rs7545208	0,94	(0,81- 1,10)	0,4925			
rs10789152	0,94	(0,66- 1,34)	0,7484			
rs11588217	1,05	(0,73- 1,49)	0,7766			
rs17126246	0,90	(0,62- 1,29)	0,5822			
Expo. x rs7545208	0,43	(0,16- 1,19)	0,1063	0,34	(0,14- 0,82)	0,0172
Expo. x rs7526950	0,55	(0,13- 2,21)	0,4046	0,58	(0,15- 2,21)	0,4254
Expo. x rs705551	0,35	(0,11- 1,13)	0,0806	0,34	(0,11- 1,03)	0,0568
Expo. x rs705540	4,51	(1,12- 18,0)	0,0330	4,17	(1,11- 15,6)	0,0336
Expo. x rs2806532	0,72	(0,23- 2,25)	0,5737			
Expo. x rs2029868	15,3	(1,66- 142)	0,0160	22,6	(4,68- 109)	0,0001
Expo. x rs17126246	0,55	(0,04- 7,43)	0,6532			
Expo. x rs11588217	4,82	(0,43- 53,6)	0,2004	2,84	(0,50- 16,0)	0,2359
Expo. x rs10789152	0,02	(<0,01- 0,21)	0,0010	0,01	(<0,01- 0,11)	0,000043

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	1,9587	9	0,9921	1	0,6134
Interaktion(en) GxE	36,6912	9	0,00003	7	$3{,}176 \times 10^{-6}$
Gemeinsamer Effekt (joint)	40,5512	18	0,0018	8	$7{,}505 \times 10^{-6}$

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.2 Modellschätzung LD-Block Nr. 5078 (Chr. 1q25.3,intergenetischer Bereich)

Der LD-Block Nr. 2271 im Chromosomabschnitt 1p31.3 liegt im intergenetischen Bereich. Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 3 der 7 Marker verbleiben im AIC-Modell (Tabelle 41). Dabei zeigen die Marker rs10911725 (OR=0,09; 95%-CI: 0,02-0,3), rs10737261 (OR=0,37; 95%-CI: 0,1-1,2) und chr1_185386615_C_T (OR=0,57; 95%-CI: 0,2-1,5) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR>47 und damit wesentlich höher als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risikoerhöhende bzw. –senkende Marker diskutierbar. Zusammen betrachtet kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker des LD-Blockes Nr. 5078 geschlossen werden.

Tabelle 41 Modellschätzung: LD-Block Nr. 5078

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,73	(2,48- 3,00)	$2{,}0 \times 10^{-95}$	2,73	(2,48- 3,00)	$1{,}2 \times 10^{-95}$
Strahlenexposition	20,4	(0,03- >999)	0,3522	47,4	(4,6- 493)	0,0012
chr1_185386615_C_T	1,02	(0,87- 1,20)	0,7189			
rs4651259	0,94	(0,80- 1,11)	0,5009	0,97	(0,88- 1,06)	0,5293
rs7519702	1,01	(0,87- 1,17)	0,8363			
rs10737261	1,05	(0,83- 1,33)	0,6567			
rs10798014	1,09	(0,86- 1,38)	0,4559			
rs10911725	1,05	(0,80- 1,36)	0,7127			
rs10911734	1,02	(0,82- 1,28)	0,8047			
Expo. x rs7519702	1,12	(0,51- 2,42)	0,7713			
Expo. x rs4651259	0,72	(0,24- 2,16)	0,5673			
Expo. x rs10911734	1,49	(0,48- 4,57)	0,4831			

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Expo. x rs10911725	0,09	(<0,01- 2,33)	0,1498	0,09	(0,02- 0,34)	0,0004
Expo. x rs10798014	1,27	(0,06- 23,9)	0,8703			
Expo. x rs10737261	0,52	(0,02- 10,9)	0,6744	0,37	(0,11- 1,22)	0,1057
Expo. x chr1_185386615_C_T	0,56	(0,21- 1,48)	0,2451	0,57	(0,22- 1,48)	0,2536

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	1,7001	7	0,9746	1	0,5293
Interaktion(en) GxE	22,4521	7	0,0021	3	0,0001
Gemeinsamer Effekt (joint)	24,6258	14	0,0384	4	0,0004

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.3 Modellschätzung LD-Block Nr. 33131 (Chr. 5q23.2, intergenetischer Bereich)

Der LD-Block Nr. 33131 im Chromosomabschnitt 5q23.2 liegt im intergenetischen Bereich. Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 7 der 12 Marker verbleiben im AIC-Modell (Tabelle 42). Dabei zeigen die Marker rs7728845 (OR=46; 95%-CI: 3,9- >500). rs7703080 (OR=38; 95%-CI: 2,3- >600) und rs13165542 (OR=57; 95%-CI: 2,6- >1000) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen die Marker rs7735409 (OR=0,02; 95%-CI: <0,5), rs6887967 (OR=0,04; 95%-CI: <0,5) und rs4572998 (OR=0,17; 95%-CI: 0,04-0,7) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR>4 und damit etwas höher als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risiko-erhöhende bzw. –senkende Marker diskutierbar. Zusammen betrachtet kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker des LD-Blockes Nr. 33131 geschlossen werden.

Tabelle 42 Modellschätzung: LD-Block Nr. 33131

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,74	(2,49- 3,02)	$4,9x10^{-94}$	2,74	(2,49- 3,02)	$1,1x10^{-94}$
Strahlenexposition	<0,01		0,9854	4,14	(1,82- 9,40)	0,0007
rs1972627	1,11	(0,79- 1,56)	0,5349			
rs4572998	1,00	(0,77- 1,30)	0,9671			
rs6884946	0,98	(0,77- 1,25)	0,9254			
rs6887967	0,94	(0,71- 1,26)	0,7068			
rs7703080	0,89	(0,59- 1,32)	0,5645			
rs7705033	1,48	(0,41- 5,32)	0,5437			
rs7728845	1,11	(0,76- 1,63)	0,5748			
rs7731839	1,05	(0,79- 1,38)	0,7214			
rs7735409	0,67	(0,19- 2,29)	0,5245	0,95	(0,86- 1,05)	0,3454
rs11241696	1,00	(0,85- 1,18)	0,9211			
rs12109037	0,94	(0,70- 1,25)	0,6799			
rs13165542	0,94	(0,63- 1,40)	0,7954			
Expo. x rs7735409	0,03	(<0,01- 3,62)	0,1532	0,02	(<0,01- 0,54)	0,0184
Expo. x rs7731839	>999		0,9846			
Expo. x rs7728845	>999		0,9807	46,0	(3,92- 539)	0,0023
Expo. x rs7703080	43,2	(2,55- 732)	0,0090	38,0	(2,30- 627)	0,0109
Expo. x rs6887967	0,05	(<0,01- 0,69)	0,0252	0,04	(<0,01- 0,50)	0,0117
Expo. x rs6884946	4,93	(0,79- 30,4)	0,0856	4,23	(1,19- 15,0)	0,0256

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Expo. x rs4572998	0,17	(<0,01- 8,11)	0,3692	0,17	(0,04- 0,69)	0,0129
Expo. x rs1972627	>999		0,9850			
Expo. x rs13165542	50,8	(1,56- >999)	0,0269	57,1	(2,56- >999)	0,0106
Expo. x rs12109037	1,26	(0,03- 47,3)	0,8971			
Expo. x rs11241696	1,01	(0,27- 3,72)	0,9788			

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	2,1681	12	0,9991	1	0,3454
Interaktion(en) GxE	33,0555	11	0,0005	7	0,000014
Gemeinsamer Effekt (joint)	36,7623	23	0,0344	8	0,000013

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.4 Modellschätzung LD-Block Nr. 33135 (Chr. 5q23.2; zwischen CSNK1G3 und LINCO1170)

Der LD-Block Nr. 33135 im Chromosomabschnitt 5q23.2 liegt im nicht-kodierender Bereich zwischen CSNK1G3 und LINCO1170. Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 3 der 4 Marker verbleiben im AIC-Modell (Tabelle 43). Dabei zeigt der Marker rs6891344 (OR=2,15; 95%-CI: 1,2-3,9) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen die Marker rs6895877 (OR=0,35; 95%-CI: 0,2-0,7) und rs257140 (OR=0,39; 95%-CI: 0,2-03,8) eine negative Interaktion für das jeweils seltene Allel.

Tabelle 43 Modellschätzung: LD-Block Nr. 33135

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,74	(2,49- 3,01)	$2,0x10^{-96}$	2,74	(2,49- 3,01)	$1,0x10^{-96}$
Strahlenexposition	5,57	(2,46- 12,6)	0,000038	5,50	(2,70- 11,2)	$2,571x10^{-6}$
rs257140	0,98	(0,85- 1,12)	0,7946	0,97	(0,87- 1,09)	0,7122
rs6882579	0,98	(0,82- 1,18)	0,8930			
rs6891344	1,01	(0,88- 1,16)	0,8593			
rs6895877	0,99	(0,88- 1,13)	0,9975			
Expo. x rs6895877	0,34	(0,16- 0,72)	0,0046	0,35	(0,17- 0,70)	0,0030
Expo. x rs6891344	2,12	(1,12- 4,01)	0,0204	2,15	(1,19- 3,91)	0,0112
Expo. x rs6882579	0,98	(0,36- 2,70)	0,9832			
Expo. x rs257140	0,39	(0,19- 0,80)	0,0111	0,39	(0,20- 0,78)	0,0079

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	0,2027	4	0,9952	1	0,7122
Interaktion(en) GxE	27,7497	4	0,000014	3	$2,486x10^{-6}$
Gemeinsamer Effekt (joint)	29,4426	8	0,0003	4	$6,535x10^{-6}$

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.5 Modellschätzung LD-Block Nr. 33137 (Chr. 5q23.2; zwischen CSNK1G3 und LINCO1170)

Der LD-Block Nr. 33137 im Chromosomabschnitt 5q23.2 liegt im nicht-kodierenden Bereich zwischen CSNK1G3 und LINCO1170. Für diesen LD-Block konnte sowohl im taxativen Modell, als auch

nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 4 der 5 Marker verbleiben im AIC-Modell (Tabelle 44). Dabei zeigen gleich 4 Marker (rs6876166, rs3909326, rs12514988 und s11747272; OR von 1,27-4,77) eine positive Interaktion für das jeweils seltene Allel. Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=0,64 und damit wesentlich niedriger als in vergleichbaren Modellen geschätzt. Zusammen betrachtet markieren somit die häufigeren Allele der 4 Marker eine Senkung der Fall-Wahrscheinlichkeit.

Tabelle 44 Modellschätzung: LD-Block Nr. 33137

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,73	(2,48- 3,00)	$3,9x10^{-96}$	2,73	(2,48- 3,00)	$3,7x10^{-96}$
Strahlenexposition	0,66	(0,10- 4,30)	0,6708	0,64	(0,11- 3,58)	0,6140
rs3909326	0,96	(0,78- 1,19)	0,7569			
rs6876166	0,99	(0,87- 1,14)	0,9898			
rs10052257	0,97	(0,87- 1,09)	0,6919	0,98	(0,88- 1,08)	0,7211
rs11747272	0,99	(0,86- 1,14)	0,9428			
rs12514988	0,96	(0,78- 1,17)	0,7055			
Expo. x rs6876166	1,24	(0,45- 3,39)	0,6638	1,27	(0,49- 3,30)	0,6196
Expo. x rs3909326	2,99	(0,72- 12,4)	0,1310	2,97	(0,76- 11,5)	0,1141
Expo. x rs12514988	3,82	(1,08- 13,4)	0,0368	3,76	(1,11- 12,6)	0,0321
Expo. x rs11747272	4,75	(1,79- 12,5)	0,0017	4,77	(1,83- 12,4)	0,0014
Expo. x rs10052257	0,96	(0,55- 1,67)	0,8971			

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	0,3709	5	0,9961	1	0,7211
Interaktion(en) GxE	26,0107	5	0,000089	4	0,000025
Gemeinsamer Effekt (joint)	26,8247	10	0,0028	5	0,00007

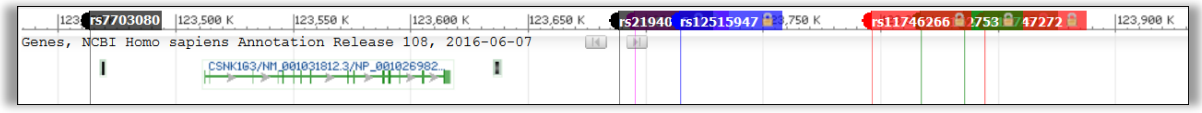
df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.6 Modellschätzung LD-Blöcke Nr. 33131-33137 (Chr. 5q23.2, nahe CSNK1G3)

Die LD-Blöcke Nr. 33131-33137 im Chromosomabschnitt 5q23.2 liegen um das Gene CSNK1G3 (Abbildung 23). Für diese LD-Blöcke konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 9 der 20 Marker verbleiben im AIC-Modell (Tabelle 45). Dabei zeigen die Marker rs11747272 (OR=2,68) und chr5_123135986_A_G (OR=2,03; 95%-CI: 0,6-6,6) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen gleich 7 Marker (OR von 0,15 bis 0,46), vor allem aber Marker rs11746266 (OR=0,27; 95%-CI: 0,1-0,7) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR>5,7 und damit wesentlich höhere als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risiko-erhöhende bzw. –senkende Marker diskutierbar. Zusammen betrachtet kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker der LD-Blöcke Nr. 33131-33137 geschlossen werden.

Abbildung 23 CSNK1G3 mit ausgewählten Markern



Die Abbildung zeigt die Lage des Gens CSNK1G3 sowie ausgewählter Marker im Bereich 123.400K bis 123.900K des Chromosom 5q23.2 gemäß GeneDB [79]

Tabelle 45 Modellschätzung: LD-Blöcke Nr. 33131-33137

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,73	(2,48- 3,00)	$4,0x10^{-94}$	2,73	(2,58- 3,00)	$3,7x10^{-95}$
Strahlenexposition	242	(0,04- >999)	0,2143	409	(5,7- >999)	0,0057
chr5_123045038_C_G	1,00	(0,81- 1,23)	0,9893			
chr5_123135986_A_G	1,18	(0,78- 1,76)	0,4214			
chr5_123170960_A_G	1,00	(0,66- 1,49)	0,9928			
rs1834887	1,00	(0,85- 1,17)	0,9534			
rs2194046	1,00	(0,78- 1,28)	0,9622			
rs6876166	1,02	(0,88- 1,17)	0,7697			
rs6889596	1,05	(0,86- 1,28)	0,6133			
rs7703080	0,96	(0,81- 1,14)	0,6960			
rs10043442	1,03	(0,85- 1,25)	0,7235			
rs10052257	0,98	(0,87- 1,11)	0,8262			
rs10478586	1,08	(0,84- 1,39)	0,5185			
rs11241696	0,95	(0,83- 1,09)	0,5260			
rs11746266	1,09	(0,87- 1,36)	0,4438			
rs11747272	1,00	(0,85- 1,16)	0,9974			
rs12109037	0,99	(0,85- 1,15)	0,9356			
rs12514988	0,91	(0,70- 1,18)	0,5100			
rs12515947	1,02	(0,81- 1,29)	0,8048			
rs12520155	1,06	(0,83- 1,36)	0,6081			
rs13162753	1,07	(0,91- 1,27)	0,3659			
rs17480969	0,86	(0,70- 1,07)	0,2013	0,93	(0,78- 1,10)	0,4175
Expo. x rs7703080	0,71	(0,17- 2,87)	0,6401	0,46	(0,17- 1,24)	0,1277
Expo. x rs6889596	0,33	(0,06- 1,69)	0,1872	0,28	(0,07- 1,16)	0,0813
Expo. x rs6876166	0,82	(0,25- 2,68)	0,7452			
Expo. x rs2194046	0,23	(0,02- 2,52)	0,2330	0,34	(0,09- 1,26)	0,1097
Expo. x rs1834887	1,56	(0,40- 6,03)	0,5183			
Expo. x rs17480969	0,72	(0,12- 4,31)	0,7222			
Expo. x rs13162753	0,43	(0,13- 1,44)	0,1733	0,40	(0,16- 1,02)	0,0555
Expo. x rs12520155	1,28	(0,13- 11,8)	0,8244			
Expo. x rs12515947	0,10	(<0,01- 1,85)	0,1251	0,15	(0,03- 0,75)	0,0212
Expo. x rs12514988	0,74	(0,10- 5,06)	0,7594			
Expo. x rs12109037	1,31	(0,49- 3,45)	0,5821			
Expo. x rs11747272	2,27	(0,72- 7,15)	0,1595	2,68	(1,27- 5,63)	0,0092
Expo. x rs11746266	0,47	(0,04- 5,33)	0,5455	0,27	(0,10- 0,72)	0,0083
Expo. x rs11241696	0,75	(0,21- 2,65)	0,6644			
Expo. x rs10478586	2,64	(0,28- 24,9)	0,3954			
Expo. x rs10052257	1,06	(0,54- 2,08)	0,8632			
Expo. x rs10043442	1,37	(0,14- 12,9)	0,7783			
Expo. x chr5_123135986_A_G	2,52	(0,17- 37,6)	0,5003	2,03	(0,62- 6,64)	0,2382
Expo. x chr5_123045038_C_G	0,18	(0,01- 2,32)	0,1896	0,29	(0,06- 1,26)	0,0996

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	4,5626	20	0,9999	1	0,4175
Interaktion(en) GxE	33,3932	19	0,0216	9	0,000078
Gemeinsamer Effekt (joint)	39,3914	39	0,4524	10	0,0001

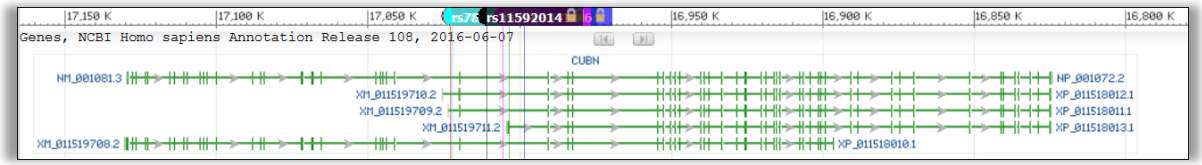
df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.7 Modellschätzung LD-Block Nr. 58899 (Chr. 10p13; CUBN)

Der LD-Block Nr. 58899 im Chromosomabschnitt 10p13 liegt innerhalb des Gens CUBN (Abbildung 24). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 7 der 10 Marker verbleiben im AIC-Modell (Tabelle 46). Dabei zeigen die Marker rs7896819 (OR=4,25), rs4748341 (OR>36) und chr10_17046641_A_T (OR=7,81; 95%-CI: 2,1-29) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen die Marker rs7922356 (OR=0,11; 95%-CI: <0,8) und rs11592014 (OR=0,01; 95%-CI: <0,11) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=0,1 und damit wesentlich niedriger als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risiko-erhöhende bzw. –senkende Marker diskutierbar. . Abgehsehen von dieser instabilen Schätzung des Radon-Haupteffekts kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker des LD-Blocks Nr. 58899 geschlossen werden.

Abbildung 24 CUBN mit ausgewählten Markern



Die Abbildung zeigt die Lage des Gens CUBN (mehrere Definitionen) sowie ausgewählter Marker im Bereich 16.800K bis 17.150K des Chromosom 10q13 gemäß GeneDB [79]

Tabelle 46 Modellschätzung: LD-Block Nr. 58899

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,72	(2,47- 2,99)	$4,8x10^{-94}$	2,72	(2,47- 2,99)	$2,8x10^{-94}$
Strahlenexposition	0,07	(<0,01- 1,40)	0,0841	0,10	(<0,01- 1,58)	0,1035
chr10_17029438_A_C	1,00	(0,77- 1,29)	0,9956			
chr10_17046641_A_T	0,93	(0,77- 1,12)	0,4914			
rs2291521	0,98	(0,83- 1,16)	0,8730			
rs2942366	0,98	(0,86- 1,11)	0,7682			
rs4748341	0,96	(0,65- 1,44)	0,8760			
rs7896819	0,94	(0,78- 1,12)	0,5172			
rs7897550	1,00	(0,78- 1,28)	0,9627			
rs7900486	1,12	(0,95- 1,31)	0,1556			
rs7922356	0,86	(0,65- 1,14)	0,3154	0,92	(0,81- 1,06)	0,2779
rs11592014	0,94	(0,65- 1,34)	0,7386			
Expo. x rs7922356	0,17	(0,02- 1,41)	0,1012	0,11	(0,01- 0,82)	0,0318
Expo. x rs7900486	2,04	(0,69- 5,96)	0,1925	2,12	(0,75- 6,00)	0,1538
Expo. x rs7897550	0,27	(0,06- 1,20)	0,0868	0,27	(0,06- 1,13)	0,0742
Expo. x rs7896819	4,69	(1,21- 18,1)	0,0250	4,25	(1,16- 15,5)	0,0289
Expo. x rs4748341	520	(24,4- >999)	0,000061	559	(36,4- >999)	$5,624x10^{-6}$
Expo. x rs2942366	1,09	(0,48- 2,47)	0,8332			
Expo. x rs2291521	1,51	(0,50- 4,54)	0,4551			
Expo. x rs11592014	0,01	(<0,01- 0,14)	0,0003	0,01	(<0,01- 0,11)	0,000028
Expo. x chr10_17046641_A_T	8,74	(2,31- 33,0)	0,0014	7,81	(2,11- 28,7)	0,0020
Expo. x chr10_17029438_A_C	0,78	(0,18- 3,35)	0,7454			

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	3,8591	10	0,9535	1	0,2779
Interaktion(en) GxE	35,2553	10	0,0001	8	0,000013
Gemeinsamer Effekt (joint)	38,4639	20	0,0078	9	0,000025

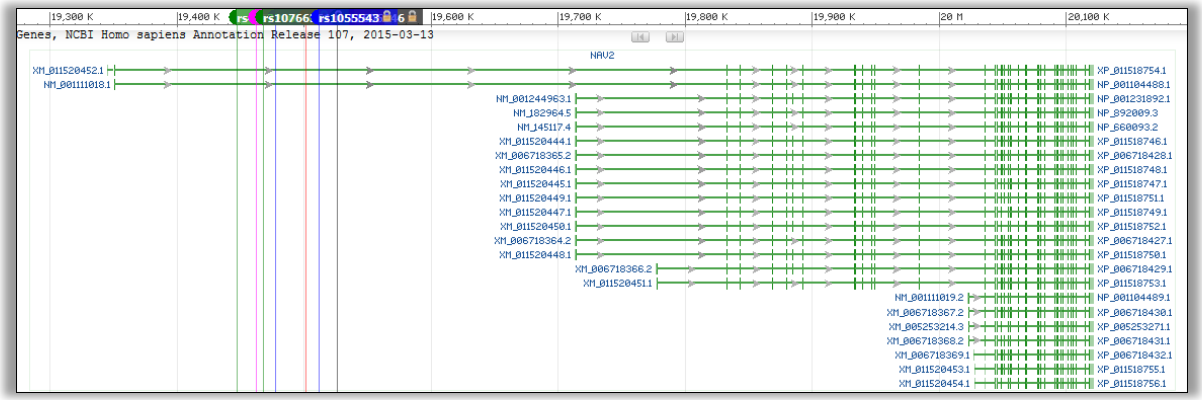
df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.8 Modellschätzung LD-Block Nr. 64068 (Chr. 11p15.1; CD163L1)

Der LD-Block Nr. 64068 im Chromosomabschnitt 11p15.11 überdeckt einen Teil des Gens NAV2 (Abbildung 25). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 9 der 16 Marker verbleiben im AIC-Modell (Tabelle 47). Dabei zeigen die Marker rs4756999 (OR=5,56; 95%-CI: 2,7-12) und rs2200568 (OR=5,63; 95%-CI: 1,1-29) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen die Marker rs2702656, rs2632019, rs2632010 und rs1055543 (OR zwischen 0,04 und 0,35)eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=0,94 und damit wesentlich niedriger als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risiko-erhöhende bzw. –senkende Marker diskutierbar. Abgehsehen von dieser instabilen Schätzung des Radon-Haupteffekts kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker des LD-Blocks Nr. 64068 geschlossen werden.

Abbildung 25 NAV2 mit ausgewählten Markern



Die Abbildung zeigt die Lage des Gens NAV2 (mehrere Definitionen) sowie ausgewählter Marker im Bereich 19.300K bis 20.100K des Chromosom 11p15.1 gemäß GeneDB [79]

Tabelle 47 Modellschätzung: LD-Block Nr. 64068

	taxatives Modell			AIC- Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,72	(2,47- 2,99)	$1{,}2\times10^{-93}$	2,73	(2,48- 3,00)	$9{,}9\times10^{-95}$
Strahlenexposition	146		0,9956	0,94	(0,07- 11,5)	0,9620
rs1055543	0,94	(0,66- 1,33)	0,7329			
rs2200568	1,02	(0,79- 1,33)	0,8240			
rs2632010	0,96	(0,74- 1,24)	0,7666			
rs2632019	0,91	(0,71- 1,16)	0,4653			
rs2632057	0,99	(0,69- 1,42)	0,9663			
rs2702656	0,81	(0,54- 1,21)	0,3141	0,92	(0,81- 1,05)	0,2618
rs2702735	1,01	(0,69- 1,47)	0,9383			
rs2729884	1,06	(0,65- 1,73)	0,8094			
rs4756999	0,95	(0,81- 1,12)	0,5916			
rs10766561	1,00	(0,77- 1,31)	0,9713			
rs10766565	0,95	(0,65- 1,41)	0,8302			
rs11025138	0,93	(0,70- 1,23)	0,6213			
rs11604718	0,96	(0,59- 1,57)	0,8974			
rs11605946	0,99	(0,69- 1,43)	0,9918			
rs11820210	0,91	(0,68- 1,24)	0,5860			
rs12271178	0,95	(0,73- 1,22)	0,7098			

	taxatives Modell			AIC- Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
expo_rs4756999	5,47	(2,19- 13,6)	0,0003	5,56	(2,67- 11,5)	$4,25x10^{-6}$
expo_rs2729884	>999		0,9643	3,37	(0,99- 11,4)	0,0513
expo_rs2702735	0,61	(0,12- 3,06)	0,5521			
expo_rs2702656	<0,01		0,9543	0,04	(<0,01- 0,55)	0,0150
expo_rs2632057	1,56	(0,34- 7,18)	0,5636			
expo_rs2632019	0,18	(0,03- 1,07)	0,0600	0,12	(0,03- 0,52)	0,0043
expo_rs2632010	0,23	(0,03- 1,85)	0,1699	0,24	(0,07- 0,78)	0,0178
expo_rs2200568	>999		0,9837	5,63	(1,08- 29,1)	0,0392
expo_rs12271178	1,33	(0,09- 18,1)	0,8299			
expo_rs11820210	739		0,9867			
expo_rs11605946	1,69	(0,33- 8,55)	0,5213	2,44	(0,92- 6,48)	0,0712
expo_rs11604718	<0,01		0,9684			
expo_rs11025138	1,38	(0,10- 17,7)	0,8016			
expo_rs10766565	<0,01		0,9658			
expo_rs10766561	>999		0,9847	3,42	(0,64- 18,3)	0,1494
expo_rs1055543	<0,01		0,9624	0,35	(0,12- 0,95)	0,0402

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

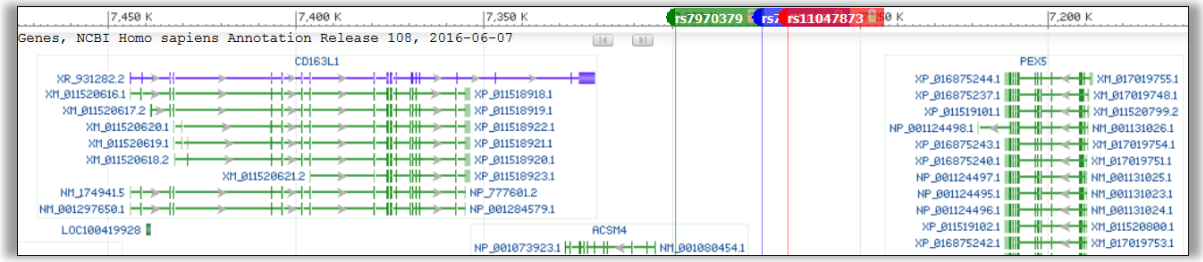
	taxatives Modell			AIC- Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	4,0358	16	0,9988	1	0,2618
Interaktion(en) GxE	32,6234	16	0,0083	9	0,0001
Gemeinsamer Effekt (joint)	37,5660	32	0,2292	10	0,0001

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.9 Modellschätzung LD-Block Nr. 68621 (Chr. 12p13.31; CD163L1/ACSM4, PEX5)

Der LD-Block Nr. 68621 im Chromosomabschnitt 12p13.31 liegt im nicht-kodierenden Bereich zwischen den Genen CD163L1/ACSM4 und PEX5 (Abbildung 26). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 2 der 3 Marker verbleiben im AIC-Modell (Tabelle 48). Dabei zeigt der Marker rs7970379 (OR=14,3; 95%-CI: 4,3-48) eine positive Interaktion für das jeweils seltene Allel.

Abbildung 26 CD163L1, ACSM4 und PEX5 mit ausgewählten Markern



Die Abbildung zeigt die Lage der Gene CD163L1 und PEX5 (mehrere Definitionen) sowie ausgewählter Marker im Bereich 7.150K bis 7.500K des Chromosom 12p12.31 gemäß GeneDB [79]

Tabelle 48 Modellschätzung: LD-Block Nr. 68621

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,72	(2,47- 2,99)	$3,8x10^{-95}$	2,72	(2,47- 2,99)	$3,7x10^{-95}$
Strahlenexposition	2,16	(1,29- 3,64)	0,0034	2,19	(1,31- 3,66)	0,0028
rs7485862	0,99	(0,89- 1,09)	0,8588			
rs7970379	0,94	(0,78- 1,13)	0,5570			

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
rs11047873	0,89	(0,67- 1,18)	0,4244	0,89	(0,68- 1,17)	0,4261
Expo. x rs7970379	15,1	(4,47- 51,4)	0,000013	14,3	(4,28- 47,9)	0,000015
Expo. x rs7485862	1,31	(0,77- 2,22)	0,3189	1,29	(0,77- 2,18)	0,3267

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	0,9862	3	0,8046	1	0,4261
Interaktion(en) GxE	19,0547	2	<.0001	2	0,000086
Gemeinsamer Effekt (joint)	19,7020	5	0,0014	3	0,0002

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.10 Modellschätzung LD-Blöcke Nr. 68621-68623 (Chr. 12p13.31; CD163L1, LOC101927882, ACSM4)

Der LD-Block Nr. 68621 im Chromosomabschnitt 12p13.31 liegt im nicht-kodierenden Bereich zwischen den Genen CD163L1/ ACSM4 und PEX5 (Abbildung 26). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 4 der 6 Marker verbleiben im AIC-Modell (Tabelle 49). Dabei zeigen alle 4 Marker rs7970379 rs7485862 rs7302538 und rs10844153 (OR zwischen 2,95 und 3,40) eine positive Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=0,32 und damit wesentlich niedriger als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risiko-erhöhende bzw. –senkende Marker diskutierbar. Abgehsehen von dieser instabilen Schätzung des Radon-Haupteffekts kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker des LD-Blocks Nr. 68621 geschlossen werden.

Tabelle 49 Modellschätzung: LD-Blöcke Nr. 68621-68623

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,72	(2,48- 3,00)	$5,1x10^{-95}$	2,72	(2,47- 2,99)	$4,1x10^{-95}$
Strahlenexposition	0,30	(0,05- 1,70)	0,1751	0,32	(0,05- 1,77)	0,1944
rs7302538	0,98	(0,76- 1,27)	0,9002			
rs7485862	0,97	(0,85- 1,10)	0,6891			
rs7970379	0,94	(0,75- 1,18)	0,6152			
rs10844153	0,96	(0,85- 1,09)	0,5913			
rs11047873	0,89	(0,62- 1,28)	0,5550	0,89	(0,68- 1,18)	0,4427
rs11051842	0,98	(0,82- 1,17)	0,8802			
Expo. x rs7970379	3,54	(0,29- 43,4)	0,3215	3,40	(0,28- 41,2)	0,3352
Expo. x rs7485862	3,12	(1,16- 8,37)	0,0239	3,21	(1,25- 8,23)	0,0152
Expo. x rs7302538	16,7	(1,12- 250)	0,0409	15,8	(1,08- 232)	0,0436
Expo. x rs11051842	1,30	(0,41- 4,09)	0,6430			
Expo. x rs10844153	3,08	(1,26- 7,51)	0,0130	2,95	(1,23- 7,08)	0,0149

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	1,2449	6	0,9746	1	0,4427
Interaktion(en) GxE	24,2862	5	0,0002	4	0,000084
Gemeinsamer Effekt (joint)	25,1796	11	0,0086	5	0,0002

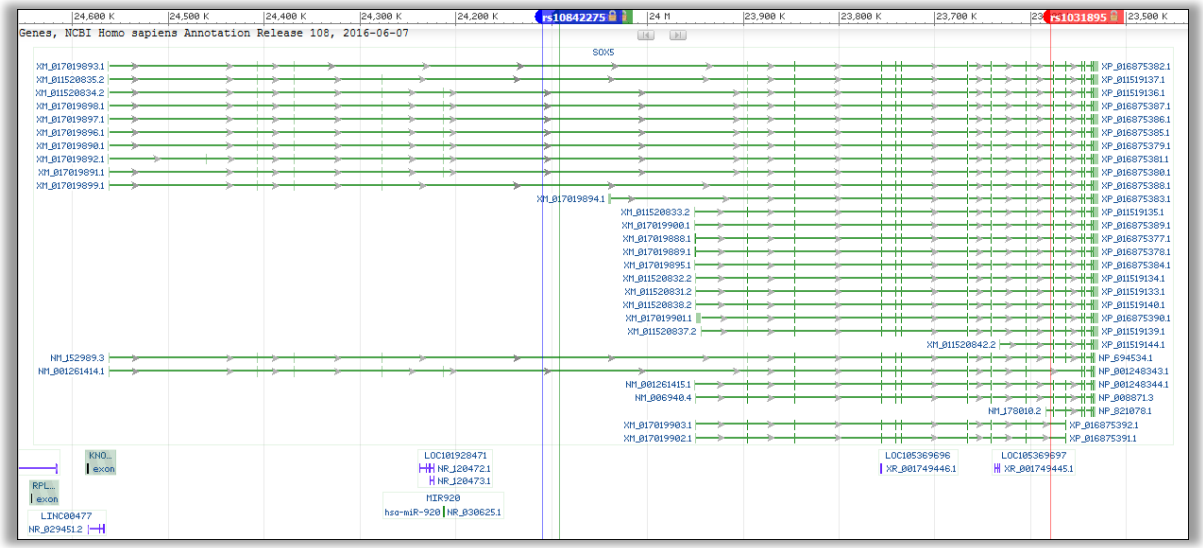
df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.11 Modellschätzung LD-Block Nr. 69267 (Chr. 12p12.1; SOX5, MIR920)

Der LD-Block Nr. 69267 im Chromosomabschnitt 12p12.1 liegt je nach Definition innerhalb oder in der Nähe des Gens SOX5 (Abbildung 27). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 4 der 14 Marker verbleiben im AIC-Modell (Tabelle 50). Dabei zeigen die Marker rs7978583, rs10842275, chr12_24281450_A_G und chr12_24252977_C_T (OR zwischen 0,15 und 0,48) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=9,58 und damit wesentlich höhere als in vergleichbaren Modellen geschätzt. Zusammen betrachtet markieren somit die häufigeren Allele der 4 Marker eine Risikosteigerung.

Abbildung 27 SOX5 mit ausgewählten Markern



Die Abbildung zeigt die Lage des Gens SOX5 (mehrere Definitionen) sowie ausgewählter Marker im Bereich 23.500K bis 24.600K des Chromosom 12p12.1 gemäß GeneDB [79]

Tabelle 50 Modellschätzung: LD-Block Nr. 69267

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,72	(2,47- 2,99)	2,1 x10^{-94}	2,72	(2,47- 2,99)	2,2 x10^{-95}
Strahlenexposition	30,5	(1,74- 536)	0,0192	9,58	(5,80- 15,8)	1,0 x10^{-18}
chr12_24207780_C_G	0,96	(0,70- 1,32)	0,8279			
chr12_24252977_C_T	0,99	(0,74- 1,33)	0,9791			
chr12_24281450_A_G	0,80	(0,48- 1,33)	0,4040			
chr12_24281623_C_T	0,92	(0,77- 1,10)	0,4085			
rs1498879	0,91	(0,73- 1,13)	0,3984			
rs2030130	1,01	(0,87- 1,17)	0,8312			
rs7136388	1,09	(0,87- 1,37)	0,4136			
rs7952778	0,92	(0,75- 1,14)	0,4852	0,94	(0,86- 1,03)	0,1976
rs7978583	0,89	(0,71- 1,12)	0,3591			
rs10505933	1,08	(0,85- 1,35)	0,5070			
rs10734732	1,04	(0,82- 1,31)	0,7346			
rs10842275	1,01	(0,82- 1,25)	0,8536			
rs11047241	0,95	(0,68- 1,32)	0,7678			
rs11831634	1,01	(0,81- 1,25)	0,9004			

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Expo. x rs7978583	0,43	(0,07- 2,66)	0,3704	0,48	(0,20- 1,16)	0,1049
Expo. x rs7952778	0,67	(0,13- 3,28)	0,6282			
Expo. x rs7136388	0,70	(0,14- 3,42)	0,6661			
Expo. x rs2030130	0,61	(0,28- 1,30)	0,2032			
Expo. x rs1498879	0,54	(0,10- 2,98)	0,4885			
Expo. x rs11831634	1,42	(0,26- 7,66)	0,6790			
Expo. x rs11047241	0,58	(0,10- 3,18)	0,5388			
Expo. x rs10842275	0,46	(0,12- 1,75)	0,2589	0,42	(0,16- 1,05)	0,0656
Expo. x rs10734732	0,87	(0,15- 5,05)	0,8794			
Expo. x rs10505933	0,87	(0,17- 4,38)	0,8740			
Expo. x chr12_24281623_C_T	0,60	(0,24- 1,51)	0,2817			
Expo. x chr12_24281450_A_G	0,28	(0,02- 2,70)	0,2722	0,15	(0,01- 1,25)	0,0802
Expo. x chr12_24252977_C_T	0,19	(0,03- 1,12)	0,0680	0,34	(0,10- 1,09)	0,0702
Expo. x chr12_24207780_C_G	1,46	(0,27- 7,78)	0,6506			

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	6,1831	14	0,9617	1	0,1976
Interaktion(en) GxE	26,3799	14	0,0232	4	0,000071
Gemeinsamer Effekt (joint)	34,8384	28	0,1746	5	0,000038

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.12 Modellschätzung LD-Blöcke Nr. 69250-69269 (Chr. 12p12.1; SOX5)

Die 19 LD-Blöcke Nr. 69250-69269 um das Gen SOX5 (12p12.1- Abbildung 27) umfassen 97 typisierte Marker. Die meisten davon stehen mit mindestens einem anderen in sehr starkem LD zueinander. Daher werden nur 16 Marker in das Schätzmodell aufgenommen.

Für diese LD-Blöcke konnte nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden, nicht jedoch im taxativen Modell. Die Interaktionsterme von 3 der 16 Marker verbleiben im AIC-Modell (Tabelle 51). Dabei zeigen die Marker rs1031895 (OR=1,45; 95%-CI: 0,9-2,4), chr12_24207780_C_G (OR=1,65; 95%-CI: 1-2,7) und chr12_23788075_A_G (OR=1,34; 95%-CI: 0,8-2,1) eine positive Interaktion für das jeweils seltene Allel.

Tabelle 51 Modellschätzung: LD-Blöcke Nr. 69267-69262

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	1,40	(0,99- 1,98)	0,0522	2,71	(2,46- 2,98)	$1,2x10^{-94}$
Strahlenexposition	0,23	(<0,01- 44,6)	0,5904	1,23	(0,51- 2,94)	0,6380
chr12_23782727_G_T	0,57	(0,06- 4,84)	0,6117			
chr12_23788075_A_G	0,71	(0,11- 4,35)	0,7146			
chr12_23826272_A_G	<0,01		0,9844			
chr12_23889474_A_G	0,97	(0,31- 3,00)	0,9595			
chr12_23969761_A_G	0,49	(0,14- 1,69)	0,2608			
chr12_24025701_A_G	1,10	(0,15- 8,01)	0,9223			
chr12_24185456_A_C	<0,01		0,9486	0,95	(0,81- 1,12)	0,5915
chr12_24207780_C_G	0,79	(0,27- 2,28)	0,6670			
chr12_24252977_C_T	1,56	(0,20- 11,8)	0,6665			
chr12_24295876_A_G	0,80	(0,29- 2,15)	0,6647			
rs1031895	0,51	(0,17- 1,55)	0,2399			
rs4301870	0,68	(0,18- 2,48)	0,5622			
rs9971691	0,86	(0,20- 3,75)	0,8494			
rs10771005	1,03	(0,39- 2,76)	0,9394			
rs16926713	0,69	(0,06- 7,85)	0,7698			
rs17477268	0,53	(0,07- 3,73)	0,5306			
Expo. x rs9971691	0,76	(0,14- 3,99)	0,7460			
Expo. x rs4301870	1,37	(0,32- 5,83)	0,6652			
Expo. x rs17477268	0,98	(0,12- 8,11)	0,9901			
Expo. x rs16926713	0,46	(0,01- 11,0)	0,6355			
Expo. x rs10771005	0,77	(0,25- 2,36)	0,6529			
Expo. x rs1031895	2,46	(0,72- 8,42)	0,1490	1,45	(0,89- 2,36)	0,1331
Expo. x chr12_24295876_A_G	1,40	(0,45- 4,29)	0,5499			
Expo. x chr12_24252977_C_T	0,38	(0,03- 4,15)	0,4324			
Expo. x chr12_24207780_C_G	1,73	(0,53- 5,64)	0,3567	1,65	(1,03- 2,65)	0,0369
Expo. x chr12_24185456_A_C	>999		0,9540			
Expo. x chr12_24025701_A_G	1,41	(0,16- 12,0)	0,7536			
Expo. x chr12_23969761_A_G	2,54	(0,66- 9,73)	0,1734			
Expo. x chr12_23889474_A_G	0,77	(0,21- 2,74)	0,6875			
Expo. x chr12_23826272_A_G	0,30		0,9987			
Expo. x chr12_23788075_A_G	1,68	(0,21- 12,9)	0,6176	1,34	(0,83- 2,14)	0,2198
Expo.chr12_23782727_G_T	1,15	(0,10- 12,5)	0,9055			

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	5,0982	16	0,9952	1	0,7764
Interaktion(en) GxE	9,2046	16	0,9047	3	0,0746
Gemeinsamer Effekt (joint)	20,8263	32	0,9354	4	0,1956

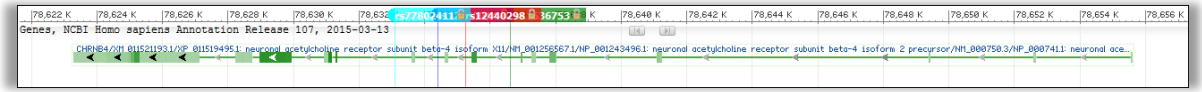
df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.13 Modellschätzung LD-Block Nr. 82003 (Chr.15q25.1, CHRNB4)

Der LD-Block Nr. 82003 im Chromosomabschnitt 15q25.1.2 liegt innerhalb des Gens CHRNB4 (Abbildung 28). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Der Interaktionsterm von einem der 4 Marker verbleibt im AIC-Modell (Tabelle 52). Dabei zeigt der Marker rs12440014 (OR=0,33; 95%-CI:) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition mit OR=4,58 und damit wesentlich höher als in vergleichbaren Modellen geschätzt. Abgesehen von dieser instabilen Schätzung des Radon-Haupteffekts kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker des LD-Blocks Nr. 82003 geschlossen werden.

Abbildung 28 CHRNB4 mit ausgewählten Markern



Die Abbildung zeigt die Lage des Gens CHRNB4 sowie ausgewählter Marker im Bereich 78.622K bis 78.655K des Chromosom 15q25.1 gemäß GeneDB [79]

Tabelle 52 Modellschätzung: LD-Block Nr. 82003

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,71	(2,46- 2,98)	$1,2x10^{-94}$	2,72	(2,47- 2,98)	3,8x10-96
Strahlenexposition	4,51	(3,06- 6,63)	$2,2x10^{-08}$	4,58	(3,15- 6,66)	1,5x10-15
rs11636753	0,98	(0,89- 1,08)	0.7428			
rs12440014	0,97	(0,87- 1,08)	0.6539	0,98	(0,88- 1,09)	0,7700
rs12440298	0,98	(0,78- 1,23)	0.9086			
rs77802411	1,06	(0,78- 1,43)	0.7006			
Expo. x rs77802411	1,07	(0,20- 5,73)	0.9321			
Expo. x rs12440298	1,32	(0,30- 5,81)	0.7076			
Expo. x rs12440014	0,33	(0,16- 0,69)	0.0031	0,33	(0,16- 0,68)	0,0028

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	0,4759	4	0,9758	1	0,7700
Interaktion(en) GxE	8,9752	3	0,0296	1	0,0028
Gemeinsamer Effekt (joint)	10,0091	7	0,1881	2	0,0087

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.14 Modellschätzung LD-Blöcke Nr. 82002-82008 (Chr.15q25.1, CHRNA3, CHRNB4)

Die LD-Blöcke Nr. 82002-82008 im Chromosomabschnitt 15q25.1 liegen um die Gene CHRNA3 und CHRNB4 (Abbildung 29). Für diese LD-Blöcke konnte nicht im taxativen Modell, jedoch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 5 der 19 Marker verbleiben im AIC-Modell (Tabelle 53). Dabei zeigen die Marker rs76912891 (OR=3,7; 95%-CI: 1,3-11) und rs17487514 (OR=2,29; 95%-CI: 1,3-4,2) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen die Marker rs12440014 (OR=0,21; 95%-CI: 0,08-0,6), rs12437528 (OR=0,19; 95%-CI: <5,3) und chr15_78923987_C_T (OR=0,78; 95%-CI: 0,54-1,3) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=3,98 und damit ein wenig höher als in vergleichbaren Modellen geschätzt, wobei der Marker rs73465097 einen nicht-signifikanten Haupteffekt der Größe OR=0,78 markiert. Abgesehen von dieser instabilen Schätzung des Radon-Haupteffekts kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker der LD-Blöcke Nr. 82002-82008 geschlossen werden.

Abbildung 29 CHRB4 mit ausgewählten Markern

Die Abbildung zeigt die Lage des Gens CHRB4 sowie ausgewählter Marker im Bereich 78.680K bis 78.625K des Chromosom 15q25.1 gemäß GeneDB [79]

Tabelle 53 Modellschätzung: LD-Blöcke Nr. 82002-82008

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,66	(2,41- 2,94)	$1,8x10^{-85}$	2,66	(2,42- 2,94)	$1,1x10^{-86}$
Strahlenexposition	4,30	(0,56- 32,5)	0,1577	3,98	(1,95- 8,14)	0,0001
rs11636753	0,93	(0,76- 1,14)	0,5122			
rs12437528	1,13	(0,76- 1,68)	0,5426			
rs12440014	0,95	(0,78- 1,16)	0,6494			
rs12440298	0,97	(0,75- 1,25)	0,8534			
rs12594247	0,94	(0,77- 1,16)	0,6180			
rs12900519	0,94	(0,74- 1,20)	0,6643			
rs17487514	1,01	(0,87- 1,17)	0,8657			
rs60445394	0,83	(0,42- 1,61)	0,5855			
rs72743168	0,95	(0,76- 1,18)	0,6674			
rs73465097	0,77	(0,43- 1,38)	0,3882	0,78	(0,47- 1,29)	0,3476
rs74552499	1,09	(0,39- 3,02)	0,8537			
rs75106522	0,99	(0,71- 1,40)	0,9959			
rs75262975	0,79	(0,33- 1,84)	0,5890			
rs76152270	0,93	(0,60- 1,44)	0,7591			
rs76912891	1,03	(0,81- 1,31)	0,7706			
rs76943320	0,75	(0,28- 1,99)	0,5687			
rs77802411	1,05	(0,75- 1,46)	0,7672			
rs77823196	1,09	(0,53- 2,24)	0,7934			
rs79345755	0,96	(0,48- 1,90)	0,9106			
expo_rs77802411	1,78	(0,25- 12,4)	0,5615			
expo_rs76912891	4,67	(1,20- 18,1)	0,0258	3,73	(1,25- 11,1)	0,0181
expo_rs76152270			0,9809			
expo_rs75106522			0,9586			
expo_rs72743168	0,92	(0,34- 2,48)	0,8755			
expo_rs17487514	2,48	(1,14- 5,40)	0,0216	2,29	(1,25- 4,21)	0,0073
expo_rs12900519			0,9597			
expo_rs12594247	0,73	(0,18- 2,90)	0,6652			
expo_rs12440298	4,92	(0,43- 56,0)	0,1983			
expo_rs12440014	0,26	(0,07- 0,88)	0,0310	0,21	(0,08- 0,55)	0,0013
expo_rs12437528	0,15	(<0,01- 4,73)	0,2820	0,19	(<0,01- 5,27)	0,3338
expo_rs11636753	0,98	(0,34- 2,81)	0,9842			
expo_chr15_78923987_C_T	0,51	(0,19- 1,39)	0,1929	0,78	(0,47- 1,29)	0,3476

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	4,6016	20	0,9999	1	0,3476
Interaktion(en) GxE	17,0624	13	0,1965	5	0,0051
Gemeinsamer Effekt (joint)	22,9101	33	0,9054	6	0,0070

df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.15 ModellschätzungLD-Block Nr. 82566 (Chr.15q26.1, ST8SIA2, snoU109)

Der LD-Block Nr. 82566 im Chromosomabschnitt 15q26.1 liegt im Bereich nahe der Gene C15orf32, ST8SIA2 und snoU109 (Abbildung 30). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 8 der 11 Marker verbleiben im AIC-Modell (Tabelle 42). Dabei zeigen die Marker rs3931230 (OR=5,7; 95%-CI: 0,9-35) und rs3848153 (OR=5,0; 95%-CI: 0,9-29) eine positive Interaktion für das jeweils seltene Allel. Im Gegenzug zeigen die Marker rs2045268 (OR=0,69; 95%-CI: 0,4-

1,3), chr15_93006740_A_G (OR=0,61; 95%-CI: 0,3-1,1) und chr15_92997155_A_G (OR=0,72; 95%-CI: 0,3-1,9) eine negative Interaktion für das jeweils seltene Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=0,08 und damit wesentlich niedriger als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risiko-erhöhende bzw. –senkende Marker diskutierbar. Zusammen betrachtet kann aber auf eine Stratifikation der Fall-Wahrscheinlichkeit unter Radon-Exposition durch die Marker des LD-Blocks Nr. 82566 geschlossen werden.

Abbildung 30 ST8SIA2 mit ausgewählten Markern

Die Abbildung zeigt die Lage der Gene ST8SIA2 und C15orf32 (mehrere Definitionen) sowie ausgewählter Marker im Bereich 92.390K bis 92.500K des Chromosom 15q26.1 gemäß GeneDB [79]

Tabelle 54 Modellschätzung: LD-Block Nr. 82566

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,66	(2,41- 2,93)	$3{,}1\times10^{-87}$	2,67	(2,43- 2,95)	$1{,}5\times10^{-88}$
Strahlenexposition	0,07	(<0,01- 3,65)	0,1927	0,08	(<0,01- 3,64)	0,1970
chr15_92987079_A_T	0,98	(0,65- 1,47)	0,9490			
chr15_92997155_A_G	0,97	(0,87- 1,08)	0,6467			
chr15_93006088_A_G	0,99	(0,83- 1,19)	0,9796			
chr15_93006740_A_G	0,97	(0,86- 1,10)	0,7160			
rs1455773	1,00	(0,84- 1,18)	0,9972			
rs1455777	1,01	(0,67- 1,52)	0,9539			
rs2045268	1,00	(0,83- 1,20)	0,9626			
rs2168351	0,96	(0,80- 1,15)	0,6958	0,97	(0,88- 1,06)	0,5377
rs3848153	0,93	(0,73- 1,18)	0,5740			
rs3931230	0,93	(0,70- 1,22)	0,6068			
rs66861609	1,01	(0,89- 1,14)	0,8418			
Expo. x rs66861609	2,17	(1,11- 4,24)	0,0231	2,27	(1,19- 4,33)	0,0127
Expo. x rs3931230	5,96	(0,83- 42,7)	0,0758	5,66	(0,91- 35,0)	0,0621
Expo. x rs3848153	5,42	(0,91- 32,2)	0,0628	5,01	(0,86- 29,1)	0,0728
Expo. x rs2168351	1,23	(0,44- 3,40)	0,6792	1,31	(0,66- 2,58)	0,4325
Expo. x rs2045268	0,73	(0,25- 2,09)	0,5608	0,69	(0,36- 1,29)	0,2488
Expo. x rs1455777	1,10	(0,04- 27,8)	0,9506			
Expo. x rs1455773	1,30	(0,52- 3,26)	0,5642	1,37	(0,57- 3,28)	0,4791
Expo. x chr15_93006740_A_G	0,62	(0,32- 1,18)	0,1466	0,61	(0,33- 1,14)	0,1277
Expo. x chr15_93006088_A_G	0,73	(0,27- 1,97)	0,5480	0,72	(0,27- 1,89)	0,5126
Expo. x chr15_92997155_A_G	0,87	(0,45- 1,65)	0,6738			
Expo. x chr15_92987079_A_T	0,97	(0,04- 23,1)	0,9871			

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	1,4804	11	0,9996	1	0,5377
Interaktion(en) GxE	23,2919	11	0,0161	8	0,0018
Gemeinsamer Effekt (joint)	26,0518	22	0,2495	9	0,0033

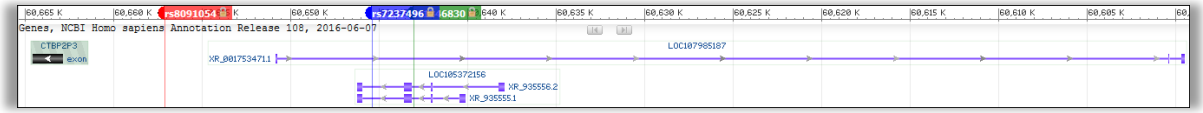
df: Freiheitsgrade (degree of freedom) entspricht der Anzahl geschätzter Effekte

4.10.16 ModellschätzungLD-Block Nr. 91734 (Chr. 18q21.32; LOC107985187, LOC105372156, RP11-325K19.1)

Der LD-Block Nr. 91734 im Chromosomabschnitt 18q21.32 liegt nahe oder innerhalb der Gene LOC107985187 und LOC105372156 (Abbildung 31). Für diesen LD-Block konnte sowohl im taxativen Modell, als auch nach Variablenselektion eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 3 der 4 Marker verbleiben im AIC-Modell (Tabelle 55). Dabei zeigen die Marker rs8091054 (OR=0,42; 95%-CI: 0,06-2,8), rs7237496 (OR=0,27; 95%-CI: 0,04-1,7) und rs1346830 (OR=0,08; 95%-CI: 0,03-0,21) eine negative Interaktion für das jeweils seltene Allel. Der gemeinsame Interaktionseffekt GxE aller drei Marker ist mit $p=1{,}207 \times 10^{-6}$, unter Korrektur nach Bonferroni, suggestives signifikant (siehe Abschnitt 4.3 Genomische Blockstruktur).

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR>43 und damit wesentlich höher als in vergleichbaren Modellen geschätzt. Zusammen betrachtet markieren somit die häufigeren Allele der 3 Marker eine Risikosteigerung.

Abbildung 31 LOC107985187 mit ausgewählten Markern



Die Abbildung zeigt die Lage der Gene LOC107985187und LOC105372156 sowie ausgewählter Marker im Bereich 60.605K bis 60.665K des Chromosom 18q21.32 gemäß GeneDB [79]

Tabelle 55 Modellschätzung: LD-Block Nr. 91734

	taxatives Modell			AIC-Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	2,73	(2,48- 3,00)	$6{,}8 \times 10^{-96}$	2,72	(2,48- 2,99)	$6{,}2 \times 10^{-96}$
Strahlenexposition	197	(40,9- 950)	$4{,}5 \times 10^{-11}$	198	(43,0- 916)	$1{,}2 \times 10^{-11}$
rs1346830	1,00	(0,44- 2,30)	0,9841			
rs7237496	1,02	(0,74- 1,39)	0,8847	1,01	(0,92- 1,11)	0,7416
rs8091054	0,98	(0,69- 1,39)	0,9422			
rs11659206	0,98	(0,43- 2,24)	0,9756			
Expo. x rs8091054	0,42	(0,06- 2,95)	0,3888	0,42	(0,06- 2,83)	0,3741
Expo. x rs7237496	0,27	(0,04- 1,71)	0,1661	0,27	(0,04- 1,68)	0,1633
Expo. x rs1346830	0,09	(0,03- 0,22)	$2{,}0 \times 10^{-07}$	0,08	(0,03- 0,21)	$8{,}9 \times 10^{-08}$

taxatives Modell: Modell mit Haupt- und Interaktionseffekt für jeden Marker eines LD-Blocks;
AIC-Modell: Modell nach Marker-Auswahl gemäß AIC-Kriterium; [1] Odds-Ratio mit 95%-Konfidenzintervall

	taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	0,1119	4	0,9985	1	0,7416
Interaktion(en) GxE	28,6744	3	$2{,}621 \times 10^{-6}$	3	$1{,}207 \times 10^{-6}$
Gemeinsamer Effekt (joint)	30,3142	7	0,000083	4	$4{,}253 \times 10^{-6}$

df: Freiheitsgrade (degree of freedom) entspricht der **Anzahl geschätzter Effekte**

4.11 Gen-Set-Analyse GSA (AP 3)

4.11.1 Methode: Gen-Set Enrichment Analyse (GSEA)

In den vergangenen Jahren wurde eine Vielzahl an Methoden zur Gen-Set-Analyse (GSA) vorgeschlagen. Diese ergänzen die gängigen Methoden zur Analyse von genomweiten Assoziationsstudien (GWAS), indem die gemeinsame Assoziation von Genen, bzw. den zugeordneten Markern, die einer sinnvollen, vordefinierten Gruppe (den Gen-Sets GS) angehören, untersucht wird.[80-86]

Für eine Gen-Set-Analyse sind eine Reihe von Zuordnungen notwendig, die auf Basis von öffentlich zugänglichen Datenbanken getroffen wurden.

(a) Die jeweilige Position der Marker am Genom wurde dem Mapping-File des GAME-ON Konsortiums (Stand 8.4.2015, siehe Bericht des BfS-Projekt 3614S10014 – AP2: Qualitätssicherung der Typisierung des Projekts) entnommen. Diese Positionsangaben entsprechen dem *„human assembly GRCh37.p13"* des *Genome Reference Consortium.*

(b) Die Zuordnung der Marker zu Genen (MtG: marker to gene assignment) erfolgte gemäß *ENSEMBL*[87].

(c) Die Zuordnung von Genen zu Gen-Sets (GtP: Gene to Pathway assignment) erfolgt gemäß *Gene Ontology* (GO)[88] und *Human Genome Nomenclature Committee* (HGNC)[89] (siehe Kapitel 4.11.2).

In dieser Untersuchung wurde das Verfahren der *Gene-Set Enrichment Analyse* (GSEA) von Subramanian, et al. angewandt.[90,91]

Für die GSEA selbst wird zunächst eine gegebene Liste von Markern/LD-Blöcken nach den in der Einzel-/Multimarker-Assoziationsanalyse erzielten p-Werten den Rängen nach geordnet. Der Grad der "Anreicherung an Signifikanz" („enrichment") wird dabei als Enrichment-Score (ES) quantifiziert, indem diese geordnete Liste sequenziell – von Marker zu Marker - durchwandert wird. Zu Anfang wird eine *kumulative Summe der Signifikanz C* auf null gesetzt. Diese wird erhöht, wenn ein Marker dem untersuchten Gen-Set GS angehört, andernfalls wird C verringert. Die Schrittgröße der Zunahme bzw. der Abnahme wird so gewählt, dass C am Ende der Liste wieder null erreicht und stets im Wertebereich zwischen -1 und +1 liegt. Durch diese Normalisierung wird eine Vergleichbarkeit von Gen-Sets unterschiedlicher Größe gewährleistet.

Ist ein Gen-Set mit Markern mit niedrigen p-Werten angereichert, wird C sehr rasch und weit ansteigen. Sind die Marker/LD-Blöcke des GS nicht mit einer Zielerkrankung assoziiert, folgt C einem Zufallspfad um den Wert null. Der Enrichment-Score (ES) selbst ist definiert als das Maximum von C und kann als gewichtete Kolmogorov-Smirnov-Statistik interpretiert werden. GSEA ist daher auch ein Test, ob die Verteilung der, dem GS zugeordneten p-Werte einer Gleichverteilung folgt. Um Abhängigkeiten von Markern in LD zu vermeiden, basiert die durchgeführte GSA auf den p-Werten je LD-Block aus der Multimarker-Assoziationsanalyse.

Der für eine GS erzielte ES wird dann in einem einseitigen, statistischen Test auf dem Signifikanzniveau von 5% getestet. Die Verteilung des ES unter der Null-Hypothese wird durch eine Monte-Carlo-Simulation mit 5000 Wiederholungen je GS generiert. Die GSEA testet dabei eine sogenannte „kompetitive Null-Hypothese", bei der angenommen wird, dass die Gene des untersuchten Gen-Sets dasselbe Ausmaß an kumulierte Assoziation mit dem betrachteten Phänotyp aufweisen, wie alle verbleibenden, vermessenen Gene des Genoms.

Die GSEA definiert jene Gene, deren Marker/LD-Blöcke vor dem Maximum der *kumulative Summe der Signifikanz* C rangieren (und daher zum Enrichment-Score ES beitragen), als „significance driving genes"[92] oder „leading-edge subset"[90]. Damit werden also jene Gene innerhalb des Gen-Sets identifiziert, die zur kumulativen Assoziation wesentlich beitragen.
Bevor eine GSEA durchgeführt werden kann, müssen die zu untersuchenden Gene-Sets definiert werden.

4.11.2 Auswahl von Gen-Sets für die Gen-Set Analyse (AP 3.1)

Lungenkrebs durchläuft einen vielschichtigen Prozess (Tumorgenese) von der initialen Dysbalance zwischen Zellproliferation und Zelltod bis zum diagnostizierbaren Tumor. Dieser Prozess wird durch genetische und epigenetische Varianten begünstigt oder ist durch Veränderungen eben solcher charakterisiert.[93,94] Durch umfangreiche molekulargenetische Studien von spezifischen Genen und molekulargenetische Wirkmechanismen konnten zahlreichen genetische und epigenetische Veränderungen im Tumorgewebe von Lungenkrebspatienten gefunden werden (>20 Veränderungen je Tumor).[12,95,96] Diese „*progression-associated genes and pathways*“, die am Verlauf der Tumorgenese beteiligt sind, sind dabei von „*susceptibility genes*“ zu unterschieden. Letztere sind die Träger jener genetischen Prädisposition, die eine Person von dessen Eltern vererbt bekam und die das Risiko einer Erkrankung beeinflussen. Belegt ist, dass sich *susceptibility und progression-associated* Faktoren untereinander, aber auch zwischen histologischen Subtypen des Lungenkrebses unterschieden. Des Weiteren gilt Rauchen als der wichtigste Risikofaktor des Lungenkrebses. Ihm können (~85%) der nicht-kleinzelligen Lungentumore (NSCLC:) und ~99% der kleinzelligen Lungentumore (SCLC) zugeordnet werden. Tabakrauch besteht aus über 7.000 Substanzen, von denen mehr als 70 als krebserregend eingestuft werden. Alleine 20 davon führen zu genetischen Veränderungen durch DNA-Aduktbildungsreaktionen („*DNA adduct formation*“). Andere exogene Risikofaktoren sind zum Beispiel die in diesem Projekt im Fokus stehende Exposition mit Radon/Strahlung, sowie Arsen oder Asbest.[27,97] Neben dem Rauchen gilt Radon als die zweithäufigste Ursache für Lungenkrebs und wird als Ursache von etwa 10% aller Lungenkrebsfälle angesehen.[5,27] Neben Veränderungen der Epigenetik und der Stabilität des Genoms führen diese exogenen Faktoren, soweit bekannt, auch zu DNA-Strangbrüchen oder zur Aktivierung onkogenetischer Wirkmechanismen.[97] Alle Zerfallsprodukte des natürlich vorkommenden Gases Radon sind radioaktiv. Als gesundheitsgefährdend gilt das emittieren von α-Teilchen, die in der Lage sind eine hohe lineare Energie auf andere Moleküle zu übertragen und diese dabei zu schädigen. Dabei kann eine Exposition durch α-Teilchen zu folgenden Schäden bzw. Zellreaktionen führen:

a) Zerstörung des empfindlichen DNA-Moleküls („*DNA-damage*“) durch komplexe DNA-Schädigung
b) Zerstörung von Zellen durch einen eingeleiteten oder ausgelösten Zelltod (Apoptose; Nekrose)
c) die Veränderung der DNA-Struktur.

Darüber hinaus kann Strahlung zu *de novo*-Mutationen in der DNA führen, die an Nachkommen vererbt werden.[98]

Die hier durchgeführte Fall-Kontroll-Studie zielt primär auf die Entdeckung von „*susceptibility genes*“ ab. Dennoch werden für die Auswahl zu testender Gen-Sets (AP 3.1) auch die „*progression-associated genes and pathways“* weiter betrachtet. Vergleichende Untersuchungen der Gen-Expression zwischen Tumor- und gesundem Gewebe liefern zwar ein umfassendes Profil der genetischen Veränderungen am Ende der Tumorgenese, eine Abgrenzung der biologisch relevante „*driver mutations*“ von denen der großen Mehrheit der „*passenger mutations*“ kann aber nicht erfolgen. [93] Es darf aber – im Sinne einer wissenschaftlichen Arbeitshypothese - vermutet werden, dass zwischen „*susceptibility genes*“ und „*driver mutations*“ eine Verbindung bestehen kann.

Die Auswahl zu testender Gen-Sets beruht daher auf drei Aspekten:

1. Gen-Sets (HGNC Genfamilien und GO-Begriffe), definiert durch in der GWA-Analyse auffälligen Gene bzw. LD-Blöcke
2. Gen-Sets (HGNC Genfamilien), definiert durch publizierte genetische Interaktionen mit einer Radon-Exposition hinsichtlich Lungenkrebs
3. Gen-Sets, Signalwege (GO-Begriffe), definiert durch *progression-associated genes, falls* diese durch bekannte Wirkmechanismen einer Radon- bzw. Strahlungsbelastung plausibel erscheinen

4.11.3 Gen-Sets (HGNC Genfamilien und GO-Begriffe), definiert durch in der GWA-Analyse auffällige Gene bzw. LD-Blöcke

Durch die GWA fielen einige Markern/LD-Blöcken durch eine genomweite signifikante oder suggestive GxE-Interaktion auf (siehe Kapitel 4.10). Im Folgenden werden jene Gene aufgelistet, die in oder Nähe dieser Markern/LD-Blöcken liegen.

4.11.3.1 UBE2U (LD-Block 2271; Chromosom 1p31.3)

„In eukaryotic cells the stability and function of many proteins are regulated by the addition of ubiquitin or ubiquitin-like peptides. This process is dependent upon the sequential action of an E1-activating enzyme, an E2-conjugating enzyme, and an E3 ligase.“[99] Das Gen UBE2U kodiert das Enzym *ubiquitin conjugating E2 U*. Es ist Teil der Genfamilie *Ubiquitin conjugating enzymes E2* (http://www.genenames.org/cgi-bin/genefamilies/set/102; **HGNC:102**) mit 41 Mitgliedern. Diese Familie enthält auch das Gen UBE2N (12q22), für das eine Assoziation zu Lungenkrebs unter Kaukasiern berichtet wurde.[100] UBE2N wie drei weitere UBE2-Gene werden auch DNA-Reparatur-Mechanismen zugeordnet (z.B. **GO:0006282** *regulation of DNA repair*), nicht jedoch UBE2U.

Das Gen UBE2U wird insgesamt 10 GO-Begriffen zugeordnet:

- GO:0000209 protein polyubiquitination
- GO:0000790 nuclear chromatin
- GO:0005524 ATP binding
- GO:0005737 cytoplasm
- GO:0006281 DNA repair
- GO:0016574 histone ubiquitination
- GO:0031625 ubiquitin protein ligase binding
- GO:0033503 HULC complex
- GO:0043161 proteasome-mediated ubiquitin-dependent protein
- GO:0061630 ubiquitin protein ligase activity

Das Gen **UBE2U** wurde der HGNC Genfamilie der *„Ubiquitin conjugating enzymes E2“* (**HGNC:102**; http://www.genenames.org/cgi-bin/genefamilies/set/102) zugeordnet.

4.11.3.2 CSNK1G3 (LD-Blöcke 33135/3313 ; Chromosom 5q23.2)

Das Gen CSNK1G3 (casein kinase 1 gamma 3) kodiert ein Mitglied der Familie der „serine/threonine protein kinases“, die Kaseine und andere säurehaltige Proteine phosphorylieren.[79]

Das Gen CSNK1G3 wurde keiner HGNC Genfamilie zugeordnet.
Das Gen CSNK1G3 wird insgesamt 9 GO-Begriffen zugeordnet:

- GO:0004672 protein kinase activity
- GO:0004674 protein serine/threonine
- GO:0005524 ATP binding
- GO:0006464 cellular protein modification
- GO:0006897 endocytosis
- GO:0007165 signal transduction
- GO:0008360 regulation of cell shape
- GO:0016055 Wnt signalling pathway
- GO:0018105 peptidyl-serine phosphorylation

4.11.3.3 LINC01170 (LD-Blöcke 33135/3313; Chromosom 5q23.2)

LINC01170 ist ein langes intergenetisches und damit keine Protein kodierendes RNA-Gen.[79]
Das Gen LINC01170 wird weder einer HGNC Genfamilie noch einem GO-Begriff zugeordnet.

4.11.3.4 CUBN (LD-Block 58899; Chromosom 10p13)

Das Protein *Cubilin*, das durch das Gen CUBN kodiert wird, ist Bestandteil der Zellmembran mehrerer Gewebe und agiert als Rezeptor für den *„intrinsic factor-vitamin B12 complexe*."[79] Cubilin ist eines von zwei Hauptproteinen, die an der Endozytose (der Aufnahme von Flüssigkeit oder Partikeln aus der Umgebung einer Zelle) beteiligt sind.

Das Gen CUBN wurde keiner HGNC Genfamilie zugeordnet.
Das Gen CUBN wird insgesamt 26 GO-Begriffen zugeordnet:

- GO:0001894 tissue homeostasis
- GO:0004872 receptor activity
- GO:0005215 transporter activity
- GO:0005509 calcium ion binding
- GO:0005765 lysosomal membrane
- GO:0005783 endoplasmic reticulum
- GO:0005794 Golgi apparatus
- GO:0005905 clathrin-coated pit
- GO:0006898 receptor-mediated endocytosis
- GO:0008203 cholesterol metabolic process
- GO:0009235 cobalamin metabolic process
- GO:0010008 endosome membrane
- GO:0015031 protein transport
- GO:0015889 cobalamin transport
- GO:0016020 membrane
- GO:0016324 apical plasma membrane
- GO:0030139 endocytic vesicle
- GO:0031232 extrinsic component of external side of plasma membrane
- GO:0031419 cobalamin binding
- GO:0031526 brush border membrane
- GO:0042157 lipoprotein metabolic process
- GO:0042359 vitamin D metabolic process
- GO:0042803 protein homodimerization
- GO:0042953 lipoprotein transport
- GO:0043202 lysosomal lumen
- GO:0070062 extracellular exosome

4.11.3.5 CD163L1 (LD-Block 68623; Chromosom 12p13.31)

Das Gen CD163L1 kodiert ein Protein aus der "*scavenger receptor cysteine-rich (SRCR)*" Superfamilie (**HGNC:1253**; http://www.genenames.org/cgi-bin/genefamilies/set/1253). Diese Familie ist durch eine 100-110 Aminosäure SRCR-Domäne definiert, die Protein-Protein Interaktionen und Ligand-Bindungen vermittelt.[89]

Das Gen **CD163L1** wird insgesamt 3 GO-Begriffen zugeordnet:

- GO:0005044 scavenger receptor activation
- GO:0005576 extracellular region
- GO:0006898 receptor-mediated endocytosis

Das Gen **CD163L1** wurde der HGNC Genfamilien der *„Scavenger receptors (SCAR)"* (**HGNC:1253**; http://www.genenames.org/cgi-bin/genefamilies/set/1253) zugeordnet.

4.11.3.6 LOC101927882 (LD-Block 68623; Chromosom 12p13.31)

Das Gen LOC101927882 ist ein nicht-kodierendes Pseudogen ohne bekannte Funktion.[79]
Das Gen LOC101927882 wird weder einer HGNC Genfamilie noch einem GO-Begriff zugeordnet.

4.11.3.7 ACSM4 (LD-Block 68623; Chromosom 12p13.31)

Das Gen ACSM4 kodiert ein Protein aus der *acyl-CoA synthetase medium-chain* Genfamilie (**HGNC:40**; http://www.genenames.org/cgi-bin/genefamilies/set/40).

Das Gen ACSM4 wird insgesamt 9 GO-Begriffen zugeordnet:

- GO:0003996 acyl-CoA ligase activity
- GO:0004321 fatty-acyl-CoA synthase activity
- GO:0005524 ATP binding
- GO:0005759 mitochondrial matrix
- GO:0006633 fatty acid biosynthetic process
- GO:0006637 acyl-CoA metabolic process
- GO:0015645 fatty acid ligase activity
- GO:0046872 metal ion binding
- GO:0047760 butyrate-CoA ligase activity

Das Gen **ACSM4** wurde der HGNC Genfamilie der *„Acyl-CoA synthetase family (ACS)"* (**HGNC:40**; http://www.genenames.org/cgi-bin/genefamilies/set/40) zugeordnet.

4.11.3.8 SOX5 (LD-Block 69267; Chromosom 12p12.1)

Das Gen SOX5 kodiert ein Protein aus der "sex determining region Y (SRY-related HMG-box)" und ist der Genfamilie *„SRY-boxes (SOX)"* (**HGNC:757**; http://www.genenames.org/cgi-bin/genefamilies/set/757).

Das Gen SOX5 wird insgesamt 9 GO-Begriffen zugeordnet (keiner davon wurde zuvor gelistet):

- GO:0003677 DNA binding
- GO:0003700 transcription factor activity, sequence-specific DNA binding
- GO:0006355 regulation of transcription, DNA-templated
- GO:0006366 transcription from RNA polymerase II promoter
- GO:0032332 positive regulation of chondrocyte differentiation
- GO:0055059 asymmetric neuroblast division
- GO:0061036 positive regulation of cartilage development
- GO:0071560 cellular response to transforming growth factor beta stimulus
- GO:2000741 positive regulation of mesenchymal stem cell differentiation

4.11.3.9 MIR920 (LD-Block 69267, Chromosom 12p12.1)

Das Gen MIR920 ist eine kurze, nicht-kodierende *microRNA*-Sequenz, das in die post-transitionale Regulierung von Genexpression involviert ist. Es ist Bestandteil der Genfamilie der „m*icroRNAs"* (**HGNC:476**; http://www.genenames.org/cgi-bin/genefamilies/set/476), die aus 1776 kurzen Genen besteht (siehe auch Kapitel 4.11.5.1).

Das Gen MIR920 wurde keinem GO-Begriffe zugeordnet.

4.11.3.9.1 LOC107985187 (LD-Block 91734; Chromosom 18q21.32)

Das Gen LOC107985187 ist ein nicht-kodierendes Pseudogen ohne bekannte Funktion. Es wird weder einer HGNC Genfamilie noch einem GO-Begriff zugeordnet.

4.11.4 Gen-Sets (HGNC *Genfamilien*), definiert durch publizierte genetische Interaktionen mit einer Radon-Exposition hinsichtlich Lungenkrebs

4.11.4.1 SIRT1

SIRT1 wurde als *Susceptibility*-Gen für ein *Plattenepithelkarzinom* in der Lunge („squamous cell lung carcinoma") in der Kohorte ehemaliger Uranbergarbeiter des Colorado-Plateaus identifiziert.[22]

SIRT1 reguliert wichtige biologische Prozesse: Sirtuine in Bakterien, Archaea und bei Eukaryoten. Es konnte eine Funktion als Tumor-*Suppressor* bei radon-induziertem Krebs unter Bergarbeitern für SIRT1 belegt werden. Jedoch ist noch nicht geklärt, ob SIRT1 gegenüber Krebs im Allgemeinen als Tumor-*Promoter* oder als Tumor-*Suppressor* agiert.[101,102] Der Sirtuin-Familie wird dabei eine zentrale Rolle in der Regulation einer Mehrzahl an molekular-genetischen Signalpfaden zugeschreiben. Diese sind gemäß Lin and Fang, 2013 [102]:

a) SIRT1-p53 Achse: p53 wird dabei sowohl transkriptional als auch post-transkriptional von SIRT1 reguliert. Der p53-Signalweg wird gesondert betrachtet (siehe Kapitel 4.11.4.3).

b) SIRT1 und FOXO: „FOXO proteins are phylogenetically conserved and regulate key physiological functions, including cell proliferation, cell differentiation, and survival, and their dysregulation is associated with tumorigenesis".[102] SIRT1 interagiert dabei mit allen Genen der HGNC FOXO-Genfamilie *(**HGNC:508**; http://www.genenames.org/cgi-bin/genefamilies/set/508).*

c) SIRT1 und Autophagozytose: „Mechanically, SIRT1 interacts with several autophagy (ATG) proteins, such as Atg5, Atg7, and Atg8, to regulate autophagy".[102] SIRT1 interagiert dabei mit allen Genen der HGNC ATG-Genfamilie *(**HGNC:1022;** http://www.genenames.org/cgi-bin/genefamilies/set/1022).*

d) SIRT1 im TGF-β Signalweg: „SIRT1 has been demonstrated to participate in the regulation of TGF-β signalling by interacting with and deacetylating Smad7 and Smad3".[102] SIRT1 interagiert dabei mit Genen der HGNC SMAD-Genfamilie *(**HGNC:750;** http://www.genenames.org-/cgi-bin/genefamilies/set/750).*

e) SIRT1 im Wnt/β –Signalweg (siehe Kapitel 4.11.5.3 Abschnitt Wnt/β-catenin Signalweg*)*

g) SIRT1 und die AP-1 transkriptionale Aktivierung: SIRT1 unterdrückt die transkriptionale Aktivität von AP-1 (aktueller Genname: JUN) in Immunzellen und während der Tumorgenese. → Die Genfamilie „AP-1 transcription factor" besthe dabei aus zwei kleinen Unterfamilien: SIRT1 interagiert dabei mit 4 Genen der *„Fos transcription factor family"* (**HGNC:1256** http://www.genenames.org/cgi-bin/genefamilies/set/1256) und den 3 Genen der *"Jun transcription factor family"* (**HGNC:1257:** http://www.genenames.org/cgi-bin/genefamilies-/set/1257).

Das Gen **SIRT1** wurde ferner der HGNC Genfamilie der *„Sirtuins (SIRT)"* zugeordnet (**HGNC:938:** http://www.genenames.org/cgi-bin/genefamilies/set/938).

4.11.4.2 GSTM, GSTT und EPHX1

Ruano-Ravina, et al., 2014 [19] haben ein höheres Lungenkrebsrisiko bei gleicher Innenraum-Radon-Belastung (bei - im Vergleich zur Exposition der Wismut-Bergarbeiter - niedrigen Dosen von 51-147 Bq/m³ bzw. >147 Bq/m³) bei Absenz von GSTM1 oder GSTT1 beobachtet. Kein klarer Effekt wurde für EPHX1 beobachtet, aber auch keine additive GxE-Interaktion für eines der drei Gene belegt. Diese Ergebnisse werden durch die Ergebnisse von Bonner, et al., 2006 [103] unterstützt. *„A plausible explanation is based on the knowledge that high-linear energy transfer α-particle, like those emitted by inhaled radon, causes DNA damage by the metabolic generation of reactive oxygen species (ROS), which in turn, in the absence of regulatory enzymes such as glutathione S-transferase, induce signalling pathways that could preclude cancer appearance."*[19]
Eine ähnliche GxE-Interaktion hinsichtlich des Lungenkrebses wurde für GSTM1 und Passivrauchen belegt.[103]
Jedoch sollte die Möglichkeit des „Confounding" durch ähnliche Effekte hinsichtlich Tabak-Teer oder entzündlichen Lungenerkrankungen in Betracht gezogen werden. *„The expression level of GSTs has been shown to be a factor that determines the cellular sensitivity to a broad spectrum of*

toxic chemicals. However, the regulation of the expression of the GST gene families is complex, as they exhibit sex-, age-, tissue-, and species-specific patterns of expression.“[104] GSTs sind in der Entwicklung vieler Lungenkrankheiten (Asthma, COPD, Entzündungen) involviert.
Die Gene **GSTT1, GSTT2** und **GSTM1-GSTM5** wurden der HGNC Genfamilie der „*Glutathione S-transferases (GST)*“ (**HGNC:564:** http://www.genenames.org/cgi-bin/genefamilies/set/546) zugeordnet.

4.11.4.3 P53

Mutationen im Gen p53 (TP53) wurden in Lungenkrebsgewebe von radon-exponierten Bergarbeitern aus New Mexico und in schwedischen LK-Patienten mit langjähriger Innenraum-Radon-Exposition gefunden.[23-25] „*Various studies have examined the role of mutations of the p53 and p16 tumor suppressor loci, but no particular locus has thus far been proven to be predominant*“.[27]

Das Gene **p53** ist im **GO:0072331** *signal transduction by p53 class mediator* enthalten.
Das Gene **p53** wurde keiner Genfamilie zugeordnet.

4.11.4.4 CDKN2A und MGMT

Exposition gengenüber Radon über längere Zeit wurde unter chinesischen Bergarbeitern als mit verstärkter DNA-Methylierung am Promoter der Gene CDKN2A und MGMT assoziiert beobachtet.[105] Interessanterweise wurde für eine Exposition gegenüber Plutonium, dessen Wirkung ebenfalls über alpha-Teilchen vermittelt wird, eine Inaktivierung von CDKN2A durch abgeschwächte DNA-Methylierung beobachtet.[21]

Das Gen **CDKN2A**: *cyclin dependent kinase inhibitor 2A* (alternativer Name: CDK4 inhibitor p16-INK4) kodiert zwei Proteine: p16 (oder p16INK4a) und p14arf. (GeneCards: http://www.genecards.org/cgi-bin/carddisp.pl?gene=CDKN2A). Beide Proteine sind involviert in die Regulierung des Zell-Zyklus und haben eine tumor-suppressive Wirkung. Dennoch wurde das Gen **CDKN2A** keinem GO-Begriff zugeordnet.

O(6)-methylguanine-DNA methyltransferase (**MGMT**) ist eine der bedeutendsten DNA-Reparatur Enzyme in der Abwehr von häufig auftretenden Karzinogene wie Alkylate oder Tabak.[106]

Das Gen **MGMT** ist im **GO:0006307** „*DNA dealkylation involved in DNA repair*“ enthalten.
Weder **CDKN2A** noch **MGMT** wurden einer Genfamilie zugeordnet.

4.11.4.5 Interleukin 6 (IL6)

Das Gen IL6 kodiert ein Zytokin, das bei entzündlichen Prozessen und der Bildung von B-Zellen eine aktive Rolle spielt. Eine Assoziation zwischen IL6 und Lungenkrebs (Plattenspielepithel) wurde in der *Saccomanno Uran-Bergarbeiter Kohorte* beobachtet und in den Fall-Kontroll-Studien GENEVA (Gene Environment Association Studies) und PLCO (Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial) repliziert.[20]

Das Gen **IL6** wurde sowohl der HGNC Genfamilie der Interleukine (43 Gene; **HGNC:1264** http://www.genenames.org/cgi-bin/genefamilies/set/1264) als auch der HGNC Genfamilie der Interleukin 6 Zytokin-Familie (7 Gene; **HGNC:598:** http://www.genenames.org/cgi-bin/genefamilies/set/598) zugeordnet.

4.11.5 Gen-Sets (GO-Begriffe), definiert durch progressionsassoziierte Gene, falls diese durch bekannte Wirkmechanismen einer Radon- bzw. Strahlungsbelastung plausible erscheinen

Da Strahlung zur Schädigung der DNA an jeglicher Position führen kann, gelten alle jene Gene als Kandidaten für „*susceptibility genes*“ des Lungenkrebses, die an der Metabolisierung von Karzino-

genen oder in der DNA-Reparatur beteiligt sind.[107,108] Aus diesem Grund werden die folgenden Gen-Sets (GO-Begriffe) in die Gen-Set-Analyse aufgenommen:

- GO:0036473 cell death in response to oxidative stress
- GO:0070265 necrotic cell death
- GO:0006915 apoptotic process
- GO:0097468 programmed cell death in response to reactive oxygen species
- GO:0097300 programmed necrotic cell death
- GO:0006281 DNA repair
- GO:0007165 signal transduction

4.11.5.1 Epigenetische Veränderungen mit Bezug zu Lungenkrebs

„Epigenetic modifications refer to a number of molecular mechanisms that regulate gene expression without changing the DNA sequence. These include the following: 1) alteration of the methylation status of DNA within CpG islands, with hyper-methylation of CpG island promoters of tumour suppressor genes leading to their silencing.“ [95] Eine herausragende Rolle scheinen dabei den Gen let-7 zuzukommen (zu mindestens hinsichtlich tabakrauchbedingtem Adenokarzinom). Die Gene der miRNA-Genfamilie let-7 von hemmt die Expression des RAS-Proteins und reguliert die Expression andere Gene des Zell-Zyklus sowie von *„DNA damage response genes*". Let-7 selbst zeigt im Vergleich zu gesunden Gewebe eine verminderte Expression in Lungenkrebsgewebe. Dabei konnten Expressionsmuster der Gene Let-7a-1 und Let-7f-1 identifiziert werden, die bei geringer Expression mit einer verkürzten Überlebenszeit korrelieren.[95,109]

In die Gen-Set-Analyse wird daher die HGNC Genfamilie der *„MicroRNAs (MIR)*“ - beschränkt auf let-7-Gene - aufgenommen (**HGNC:476:** http://www.genenames.org/cgi-bin/genefamilies/set/476).

4.11.5.1.1 Anomalitäten im Signalweg des „growth-stimulatory signalling“

„Genetic abnormalities linked to the risk of lung cancer should be regarded in the context of signalling pathways having changed their main function. … Most stimulatory signalling pathways are led by oncogenes, which drive cells towards a malignant phenotype, proliferation and escape from apoptosis“.[95] Brambilla et al. listen folgende Signalwege mit Bezug zu Lungenkrebs auf:

4.11.5.1.2 Signalweg des „epidermal growth factor receptor signalling“

In die Gen-Set-Analyse werden 9 GO-Begriffe durch ihren Bezug zum *„epidermal growth factor receptor* (EGRF)“-Signalweg [93-95,97] aufgenommen:

- GO:0000165 MAPK cascade [96]
- GO:0038127 ERBB signalling pathway
- GO:0007173 epidermal growth factor receptor signalling pathway
- GO:0038128 ERBB2 signalling pathway
- GO:0038129 ERBB3 signalling pathway
- GO:0038130 ERBB4 signalling pathway
- GO:1901185 negative regulation of ERBB signalling pathway
- GO:1901186 positive regulation of ERBB signalling pathway
- GO:1901184 regulation of ERBB signalling pathway

4.11.5.1.3 „Ras/mitogen-activated protein“-Kinase und der PI3K/Akt-Signalweg

In die Gen-Set-Analyse werden der folgende GO-Begriff durch seinen Bezug zum *„Insulin/PI3K/PTEN/AKT mTOR signalling arm*“ [93-96] aufgenommen:

- GO:0038201 TOR complex

4.11.5.1.4 ALK: „Anaplastic lymphoma kinase fusion“ Protein

In die Gen-Set-Analyse wird der folgende GO-Begriff durch seinen Bezug zum *„Anaplastic lymphoma kinase“* aufgenommen:

- GO:0007169 transmembrane receptor protein tyrosine kinase signalling pathway

Das zentrale Gen dieses GO-Begriffs ist ALK. In die Gen-Set-Analyse wird daher auch die **HGNC** Genfamilie *„Receptor Tyrosine Kinases“* aufgenommen (**HGNC:321:** http://www.genenames.org/cgi-bin/genefamilies/set/321).

4.11.5.1.5 Thyroid transcription factor 1 (NKX2-1 ; alternativer Name TITF1)

Das Gen NKX2-1 ist Teil der HOX-Genfamilie (314 *homeotic genes* [110]), speziell der 69 Gene umfassenden NKL *subclass homeoboxes and pseudogenes*. *„Many reports have shown that the protein products of HOX genes also play key roles in the development of cancers. Based on our review of the literature, we found that the expression of HOX genes is not only up- or downregulated in most solid tumors but also that the expression of specific HOX genes in cancers tends to differ based on tissue type and tumor site. … Blocking the activity of HOX proteins by interfering with their binding to the PBX cofactor caused … lung cancer cells to undergo apoptosis in vitro.“* [111] Die 235 funktionalen und die 65 Pseudogene der *„homeobox“*-Genfamilie können dabei in 13 Klassen unterteilt werden (Tabelle 56). [110]

Tabelle 56 Klassifikation und Nomenklatur aller *„homeobox“*-Gene

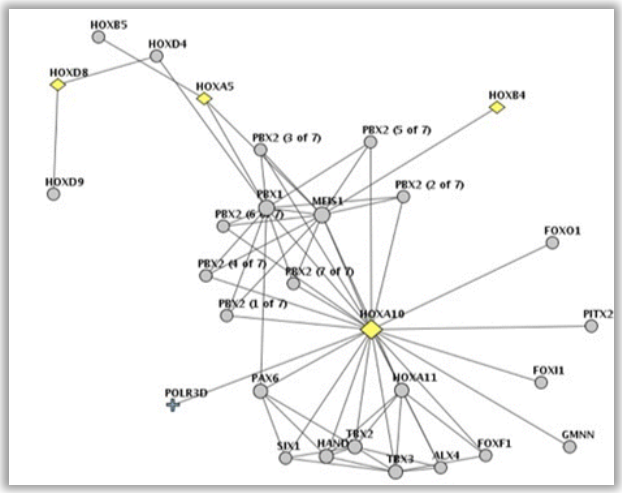
Class	Subclass	Number of gene families	Number of genes	Number of pseudogenes	HGNC Genfamilie Nr.
ANTP	HOXL	14	52	0	518
	NKL	23	48	19[b]	519
PRD	PAX	3	7[a]	0	521
	PAXL	28	43	24[c, d]	
LIM		6	12	0	522
POU		7	16	8[e]	523
HNF		2	3	0	524
SINE		3	6	0	525
TALE		6	20	10[f]	526
CUT		3	7	3[g]	527
PROS		1	2	0	528
ZF		5	14	1[h]	529
CERS		1	5[i]	0	530
Totals		102	235 [a]	65 [b-h]	

gemäß Table 1 in [110]

In die Gen-Set-Analyse werden daher 12 HGNC definierten HOX-Genfamilien aufgenommen (HGNC:518 bis HGNC:530, nicht HGNC:520).

„…Because many HOX genes were found to be aberrantly expressed in multiple different cancer types, we [Bhatlekar et al.] *speculated that there may be similar HOX-related regulatory networks that become dysregulated during cancer development. To investigate this possibility, we used bioinformatic analysis to identify possible regulatory gene networks for HOX genes that are aberrantly expressed in …* [the] *lung.“* (Abbildung 32).[111]
Die Gene dieses Netzwerks werden als gesondertes, publikationsbasiertes Gen-Set (LK-HOX-Familie) in die Gen-Set-Analyse aufgenommen.

Abbildung 32 Netzwerk aus regulatorischen HOX-Genen in soliden Lungentumoren



gemäß Figure 2 in [111]: Predicted regulatory gene networks for HOX genes in solid tumors: lung cancer

Tabelle 57 HOX -Gene in regulatorischen Netzwerken bezüglich Lungenkrebs

Gene name	Description
HOXD9, HOXD8, HOXB5, HOXD4, HOXA5, HOXB4, HOXA10, HOXA11	
Non-HOX genes	
PBX1	Pre-B-cell leukemia homeobox 1
MEIS1	Meis homeobox 1
FOXB1	Forkhead box B1
ALX4	ALX homeobox 4
PBX2	Pre-B-cell leukemia homeobox 2
PAX6	Paired box 6
POLR3D	Polymerase (RNA) III (DNA directed) polypeptide D
SIX1	SIX homeobox 1
HAND	Heart and neural crest derivatives
TBX2	T-box 2
TBX3	T-box 3
GMN	Protein GMN
PITX2	Paired-like homeodomain 2
FOXO1	Forkhead box O1
FOXI1	Forkhead box I1
FOXI1	Forkhead box I1

gemäß Figure 2 in [111]:

4.11.5.1.6 MYC family

MYC reguliert die Expression von etwa 15 % aller menschlichen Gene durch die Bindung von Enhancer Box Sequenzen (E-boxes) und durch die Rekrutierung von Histon-Acetyltransferasen.[95] *„The MYC family of proteins is a group of basic-helix-loop-helix-leucine zipper transcription factors that feature prominently in cancer. Overexpression of MYC is observed in the vast majority of human malignancies and promotes an extraordinary set of changes that impact cell proliferation, growth, metabolism, DNA replication, cell cycle progression, cell adhesion, differentiation, and metastasis.“* Jedoch *„… the coding sequence of MYC does not need to be changed in order for its oncogenic potential to be unleashed.“*[112]

MYC ist ein einzelnes Gen, das keiner spezifischen Genfamilie zugeordnet ist. Es ist im „*canonical Wnt signalling*“-Signalweg (GO:0060070, siehe Kapitle 4.11.5.3 Abschnitt Wnt/β-catenin Signalweg) enthalten.

4.11.5.2 Anomalitäten in Signalwegen der Tumor-Suppression

Die folgende Auflistung folgt der Einteilung gemäß Brambilla et al.[95]

4.11.5.2.1 p53 Signalweg

Siehe Kapitel 4.11.4.3 *P53*.

4.11.5.2.2 p16INK4/cyclin D1/Rb-Signalweg

Der p16INK4A/RB-Signalweg reguliert den Zellzyklus zwischen den sogenannten Phasen G1 to S.[93,95]

In die Gen-Set-Analyse wurde daher folgender GO-Begriff aufgenommen:

- GO:0000083 regulation of transcription involved in G1/S transition of mitotic cell cycle

4.11.5.3 Programmierter Zelltod

4.11.5.3.1 Mitochondriale Apoptose (Bax/Bcl-2)

„Bcl-2 (anti-apoptotic) and Bax (pro-apoptotic) are key factors of mitochondrial apoptosis in controlling the mitochondrial outer membrane permeabilisation, which leads to release of cytochrome C, the point of no return in the cell's commitment to apoptosis."[95]

In die Gen-Set-Analyse wurde daher folgender GO-Begriff aufgenommen:

- GO:0097345 mitochondrial outer membrane permeabilization

4.11.5.3.2 Deregulation der Todesrezeptor

Brambilla et al.[95] benennen zwei Gene des *„Death Receptor"*-Signalwegs als relevante in der Tumorgenese des Lungenkrebses: **FasL** und **E2F1**. *FasL* wird in 70% der NSCLC als *„down-regulated"* genannt. *E2F1* wird in Verbindung mit dem p53-Rb-Signalweg gebracht, aber auch als Apoptosefaktor genannt. Der *„Death Receptor"*-Signalweg wurde von *Gene Ontology* unter der Nummer GO:0008624 geführt und später in den Signalweg *GO:0097190 apoptotic signalling pathway* integriert.

In die Gen-Set-Analyse werden daher folgende GO-Begriffe aufgenommen:

- GO:0036337 Fas signalling pathway
- GO:0097190 apoptotic signalling pathway

4.11.5.3.3 Immortalisierung von Zellen und Aktivierung der Telomerase

„Telomerase is an RNA protein complex responsible for telomere repeat synthesis. Telomeres maintain genomic integrity in normal cells, and their progressive shortening during successive cell divisions induces chromosomal instability." [95,113] Einer der ersten durch GWAS identifizierten und bestätigten *„susceptibility loci"* war die genomische Region 5p15.33 die das Gen TERT enthält. TERT ist eine *„reverse transcriptase component"* der Telomerase.[114,115]

In die Gen-Set-Analyse wurde daher folgender GO-Begriff aufgenommen:

- GO: 0003720 telomerase activity

4.11.5.3.4 Wnt/β-catenin Signalweg

Ding et al. Fanden, bei einem Vergleich von gesunden und Tumorgewebe, vermehrt somatische Mutationen in Genes des *Wnt signalling -Signalweg*,.[96,97] Dabei unterscheidet man zumindest drei Wnt-Signalwege: den "*canonical/β-catenin*"-Signalweg, den "*planar cell polarity (PCP)*"-Signalweg und den "*Wnt/Ca^{2+}*"-Signalweg. Die zentrale Rolle der Wnt-Signalwege in der Funktion der Selbsterneuerung von Krebszellen und deren Differenzierung (von *„cancer stem cells"*) in Zellen unterschiedlichsten Zelltyps ist allgemein akzeptiert. *„One of the hallmarks of stem cells is their ability to*

maintain long telomeres by function of the TERT gene. TERT expression was found to be directly enhanced by binding of β-catenin to its promoter region and thereby links telomerase activity to Wnt signalling". [116]

In NSCLC wird die Wnt-Aktivität durch eine erhöhte Expression von β-catenin vermittelt.

In die Gen-Set-Analyse wurden daher folgende GO-Begriffe aufgenommen:

- GO:0060070 canonical Wnt signalling pathway
- GO:1904886 beta-catenin destruction complex disassembly (Teil von GO:0060070)
- GO:0060071 Wnt signalling pathway, planar cell polarity pathway
- GO:0007223 Wnt signalling pathway, calcium modulating pathway

4.11.6 Übersicht: Auswahl der Gen-Sets: Genfamilien und Signalwege

Insgesamt: 148 Gen-Sets wurden für die Gen-Set-Analyse ausgewählt. Darunter befinden sich 119 GO-Begriffe, 28 HGNC Genfamilien und 1 neu zusammengestellter Gen-Set (LK-HOX-Familie). 7 GO-Begriffe (GO:0009380, GO:0010213, GO:0036299, GO:0038130, GO:0046787, GO:0100026 und GO:1902113) zeigten sich hinsichtlich der hier zugeordneten Gene als identisch. Sechs dieser GO-Begriffe sind Teilaspekte der DNA-Reparatur, der siebente (GO:0038130) betrifft den *ERBB4 signalling-Signalweg*, der wiederum mit DNA-Reparatur in Verbindung steht.[117] Daher wird von diesen 7 GO-Begriffen nur GO:0009380 *excinuclease repair complex* im Weiteren berücksichtigt. Weiter 22 Gen-Sets (18 GO-Begriffe: GO:0000725, GO:0006290, GO:0033503, GO:0036337, GO:0036473, GO:0038127, GO:0043504, GO:0045004, GO:0045738, GO:0055059, GO:0072331, GO:0097196, GO:0097300, GO:0098504, GO:1901184, GO:1901186, GO:1990391, GO:2000741; und 4 HGNC Genfamilien: HGNC:1256, HGNC:1257, HGNC:524, HGNC:528) bestehen aus weniger als 5 Genen und wurden von der Gen-Set-Analyse ausgeschlossen.

Es verbleiben somit **120 Gen-Sets**, darunter 95 GO-Begriffe, 24 HGNC Genfamilien und ein publikationsbasiertes Gen-Set. Diese Gen-Sets bestehen aus 6 bis 3946 Genen (Median: 47).

Da in der Gen-Set-Analyse die p-Werte der Modelschätzung je LD-Block Eingang finden, wurde die GtP (*Gene-to-Pathway*) Annotation in eine LDtP (*LD-Block-to-Pathway*) umgeschrieben. Keines der sehr kurzen Gene der HGNC Genfamilie 476b *microRNAs LET7* konnte einem LD-Block zugeordnet werden. Dieses Gen-Set bleibt daher bei der GSA unberücksichtigt. (Anmerkung: Die HGNC Genfamilie 476 *microRNAs* besteht im Original aus 1771 Genen, kann aber 159 LD-Blöcken zugeordnet werden.*)*

Es verbleiben somit **119 Gen-Sets**, die für die Gen-Set-Analyse ausgewählt wurden, darunter 95 GO-Begriffe, 23 HGNC Genfamilien und ein publikationsbasierter Gen-Set. Diese Gen-Sets bestehen aus 5 bis 7237 LD-Blöcken (Median: 67).

Tabelle 58 Gene-to-Pathway (*GtP*)-Annotation - Übersicht

Größe der Gen-Sets (>5 Genes)	Anzahl Gene							
	6-10	10-50	51-100	>100	N	Min	Max	Median
GO-Begriff	15	27	16	37	**95**	6	3.946	55
HGNC Genfamilie	6	14	3	1	**24**	6	1.776	26
Publikationsbasierter	--	1	--	--	**1**	24	24	24
Gesamt	*21*	*42*	*19*	*38*	***120***	*6*	*3.946*	*46*

publikationsbasierter: LK-HOX-Familie

Tabelle 59 LB-Block-to-Pathway (LDtP)-Annotation - Übersicht

Größe der Gen-Sets (>5 Genes)	Anzahl LD-Blöcke								
	1-5	6-10	10-50	51-100	>100	N	Min	Max	Median

Größe der Gen-Sets (>5 Genes)	Anzahl LD-Blöcke								
	1-5	6-10	10-50	51-100	>100	N	Min	Max	Median
GO-Begriff	2	10	21	15	47	**95**	5	7.237	95
HGNC Genfamilie	1	3	9	9	1	**23**	5	159	45
publikationsbasierter	--	--	1	--	--	**1**	41	41	41
gesamt	3	13	31	24	48	**119**	5	7.237	67

publikationsbasierter: LK-HOX-Familie

4.11.7 GSA Ergebnisse (AP 3.2)

Von den 119 untersuchten Gen-Sets weisen zwei einen p-Wert kleiner 0,05 (nominale Signifikanz) auf, zwei weitere einen p-Wert knapp über 0,05 (siehe Tabelle 76).

Für das **Gen-Set *GO:0006307 „DNA dealkylation involved in DNA repair“,*** das 10 Gene mit 90 typisierten Markern umfasst, wird eine p-Wert von $p_{GS} = 0{,}0139$ erzielt (Tabelle 60). Werden, in einem methodisch alternativen Ansatz, die dem GS zugehörigen LD-Blöcke als signifikant (p≤0,05) und nicht-signifikant klassifiziert, kann mit einem einseitigen Fishers exaktem Test ein noch kleiner p-Wert (=höher Signifikanz) für eine Anhäufung interaktionstragender LD-Blöcke erzielt werden ($p_{GS,Fischer} = 0{,}0060$). 15 der 90 LD-Blöcke des GS (16%), aber nur 6.404 aller verbleibenden 90.768 LD-Blöcke in vermessenen Genom (7%), werden als signifikant klassifiziert.

Die Funktion dieses *GO*-Begriffs wird von *Gene Ontology* (GO) wie folgt beschrieben: *„The repair of alkylation damage, e.g. the removal of the alkyl group at the O6-position of guanine by O6-alkylguanine-DNA alkyltransferase (AGT)*“. GO:0006307 ist in der GO-Hierarchie sowohl eine direkte Tochter von *DNA repair* (GO:0006281) als auch von DNA dealkylation involved in DNA repair (GO:0035510) / *DNA modification* (GO:0006304).

Es wurden insgesamt 7 „driving“-Gene mit 21 „driving“-LD-Blöcken für dieses Gen-Set durch die GSEA deklariert, allen voran das Gen **FTO** (*Fat mass and obesity-associated protein*) auf Chromosom 16q12.2 (Tabelle 61, Abbildung 33). FTO wird auch als **ALKBH9** - *alpha-ketoglutarate dependent dioxygenase* – benannt. Das kodierte Protein ist eine Dioxygenase, die in die Reparatur durch oxidative Demethylierung, alkylierter DNA bzw. RNA involviert ist.[118] Der LD-Block NR. 84616 weist dabei mit p=0.0005 die größte Signifikanz einer GxE-Interaktion auf. Auch für die benachbarten LD-Blöcke Nr.84613 bis Nr. 84619 konnten mit p-Werte ≤0.05 Hinweise auf eine GxE-Interaktion beobachtet werden. Die betreffende Region erstreckt sich von 53.951.562 bis 53.995.500 auf Chromosom 16q12.2.

Für die Genfamilie **HGNC:476 *„microRNAs“,*** die gemäß Definition aus 1776 Gene besteht, aber nur durch 147 typisierten Markern abgedeckt sind, wird ein p-Wert von $p_{GS} = 0.0159$ erzielt (Tabelle 60).

MicroRNAs sind nicht-kodierende RNA-Moleküle mit einer Länge von nur etwa 20 Basenpaaren. MicroRNAs spielen eine wichtige Rolle im komplexen Netz der Genregulation, insbesondere bei der Gen-Stilllegung („Gen-Silencing“). Die Genregulation erfolgt durch Bindung der microRNAs an den 3’-UTR Bereich von Zielgenen auf der mRNA, wodurch die Translation in Proteine gehemmt oder die Gensequenz durch Zerschneiden abgebaut wird.[119]

Es wurden insgesamt 38 „driving“-Gene mit 44 „driving“-LD-Blöcken für diese Genfamilie durch die GSEA deklariert, die über mehrere Chromosomen verstreut sind (Abbildung 37). Die *microRNAs* mit den signifikantesten Einzelgen-Assoziationen sind MIR1207, MIR1208 und MIR608. Die *mircoRNA* MIR920 ist eine der Gene des bereits durch die Multimarker-Assoziationsanalyse auffälligen LD-Blocks Nr. 69267 auf Chromosom 12.12.1.

Die Gen-Sets **GO:0006637** „acyl-CoA metabolic process“ und **GO:0016020** „Membrane (cellular-component)“ bestehen aus 23 bzw. 1896 Gene und erzielten bei der GSEA p-Werte von 0,0538 bzw. 0,0558.

Die 11 „driving“-Gene des Gen-Sets GO:0006637 liegen über mehrere Chromosome verteilt. Für keine der beteiligten LD-Blöcke allein konnte auch nur eine annähernd signifikante Interaktion beobachtet werden (alle p>0,05).

Die 90 „driving“-Gene des Gen-Sets **GO:0016020** *„Membrane (cellular-component)“* liegen über alle Autosomen verteilt. Das Gen-Set kann als weitgefasster Überbegriff von 24 GO-Tochter-Begriffen angesehen werden. Die p-Werte eines Tests auf Interaktion der beteiligten LD-Blöcke streuen von ~0,01 bis $1x10^{-6}$. Das auffälligste (am signifikantesten assoziierte) Gen ist ***CUBN***, auf Chromosom 10p13. Für den entsprechenden LD-Block Nr. 58899 wurde für die GxE-Interaktion im Multimarker-Assoziationsmodell mit Variablenselektion (AIC-Modell) ein p-Wert von $1,3x10^{-5}$ erzielt. Die Funktion von CUBN wird wie folgt beschrieben: „*Cubilin (CUBN) acts as a receptor for intrinsic factor-vitamin B12 complexes.*“ Bemerkenswert ist auch, dass die Aktivität und Expression von Cubilin, angeregt von Tretinoin, in Krebszellen erhöht ist.[120]

Tabelle 60 GSEA: Übersicht der Ergebnisse (Auszug grenzwertig signifikanter Ergebnisse)

Gen-Set ID	Beschreibung	Anzahl Gene	Anzahl Marker	Anzahl „driving“-Marker	p_{GS}-Wert
GO:0006307	*DNA dealkylation involved in DNA repair*	10	90	21	0,0139
HGNC:476	*microRNAs*	1776	147	44	0,0159
GO:0006637	*acyl-CoA metabolic process*	23	36	20	0,0538
GO:0016020	*membrane (cellular-component)*	1896	5903	178	0,0558

Eine komplette Übersicht der GSEA-Ergebnisse siehe Tabelle 76 im Anhang.

4.11.7.1 GSA –Ergebnisse: GO:0006307 An DNA-Reparatur beteiligte DNA-Dealkylierung

Abbildung 33 Manhattan-Plot GO:0006307 *An DNA-Reparatur beteiligte DNA-Dealkylierung*

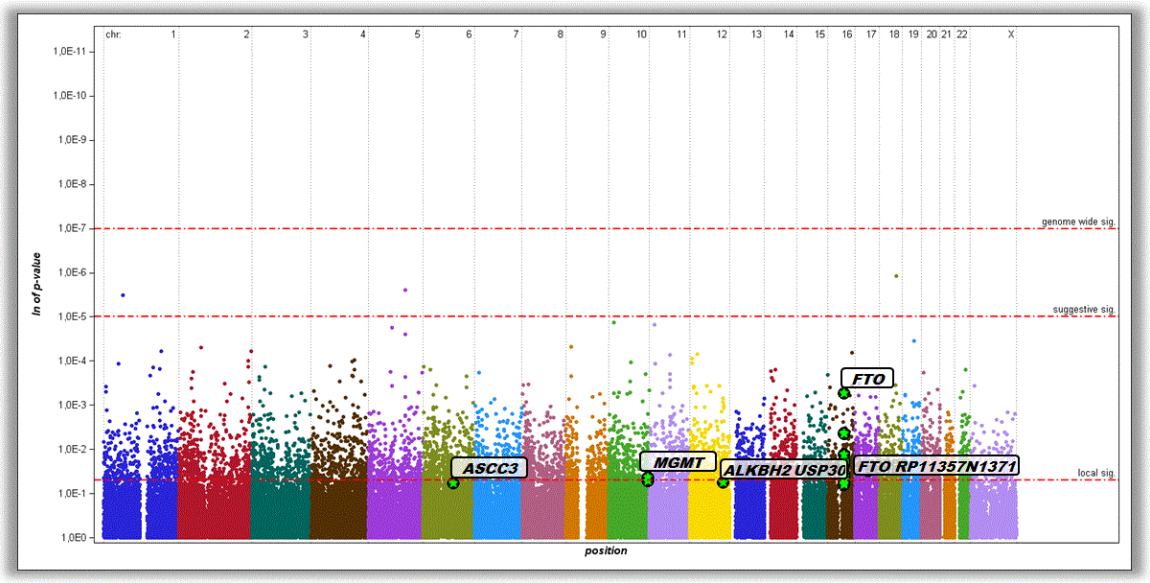


Abbildung 34 GSEA: GO:0006307 *An DNA-Reparatur beteiligte DNA-Dealkylierung*

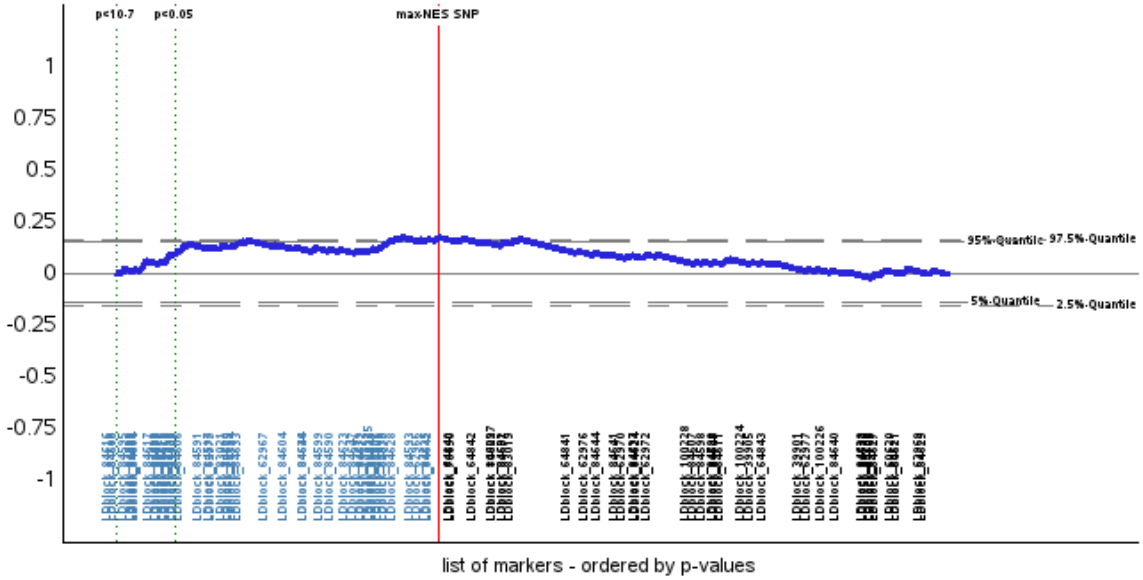


Tabelle 61 Signifikanz der „driving"-Gene des GO-Begriffs GO:0006307 *An DNA-Reparatur beteiligte DNA-Dealkylierung*

„driving"-Gen	Anzahl „driving"-LD-Blöcke in Gen	signifikantester „driving"-LD-Block	
		LD-Block Nr.	p-Wert
FTO	15	84616	0,0005
FTO RP11357N1371	1	84632	0,0411
MGMT	2	62975	0,0430
ALKBH2 USP30	1	71830	0,0562
FTO RP11357N1311	1	84635	0,0573
ASCC3	1	39903	0,0588
gesamt	**21**		

4.11.7.2 Modellschätzung LD-Blöcke Nr. 84589-84647 (Chr. 16q12.2; FTO/ALKBH9)

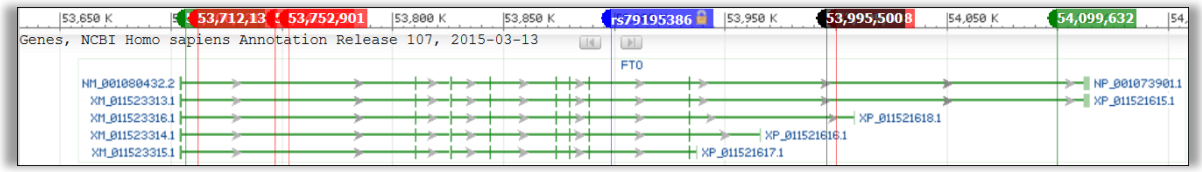
Die insgesamt 402 Marker in und um das Gen **FTO** (*Fat Mass And Obesity-Associated Protein; alternativer Name* **ALKBH9** - *alpha-ketoglutarate dependent dioxygenase)* wurden nicht weniger als 58 LD-Blöcken zugeordnet (siehe Abbildung 35). Die meisten Marker stehen mit mindestens einem anderen Marker im selben Block in sehr straken LD zueinander. Dies berücksichtigend, wurden nur 18 Marker in das Schätzmodell aufgenommen.

Für die LD-Blöcke Nr. 84589-84647 im Chromosomabschnitt 16q12.2 konnte weder im taxativen Modell (p=0,9827) noch nach Variablenselektion (p=0,8691) eine signifikante GxE-Interaktion beobachtet werden. Die Interaktionsterme von 11 der 18 Marker verbleibt im AIC-Modell, (Tabelle 62). Die in der GSEA identifizierten „driving"-LD-Blöcke Nr. 84613-84619 werden in beiden Modellan aber nur durch zwei Marker repräsentiert. Die anderen Marker der betreffenden Region wurden aufgrund ausgeprägter Korrelation der Genotypen von der Modelbildung ausgeschlossen (siehe Kapitel 4.6 Analysemodell für eine Multimarker-Assoziationsanalyse). Der Marker an Position 53.995.500 weist dabei die signifikanteste Interaktion (p=0,013) aller Marker im Modell auf. Gleichzeitig wurde für diesen Marker aber auch eine starker Haupteffekt (OR=7,9; 95%CI: 0,8-77) geschätzt, der Signifikant nur knapp verfehlt (p=0,0755).

Für 10 Marker wird ein genetischer Haupteffekt im AIC-besten Modell geschätzt, wobei beispielsweise die Marker chr16_53995500_A_G (siehe oben) und chr16_53746875_A_G (OR=2,87; 95%-CI: 0,5-15) eine Steigerung des LK-Risikos je seltenem Allel markieren, die Marker chr16_53999638_C_T (OR=0,19; 95%-CI: 0,02-1,8) und chr16_53712135_A_G (OR=0,35; 95%-CI: 0,03-3,31) eine Senkung – unabhängig einer Radon-Exposition. Die Marker an Position 53.712.135 (OR=2,11; 95%-CI: 0,2-22), 53.746.875 (OR=2,16; 95%-CI: 0,1-36), 53.752.901 (OR=2,31; 95%-CI: 0,9-8,8) und 53.999.638 (OR=4,37; 95%-CI: 0,4-45) markieren eine nicht signifikante Steigerung des LK-Risikos unter Radon-Exposition je seltenem Allel. Im Gegenzug markiert der Marker an Position 53.995.500 (OR=0,03) die einzige signifikante (p<0,05) Senkung des LK-Risikos unter Radon-Exposition je seltenem Allel.

Im selben Modell wird die Risikosteigerung durch die Strahlenexposition direkt mit OR=1,61 (Punktschätzer) und damit wesentlich niedriger als in vergleichbaren Modellen geschätzt. Dadurch ist die Einteilung in Risiko-erhöhende bzw. –senkende Marker diskutierbar. Abgesehen von dieser instabilen Schätzung des Radon-Haupteffekts kann aber auf eine Risikostratifikation unter Radon-Exposition durch die Marker des LD-Blocks Nr. 58899 geschlossen werden.

Abbildung 35 FTO/ALKBH9 mit ausgewählten Markern



Die Abbildung zeigt die Lage des Gens FTO sowie ausgewählter Marker
im Bereich 53.650K bis 54.100K des Chromosom 16q12.2 gemäß GeneDB [79]

Tabelle 62 Modellschätzung: LD-Blöcke Nr. 84589-84647

	Taxatives Modell			AIC-bestes Modell		
	Odds-Ratio[1]		p-Wert	Odds-Ratio[1]		p-Wert
Propensity Score	1,39	(0,97- 1,99)	0,0701	1,42	(0,99- 2,03)	0,0502
Strahlenexposition	1,56	(0,28- 8,56)	0,6072	1,61	(0,36- 7,17)	0,5258
chr16_53706236_AA_INDEL	0,61	(0,09- 4,06)	0,6156	0,87	(0,14- 5,16)	0,8796
chr16_53712135_A_G	0,45	(0,04- 4,83)	0,5124	0,35	(0,03- 3,31)	0,3642
chr16_53715082_C_T	>999		0,9582			
chr16_53715344_AATTT_INDEL	<0,01		0,9638			
chr16_53735955_C_T	<0,01		0,9389	0,35	(0,03- 3,31)	0,3642
chr16_53736883_G_T	2,72	(0,49- 15,0)	0,2504			
chr16_53746875_A_G	0,58	(0,03- 9,98)	0,7108	2,87	(0,53- 15,3)	0,2179
chr16_53752901_A_G	<0,01		0,9683	0,40	(0,02- 5,54)	0,4968
chr16_53770081_A_G	<0,01		0,9577	0,55	(0,11- 2,59)	0,4503
chr16_53774354_A_G	<0,01		0,9553			
chr16_53851304_A_G	0,98	(0,19- 5,05)	0,9835			
chr16_53930993_A_G	1,39	(0,43- 4,44)	0,5726	1,24	(0,43- 3,55)	0,6774
chr16_53977414_A_G	0,61	(0,03- 10,9)	0,7379	0,55	(0,04- 6,30)	0,6371
chr16_53995500_A_G	10,7	(0,78- 147)	0,0752	7,90	(0,80- 77,2)	0,0755
chr16_53999638_C_T	0,22	(0,01- 2,45)	0,2197	0,19	(0,02- 1,82)	0,1513
chr16_54051429_C_T	<0,01		0,9707			
chr16_54099632_C_T	0,21		0,9974			
rs79195386	<0,01		0,9114			
Expo. x chr16_53706236_AA_INDEL	1,94	(0,25- 14,8)	0,5223	1,25	(0,18- 8,66)	0,8146
Expo. x chr16_53712135_A_G	1,71	(0,14- 20,5)	0,6722	2,11	(0,20- 22,3)	0,5319
Expo. x chr16_53715082_C_T	<0,01		0,9456			
Expo. x chr16_53715344_AATTT_INDEL	>999		0,9615	1,15	(0,30- 4,35)	0,8340
Expo. x chr16_53735955_C_T	>999		0,9424			
Expo. x chr16_53736883_G_T	0,69	(0,08- 5,62)	0,7289	0,62	(0,07- 4,88)	0,6538
Expo. x chr16_53746875_A_G	1,51	(0,07- 31,3)	0,7889	2,16	(0,12- 36,3)	0,5913
Expo. x chr16_53752901_A_G	>999		0,9659	2,31	(0,61- 8,78)	0,2157
Expo. x chr16_53770081_A_G	>999		0,9588			
Expo. x chr16_53774354_A_G	>999		0,9559			
Expo. x chr16_53851304_A_G	1,12	(0,19- 6,63)	0,8980	1,07	(0,54- 2,11)	0,8331
Expo. x chr16_53930993_A_G	0,79	(0,21- 2,95)	0,7325	0,86	(0,25- 2,91)	0,8157
Expo. x chr16_53977414_A_G	1,25	(0,05- 30,8)	0,8884	1,57	(0,09- 25,5)	0,7488
Expo. x chr16_53995500_A_G	0,02		0,0141	0,03		0,0130
Expo. x chr16_53999638_C_T	3,84	(0,31- 46,7)	0,2905	4,37	(0,42- 45,2)	0,2160
Expo. x chr16_54051429_C_T	0,50		0,9987			
Expo. x chr16_54099632_C_T	<0,01		0,9850			
Expo. x rs79195386	>999		0,9156			

[1] Odds-Ratio mit 95%-Konfidenzintervall

	Taxatives Modell			AIC-Modell	
	χ^2	df	p-Wert	df	p-Wert
Haupteffekt(e) G	6,7848	18	0,9918	10	0,8302
Interaktion(en) GxE	7,7097	18	0,9827	11	0,8691
Gemeinsamer Effekt (joint)	17,0714	36	0,9969	21	0,9367

*4.11.7.3 GSA –Ergebnisse: HGNC:476 m*icroRNAs

Abbildung 36 Manhattan–Plot: HGNC:476 *microRNAs*

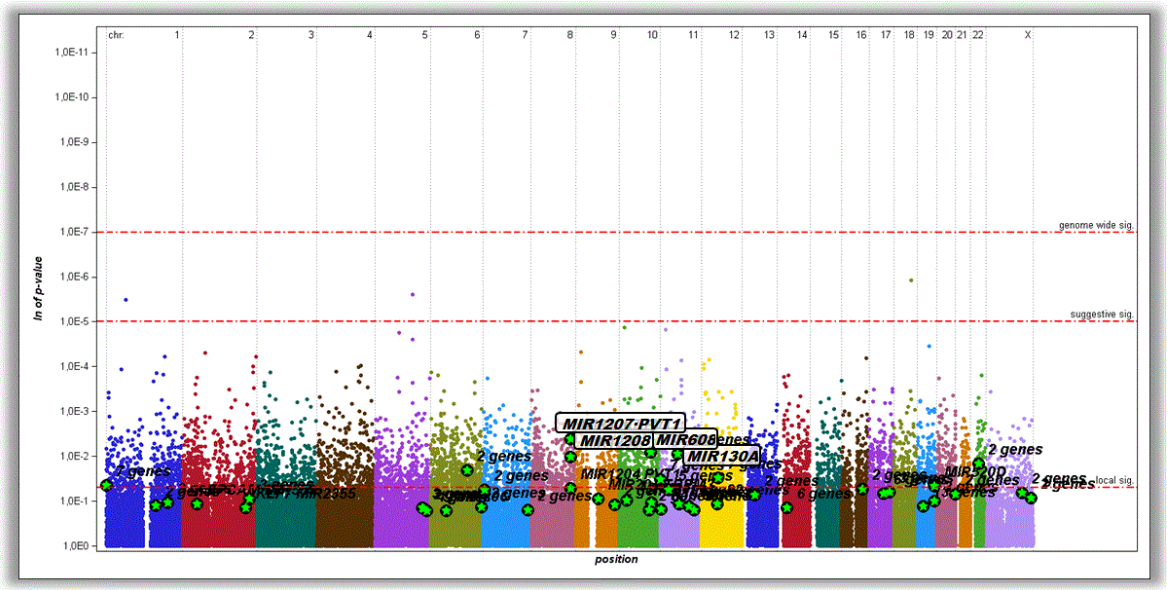


Abbildung 37 GSEA: HGNC:476 *microRNAs*

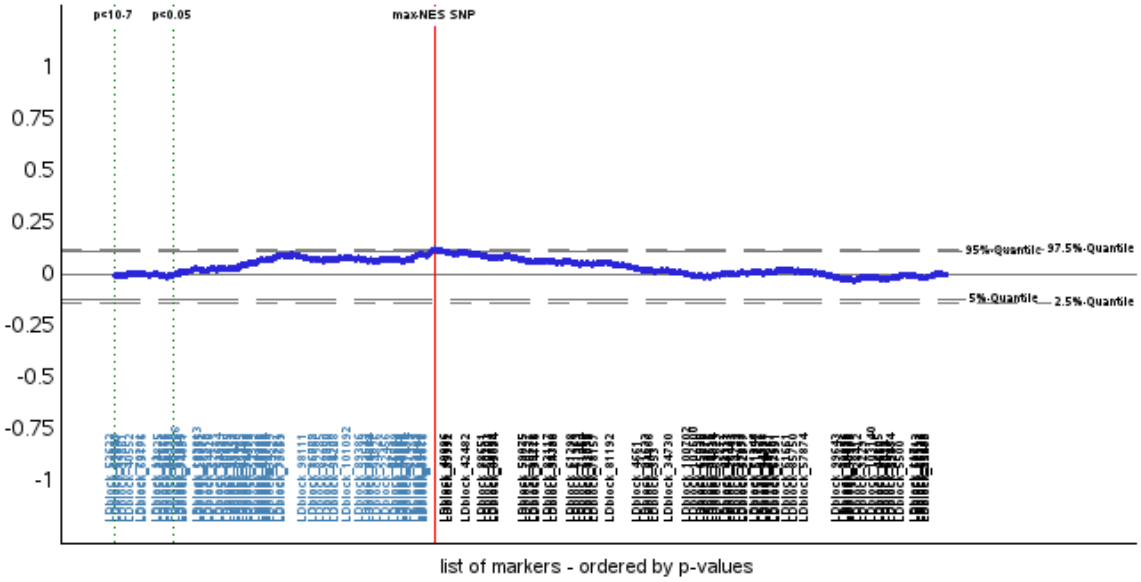


Tabelle 63 Signifikanz der „driving“-Gene der Genfamilie HGNC: 476 *microRNAs*

„driving“-Gen	Anzahl „driving“-LD-Blöcke in Gen	signifikantester „driving“-LD-Block	
		LD-Block Nr.	p-Wert
MIR1207 PVT1	1	52622	0,0041
C10orf2 FAM178A LZTS2 MIR608 MRPL43 PDZD7 RP11108L741 RP11108L771 SEMA4G	1	61633	0,0083
AP0006621 AP00066241 CLP1 MIR130A SERPING1 SLC43A1 SMTNL1 TIMM10 UBE2L6 ZDHHC5	1	65047	0,0088
MIR1208	1	52641	0,0103
MIR548J TPST2	1	99861	0,0143
FAM184A MIR548B	1	40552	0,0208
AC012531171 HOXC10 HOXC5 HOXC6 HOXC9 MIR196A2 MIR196A21	1	70201	0,0300
C11orf35 MIR210 MIR2101 MIR210HG PHRF1 RASSF7 RP11496I911	1	63176	0,0322
FAM132A MIR200A MIR429 SDF4 TNFRSF18 TTLL10 TTLL10AS1	1	11	0,0452
MIR520D	1	94825	0,0465
MIR1204 PVT1	1	52576	0,0511
MIR1538 NFAT5	1	85216	0,0530
FBXL18 MIR589	1	42976	0,0563
BDNF BDNFAS1 LIN7C MIR4454 RP11587D2141	1	64392	0,0566
ARSG MIR635 WIPI1	1	88693	0,0638
MIR1277 WDR44	1	103266	0,0648
HOXB3 HOXB4 MIR10A RP11357H14161 RP11357H14171 RP11357H1471	1	87948	0,0689
C20orf166 MIR133A2	1	97875	0,0697
FOXO1 MIR320D1	1	74051	0,0719
MIR510 MIR513A1	1	103813	0,0853
MIR204 TRPM3	1	55255	0,0874
CTDSP1 MIR26B VIL1	1	15092	0,0889
KIAA1217 MIR603 PRINS RP11183E931	1	59210	0,0960
MIR371A MIR372 NLRP12	1	94828	0,0982
MIR1307 PDCD11 USMG5	1	61694	0,1076
MIR1231 NAV1 RP1190L2031	1	5499	0,1089
AP0023801 C11orf10 C11orf9 DAGLA FADS1 FADS2 FEN1 MIR611 RP11467L20101 RP11467L2061 RP11467L2091	1	65192	0,1164
EPCAM MIR559	1	9667	0,1165
AC0081471 AC0081472 AC1400611 AC14006110 AC14006111 AC1400612 AC1400613 AC1400614 AC1400615 AC1400616 AC1400617 AC1400618 C12orf62 CERS5 FAM186A LIMA1 MIR1293 RP11411N411 RP3405J1021 RP3405J1031 RP3405J1041 orphan1	1	70015	0,1182
DENND1A MIR601	1	57313	0,1207
MIR556 NOS1AP	1	4343	0,1252
AC0920671 CILP2 CTC260F2031 GATAD2A HAPLN4 MAU2 MIR640 NCAN NDUFA13 SUGP1 TM6SF2 TSSK6 YJEFN3	1	93479	0,1281
C11orf54 KIAA1731 MIR1304 SNORA18 SNORA8 SNORD5 TAF1D	1	66325	0,1307
MIR1913 RPS6KA2	1	42483	0,1350
AP4S1 COCH MIR624 RP11829H1631 RP11829H1641 STRN3	1	76891	0,1399
CTB157D1711 GALNT10 MIR1294	1	34106	0,1404
KLF7 MIR2355	1	14578	0,1426
CTC231O1111 MIR146A	1	34327	0,1489
H19 MIR675	1	63292	0,1542
BTG4 C11orf88 MIR34B MIR34C RP11794P661	1	66975	0,1552
MIR1287 PYROXD2	1	61563	0,1571
CNTNAP2 MIR548F4	1	47394	0,1605
MIR206	1	38757	0,1613
FBLL1 MIR103A1 PANK3 RARS	1	34599	0,1658
	44		

4.11.7.4 GSA –Ergebnisse: GO:0006637 und GO:0016020

Abbildung 38 Manhattan-Plot Gen-Set Nr. 41: GO:0006637 *acyl-CoA metabolic process*

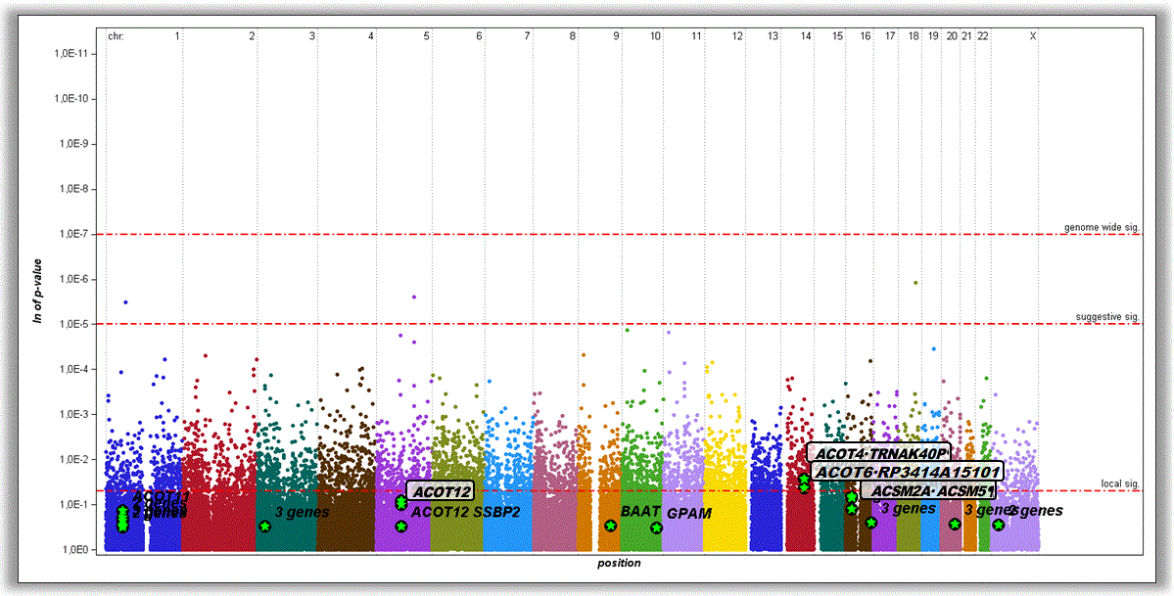
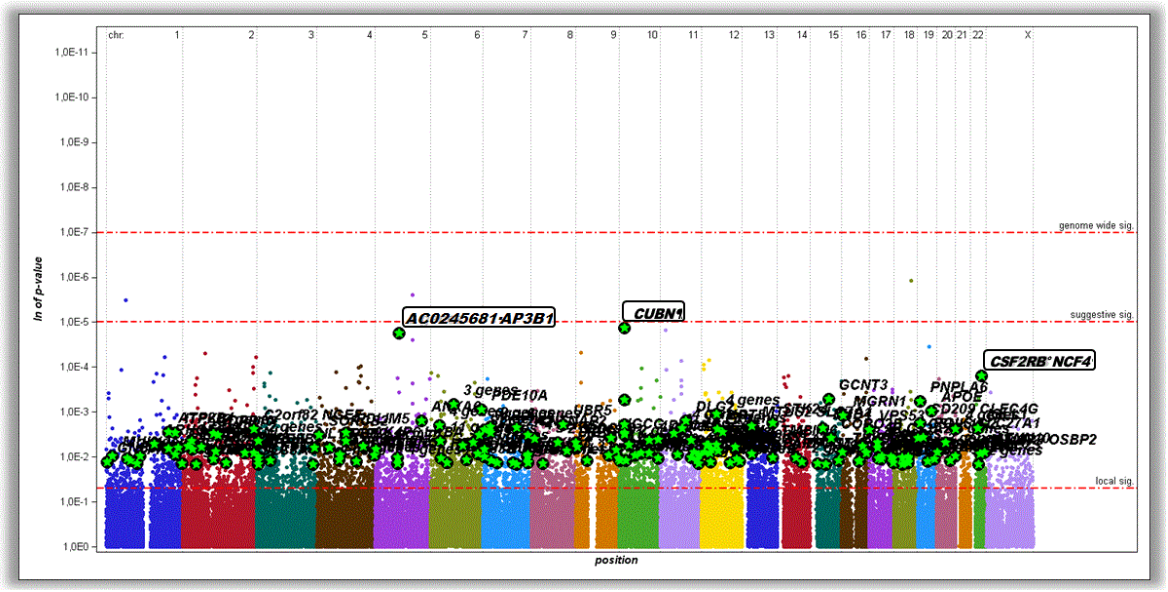


Abbildung 39 Manhattan-Plot Gen-Set Nr. 57: GO:0016020 *membrane (cellular-component)*



4.12 Voraussichtlicher Nutzen bzw. Verwertbarkeit der Ergebnisse

Eine Identifizierung von Genen, die für Strahlenempfindlichkeit prädisponieren, können weitere Einblicke in die individuelle Strahlenempfindlichkeit und in die Ätiologie des Lungenkrebs geben und damit dessen Therapiemöglichkeiten im Besonderen durch Strahlentherapie verbessern.

Ein direkter Nutzen für Prävention, Diagnose oder Therapie kann zum derzeitigen Stand des Gesamtprojekts nicht abgeleitet werden.

4.13 Fortschritte im Forschungsgebiet während der Durchführung des FE-Vorhabens

Während der Projektlaufzeit wurde ein Übersichtsarbeit zu den Fortschritten in der Erforschung der genetischen Prädisposition für Lungenkrebs in GWAS innerhalb der der vergangenen 10 Jahre von Bosse and Amos, 2017 [26] veröffentlicht. Ebenso wurde die Identifizierung von 18 für Lungenkrebs prädisponierender Loci mit genomweiter Signifikanz, darunter 10 neu Loci, durch das ILCCO/TRICL Konsortium veröffentlicht. [121]

Choi, et al., 2016 [122] veröffentlichten eine Übersichtsarbeit zu Lungenkrebs bei Nie-Rauchern verursacht durch eine Innenraumradonbelastung. Die meisten der darin erwähnten Gene sind in Tabelle 2 enthalten.

4.14 Erfolgte und geplante Veröffentlichungen

Die Ergebnisse des Gesamtprojekts wurden am 26. 7.2017 im Rahmen eines publikumsöffentlichen Vortrags am BfS in Neuherberg präsentiert. Das Gesamtprojekt wird in einem Manuskript zur Publikation in einer Fachzeitschrift zusammengefasst und zur Veröffentlichung eingereicht.

5 Erfolgskontrollbericht : Forschungsvorhaben FKZ 3615S32253

Genomweite Analyse genetisch bedingter Strahlenempfindlichkeit in Wismut-Bergarbeitern – Datenauswertung und Bewertung der Assoziationsanalysen

5.1 Beitrag der Ergebnisse zu den förderpolitischen Zielen des Förderprogramms

Kenntnisse über individuelle Strahlenempfindlichkeit sind sowohl für die Bewertung von Strahlenrisiken (Prävention) als auch für eine optimale Strahlentherapie (Diagnose und Therapie) sehr wichtig. Der Behandlungserfolg einer Strahlentherapie ist z.B. von der applizierten Dosis und vielen anderen Parametern abhängig. Eine frühe Erkennung durch *in vitro*-Tests an einfach zugänglichem Patientenmaterial ist daher höchst wünschenswert.

Ein direkter Nutzen für Prävention, Diagnose oder Therapie kann zum derzeitigen Stand des Gesamtprojekts nicht abgeleitet werden.

5.2 Wissenschaftlich-technische Ergebnisse und wesentliche Erfahrungen des Vorhabens

Durch das Forschungsvorhaben FKZ 615S32253 konnten Regionen am menschlichen Genom identifiziert werden, die das Lungenkrebsrisiko unter lebenslanger, beruflicher Radonbelastung mit determinieren. Durch das Forschungsvorhaben konnte gezeigt werden, dass genomweite Interaktionsanalysen bei der Datensätze verschiedener Studien verknüpft werden mit geeigneter statischer Modellbildung erfolgreich durchgeführt werden können.

5.3 Erfindungs-/Schutzanmeldungen, Fortschreibung des Verwertungsplans

Es sind keine Erfindungs-/Schutzanmeldungen geplant.

5.4 Wirtschaftlichen Erfolgsaussichten nach Auftragsende

Aus dem Forschungsvorhaben FKZ 615S32253 können keine direkten wirtschaftlichen Erfolgsaussichten abgeleitet werden.

5.5 Wissenschaftliche oder technische Erfolgsaussichten

Aus dem Forschungsvorhaben FKZ 615S32253 können keine direkten technischen Erfolgsaussichten abgeleitet werden.

Die Ergebnisse des Forschungsvorhaben FKZ 615S32253 tragen wesentlich zur Klärung individueller Prädisposition für Lungenkrebs unter Strahlenbelastung bei.

5.6 Wissenschaftliche und wirtschaftlichen Anschlussfähigkeit für eine mögliche notwendige nächste Phase bzw. die nächsten innovativen Schritte

Die in Tabelle 1 und Tabelle 2 gelisteten Gene sollten durch DNA-Proben der derzeit rund 250 Lungenkrebsfälle des *Wismut-Pathologiearchivs der BfS* validiert werden. Zusätzlich kann eine Schätzung der Gen-Expression (*Variant Effect Prediction*) auf Grundlage der gemessenen Genotypen erfolgen. Dazu stehen eine Reihe von Computer-Routinen und Tools zur Verfügung (z.B. bei ENSEMBL: http://www.ensembl.org/Homo_sapiens/Tools/VEP; PolyPhen2[123], SIFT[124], MutationAssessor[125], MAPP[126], AlignGVGD[127], Panther[128], CADD[129] oder GERP[130]). Miosge, et al., 2015 [131] untersuchten die Vorhersagegüte einer „transactivation“ von weniger als 50% im Vergleich zum Wildtyp, beispielhaft für PolyPhen2 bei Vorliegen mehrerer Mutationen des Gens TP53. Von 2.036 Proben

im Experiment waren ~35% „*true positive*“, ~35% „*true negative*“, ~20% „*false positive*“, ~10% „*false negative*“. Daraus lässt sich eine Sensitivität von ~78% und eine Spezifität von 63% ableiten.

Die Expressions-Schätzung ist im Vergleich zur experimentellen Expressions-Messung kostengünstiger, weniger aufwändig und kann für fast beliebig große Stichproben durchgeführt werden. Da es sich bei diesem Ansatz um eine *in silco*-Vorhersage der funktionellen Konsequenz auf Basis der DNA-Sequenz handelt, ersetzt er eine Validierung durch experimentell, *in vivo* gemessene Expression nicht.

5.7 Arbeiten, die zu keinen Lösungen geführt haben

Alle im Antrag enthaltenen Fragestellungen wurden beantwortet. Es gab keine Arbeiten, die nicht zu einem Ergebnis geführt hätten.

5.8 Präsentationsmöglichkeiten

Die Ergebnisse des Gesamtprojekts wurden am 26. 7.2017 im Rahmen eines publikumsöffentlichen Vortrags am BfS in Neuherberg präsentiert. Das Gesamtprojekt wird in einem Manuskript zur Publikation in einer Fachzeitschrift zusammengefasst und zur Veröffentlichung eingereicht.

5.9 Einhaltung der Kosten- und Zeitplanung

5.9.1 Zeitplan

Der geplant Zeitplan für das Forschungsvorhaben FKZ 615S32253 konnte nicht strikt eingehalten werden. Der Anschlussbericht wurde der Fachbegleitung erst am 19.6.2017 übergeben. Der vertraglich vereinbarte Vortrag zu den Ergebnissen das Forschungsvorhaben wurde erst am 26. 7.2017 gehalten. Die längere Projektlaufzeit war für das BfS kostenneutral.

5.9.2 Finanzplan

Die beantragten Personal- und Sachmittel wurden gemäß Vereinbarung verbraucht.

6 Anhang

6.1 Verwendete Programme

SAS software 9.4	Copyright © 2002-2012 by SAS Institute Inc., Cary. NC. USA. *www.**sas**.com/*
PLINK 1.9 beta	Shaun Purcell. Christopher Chang [132,133] https://www.cog-genomics.org/plink2
EIGENSTRAT	Price et al. [56] integrated in EIGENSOFT package http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html
ADDMIXTURE	Version 1.3.0 David H. Alexander, Suyash S. Shringarpure, John Novembre, Kenneth Lange [28] https://www.genetics.ucla.edu/software/admixture/download.html/

6.2 Originalstudien

Tabelle 64 Originalstudien

Akronym	Title der Studie	Institution	PI (principal investigator)	Land	Design	Zeitraum
CARET	The Carotene and Retinol Efficacy Trial	Fred Hutchinson Cancer Research Center (FHCRC)	J. Doherty, C. Chen	USA	Cohort	Recruitment 1985-1996
BioVU	Vanderbilt 2	Vanderbilt University	M. Aldrich	USA	Hosp. CC	2007- ongoing
HLCS	Harvard Lung Cancer Study	*Harvard School of Public Health*, Mass General Hospital	D. Christiani	USA	Hosp. CC	1992-2004
ATBC	The Alpha-Tocopherol, Beta-Carotene Cancer Prevention	National Cancer Institute (NCI)	D. Albanes	Finland	Cohort	1985-1993
PLCO	The Prostate, Lung, Colo-rectal and Ovarian Cancer Screening Trial	National Cancer Institute (NCI)	N. Caporaso	USA	Cohort	1992-2001
MSH-PMH	Mount Sinai Hospital-Princess Margaret Hospi-tal Study	Mount Sinai Hospital (MSH), Princess Margaret Hospital (PMH)	R.J. Hung, G. Liu	Canada	Hosp. CC	2008-2012
LCRI-DOD	Study of Lung Cancer in Appalachian Kentucky	Markey Cancer Center	S. Arnold	USA	Pop. CC	2012- ongoing
Tampa	Tampa Lung Cancer Study	Washington State University (WSU)	P. Lazarus	USA	Hosp. CC	1999-2003
NELCS	New England Lung Cancer Study	Dartmouth College of Medicine	A. Andrew	USA	Pop. CC	2005-2007
TLC	Total Lung Cancer: Mo-lecular Epidemiology of Lung Cancer Survival	Moffitt Cancer Center, Tampa	M.B. Schabath	USA	case only	2012-- ongoing
MEC	Multiethnic Cohort Study	University of Hawaii (USC)	L. Le Marchand	USA	Cohort	Recruitment 1993-1996
Canada	Canadian screening study	University Health Network (UHN), British Co-lumbia Cancer Agency (BCCA)	St. Lam, M.S.Tsao, G. Liu	Canada	screening cohort	2004-2011, 2008-2013
EAGLE	Environment and Genetics in Lung Cancer Study Etiology	National Cancer Institute (NCI)	M.T. Landi	Italy	Pop.CC	2002-2005
Copenhagen	Coppenhagen lung cancer study	University of Copenhagen	S. E. Bojesen	Denmark		

Akronym	Title der Studie	Institution	PI (principal investigator)	Land	Design	Zeitraum
CAPUA	Cancer de Pulmon en Asturias	University of Oviedo	A. Tardon	Spain	Hosp.CC	2002-2012
GLC	German lung cancer study	University of Göttingen, Deutsches Krebsforschungszentrum Heidelberg (DKFZ)	H. Bickeböller, A. Risch	Germany	Mixed CC	1998-2013
GLC-500K	German lung cancer study	University of Göttingen, Helmholtz Zentrum München (HMGU), DKFZ	H. Bickeböller, A. Risch, H.-E. Wichmann	Germany	Mixed CC	1998-2013
Akronym	Title der Studie	Institution	PI (principal investigator)	Land	Design	Zeitraum
Nijmegen	The Nijmegen Lung Cancer Study	Radboud University Medical Centre	L. A. Kiemeney	The Netherlands	Pop. CC	2002-2008
ReSoLucent	Resource for the Study of Lung Cancer Epidemiology in North Trent	University of Sheffield,	P. Woll	UK	Mixed CC	2005-2014
Norway	Norway Lung Cancer Study	National Institute of Occupational Health (NIOH)	A. Haugen	Norway	Pop. CC	1986-2005
LLP-2008, LLP-2013	Liverpool Lung Cancer Project	University of Liverpool	J.K. Field	UK	Cohort	1999-2007, 1999-2011
NSHDC	Northern Sweden Health and Disease Cohort	Umeå University	M. Johansson	Sweden	Cohort	1985- ongoing
Wismut	Bioproben-Bank von ehemaligen Beschäftigten der SAG/SDAG Wismut	Ruhr-Universität Bochum Bundesamt für Strahlenschutz (BfS)	B. Pesch, M. Gomolka	Germany	Sample selection	2009-2011
MDCS	The Malmö Diet and Cancer Study	Lund University	J. Manjer	Sweden	Cohort	1991-1996
Indoor-Radon	Indoor radon and lung cancer in Germany	Helmholtz Zentrum München (HMGU), Bundesamt für Strahlenschutz (BfS)	H.-E. Wichmann, L. Kreienbrock,	Germany	Pop. CC	1990-1997, 2000-2003
NICCC-LCA	Clalit National Israeli Cancer Control Center-lung cancer study	Carmel Medical Center & Technion	G. Rennert	Israel	Pop.CC	2008-ongoing
MLT	Russian Multicancer study	International Agency for Research on Cancer (IARC)	P. Brennan	Russia	Hosp. CC	??

6.3 Abbildungen und Tabellen

Tabelle 65 Alter, Geschlecht, Rauchverhalten der Studienteilnehmer

		Lungenkrebs		*Alter**	*Geschlecht*		*Rauchverhalten*			
	N	*Kontrollen*	*Fälle*	*Median*	*männlich*	*weiblich*	*Nie-Raucher*	*Ex-Raucher*	*Raucher*	*Jemals-Raucher*
gesamt	28,599	13,522	15,077	28,599	18,059	10,540	5,676	9,518	12,039	1,366
		47%	53%	63	63%	37%	20%	33%	42%	5%
Originalstudie										
Indoor-Radon	58		100%	67	100%		2%			98%
Wismut	405	99%	1%	77	100%		33%	5%	61%	1%
Wismut-Bergarbeiter gesamt	**463**	**87%**	**13%**	**76**	**100%**		**29%**	**4%**	**53%**	**13%**
GLC-550K	949	50%	50%	46	56%	44%	27%	23%	50%	
OncoArray-C§										
ATBC_1	320	56%	44%	61	100%				100%	
ATBC_2	1,363	36%	64%	58	100%				100%	
CANADA	656	67%	33%	65	43%	57%	0%	43%	57%	
CAPUA	1,399	49%	51%	68	88%	12%	17%	42%	41%	<1%
COPENHAGEN	1,823	74%	26%	64	44%	56%	27%	63%	11%	
EAGLE	3,494	49%	51%	67	79%	21%	19%	38%	43%	
CARET	1,065	49%	51%	60	67%	33%		20%	80%	
LLP-2008	200	51%	50%	69	59%	41%	18%	53%	30%	
LLP-2013	675	53%	47%	67	56%	44%	37%	47%	16%	<1%
GLC	1,014	22%	78%	47	55%	45%	13%	17%	68%	3%
HSPH	1,605	32%	68%	64	48%	52%	24%	51%	25%	
ISRAEL	1,149	44%	56%	68	63%	37%	33%	34%	33%	
LCRI-DOD	220	58%	42%	63	48%	52%	29%	32%	39%	<1%
MDCS	325	51%	49%	62	44%	56%	26%	31%	43%	
MEC	430	50%	50%	73	53%	47%	29%	44%	27%	
NELCS	329	51%	49%	62	44%	56%	25%	43%	32%	
NIJMEGEN	816	54%	46%	61	61%	39%	14%	45%	41%	
NORWAY	725	57%	43%	62	69%	31%	3%	13%	27%	57%
NSHDC	473	50%	50%	60	50%	50%	12%	28%	60%	
PLCO	2,231	40%	60%	68	61%	39%	9%	44%	47%	
RESOLUCENT	750	34%	66%	56	48%	52%	18%	27%	55%	1%
RUSSIAN_CE	2,009	51%	49%	61	67%	33%	30%	21%	49%	
TAMPA	242	60%	40%	65	67%	33%	22%			78%
TLC	419		100%	66	47%	53%	7%	60%	33%	
TORONTO	2,295	41%	59%	64	50%	50%	26%	44%	28%	2%
VANDERBILT	1,160	48%	52%	66	54%	46%	47%			53%

* Alter bei Diagnose/Interview; § OncoArray-Konsortium

Abbildung 40 Verteilung der *Working Level Months* (WLM) unter Wismut-Bergarbeitern

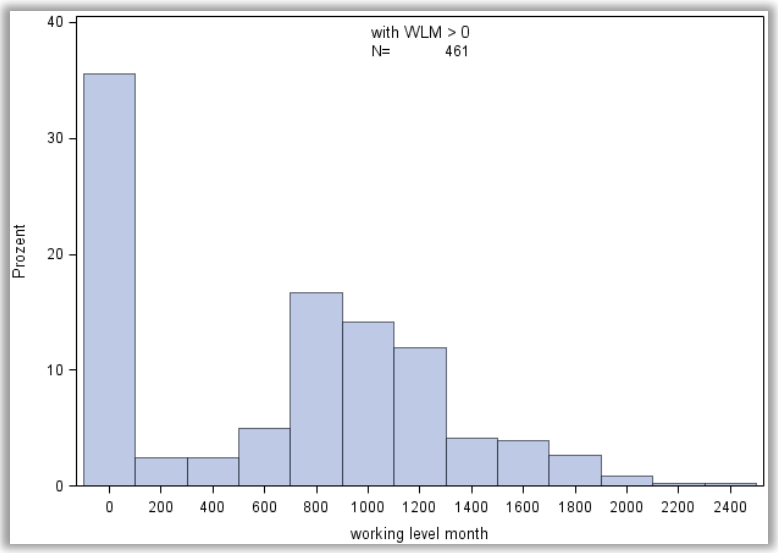


Tabelle 66 Strahlenexposition der Studienteilnehmer

	Working Level Months					
	N	*Min*	*Max*	*Median*	*Mean*	*Std*
gesamt	28.599	0	2.479	0	11	110.62
nicht-exponiert	28.291	0	46	0	0	1.49
exponiert	308	51	2.479	966	986	419.39

Std Standardaweichung

Tabelle 67 Studienort (Kontinent), Genotypisierung, Studiendesign, Fallzahl der Studienteilnehmer

	Studien-Acronym	GenoCent	Studien-Design	gesamt	Lungenkrebs	
					Kontrollen	Fälle
gesamt				28.599	13.522	15.077
Amerika	*CARET*	*CIDR*	*nested CC*	1.065	519	546
	VANDERBILT	*CIDR*	*Hosp CC*	1.160	558	602
	HLCS	*CIDR*	*Hosp CC*	1.605	512	1.093
	ATBC	*CIDR*	*nested CC*	1.683	666	1.017
	PLCO	*CIDR*	*nested CC*	2.231	885	1.346
	MSH-PMH	*CIDR*	*Clinic CC*	2.295	946	1.349
	LCRI-DOD	*CIDR*	*Pop CC*	220	128	92
	TAMPA	*CIDR*	*Hosp CC*	242	144	98
	NELCS	*CIDR*	*Pop CC*	329	169	160
	TLC	*CIDR*	*case only*	419	--	419
	MEC	*CIDR*	*nested CC*	430	217	213
	CANADA	*CIDR*	*nested CC*	656	442	214
	gesamt			**12.335**	**5.186**	**7.149**
Europa	*EAGLE*	*CIDR*	*Hosp CC*	3.494	1.702	1.792
	COPENHAGEN	*Cambridge*	*Pop CC*	1.823	1.341	482
	CAPUA	*CIDR*	*Hosp CC*	1.399	684	715
	GLC	*HMGU*	*Hosp CC*	1.014	221	793
	GLC-550K	*HMGU*	*Hosp CC*	949	478	471
	NIJMEGEN	*CIDR*	*Pop CC*	816	442	374
	RESOLUCENT	*CIDR*	*Pop CC*	750	258	492
	NORWAY	*CIDR*	*Hosp CC*	725	416	309
	LLP-2013	*CIDR*	*nested CC*	675	355	320
	NSHDC	*CIDR*	*nested CC*	473	236	237
	Wismut	*HMGU*	*nested CC*	405	402	3
	MDCS	*CIDR*	*nested CC*	325	167	158
	LLP-2008	*CIDR*	*nested CC*	200	101	99
	INDOOR-RADON	*HMGU*	*nested CC*	58	--	58
	gesamt			**13.106**	**6.803**	**6.303**
Israel	*ISRAEL*	*CIDR*	*Pop CC*	1.149	508	641
Russland	*MLD*	*CIDR*	*Hosp CC*	2.009	1.025	984

Die Beschreibung der Originalstudien siehe Anhang Tabelle 64,
GenoCent Genotypisierungs-Zentrum;

Tabelle 68 Verteilung der genomischen Subcluster je Originalstudie

		1		2		3		4		5		6	
	n	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
CEU	60	17	28%	1	1%	11	18%	7	11%	4	6%	20	33%
OncoArray-Konsortium (gesamt)	**17.531**	**3.592**	**20%**	**1.896**	**11%**	**2.448**	**14%**	**1.350**	**8%**	**490**	**3%**	**7.755**	**44%**
Kontrollen	**8.063**	**2.946**	**36%**	**1.641**	**20%**	**889**	**11%**	**1.459**	**18%**	**855**	**10%**	**36%**	**3%**
Fälle	**9.468**	**3.562**	**37%**	**1.815**	**19%**	**1.011**	**10%**	**1.702**	**17%**	**1.059**	**11%**	**37%**	**3%**
ATBC_1	**320**	69	22%	26	8%	54	17%	24	8%	7	2%	140	44%
ATBC_2	**1.363**	280	21%	145	11%	177	13%	107	8%	32	2%	622	46%
CANADA	**284**	60	21%	29	10%	34	12%	26	9%	8	3%	127	45%
CAPUA	**1.227**	251	20%	131	11%	154	13%	98	8%	37	3%	556	45%
COPENHAGEN	**804**	160	20%	95	12%	119	15%	52	6%	28	3%	350	44%
EAGLE	**2.744**	576	21%	275	10%	375	14%	233	8%	68	2%	1.217	44%
CARET	**712**	166	23%	69	10%	101	14%	47	7%	17	2%	312	44%
LLP-2008	**118**	26	22%	15	13%	19	16%	8	7%	2	2%	48	41%
LLP-2013	**373**	73	20%	47	13%	56	15%	30	8%	10	3%	157	42%
GLC	**557**	101	18%	64	11%	102	18%	39	7%	13	2%	238	43%
HLCS	**762**	155	20%	80	10%	98	13%	63	8%	26	3%	340	45%
ISRAEL	**725**	146	20%	86	12%	98	14%	64	9%	17	2%	314	43%
LCRI-DOD	**105**	21	20%	7	7%	18	17%	4	4%	3	3%	52	50%
MDCS	**144**	34	24%	16	11%	21	15%	11	8%	5	3%	57	40%
MEC	**229**	45	20%	21	9%	40	17%	22	10%	5	2%	96	42%
NELCS	145	19	13%	12	8%	17	12%	14	10%	12	8%	71	49%
NIJMEGEN	**501**	103	21%	52	10%	72	14%	45	9%	15	3%	214	43%
NORWAY	**505**	93	18%	61	12%	70	14%	33	7%	15	3%	233	46%
NSHDC	**238**	52	22%	30	13%	32	13%	27	11%	4	2%	93	39%
PLCO	**1.363**	265	19%	134	10%	185	14%	99	7%	47	3%	633	46%
RESOLUCENT	**357**	66	18%	44	12%	47	13%	19	5%	10	3%	171	48%
RUSSIAN_CE	**1.352**	289	21%	140	10%	194	14%	93	7%	40	3%	596	44%
TAMPA	**163**	26	16%	16	10%	23	14%	12	7%	3	2%	83	51%
TLC	**198**	37	19%	25	13%	26	13%	14	7%	6	3%	90	45%
TORONTO	**1152**	254	22%	133	12%	166	14%	79	7%	32	3%	488	42%
VANDERBILT	621	127	20%	79	13%	80	13%	50	8%	22	4%	263	42%
Wismut /Indoor-Radon	469	98	21%	64	14%	70	15%	37	8%	6	1%	194	41%

Die Beschreibung der Originalstudien siehe Anhang Tabelle 64.

Abbildung 41 Anteil der Studienteilnehmer an den genomischen Subclustern je Originalstudie

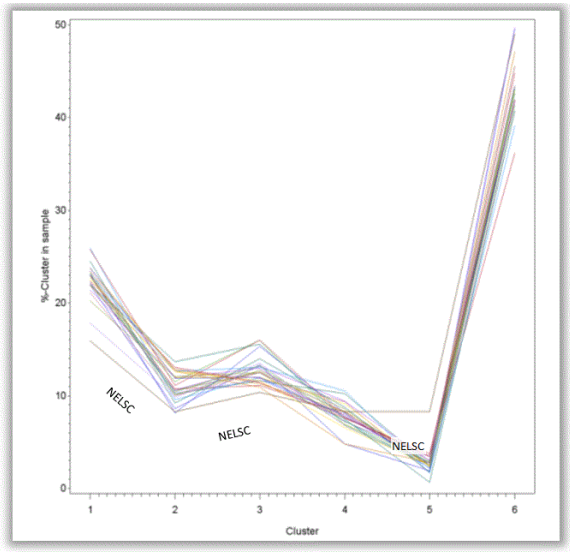


Tabelle 69 **Verteilung der *„ancestral population“*-Cluster unter mit dem OncoArray typisierten Studienteilnehmern (AIMS + Zufallsauswahl an SNPs)**

	gesamt	1		2		3		4	
	n	n	%	n	%	n	%	n	%
Gesamt (OncoArray)	17.531	6	0%	1.724	9%	15.290	87%	511	2%
Kontrollen	8.063	2	0%	765	9%	7.065	88%	231	3%
Fälle	9.468	4	0%	959	10%	8.225	87%	280	3%
ATBC_1	320			25	7%	288	90%	7	2%
ATBC_2	1.363	1	0%	133	9%	1.197	87%	32	2%
CANADA	284			28	9%	246	86%	10	3%
CAPUA	1.227			111	9%	1.077	87%	39	3%
COPENHAGEN	804			86	10%	690	85%	28	3%
EAGLE	2.744	1	0%	258	9%	2.412	87%	73	2%
CARET	712	1	0%	63	8%	631	88%	17	2%
LLP-2008	118			15	12%	101	85%	2	1%
LLP-2013	373	1	0%	43	11%	320	85%	9	2%
GLC	557	1	0%	56	10%	483	86%	17	3%
HLCS	762			74	9%	658	86%	30	3%
ISRAEL	725			81	11%	626	86%	18	2%
LCRI-DOD	105			6	5%	96	91%	3	2%
MDCS	144			14	9%	123	85%	7	4%
MEC	229			16	6%	207	90%	6	2%
NELCS	145	1	0%	11	7%	121	83%	12	8%
NIJMEGEN	501			45	8%	442	88%	14	2%
NORWAY	505			52	10%	437	86%	16	3%
NSHDC	238			26	10%	208	87%	4	1%
PLCO	1.363			125	9%	1.195	87%	43	3%
RESOLUCENT	357			37	10%	309	86%	11	3%
RUSSIAN_CE	1.352			126	9%	1.184	87%	42	3%
TAMPA	163			14	8%	145	88%	4	2%
TLC	198			25	12%	168	84%	5	2%
TORONTO	1152			122	10%	997	86%	33	2%
VANDERBILT	621			71	11%	527	84%	23	3%
Wismut /Indoor-Radon	469			61	13%	402	85%	6	1%

Die Beschreibung der Originalstudien siehe Anhang Tabelle 64.

Tabelle 70 Veränderung der Signifikanz über verschiedene Werte für ρ / p-Werte des Modells eines marginalen Haupteffekts des Genotyps (DxG-Modell) gemäß "*model averaging*"

ρ	1-ρ	Anzahl signifikante Marker im DxG-Modell	ExG-Modell	GxE-Modell	min p-Wert (H2)	signifikante SNPs
0,5	0,5	349	6.029	51	0,28315	
0,6	0,4	275	7.250	44	0,28315	
0,7	0,3	202	8.488	41	0,28315	
0,8	0,2	127	9.779	26	0,28315	
0,9	0,1	70	11.059	20	0,28315	
0,99	0,01	5	12.127	3	0,25763	
0,999	0,001	1	12.213	0	0,28315	
0,9999	0,00014	0	12.221	0	0,28315	
0,99999	$1x10^{-05}$	0	12.222	0	0,28315	
	$1x10^{-06}$	0	12.222	0	0,28315	
	$1x10^{-07}$	0	12.222	0	0,28315	
	$1x10^{-08}$	0	12.222	0	0,28315	
	$1x10^{-09}$	0	12.222	0	0,28315	
	$1x10^{-10}$	0	12.222	0	0,28315	
	$1x10^{-11}$	0	12.222	0	0,28315	
	$1x10^{-12}$	0	12.222	0	0,28315	
	$1x10^{-13}$	0	12.222	0	0,28315	
	$1x10^{-14}$	0	12.222	0	0,28315	
	$1x10^{-15}$	0	12.222	0	0,28315	
	$1x10^{-16}$	0	12.222	0	0,28315	
	$1x10^{-17}$	0	12.222	0	0,28315	
	$1x10^{-18}$	0	12.222	0	0,28315	
	$1x10^{-19}$	0	12.222	0	0,28315	
	$1x10^{-20}$	0	12.222	0	0,28315	

DxG-Modell : Modell eines marginalen Haupteffekts des Genotyps
ExG-Modell : Modell eines marginalen Haupteffekts der Strahlenexposition
GxE-Modell: Modell mit Haupteffekten und Interaktion

Tabelle 71 Veränderung der Signifikanz über verschiedene Werte für ρ / p-Werte des Modells eines marginalen Haupteffekts des Genotyps (DxG-Modell) von TRICL/ILCCO (McKay et al.)

ρ	1-ρ	Anzahl signifikante Marker im D-Modell	E-Modell	GxE-Modell	min p-Wert (H2)	signifikante SNPs
0,5	0,5	10.620	4.984	1.096	0,28315	
0,6	0,4	8.867	6.137	1.157	0,28315	
0,7	0,3	7.117	7.426	1.103	0,28315	
0,8	0,2	5.256	8.846	959	0,28315	
0,9	0,1	3.246	10.380	699	0,28315	
0,99	0,01	891	11.905	225	0,28315	
0,999	0,001	461	12.110	103	0,28315	
0,9999	0,00014	336	12.157	64	0,28315	
0,99999	$1x10^{-05}$	269	12.174	48	0,28315	
	$1x10^{-06}$	231	12.180	42	0,26992	
	$1x10^{-07}$	200	12.185	37	0,23370	
	$1x10^{-08}$	175	12.188	34	0,20448	
	$1x10^{-09}$	147	12.192	30	0,17177	
	$1x10^{-10}$	119	12.196	26	0,13905	
	$1x10^{-11}$	109	12.200	22	0,12736	
	$1x10^{-12}$	100	12.203	19	0,11685	
	$1x10^{-13}$	91	12.207	15	0,10633	
	$1x10^{-14}$	77	12.207	15	0,08997	
	$1x10^{-15}$	56	12.208	14	0,06543	
	$1x10^{-16}$	33	12.216	6	0,03856	rs12440014
	$1x10^{-17}$	0	12.222	0	0,28315	
	$1x10^{-18}$	0	12.222	0	0,28315	
	$1x10^{-19}$	0	12.222	0	0,28315	
	$1x10^{-20}$	0	12.222	0	0,28315	

DxG-Modell : Modell eines marginalen Haupteffekts des Genotyps
ExG-Modell : Modell eines marginalen Haupteffekts der Strahlenexposition
GxE-Modell: Modell mit Haupteffekten und Interaktion

Tabelle 72 Veränderung der Signifikanz gemäß Hyprid-2-step Verfahren nach Murcray über verschiedene Werte für ρ:
p-Werte eines taxativen Multimarker-Models
marginalen Haupteffekts des Genotyps gemäß DxG-Modell

ρ	1-ρ	Anzahl signifikante Marker im D-Modell	E-Modell	GxE-Modell	min p-Wert (H2)	signifikante LD-Block
0.5000	5.0E-01	2,718	5,219	180	0.2726	
0.6000	4.0E-01	2,171	6,013	163	0.2726	
0.7000	3.0E-01	1,667	6,824	142	0.2726	
0.8000	2.0E-01	1,104	7,575	110	0.2726	
0.9000	1.0E-01	594	8,317	66	0.2726	
0.9900	1.0E-02	74	9,007	9	0.2721	
0.9990	**1.0E-03**	**24**	**9,076**	**2**	**0.2715**	
0.9999	1.0E-04	9	9,087	2	0.2716	
1.0000	1.0E-05	6	9,087	2	0.2715	
1.0000	1.0E-06	5	9,087	2	0.2715	
1.0000	1.0E-07	2	9,087	2	0.2715	
1.0000	1.0E-08	2	9,088	1	0.2715	
1.0000	1.0E-09	1	9,088	1	0.2715	
1.0000	1.0E-10	0	9,089	0	0.2715	
1.0000	1.0E-11	0	9,089	0	0.2715	
1.0000	1.0E-12	0	9,089	0	0.2715	
1.0000	1.0E-13	0	9,089	0	0.2715	
1.0000	1.0E-14	0	9,089	0	0.2715	
1.0000	1.0E-15	0	9,089	0	0.2715	
1.0000	1.0E-16	0	9,089	0	0.2715	
1.0000	1.0E-17	0	9,089	0	0.2715	
1.0000	1.0E-18	0	9,089	0	0.2715	
1.0000	1.0E-19	0	9,089	0	0.2715	
1.0000	1.0E-20	0	9,089	0	0.2715	

Abbildung 42 Veränderung der Signifikanz gemäß Hyprid-2-step Verfahren nach Murcray über verschiedene Werte für ρ:
p-Werte eines taxativen Multimarker-Models
marginalen Haupteffekts des Genotyps gemäß DxG-Modell

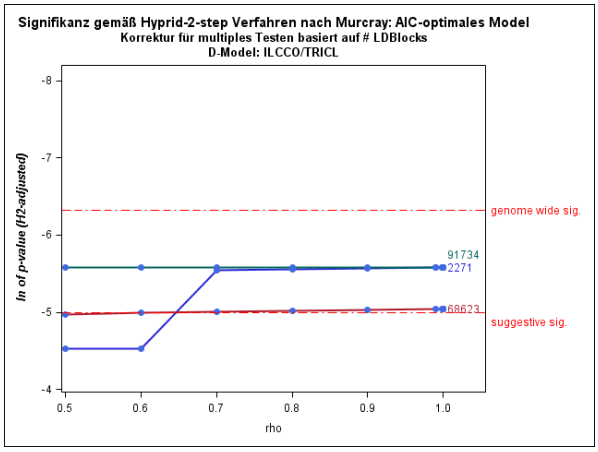


Tabelle 73 Veränderung der Signifikanz gemäß Hyprid-2-step Verfahren nach Murcray über verschiedene Werte für ρ:
p-Werte eines AIC-optimalen Multimarker-Models
marginalen Haupteffekts des Genotyps gemäß DxG-Modell

ρ	1-ρ	Anzahl signifikante Marker im D-Modell	E-Modell	GxE-Modell	min p-Wert (H2)	signifikante LD-Block
0.5000	5.0E-01	5,849	9,838	1,133	0.0265	91734
0.6000	4.0E-01	4,673	11,372	1,066	0.0250	91734
0.7000	3.0E-01	3,555	12,938	948	0.0239	91734
0.8000	2.0E-01	2,423	14,502	738	0.0230	91734
0.9000	1.0E-01	1,311	16,091	426	0.0222	91734
0.9900	1.0E-02	161	17,568	56	0.0215	91734
0.9990	1.0E-03	30	17,718	8	0.0214	91734
0.9999	1.0E-04	10	17,735	2	0.0214	91734
1.0000	1.0E-05	3	17,738	1	0.0214	91734
1.0000	1.0E-06	1	17,738	1	0.0214	91734
1.0000	1.0E-07	0	17,738	1	0.0214	91734
1.0000	1.0E-08	0	17,738	1	0.0214	91734
1.0000	1.0E-09	0	17,738	1	0.0214	91734
1.0000	1.0E-10	0	17,739	0	0.0214	91734
1.0000	1.0E-11	0	17,739	0	0.0214	91734
1.0000	1.0E-12	0	17,739	0	0.0214	91734
1.0000	1.0E-13	0	17,739	0	0.0214	91734
1.0000	1.0E-14	0	17,739	0	0.0214	91734
1.0000	1.0E-15	0	17,739	0	0.0214	91734
1.0000	1.0E-16	0	17,739	0	0.0214	91734
1.0000	1.0E-17	0	17,739	0	0.0214	91734
1.0000	1.0E-18	0	17,739	0	0.0214	91734
1.0000	1.0E-19	0	17,739	0	0.0214	91734
1.0000	1.0E-20	0	17,739	0	0.0214	91734

Abbildung 43 Veränderung der Signifikanz gemäß Hyprid-2-step Verfahren nach Murcray über verschiedene Werte für ρ:
p-Werte eines AIC-optimalen Multimarker-Models
marginalen Haupteffekts des Genotyps gemäß DxG-Modell

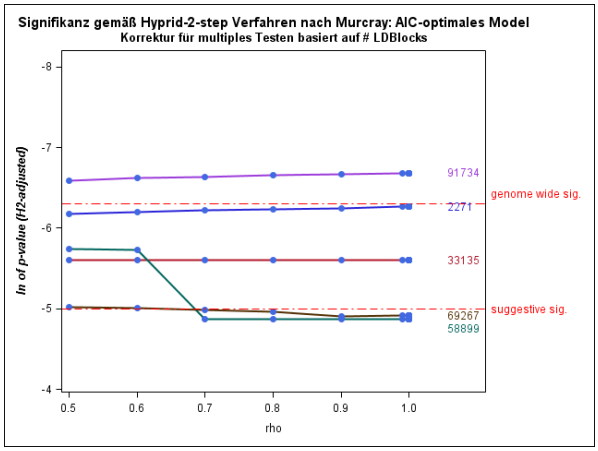


Tabelle 74 Veränderung der Signifikanz gemäß Hyprid-2-step Verfahren nach Murcray über verschiedene Werte für ρ:
p-Werte eines AIC-optimalen Multimarker-Models
marginalen Haupteffekts des Genotyps gemäß TRICL/ILCCO (McKay et al.)

		Anzahl signifikante Marker im			min p-Wert	signifikante
ρ	1-ρ	D-Modell	E-Modell	GxE-Modell	(H2)	LD-Block
0.5000	**5.0E-01**	**10,167**	**9,041**	**1,930**	**0.0265**	91734
0.6000	4.0E-01	8,469	10,543	1,895	0.0250	91734
0.7000	3.0E-01	6,755	12,145	1,741	0.0239	91734
0.8000	2.0E-01	4,981	13,798	1,442	0.0230	91734
0.9000	1.0E-01	3,053	15,493	1,024	0.0222	91734
0.9900	1.0E-02	800	17,264	360	0.0215	91734
0.9990	1.0E-03	378	17,514	212	0.0214	91734
0.9999	**1.0E-04**	**256**	**17,567**	**170**	**0.0214**	91734
1.0000	1.0E-05	203	17,598	141	0.0214	91734
1.0000	1.0E-06	176	17,619	120	0.0214	91734
1.0000	1.0E-07	148	17,629	110	0.0214	91734
1.0000	1.0E-08	130	17,644	95	0.0214	91734
1.0000	1.0E-09	112	17,655	84	0.0214	91734
1.0000	1.0E-10	92	17,667	72	0.0214	91734
1.0000	1.0E-11	83	17,672	67	0.0214	91734
1.0000	1.0E-12	71	17,674	65	0.0214	91734
1.0000	1.0E-13	60	17,677	62	0.0214	91734
1.0000	1.0E-14	52	17,686	53	0.0214	91734
1.0000	1.0E-15	43	17,702	37	0.0214	91734
1.0000	1.0E-16	27	17,722	17	0.0214	91734
1.0000	1.0E-17	0	17,739	0	0.0214	91734
1.0000	1.0E-18	0	17,739	0	0.0214	91734
1.0000	1.0E-19	0	17,739	0	0.0214	91734
1.0000	1.0E-20	0	17,739	0	0.0214	91734

Abbildung 44 Veränderung der Signifikanz gemäß Hyprid-2-step Verfahren nach Murcray über verschiedene Werte für ρ:
p-Werte eines AIC-optimalen Multimarker-Models
marginalen Haupteffekts des Genotyps gemäß TRICL/ILCCO (McKay et al.)

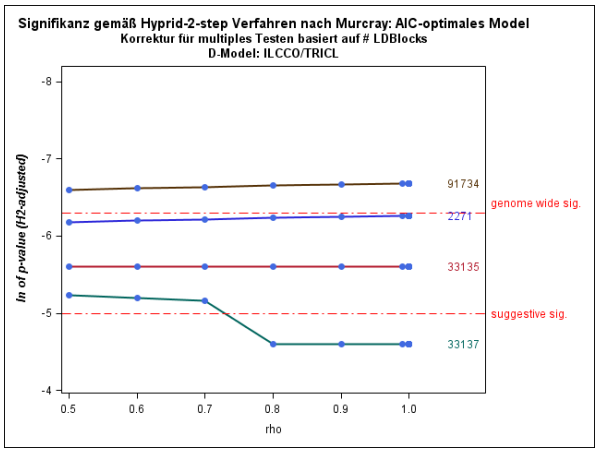


Tabelle 75 Regionen mit ausgeprägtem LD, PC-SNP Korrelation oder bekannter Assoziation zu Lungenkrebs

Chromosom	Position von	bis	ausgeprägtes LD oder Korrelation	Assoziation mit Lungenkrebs
1	78.300	78.700		1p31
1	8.500.000	9.000.000	PC-SNP Korrelation	
1	42.000.000	52.000.000	ausgeprägtes LD	
1	110.000.000	120.000.000	ausgeprägtes LD	
1	182.000.000	195.000.000	ausgeprägtes LD	
1	200.000.000	240.000.000	PC-SNP Korrelation	
2	8.000.000	8.500.000	PC-SNP Korrelation	
2	69.000.000	69.500.000	PC-SNP Korrelation	
2	86.000.000	100.500.000	ausgeprägtes LD	
2	111.500.000	143.000.000	ausgeprägtes LD	
2	160.000.000	230.000.000	ausgeprägtes LD	
3	4.000.000	4.500.000	PC-SNP Korrelation	
3	21.000.000	25.500.000	ausgeprägtes LD	
3	47.500.000	50.000.000	ausgeprägtes LD	
3	58.000.000	68.000.000	ausgeprägtes LD	
3	75.500.000	76.550.000	ausgeprägtes LD	
3	83.500.000	87.000.000	ausgeprägtes LD	
3	89.000.000	97.500.000	ausgeprägtes LD	
3	108.000.000	140.000.000	PC-SNP Korrelation	
3	189.200.000	189.400.000		3q28
4	9.600.000	9.800.000		4p16
4	20.000.000	26.000.000	PC-SNP Korrelation	
4	75.000.000	123.000.000	PC-SNP Korrelation	
5	1.200.000	6.000.000		5p15_TERT
5	1.200.000	6.000.000		5p15
5	41.000.000	52.500.000	ausgeprägtes LD	
5	71.000.000	100.500.000	ausgeprägtes LD	
5	129.000.000	132.000.000	ausgeprägtes LD	
5	135.500.000	138.500.000	ausgeprägtes LD	
6	14.000.000	20.000.000	PC-SNP Korrelation	
6	25.500.000	33.500.000	ausgeprägtes LD	
6	31.200.000	31.600.000		6p21
6	31.600.000	46.000.000		6p21_BAG6
6	57.000.000	86.000.000	ausgeprägtes LD	
6	106.000.000	118.000.000	ausgeprägtes LD	
6	138.000.000	139.000.000	ausgeprägtes LD	
6	139.000.000	142.500.000	ausgeprägtes LD	
6	167.200.000	167.600.000		6q27
7	50.000.000	72.000.000	ausgeprägtes LD	
7	111.000.000	140.000.000	PC-SNP Korrelation	
8	8.000.000	12.000.000	ausgeprägtes LD	
8	27.000.000	30.000.000	ausgeprägtes LD	
8	27.200.000	27.600.000		8p21
8	32.200.000	32.600.000		8p12
8	43.000.000	60.000.000	ausgeprägtes LD	
8	112.000.000	115.000.000	ausgeprägtes LD	
8	94.000.000	95.000.000	PC-SNP Korrelation	
9	20.000.000	22.200.000		9p21
9	77.500.000	125.000.000		9q31
10	2.000.000	9.000.000	ausgeprägtes LD	
10	37.000.000	43.000.000	ausgeprägtes LD	
10	90.000.000	107.000.000		10q24
11	7.000.000	59.000.000	ausgeprägtes LD	
11	57.200.000	57.400.000		11q12
11	87.500.000	90.500.000	ausgeprägtes LD	
11	118.000.000	118.200.000		11q23
12	800.000	1.200.000		12p13
12	23.000.000	58.000.000	ausgeprägtes LD	
12	109.500.000	128.000.000	ausgeprägtes LD	
13	32.800.000	50.000.000		13q13
14	28.000.000	70.000.000	PC-SNP Korrelation	

Chromosom	Position von	bis	ausgeprägtes LD oder Korrelation	Assoziation mit Lungenkrebs
15	47.400.000	47.600.000		15q21
15	49.200.000	51.000.000		15q21
15	78.600.000	79.000.000		15q25_CHRNA3
16	12.000.000	23.000.000	PC-SNP Korrelation	
17	46.000.000	54.000.000	PC-SNP Korrelation	
18	4.000.000	5.000.000	PC-SNP Korrelation	
19	41.200.000	41.400.000		19q13
20	18.000.000	34.500.000	ausgeprägtes LD	
20	59.000.000	62.400.000		20q13
21	19.000.000	27.000.000	PC-SNP Korrelation	
22	29.000.000	29.200.000		22q12

Region mit bekannter Assoziation zu Lungenkrebs oder ausgeprägtem LD wurde definiert gemäß: [7-14]

Tabelle 76 Für die GSEA ausgewählte Gen-Sets und erzielte p-Werte

Kennung	ausgeschlossen	Offizielle Beschreibung	p-Wert
GO:0000012		single strand break repair	0,9204
GO:0000083		regulation of transcription involved in G1/S transition of mitotic cell cycle	0,9811
GO:0000165		MAPK cascade	1,0000
GO:0000209		protein polyubiquitination	0,5889
GO:0000725	n≤5 Gene	recombinational repair	
GO:0000726		non-recombinational repair	0,6126
GO:0000731		DNA synthesis involved in DNA repair	0,4323
GO:0000790		nuclear chromatin	1,0000
GO:0001894		tissue homeostasis	0,2649
GO:0003677		DNA binding	0,7061
GO:0003700		transcription factor activity, sequence-spec	0,5544
GO:0003720		telomerase activity (50 gene)	0,9741
GO:0003996		acyl-CoA ligase activity	1,0000
GO:0004321		fatty-acyl-CoA synthase activity	1,0000
GO:0004672		protein kinase activity	1,0000
GO:0004674		protein serine/threonine	0,6434
GO:0004872		receptor activity	0,6898
GO:0005044		scavenger receptor activi	0,1295
GO:0005215		transporter activity	0,9522
GO:0005509		calcium ion binding	0,9831
GO:0005524		ATP binding	0,2849
GO:0005576		extracellular region	0,5145
GO:0005737		cytoplasm	0,4482
GO:0005759		mitochondrial matrix	0,6120
GO:0005765		lysosomal membrane	0,9082
GO:0005783		endoplasmic reticulum	0,5100
GO:0005794		Golgi apparatus	0,6474
GO:0005905		clathrin-coated pit	0,0777
GO:0006281		DNA repair	1,0000
GO:0006282		regulation of DNA repair	1,0000
GO:0006284		base-excision repair	0,9087
GO:0006289		nucleotide-excision repair	0,7461
GO:0006290	n≤5 Gene	pyrimidine dimer repair	
GO:0006298		mismatch repair	0,9314
GO:0006301		postreplication repair	0,6675
GO:0006302		double-strand break repair	0,8340
GO:0006303		double-strand break repair via nonhomologous end joining	0,7170
GO:0006307		DNA dealkylation involved in DNA repair	0,0139
GO:0006355		regulation of transcription, DNA-templated	0,9044
GO:0006366		transcription from RNA polymerase II promote	0,5979
GO:0006464		cellular protein modifica	0,7365
GO:0006633		fatty acid biosynthetic process	0,8846
GO:0006637		acyl-CoA metabolic process	0,0538
GO:0006897		endocytosis	0,6534
GO:0006898		receptor-mediated endocyt	1,0000
GO:0006915		apoptotic process	0,9821
GO:0007165		signal transduction	0,4701
GO:0007169		transmembrane receptor protein tyrosine kinase signalling pathway	0,8028
GO:0007173		epidermal growth factor receptor signalling pathway	0,2550
GO:0007223		Wnt signalling pathway, calcium modulating pathway	0,8738
GO:0008203		cholesterol metabolic pro	0,6295
GO:0008360		regulation of cell shape	0,5113
GO:0009235		cobalamin metabolic proce	0,6574
GO:0009380		excinuclease repair complex	0,3406
GO:0010008		endosome membrane	0,7332
GO:0010213	=GO: 0009380	non-photoreactive DNA repair	
GO:0015031		protein transport	0,6618
GO:0015645		fatty acid ligase activity	1,0000
GO:0015889		cobalamin transport	0,8486
GO:0016020		membrane	0,0558
GO:0016055		Wnt signalling pathway	1,0000
GO:0016324		apical plasma membrane	1,0000
GO:0016574		histone ubiquitination	0,1434
GO:0018105		peptidyl-serine phosphory	0,8765
GO:0030139		endocytic vesicle	0,7497
GO:0031232		extrinsic component of ex	0,0817
GO:0031419		cobalamin binding	0,6474
GO:0031526		brush border membrane	0,5951
GO:0031625		ubiquitin protein ligase binding	0,9470
GO:0032332		positive regulation of chondrocyte different	0,5697
GO:0033503	n≤5 Gene	HULC complex	
GO:0036297		interstrand cross-link repair	0,8568
GO:0036299	=GO: 0009380	non-recombinational interstrand cross-link repair	
GO:0036337	n≤5 Gene	Fas signalling pathway (3 Gene)	
GO:0036473	n≤5 Gene	cell death in response to oxidative stress	

Kennung	ausgeschlossen	Offizielle Beschreibung	p-Wert
GO:0038127	n≤5 Gene	ERBB signalling pathway	
GO:0038128		ERBB2 signalling pathway	0,9582
GO:0038129		ERBB3 signalling pathway	0,4044
GO:0038130	=GO: 0009380	ERBB4 signalling pathway	
GO:0038201		TOR complex	0,9064
GO:0042157		lipoprotein metabolic pro	0,6733
GO:0042359		vitamin D metabolic proce	0,7351
GO:0042803		protein homodimerization	0,8167
GO:0042953		lipoprotein transport	0,8267
GO:0043161		proteasome-mediated ubiquitin-dependent prot	0,8163
GO:0043202		lysosomal lumen	0,9265
GO:0043504	n≤5 Gene	mitochondrial DNA repair	
GO:0045002		double-strand break repair via single-strand annealing	0,1574
GO:0045004	n≤5 Gene	DNA replication proofreading	
GO:0045738	n≤5 Gene	negative regulation of DNA repair	
GO:0045739		positive regulation of DNA repair	0,9980
GO:0046787	=GO: 0009380	viral DNA repair	
GO:0046872		metal ion binding	0,9658
GO:0047760		butyrate-CoA ligase activity	1,0000
GO:0051103		DNA ligation involved in DNA repair	0,9975
GO:0055059	n≤5 Gene	asymmetric neuroblast division	
GO:0060070		canonical Wnt signalling pathway	0,9689
GO:0060071		Wnt signalling pathway, planar cell polarity pathway	0,9841
GO:0061036		positive regulation of cartilage development	0,5718
GO:0061630		ubiquitin protein ligase activity	0,8098
GO:0070062		extracellular exosome	0,7867
GO:0070265		necrotic cell death	0,5159
GO:0070914		UV-damage excision repair	0,9709
GO:0071560		cellular response to transforming growth fac	0,7550
GO:0072331	n≤5 Gene	signal transduction by p53 class mediator	
GO:0097190		apoptotic signalling pathway (1012 gene)	0,6651
GO:0097196	n≤5 Gene	Shu complex	
GO:0097300	n≤5 Gene	programmed necrotic cell death.	
GO:0097345		mitochondrial outer membrane permeabilization (66 gene)	0,5416
GO:0097468		programmed cell death in response to reactive oxygen species	0,4124
GO:0098504	n≤5 Gene	DNA 3' dephosphorylation involved in DNA repair	
GO:0100026	=GO: 0009380	positive regulation of DNA repair by transcription from RNA polymerase II promoter	
GO:1901184	n≤5 Gene	regulation of ERBB signalling pathway	
GO:1901185		negative regulation of ERBB signalling pathway	0,5527
GO:1901186	n≤5 Gene	positive regulation of ERBB signalling pathway	
GO:1902113	=GO: 0009380	nucleotide phosphorylation involved in DNA repair	
GO:1904886		beta-catenin destruction complex disassembly	0,9398
GO:1990391	n≤5 Gene	DNA repair complex	
GO:2000741	n≤5 Gene	positive regulation of mesenchymal stem cell	
HGNC:102		Ubiquitin conjugating enzymes E2	0,6487
HGNC:1022		ATG gene-family	0,2550
HGNC:1253		Scavenger receptors	0,3745
HGNC:1256	n≤5 Gene	FOS gene-family	
HGNC:1257	n≤5 Gene	JUN gene-family	
HGNC:1264		IL6 gene-family	1,0000
HGNC:40		Acyl-CoA synthetase famil	0,6581
HGNC:476		microRNAs	0,0159
HGNC:476b	keine annotierte LD-Blöcke	miRNA gene-family (beschränken auf LET7-Gene)	1,0000
HGNC:496		CDK gene-family	0,7716
HGNC:508		FOXO gene-family	0,1633
HGNC:518		ANTP/HOXL subclass homeoboxes gene-family	0,3765
HGNC:519		ANTP/NKL subclass homeoboxes and pseudogenes gene-family	0,7112
HGNC:521		PRD//PAX+PAXL subclass homeoboxes gene-family	0,7450
HGNC:522		LIM subclass homeoboxes gene-family	0,1992
HGNC:523		POU subclass homeoboxes gene-family	0,7676
HGNC:524	n≤5 Gene	HNF subclass homeoboxes gene-family	
HGNC:525		SINE subclass homeoboxes gene-family	1,0000
HGNC:526		TALE subclass homeoboxes gene-family	0,6920
HGNC:527		CUT subclass homeoboxes gene-family	0,6495
HGNC:528	n≤5 Gene	PROS/PROX subclass homeoboxes gene-family	
HGNC:529		ZF subclass homeoboxes gene-family	0,9566
HGNC:530		CERS subclass homeoboxes gene-family	0,9885
HGNC:567		Glutathione S-transferases	0,7652
HGNC:598		Interferons IFN gene-family	,
HGNC:750		SMAD gene-family	0,7218
HGNC:757		SRY-boxes	0,9263
HGNC:938		SIRT gene-family	0,8095
publikationsbasierte		Homeoboxes-Gene in regulatorischen Netzwerken Lungenkrebs	0,4402

gesamt: 148 Gen-Sets: davon 119 GO-Begriffe, 28 HGNC: Genfamilien und 1 neu zusammengestellt

7 Referenzen

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin* 2016;**66**(1):7-30.
2. Barnes B, Kraywinkel K, Nowossadeck E, Schönfeld I, Starker A, Wienecke A, Wolf U. Bericht zum Krebsgeschehen in Deutschland 2016. Berlin: Zentrum für Krebsregisterdaten im Robert Koch-Institut, 2016.
3. Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, Deo H, Falk R, Forastiere F, Hakama M, Heid I, Kreienbrock L, Kreuzer M, Lagarde F, Makelainen I, Muirhead C, Oberaigner W, Pershagen G, Ruano-Ravina A, Ruosteenoja E, Rosario AS, Tirmarche M, Tomasek L, Whitley E, Wichmann HE, Doll R. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *BMJ* 2005;**330**(7485):223.
4. Grosche B, Kreuzer M, Kreisheimer M, Schnelzer M, Tschense A. Lung cancer risk among German male uranium miners: a cohort study, 1946-1998. *Br J Cancer* 2006;**95**(9):1280-7.
5. *Health Effects of Exposure to Radon: BEIR VI*. Washington (DC), 1999.
6. The International Lung Cancer Consortium (ILCCO); Transdisciplinary Research in Cancer of the Lung (TRICL). http://ilcco.iarc.fr/; http://u19tricl.org/.
7. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, Qureshi M, Dong Q, Gu X, Chen WV, Spitz MR, Eisen T, Amos CI, Houlston RS. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat.Genet* 2008;**40**(12):1407-1409.
8. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-de-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcatova I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008;**452**(7187):633-637.
9. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008;**40**(5):616-622.
10. Truong T, Hung RJ, Amos CI, Wu X, Bickeboller H, Rosenberger A, Sauter W, Illig T, Wichmann HE, Risch A, Dienemann H, Kaaks R, Yang P, Jiang R, Wiencke JK, Wrensch M, Hansen H, Kelsey KT, Matsuo K, Tajima K, Schwartz AG, Wenzlaff A, Seow A, Ying C, Staratschek-Jox A, Nurnberg P, Stoelben E, Wolf J, Lazarus P, Muscat JE, Gallagher CJ, Zienolddiny S, Haugen A, van der Heijden HF, Kiemeney LA, Isla D, Mayordomo JI, Rafnar T, Stefansson K, Zhang ZF, Chang SC, Kim JH, Hong YC, Duell EJ, Andrew AS, Lejbkowicz F, Rennert G, Muller H, Brenner H, Le Marchand L, Benhamou S, Bouchardy C, Teare MD, Xue X, McLaughlin J, Liu G, McKay JD, Brennan P, Spitz MR. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J Natl.Cancer Inst.* 2010;**102**(13):959-971.
11. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeboller H, Risch A, McKay JD, Wang Y, Dai J, Gaborieau V, McLaughlin J, Brenner D, Narod SA, Caporaso NE, Albanes D, Thun M, Eisen T, Wichmann HE, Rosenberger A, Han Y, Chen W, Zhu D, Spitz M, Wu X, Pande M, Zhao Y, Zaridze D, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Krokan HE, Gabrielsen ME, Skorpen F, Vatten L, Njolstad I, Chen C, Goodman G, Lathrop M, Benhamou S, Vooder T, Valk K, Nelis M, Metspalu A, Raji O, Chen Y, Gosney J, Liloglou T, Muley T, Dienemann H, Thorleifsson G, Shen H, Stefansson K, Brennan P, Amos CI, Houlston R, Landi MT. Influence of common

genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet* 2012;**21**(22):4980-95.

12. Brennan P, Hainaut P, Boffetta P. Genetics of lung-cancer susceptibility. *Lancet Oncol* 2011;**12**(4):399-408.
13. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, Zong X, Laplana M, Wei Y, Han Y, Lloyd A, Delahaye-Sourdeix M, Chubb D, Gaborieau V, Wheeler W, Chatterjee N, Thorleifsson G, Sulem P, Liu G, Kaaks R, Henrion M, Kinnersley B, Vallee M, LeCalvez-Kelm F, Stevens VL, Gapstur SM, Chen WV, Zaridze D, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Krokan HE, Gabrielsen ME, Skorpen F, Vatten L, Njolstad I, Chen C, Goodman G, Benhamou S, Vooder T, Valk K, Nelis M, Metspalu A, Lener M, Lubinski J, Johansson M, Vineis P, Agudo A, Clavel-Chapelon F, Bueno-de-Mesquita HB, Trichopoulos D, Khaw KT, Johansson M, Weiderpass E, Tjonneland A, Riboli E, Lathrop M, Scelo G, Albanes D, Caporaso NE, Ye Y, Gu J, Wu X, Spitz MR, Dienemann H, Rosenberger A, Su L, Matakidou A, Eisen T, Stefansson K, Risch A, Chanock SJ, Christiani DC, Hung RJ, Brennan P, Landi MT, Houlston RS, Amos CI. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* 2014;**46**(7):736-41.
14. Fehringer G, Liu G, Pintilie M, Sykes J, Cheng D, Liu N, Chen Z, Seymour L, Der SD, Shepherd FA, Tsao MS, Hung RJ. Association of the 15q25 and 5p15 lung cancer susceptibility regions with gene expression in lung tumor tissue. *Cancer Epidemiol Biomarkers Prev* 2012;**21**(7):1097-104.
15. Chen LS, Hung RJ, Baker T, Horton A, Culverhouse R, Saccone N, Cheng I, Deng B, Han Y, Hansen HM, Horsman J, Kim C, Lutz S, Rosenberger A, Aben KK, Andrew AS, Breslau N, Chang SC, Dieffenbach AK, Dienemann H, Frederiksen B, Han J, Hatsukami DK, Johnson EO, Pande M, Wrensch MR, McLaughlin J, Skaug V, van der Heijden HF, Wampfler J, Wenzlaff A, Woll P, Zienolddiny S, Bickeboller H, Brenner H, Duell EJ, Haugen A, Heinrich J, Hokanson JE, Hunter DJ, Kiemeney LA, Lazarus P, Le Marchand L, Liu G, Mayordomo J, Risch A, Schwartz AG, Teare D, Wu X, Wiencke JK, Yang P, Zhang ZF, Spitz MR, Kraft P, Amos CI, Bierut LJ. CHRNA5 Risk Variant Predicts Delayed Smoking Cessation and Earlier Lung Cancer Diagnosis-A Meta-Analysis. *J Natl Cancer Inst* 2015;**107**(5).
16. Bierut LJ. Convergence of genetic findings for nicotine dependence and smoking related diseases with chromosome 15q24-25. *Trends Pharmacol Sci* 2010;**31**(1):46-51.
17. Pesch B, Johnen G, Lehnert M. Aufbau einer Bioproben-Bank von ehemaligen Beschäftigten der SAG / SDAG Wismut – Pilotstudie. *Ressortforschungsberichte zur kerntechnischen Sicherheit und zum Strahlenschutz* BfS - Bundesamt für Strahlenschutz, 2015; urn:nbn:de:0221-2015102213745.
18. Brüske-Hohlfeld I, Rosario AS, Wolke G, Heinrich J, Kreuzer M, Kreienbrock L, Wichmann HE. Lung cancer risk among former uranium miners of the WISMUT Company in Germany. *Health Phys* 2006;**90**(3):208-16.
19. Ruano-Ravina A, Pereyra MF, Castro MT, Perez-Rios M, Abal-Arca J, Barros-Dios JM. Genetic susceptibility, residential radon, and lung cancer in a radon prone area. *J Thorac Oncol* 2014;**9**(8):1073-80.
20. Leng S, Thomas CL, Snider AM, Picchi MA, Chen W, Willis DG, Carr TG, Krzeminski J, Desai D, Shantu A, Lin Y, Jacobson MR, Belinsky SA. Radon Exposure, IL-6 Promoter Variants, and Lung Squamous Cell Carcinoma in Former Uranium Miners. *Environ Health Perspect* 2016;**124**(4):445-51.
21. Belinsky SA, Klinge DM, Liechty KC, March TH, Kang T, Gilliland FD, Sotnic N, Adamova G, Rusinova G, Telnov V. Plutonium targets the p16 gene for inactivation by promoter hypermethylation in human lung adenocarcinoma. *Carcinogenesis* 2004;**25**(6):1063-7.
22. Leng S, Picchi MA, Liu Y, Thomas CL, Willis DG, Bernauer AM, Carr TG, Mabel PT, Han Y, Amos CI, Lin Y, Stidley CA, Gilliland FD, Jacobson MR, Belinsky SA. Genetic variation in SIRT1

affects susceptibility of lung squamous cell carcinomas in former uranium miners from the Colorado plateau. *Carcinogenesis* 2013;**34**(5):1044-50.
23. Vahakangas KH, Samet JM, Metcalf RA, Welsh JA, Bennett WP, Lane DP, Harris CC. Mutations of p53 and ras genes in radon-associated lung cancer from uranium miners. *Lancet* 1992;**339**(8793):576-80.
24. Taylor JA, Watson MA, Devereux TR, Michels RY, Saccomanno G, Anderson M. p53 mutation hotspot in radon-associated lung cancer. *Lancet* 1994;**343**(8889):86-7.
25. Yngveson A, Williams C, Hjerpe A, Lundeberg J, Soderkvist P, Pershagen G. p53 Mutations in lung cancer associated with residential radon exposure. *Cancer Epidemiol Biomarkers Prev* 1999;**8**(5):433-8.
26. Bosse Y, Amos CI. A decade of GWAS results in lung cancer. *Cancer Epidemiol Biomarkers Prev* 2017.
27. Sethi TK, El-Ghamry MN, Kloecker GH. Radon and lung cancer. *Clin Adv Hematol Oncol* 2012;**10**(3):157-64.
28. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009;**19**(9):1655-64.
29. Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsmar C, Grabowski G, Martin LJ, Khurana Hershey GK, Chakorborty R, Baye TM. Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics* 2011;**12**:622.
30. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 2003;**73**(6):1402-22.
31. Huckins LM, Boraska V, Franklin CS, Floyd JA, Southam L, Gcan, Wtccc, Sullivan PF, Bulik CM, Collier DA, Tyler-Smith C, Zeggini E, Tachmazidou I, Gcan, Wtccc. Using ancestry-informative markers to identify fine structure across 15 populations of European origin. *Eur J Hum Genet* 2014;**22**(10):1190-200.
32. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* 2009;**30**(1):69-78.
33. Setsirichok D, Piroonratana T, Assawamakin A, Usavanarong T, Limwongse C, Wongseree W, Aporntewan C, Chaiyaratana N. Small Ancestry Informative Marker panels for complete classification between the original four HapMap populations. *Int J Data Min Bioinform* 2012;**6**(6):651-74.
34. *SAS/STAT 9.2 User's Guide*. Vol. Second Edition, 2009.
35. Leuraud K, Schnelzer M, Tomasek L, Hunter N, Timarche M, Grosche B, Kreuzer M, Laurier D. Radon, smoking and lung cancer risk: results of a joint analysis of three European case-control studies among uranium miners. *Radiat Res* 2011;**176**(3):375-87.
36. Kreuzer M, Grosche B, Schnelzer M, Tschense A, Dufey F, Walsh L. Radon and risk of death from cancer and cardiovascular diseases in the German uranium miners cohort study: follow-up 1946-2003. *Radiat Environ Biophys* 2010;**49**(2):177-85.
37. Kreuzer M, Fenske N, Schnelzer M, Walsh L. Lung cancer risk at low radon exposure rates in German uranium miners. *Br J Cancer* 2015;**113**(9):1367-9.
38. *WHO Handbook on Indoor Radon: A Public Health Perspective*. Geneva, 2009.
39. Richardson DB, Kaufman JS. Estimation of the relative excess risk due to interaction and associated confidence bounds. *Am J Epidemiol* 2009;**169**(6):756-60.
40. Zablotska LB, Lane RS, Frost SE. Mortality (1950-1999) and cancer incidence (1969-1999) of workers in the Port Hope cohort study exposed to a unique combination of radium, uranium and gamma-ray doses. *BMJ Open* 2013;**3**(2).
41. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011;**46**(3):399-424.

42. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *Am J Epidemiol* 2011;**174**(5):613-20.
43. Day AG. Why the Propensity for Propensity Scores? *Crit Care Med* 2015;**43**(9):2024-6.
44. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. In: University H, ed, 2016;32.
45. Woodward M. *Epidemiology study design and data analysis*. Texts in statistical science. Vol. 2nd ed. Boca Raton, Fla: Chapman & Hall/CRC Press, 2005.
46. Power RA, Cohen-Woods S, Ng MY, Butler AW, Craddock N, Korszun A, Jones L, Jones I, Gill M, Rice JP, Maier W, Zobel A, Mors O, Placentino A, Rietschel M, Aitchison KJ, Tozzi F, Muglia P, Breen G, Farmer AE, McGuffin P, Lewis CM, Uher R. Genome-wide association analysis accounting for environmental factors through propensity-score matching: application to stressful live events in major depressive disorder. *Am J Med Genet B Neuropsychiatr Genet* 2013;**162B**(6):521-9.
47. Morgan SL, Todd JJ. A Diagnostic Routine for the Detection of consequential Heterogeneity of Causal Effects. *Sociological Methodology* 2008;**38**(1):231-281.
48. Mansson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am J Epidemiol* 2007;**166**(3):332-9.
49. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001;**69**(1):1-14.
50. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003;**4**(8):587-97.
51. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibrium in the human genome. *Nature* 2001;**411**(6834):199-204.
52. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;**21**(2):263-265.
53. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. *Science* 2002;**296**(5576):2225-9.
54. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* 2016;**32**(2):283-5.
55. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS.Genet* 2006;**2**(12):e190.
56. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**(8):904-9.
57. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD, Goldstein DB, Reich D. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 2008;**83**(1):132-5; author reply 135-9.
58. McKay JD, Hung RJ, Amos CI, ILCCO/TIRICL-Consortium. Characterizing the genetic architecture of lung cancer susceptibility through the Oncoarray. *Nature Genetics* 2017(in press).
59. Devlin B, Bacanu SA, Roeder K. Genomic Control to the extreme. *Nat Genet* 2004;**36**(11):1129-30; author reply 1131.
60. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M, Magi R, Madden PA, Heath AC, Nyholt DR, Martin NG, Montgomery GW, Frayling TM, Hirschhorn JN, McCarthy MI, Goddard ME, Visscher PM, Consortium G. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* 2011;**19**(7):807-12.

61. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* 2007;**63**(2):111-9.
62. Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol* 2011;**35**(3):201-10.
63. Buckland ST, Burnham KP, Augustin NH. Model Selection: An Integral Part of Inference. *Biometrics* 1997;**53**(2):15.
64. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002;**21**(16):2409-19.
65. Kreuzer M, Heinrich J, Wolke G, Schaffrath RA, Gerken M, Wellmann J, Keller G, Kreienbrock L, Wichmann HE. Residential radon and risk of lung cancer in Eastern Germany. *Epidemiology* 2003;**14**(5):559-568.
66. Wichmann HE, Rosario AS, Heid IM, Kreuzer M, Heinrich J, Kreienbrock L. Increased lung cancer risk due to residential radon in a pooled and extended analysis of studies in Germany. *Health Phys.* 2005;**88**(1):71-79.
67. Kreuzer M, Grosche B, Brachner A, Martignoni K, Schnelzer M, Schopka HJ, Bruske-Hohlfeld I, Wichmann HE, Burkart W. The German uranium miners cohort study: feasibility and first results. *Radiat Res* 1999;**152**(6 Suppl):S56-8.
68. Schubauer-Berigan MK, Daniels RD, Pinkerton LE. Radon exposure and mortality among white and American Indian uranium miners: an update of the Colorado Plateau cohort. *Am J Epidemiol* 2009;**169**(6):718-30.
69. Walsh L, Tschense A, Schnelzer M, Dufey F, Grosche B, Kreuzer M. The influence of radon exposures on lung cancer mortality in German uranium miners, 1946-2003. *Radiat Res* 2010;**173**(1):79-90.
70. Fugger L, McVean G, Bell JI. Genomewide association studies and common disease--realizing clinical utility. *N Engl J Med* 2012;**367**(25):2370-1.
71. Ziegler A. Genome-wide association studies: quality control and population-based measures. *Genet Epidemiol.* 2009;**33 Suppl 1**:S45-S50.
72. Ziegler A, Sun YV. Study designs and methods post genome-wide association studies. *Human genetics* 2012;**131**(10):1525-1531.
73. Ikegawa S. A short history of the genome-wide association study: where we were and where we are going. *Genomics Inform* 2012;**10**(4):220-5.
74. Igl BW, Konig IR, Ziegler A. What do we mean by 'replication' and 'validation' in genome-wide association studies? *Hum Hered* 2009;**67**(1):66-8.
75. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010;**11**(6):446-50.
76. Maxwell CA, Moreno V, Sole X, Gomez L, Hernandez P, Urruticoechea A, Pujana MA. Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. *Mol Cancer* 2008;**7**:4.
77. Knippschild U, Wolff S, Giamas G, Brockschmidt C, Wittau M, Wurl PU, Eismann T, Stoter M. The role of the casein kinase 1 (CK1) family in different signaling pathways linked to cancer development. *Onkologie* 2005;**28**(10):508-14.
78. van Wijk SJ, Timmers HT. The family of ubiquitin-conjugating enzymes (E2s): deciding between life and death of proteins. *FASEB J* 2010;**24**(4):981-93.
79. dbGene.
80. Fehringer G, Liu G, Briollais L, Brennan P, Amos CI, Spitz MR, Bickeboller H, Wichmann HE, Risch A, Hung RJ. Comparison of pathway analysis approaches using lung cancer GWAS data sets. *PLoS One* 2012;**7**(2):e31816.
81. Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet* 2012;**28**(7):323-32.

82. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 2012;**8**(2):e1002375.
83. Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* 2011;**98**(1):1-8.
84. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010;**11**(12):843-54.
85. Fridley BL, Biernacka JM. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet* 2011;**19**(8):837-843.
86. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 2009;**10**.
87. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kahari AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SM, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P. Ensembl 2015. *Nucleic Acids Res* 2014.
88. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25-29.
89. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* 2015;**43**(Database issue):D1079-85.
90. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc.Natl.Acad.Sci.U.S.A* 2005;**102**(43):15545-15550.
91. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 2007;**23**(23):3251-3253.
92. Rosenberger A, Friedrichs S, Amos CI, Brennan P, Fehringer G, Heinrich J, Hung RJ, Muley T, Muller-Nurasyid M, Risch A, Bickeboller H. META-GSA: Combining Findings from Gene-Set Analyses across Several Genome-Wide Association Studies. *PLoS One* 2015;**10**(10):e0140179.
93. Cooper WA, Lam DC, O'Toole SA, Minna JD. Molecular biology of lung cancer. *J Thorac Dis* 2013;**5 Suppl 5**:S479-90.
94. El-Telbany A, Ma PC. Cancer genes in lung cancer: racial disparities: are there any? *Genes Cancer* 2012;**3**(7-8):467-80.
95. Brambilla E, Gazdar A. Pathogenesis of lung cancer signalling pathways: roadmap for therapies. *Eur Respir J* 2009;**33**(6):1485-97.
96. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba, II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK.

Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 2008;**455**(7216):1069-75.
97. Hubaux R, Becker-Santos DD, Enfield KS, Lam S, Lam WL, Martinez VD. Arsenic, asbestos and radon: emerging players in lung tumorigenesis. *Environ Health* 2012;**11**:89.
98. Adewoye AB, Lindsay SJ, Dubrova YE, Hurles ME. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nat Commun* 2015;**6**:6684.
99. Markson G, Kiel C, Hyde R, Brown S, Charalabous P, Bremm A, Semple J, Woodsmith J, Duley S, Salehi-Ashtiani K, Vidal M, Komander D, Serrano L, Lehner P, Sanderson CM. Analysis of the human E2 ubiquitin conjugating enzyme protein interaction network. *Genome Res* 2009;**19**(10):1905-11.
100. Kazma R, Babron MC, Gaborieau V, Genin E, Brennan P, Hung RJ, McLaughlin JR, Krokan HE, Elvestad MB, Skorpen F, Anderssen E, Vooder T, Valk K, Metspalu A, Field JK, Lathrop M, Sarasin A, Benhamou S. Lung cancer and DNA repair genes: multilevel association analysis from the International Lung Cancer Consortium. *Carcinogenesis* 2012;**33**(5):1059-64.
101. Deng CX. SIRT1, is it a tumor promoter or tumor suppressor? *Int J Biol Sci* 2009;**5**(2):147-52.
102. Lin Z, Fang D. The Roles of SIRT1 in Cancer. *Genes Cancer* 2013;**4**(3-4):97-104.
103. Bonner MR, Bennett WP, Xiong W, Lan Q, Brownson RC, Harris CC, Field RW, Lubin JH, Alavanja MC. Radon, secondhand smoke, glutathione-S-transferase M1 and lung cancer among women. *Int J Cancer* 2006;**119**(6):1462-7.
104. Wu W, Peden D, Diaz-Sanchez D. Role of GSTM1 in resistance to lung inflammation. *Free Radic Biol Med* 2012;**53**(4):721-9.
105. Su S, Jin Y, Zhang W, Yang L, Shen Y, Cao Y, Tong J. Aberrant promoter methylation of p16(INK4a) and O(6)-methylguanine-DNA methyltransferase genes in workers at a Chinese uranium mine. *J Occup Health* 2006;**48**(4):261-6.
106. Gu C, Lu J, Cui T, Lu C, Shi H, Xu W, Yuan X, Yang X, Huang Y, Lu M. Association between MGMT promoter methylation and non-small cell lung cancer: a meta-analysis. *PLoS One* 2013;**8**(9):e72633.
107. Brambilla E. [Epigenetic modifications in lung cancer]. *Ann Pathol* 2009;**29 Spec No 1**:S31-3.
108. Hornhardt S, Rossler U, Sauter W, Rosenberger A, Illig T, Bickeboller H, Wichmann HE, Gomolka M. Genetic factors in individual radiation sensitivity. *DNA Repair (Amst)* 2014;**16**:54-65.
109. Takamizawa J, Konishi H, Yanagisawa K, Tomida S, Osada H, Endoh H, Harano T, Yatabe Y, Nagino M, Nimura Y, Mitsudomi T, Takahashi T. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res* 2004;**64**(11):3753-6.
110. Holland PW, Booth HA, Bruford EA. Classification and nomenclature of all human homeobox genes. *BMC Biol* 2007;**5**:47.
111. Bhatlekar S, Fields JZ, Boman BM. HOX genes and their role in the development of human cancers. *Journal of Molecular Medicine* 2014;**92**(8):811-823.
112. Tansey WP. Mammalian MYC Proteins and Cancer. *New Journal of Science* 2014;**2014**:27.
113. Jafri MA, Ansari SA, Alqahtani MH, Shay JW. Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome Medicine* 2016;**8**:69.
114. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, Consonni D, Pesatori AC, Wacholder S, Thun M, Diver R, Oken M, Virtamo J, Albanes D, Wang Z, Burdette L, Doheny KF, Pugh EW, Laurie C, Brennan P, Hung R, Gaborieau V, McKay JD, Lathrop M, McLaughlin J, Wang Y, Tsao MS, Spitz MR, Wang Y, Krokan H, Vatten L, Skorpen F, Arnesen E, Benhamou S, Bouchard C, Metspalu A, Vooder T, Nelis M, Valk K, Field JK, Chen C, Goodman G, Sulem P, Thorleifsson G, Rafnar T, Eisen T, Sauter W, Rosenberger A, Bickeboller H, Risch A, Chang-Claude J, Wichmann HE, Stefansson K, Houlston R, Amos CI, Fraumeni JF, Jr., Savage SA, Bertazzi PA, Tucker MA, Chanock S, Caporaso NE. A genome-wide association study of lung cancer

identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* 2009;**85**(5):679-91.

115. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, McLaughlin J, Shepherd F, Montpetit A, Narod S, Krokan HE, Skorpen F, Elvestad MB, Vatten L, Njolstad I, Axelsson T, Chen C, Goodman G, Barnett M, Loomis MM, Lubinski J, Matyjasik J, Lener M, Oszutowska D, Field J, Liloglou T, Xinarianos G, Cassidy A, Vineis P, Clavel-Chapelon F, Palli D, Tumino R, Krogh V, Panico S, Gonzalez CA, Ramon QJ, Martinez C, Navarro C, Ardanaz E, Larranaga N, Kham KT, Key T, Bueno-de-Mesquita HB, Peeters PH, Trichopoulou A, Linseisen J, Boeing H, Hallmans G, Overvad K, Tjonneland A, Kumle M, Riboli E, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. Lung cancer susceptibility locus at 5p15.33. *Nat.Genet* 2008;**40**(12):1404-1406.
116. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. *Oncogene* 2016.
117. Gilmore-Hebert M, Ramabhadran R, Stern DF. Interactions of ErbB4 and Kap1 connect the growth factor and DNA damage response pathways. *Mol Cancer Res* 2010;**8**(10):1388-98.
118. Gerken T, Girard CA, Tung YC, Webby CJ, Saudek V, Hewitson KS, Yeo GS, McDonough MA, Cunliffe S, McNeill LA, Galvanovskis J, Rorsman P, Robins P, Prieur X, Coll AP, Ma M, Jovanovic Z, Farooqi IS, Sedgwick B, Barroso I, Lindahl T, Ponting CP, Ashcroft FM, O'Rahilly S, Schofield CJ. The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* 2007;**318**(5855):1469-72.
119. Hubaux R, Becker-Santos DD, Enfield KS, Lam S, Lam WL, Martinez VD. MicroRNAs As Biomarkers For Clinical Features Of Lung Cancer. *Metabolomics (Los Angel)* 2012;**2**(3):1000108.
120. Chlon TM, Taffany DA, Welsh J, Rowling MJ. Retinoids modulate expression of the endocytic partners megalin, cubilin, and disabled-2 and uptake of vitamin D-binding protein in human mammary cells. *J Nutr* 2008;**138**(7):1323-8.
121. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, Byun J, Dunning A, Pooley KA, Qian DC, Ji X, Liu G, Timofeeva MN, Bojesen SE, Wu X, Le Marchand L, Albanes D, Bickeboller H, Aldrich MC, Bush WS, Tardon A, Rennert G, Teare MD, Field JK, Kiemeney LA, Lazarus P, Haugen A, Lam S, Schabath MB, Andrew AS, Shen H, Hong YC, Yuan JM, Bertazzi PA, Pesatori AC, Ye Y, Diao N, Su L, Zhang R, Brhane Y, Leighl N, Johansen JS, Mellemgaard A, Saliba W, Haiman CA, Wilkens LR, Fernandez-Somoano A, Fernandez-Tardon G, van der Heijden HFM, Kim JH, Dai J, Hu Z, Davies MPA, Marcus MW, Brunnstrom H, Manjer J, Melander O, Muller DC, Overvad K, Trichopoulou A, Tumino R, Doherty JA, Barnett MP, Chen C, Goodman GE, Cox A, Taylor F, Woll P, Bruske I, Wichmann HE, Manz J, Muley TR, Risch A, Rosenberger A, Grankvist K, Johansson M, Shepherd FA, Tsao MS, Arnold SM, Haura EB, Bolca C, Holcatova I, Janout V, Kontic M, Lissowska J, Mukeria A, Ognjanovic S, Orlowski TM, Scelo G, Swiatkowska B, Zaridze D, Bakke P, Skaug V, Zienolddiny S, Duell EJ, Butler LM, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* 2017;**49**(7):1126-1132.
122. Choi JR, Park SY, Noh OK, Koh YW, Kang DR. Gene mutation discovery research of non-smoking lung cancer patients due to indoor radon exposure. *Ann Occup Environ Med* 2016;**28**:13.
123. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**(4):248-9.
124. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**(7):1073-81.

125. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;**39**(17):e118.
126. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 2005;**15**(7):978-86.
127. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* 2006;**34**(5):1317-25.
128. Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A* 2004;**101**(43):15398-403.
129. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**(3):310-5.
130. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005;**15**(7):901-13.
131. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Liang R, Zhang Y, Lyon S, Beutler B, Whittle B, Bertram EM, Enders A, Goodnow CC, Andrews TD. Comparison of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci U S A* 2015;**112**(37):E5189-98.
132. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 2015;**4**(1):1-16.
133. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**(3):559-75.

| Verantwortung für Mensch und Umwelt |

Kontakt:
Bundesamt für Strahlenschutz
Postfach 10 01 49
38201 Salzgitter
Telefon: + 49 30 18333 – 0
Telefax: + 49 30 18333 – 1885
Internet: www.bfs.de
E-Mail: ePost@bfs.de
Gedruckt auf Recyclingpapier aus 100 % Altpapier.

Bundesamt für Strahlenschutz