Gene-Set Meta-Analysis
to Discover Molecular-Biological Pathways
Associated to Lung Cancer

Dissertation
zur Erlangung des humanwissenschaftlichen Doktorgrades in der Medizin
der Georg-August-Universität Göttingen

vorgelegt von

Mag. Albert Rosenberger
aus Wien

Göttingen, 2017
Betreuungsausschuss

Erstbetreuer/in:  
Prof. Dr. Heike Bickeböller,  
Institut für Genetische Epidemiologie,  
Universitätsmedizin Göttingen  
Georg-August Universität Göttingen

Weitere Betreuer/innen:  
Prof. Dr. Dieter Kube  
Klinik für Hämatologie und Onkologie,  
Universitätsmedizin Göttingen  
Georg-August Universität Göttingen

Prof. Dr. Tim Friede  
Institut für Medizinische Statistik  
Universitätsmedizin Göttingen  
Georg-August Universität Göttingen

Tag der mündlichen Prüfung:  7.6.2017
Affidavit

Here I declare that my doctoral thesis entitled "Gene-Set Meta-Analysis to Discover Molecular-Biological Pathways Associated to lung cancer" has been written independently with no other sources and aids than quoted.

Albert Rosenberger

Göttingen, March 2017
Acknowledgment

On the road to complete this thesis I have been supported by numerous people and tons of chocolate.

I wish to thank my supervisor Professor Heike Bickeböller for her constant support and encouragement. I wish to thank Andrew Entwistle who vamped up the manuscripts into proper English before being published. I wish to thank Melanie Sohns and Stefanie Friedrichs who supported me in accomplishing some of the calculations. I wish to thanks the members of my thesis committee, Professor Friede and Professor Kube.

I wish to thank all members of the International Lung Cancer Consortium (ILCCO) for sharing their data, without which the whole project wouldn’t have been able to be carried out. I would like to mention especially Christopher Amos and Rayjean Hung for the coordination of ILCCO and their teams for compiling harmonized data sets. I am also indebted to the co-authors for the joint work on my papers.

I wish to thank all participants in the studies included in this work, knowing that it is not something to be taken for granted, to fill questionnaires, undergo extra-examinations or to give blood. To those being diagnosed with lung cancer, I hope you survived and enjoy your life in health.

Finally, I wish to thank my wife and children for not asking why this all took so long until being completed.
Abstract

During the last two decades the search for risk factors of many human diseases turned into a new direction, since it got feasible to gauge the human genome. Genome-wide association studies (GWAS) were carried out to identify genes or narrow genomic regions responsible for the susceptibility to health problems. Lung cancer is one of these conditions. Lung cancer is a major public health problem - worldwide. During the whole lifetime, one out of 14 men and one out of 17 women will develop an invasive lung or bronchial cancer.\(^1\) Moreover, only one or two out of 5 patients survive the first 5 years after being diagnosed.\(^2\)

The International Lung Cancer Consortium (ILCCO), overlapping with the Transdisciplinary Research in Cancer of the Lung (TRICL, another international group of cooperating lung cancer researchers) was established in 2004 with the aim of sharing comparable data from ongoing lung cancer case-control and cohort studies. The participating studies are from different geographical areas and ethnicities. On the basis of genomic data being shared within ILCCO, it was possible to identify and verify the existence of genomic risk loci for lung cancer in European population at chromosomes 5p15.33, 6p21-22 and 15q25.\(^3\text{-}10\)

However, the applied genome-wide association studies suffer from several drawbacks. One thereof is the ignorance of the complexity of molecular-biological mechanisms. Gene-set analyses methods (GSA) were proposed as complementing approaches in the investigation of the genetic basis of diseases using GWAS results to overcome this disadvantage. These aim to discover a joint association of the markers belonging to the genes of considered biological pathways (denoted as gene-sets (GS)) with a disease of interest. Moreover, even if GSAs based on different but comparable studies successfully identify joint association for the same GS, one cannot simply consider this as replicated findings. The pattern of associations of the concerned markers needs to be taken into account.

I have proposed the quantitative approach META-GSA to combine results from GSAs, respectively \(p_{GS}\)-values of GSs, by incorporating concordance of single-marker association patterns between studies, relevant for the GS of interest.\(^11\)

This new method has been applied to the data of ILCCO/TRICL. A pathway currently marked as specific to systemic lupus erythematosus was discovered as being significantly implicated in lung cancer. The gene region 6p21-22 in this pathway appears to be more extensively associated with lung cancer than previously assumed.\(^3\text{-}6,8\) Given wide-stretched linkage disequilibrium in and around the area \(APOM/BAG6/MSH5\), there is currently not enough information or evidence to conclude whether the potential pleiotropy of lung cancer and systemic lupus erythematosus is spurious, biological, or mediated.\(^13\) Further research on this pathway and gene region will be necessary.
Zusammenfassung


Ich habe den quantitativen Ansatz META-GSA entwickelt, um Ergebnisse von GSAs quantitativ valide zu kombinieren, indem die Konkordanz von Einzelmarker-Assoziationsmustern zwischen den Studien adäquat berücksichtigt wird.

# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR</td>
<td>relative risk</td>
</tr>
<tr>
<td>OR</td>
<td>odds ratio</td>
</tr>
<tr>
<td>95%-CI</td>
<td>95% confidence interval</td>
</tr>
<tr>
<td>ES</td>
<td>enrichment score</td>
</tr>
<tr>
<td>NES</td>
<td>normalized enrichment score</td>
</tr>
<tr>
<td>PDR</td>
<td>directed reversed p-value</td>
</tr>
<tr>
<td>GS</td>
<td>gene-set (of interest)</td>
</tr>
<tr>
<td>GS’</td>
<td>remaining genes in the genome</td>
</tr>
<tr>
<td>GSA</td>
<td>gene-set analysis</td>
</tr>
<tr>
<td>GWAS</td>
<td>genome-wide association studies</td>
</tr>
<tr>
<td>pooledGWAS-GSA</td>
<td>Alternative approach to META-GSA: The combining of studies is performed at the level of markers, followed by a single GSA performed on the pooled marker-specific associations.</td>
</tr>
<tr>
<td>MtG</td>
<td>annotation of markers to genes</td>
</tr>
<tr>
<td>GtP</td>
<td>annotation of genes to gene-sets/pathways</td>
</tr>
<tr>
<td>LC</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>SLE</td>
<td>Systemic Lupus Erythematosus</td>
</tr>
<tr>
<td>ILCCO</td>
<td>International Lung Cancer Consortium</td>
</tr>
<tr>
<td>TRICL</td>
<td>Transdisciplinary Research in Cancer of the Lung</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the curve</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>ALIGATOR</td>
<td>Association L1st Go AnnoTatOR</td>
</tr>
<tr>
<td>GSEA</td>
<td>Gene-Set Enrichment Analysis</td>
</tr>
</tbody>
</table>
# Table of contents

1 Introduction ........................................................................................................................................... 9  
1.1 Lung cancer – and its risk factors ........................................................................................................ 9  
1.2 Genome-wide association studies ....................................................................................................... 11  
1.3 Drawbacks of genome-wide association studies ............................................................................. 12  
2 Gene-set analysis (GSA) ..................................................................................................................... 14  
2.1 Methods for the self-contained null-hypothesis .......................................................................... 14  
2.2 Methods for the competitive null-hypothesis ............................................................................... 17  
3 Summaries ............................................................................................................................................... 19  
3.1 META-GSA: A novel method to synthesise statistical evidence of several gene-set analyses .......................................................... 19  
3.2 Gene-set analysis with respect to lung cancer: A meta-analysis .................................................. 21  
4 Discussion ............................................................................................................................................... 22  
A References ........................................................................................................................................... 25  
B Curriculum Vitae .................................................................................................................................. 32  
Publications ............................................................................................................................................. 33
1 Introduction

1.1 Lung cancer – and its risk factors

Cancer is a major public health problem all over the globe. About 14.1 million new cancer cases, 8.2 million deaths by cancer and 32.6 million people living with cancer (within 5 years of diagnosis) were counted for 2012 worldwide. In developed countries about 482,600 new male and 241,700 female cases per anno for cancer of the lung or bronchus (LC) have been estimated. This is in men the second largest, respectively in women third largest, cancer incidence. In Germany 34,690 males and 18,810 women have been newly diagnosed with LC in 2013. 29,708 men and 15,140 women died from LC in the same year. During the whole lifetime, one out of 14 men and one out of 17 women will develop an invasive lung or bronchial cancer (estimated for the United States of America, 2010 to 2012). Furthermore, once diagnosed with lung cancer, one has to face a poor prognosis. According to the “Zentrum für Krebsregisterdaten”, 84% of men 79% women will die within the first 5 years after being diagnosed with LC. Due to the combination of high incidence and poor prognosis, LC is the most common cause of cancer death worldwide. However, standardised incidence and prevalence of LC for men is decreasing since the mid of the 1980s, in Germany as in other devolved countries. On average between 2003 and 2013 the incidence declined by -1.2% (percentage points) per anno, the prevalence by -1.5%. In contrast, there is a continuously increasing trend in incidence and prevalence for women, since the beginning of the data collection. On average between 2003 and 2013 the incidence raised by +3.1% and the prevalence by +2.7%, each year.

Lifelong tobacco smoking remains the predominant cause of LC, even in former-smokers. It is known today that tobacco smoke is a mixture of more than 5,000 chemicals. It is toxic and carcinogenic. At least 98 of these components are, once inhaled, hazardous for human health. The first relation between smoking and LC was drawn in 1939, considering a German case-series of 96 LC-patients. The first scientific evidence was given by Doll and Hill in 1952, comparing 1.465 LC cases and 1.465 matched controls, but without calculating any excess risk estimates for smokers. By summarizing the results of a total of 287 studies estimating the association between smoking and LC, all published between 1950 and 1995, an overall relative risk of RR=5.50 (95%-CI 5.07-5.96) was estimated for (ever-) smokers and an RR=8.43 (95%-CI: 7.63-9.31) for current smokers, each compared to never-smokers. The relative risk was stronger for squamous LC (current smoking RR=16.91, 95%-CI: 13.14-21.76) than adeno LC (RR=4.21, 95%-CI: 3.32-5.34). LC only occurs in one out of 10 heavy smokers.

The rate of male smokers in Germany has dropped since the 1990s (1992: 37% of men smoked; 2013: 29% of men smoked). Even more, for German male smokers the 30-day prevalence of smoking more than 20 cigarettes a day fell from 27% to 2% between 1980 and 2013. Both changes act as the main explanation for the reduction of incidence and prevalence of LC. In contrast, about one out of 5 German women is smoking, without any change
of this rate in the last decades. (1992: 22% of women smoked; 2013: 20% of women smoked). There also exists a negative trend in the amount of smoked cigarettes a day in German women (as in men). For female smokers the 30-day prevalence of smoking more than 20 cigarettes per day fell from 16% to 1.4% between 1980 and 2013.

It is estimated that 10–29% of LC cases are attributable to factors other than smoking, representing between 16,000 and 24,000 LC deaths annually in the USA alone. Exposure to radon, a naturally occurring radioactive gas, is regarded as the second frequent cause of LC. About 10% of LC cases are attributed to it. Exposure to other environmental factors, like exhaust fumes of diesel engines, arsenic, asbestos, crystalline silica and some more chemical substances were found to be further risk factors for LC. An increased risk of developing LC has also been observed in patients with other diseases, such as COPD, pneumonia, tuberculosis, or the autoimmune disorder systemic lupus erythematosus (SLE). To my knowledge, the only identified factor lowering the risk for LC is the intake of soy food or soy products.

Other lifestyle factors had been discussed to lower the risk of LC, these are e.g. physical activity, consumption of fruits, red mead, tea, beer and wine or supplements of vitamins or minerals. With the exception of some particular subgroups, like non-smoking women drinking black tea every day, for none of these factors a convincing evidence of preventing lung cancer in general is given. Although relative risk estimates from meta-analyses of RR=0.9 or similar are reported, the observed associations between lung cancer and dietary factors or physical activity are hard to disentangle from cigarette smoking.

Familial aggregation of lung cancer was identified as a further risk factor, already discovered before the turn of the century, persistent even when corrected for smoking. This indicates the existence of a genetic component which is relevant in the aetiology of LC. Today (27th February 2017) PubMed lists 222,009 scientific publications assigned to the key "Lung neoplasms/genetics", some dating back to the 1960s. In one of the early enlisted publications, "a negro family is presented in which all four of eight siblings older than 50 have developed carcinoma of the lung. Attention is drawn to the possible genetic and environmental factors which are etiologically related to cancer of the human lung." Although at this time no facilities for genotyping were available, it was concluded that "the inter-relationship of smoking and genetic factors may explain the very high incidence of lung carcinoma in this family". Investigations from Germany showed a 2.6-fold increased lung cancer risk in young patients (OR, 95% CI 1.6–6.0) if a first degree relative had been diagnosed with LC cancer and a 5.6-fold increased risk (OR, 95% CI 0.7-46.9) if a parent or sibling was affected with LC at young age, too. Even for non-smokers aged 40 to 59, an increase of the lung cancer risk up to 6-fold was seen in the presence of lung cancer in a first-degree relative.

In 1990 the results of a segregation analyses performed on 337 families, each ascertained through a lung cancer patient indicated that early onset LC can be caused by a Mendelian codominant inheritance of a rare major autosomal gene. Segregation at this putative locus
could account for about 2/3 of the cumulative incidence of lung cancer in individuals up to age 50.\textsuperscript{47,48} It was not until 14 years later that the first evidence for linkage of a lung cancer susceptibility locus was published. The putative locus was localized to a region on 6q23–25.\textsuperscript{49}

This provides evidence that genetic factors contribute to the susceptibility of LC. During the last decade several genetic variants have been identified as associated to lung cancer or to a specific histological sub-type by genome-wide association studies (GWASs), candidate gene or pathway studies. Genomic risk loci in European population were identified at chromosomes e.g. 5p15.33, 6p21-22 and 15q25.\textsuperscript{3-10} Most of these variants could only be detected by combining several GWASs meta-analytically within the International Lung Cancer Consortium (ILCCO) or the Transdisciplinary Research in Cancer of the Lung (TRICL), and hence increasing the sample size.

1.2 Genome-wide association studies

In 1909 Wilhelm Johannsen coined the word “gene” to label the Mendelian unit of heredity. He also introduced the terms “genotype” and “phenotype” to discriminate between an individual genetic traits and its physically or mental appearance. But only after in 1953, Watson and Crick reached their ground-breaking conclusion that the deoxyribonucleic acid molecule (DNA) exists in the form of a three-dimensional double helix, the basis to investigate the genomic contribution to human life was given. Since then, medical and biologic scientists turn their attention to the human DNA and its differences between individuals or populations. The so-called “genomic variations” in the human genome can be of different forms, including single nucleotide polymorphisms (SNPs) or substitutions, tandem repeats, insertions or deletions (indels) or copy number variations (CNVs) or other chromosomal rearrangements. Genetic variations can be of diverse sizes; from single nucleotides to several mega bases. Owing to their inherent features, variations of larger size like e.g. tandem repeats have first been used in linkage studies, aiming to locate the chromosomal regions harbouring the mutations or genes for monogenic or familial disorders or quantitative with high penetrance traits.\textsuperscript{50} The drawback of such family-based approaches can mainly be attributed to their low statistical power, when several genomic variations, each with small effect sizes and/or low penetrance, contribute to the heritability of the trait of interest (e.g. a disease status). In such a situation family-based approaches would require impractically large sample sizes in order to detect genetic risk factors.\textsuperscript{51} Most diseases are believed to have such a complex genetic architecture.\textsuperscript{52}

Starting with the decoding of the human genome sequence in 2003 and the availability of high-throughput genotyping facilities, the focus of genetic epidemiologists shifted towards genetic variations commonly occurring in the entire population, rather than in selected families. They also started to investigate the whole sequence of the DNA, rather than pre-selected candidate genes. “Association studies using common allelic variants are cheaper and simpler than the complete resequencing of candidate genes, and have been proposed as a powerful means of identifying the common variants that underlie complex traits. In their
For association studies, typically log-linear models were assumed, which are robust against the majority of all possible genetic models. There is also no need to make assumptions about the genomic location of the causal variants. Association studies make use of the principle of linkage disequilibrium (LD) at the population level. LD simply quantifies the probability of the alleles of two loci (e.g. disease and marker loci) to appear on the same gamete more often together than expected by random. One of the simplest measures of LD is:

\[ D = p_{AB} - p_A \cdot p_B \]

where \( p_A \) is the frequency of allele A at the first locus, \( p_B \) is the frequency of allele B at the second locus, and \( p_{AB} \) is the frequency of the haplotype AB (the joint presence of A and B).

Put simply, LD is the non-random association of alleles of loci. “When evolutionary forces can be ignored, including marker and disease locus mutation, any decay in disequilibrium is due solely to recombination” and hence applies to adjacent markers. This assumption is necessary to relate an observed association between a marker locus and a trait of interest (e.g. a disease status) to a causal genomic disease locus in LD to the marker locus. However, a useful level of LD between two loci is unlikely to extend beyond an average distance of 3 kb, aside from some genomic regions with extended LD-patterns.

A genome-wide association study (GWAS) is defined as an association approach that surveys most of the genome for causal genetic variants. This is accomplished by genotyping the DNA of each study participant at nowadays at least 300,000 loci. Common SNPs, defined by a minor allele frequency >5%, are usually investigated. The density of the SNPs chosen as genomic markers is intended to be high enough to contain (or cover) almost all the genome by LD. Association to the trait of interest is usually estimated and statistically tested marker by marker. “The genome-wide association approach therefore represents an unbiased yet fairly comprehensive option that can be attempted even in the absence of convincing evidence regarding the function or location of the causal genes.” With advances in genotyping technologies and the assembling of millions of SNPs to a reference sequence of the human genome, GWAS became affordable and popular for the investigation of genomic risk factors for common complex diseases.

1.3 Drawbacks of genome-wide association studies

GWASs provide the opportunity to identify single markers or narrow genomic regions which are associated to a disease using genotypes of thousands of SNPs throughout the whole genome. However, these analyses are not free of drawbacks. Two major aspects are a) missing heritability and b) low predictive ability.

Discussing the performance of published GWAS in 2010, Eichler et al. stated: “Although recent genome-wide studies have provided valuable insights into the genetic basis of human disease, they have explained relatively little of the heritability of most complex traits, and the
variants identified through these studies have small effect sizes. This drawback, noted as missing heritability, has been related to a variety of aspects in the design and analysis of GWAS. Issues like incomplete coverage; disregard of rare variants or CNVs; genes that map to regions of copy-number polymorphic (CNP) duplications, opposing effects of selection, population history, migration and mutation rates or population stratification have been deemed to cause missing heritability. A further weaknesses is the often inadequate or completely neglected modelling of epigenetic effects, gene x environment interaction or gene x gene interaction. But exactly such interdependencies can be expected, taking into account the complexity of molecular-biological mechanisms. Yet, they require very large sample sizes to be discovered.

Given the limited ability to identify susceptibility loci, it is not surprising that GWAS results are often not sufficient to distinguish between individuals with low and high genetic disease risk (noted as low predictive ability). For breast cancer the “Gail model for prediction” achieves an AUC (area under the curve) of 58%, without regarding known genetic risk factors. After incorporating 10 relevant genetic variants identified by GWAS the AUC only increased to 61.8%. For prostate cancer an AUC of 86.2% for the prediction with the PSA (prostate specific antigen) alone was reported. After adding 33 genetic variants into the model the AUC increased only to 87.2%. For lung cancer a prediction model, mainly based on smoking history in a Chinese population, reached an AUC of 61.9%, after adding 5 relevant SNPs the AUC increased to only 63.9%.

It was pointed out that GWAS are neither intended to explain all genetic variation nor to find appropriate prediction models, but to observe associations between single loci and complex traits. However, the interplay of genes in the aetiology of the considered phenotype remains unconsidered in GWAS, but can be important regarding the biological nature of the trait. The joint consideration of molecular-biologically meaningful sets of markers, respectively genes, has therefore been proposed as an additional approach to reveal genetic risk factors or pinpoint to involved molecular mechanisms. A further part of the “missing heritability” might be explained in this way. Simultaneously considering related markers can provide a boost of power and uncover genes that are relevant in the aetiology of a disease but with low effects. To jointly analyse of several markers, respectively genes, allow researchers to better explore the multifaceted genetic architectures of complex diseases.
2 Gene-set analysis (GSA)

Several methods for gene-set analysis (GSA) were proposed as complementing approaches to the investigation of the genetic basis of diseases using GWAS results or including hundreds of markers in single, very large statistical models. All approaches have been developed to investigate pre-specified biological pathways, gene networks or gene families (further denoted as gene-sets (GS)). The annotation of markers to genes (MtG), respectively genes to gene-sets/pathways (GtP) can be obtained from public accessible data bases. For example dbSNP or ENSEMBL can be used for MtG assignment; KEGG, GO or DAVID can be used for GtP assignment.

The GSA approaches can be grouped according to the null-hypothesis being tested. In the context of GWAS these are:

Self-contained null-hypothesis (Q1) — The genes in a gene-set are not associated with the disease phenotype;

Competitive null-hypothesis (Q2) — The genes in a gene-set show the same magnitude of associations with the disease phenotype compared with genes in the rest of the genome.

2.1 Methods for the self-contained null-hypothesis

Some methods for testing the self-contained null-hypothesis (Q1) need the original genotype data as input and build a common statistical model for all markers of a considered GS. They are time- and storage-intensive. Comprehensive comparisons of these methods outline the pros and cons, but without pointing to one approach as the overall best. Two methods are exemplarily introduced here, to demonstrate the methodical challenges in the way to define such a common model:

The Network-Based Kernel Machine Test makes use of a semi-parametric logistic regression model for the probability of being a case, including all markers belonging to genes of a considered GS. Genetic effects are modelled non-parametrically. Environmental effects are modelled parametrically. The fitted model is of the form:

\[
\logit(p(y_i = 1)) = x_i^T \beta + h(z_i)
\]

where \( y_i \) is the case-control indicator (control: \( y_i = 0 \), case: \( y_i = 1 \)) for \( i = 1, ..., n \) individuals. The vector \( \beta \) represents the intercept and regression coefficient terms related to the environmental covariates \( x_i \) for the \( i \)th individual. The variable \( z_i \) denotes the genotype vector of selected marker for the \( i \)th individual.

The nonparametric, unknown centred smooth function \( h \) describes how the risk of being affected by the disease depends on the observed genotypes and can take the form:

\[
h(z_i) = \sum_{j=1}^{n} \alpha_i K(z_i, z_j)
\]
where $K$ (the so called kernel) can be understood as measuring the similarity between the individuals $i$ and $j$ based on their genotypes.

“... [F]or GWAS, the linear kernel $K(z_i, z_j) = z_i^T z_j$, was probably the most frequently applied kernel. Using this kernel in the logistic kernel machine test, is equivalent to using a logistic regression with a linearly defined random effect for all SNPs”. $^{81}$ This kernel is a special case of the more general defined d-th polynomial kernel:

$$K(z_i, z_j) = (z_i^T z_j + \rho)^d$$

where $\rho$ and $d$ are tuning parameters. $^{82}$ However, both kernels fail in case of gene-gene interaction within the GS. Alternative kernels have been defined and successfully applied in the context of GWAS$^{80,83-85}$. For instance the identity-by-state (IBS) kernel rest upon the proportion of alleles shared between two individuals $i$ and $j$. The IBS kernel is defined as:

$$K(z_i, z_j) = \sum_{l=1}^{n_m} 2I(z_{il} = z_{jl}) + I(|z_{il} - z_{jl}| = 1) \frac{1}{2n_m}$$

where $I$ denotes an indicator function taking the values 0 or 1 and $n_m$ is the number of markers belonging to GS. This kernel has been shown to be more robust in case of non-linearity of genotype effects than the linear kernel. $^{85}$

For the upper mentioned Network-Based Kernel Machine Test$^{80}$ the matrix $K$ is built as product of the observed genotypes, the MtG assignment and a weighting of genes according their importance within the GS. The network-based kernel is defined as:

$$K(z_i, z_j) = Z A N A^T Z$$

where $Z$ is a matrix containing the genotype data per individual (coded in trinary fashion – 0,1 or 2; respectively in-between values for imputed genotypes; dimension: number of individuals times number of markers in GS) and $A$ is a matrix containing the MtG assignment (dimension: number of markers in GS times number of genes in GS). The network structure of a considered GS is converted to an undirected adjacency matrix $N$ (dimension: number of genes in the path - squared) with all diagonal elements equal 1, due to “self-interaction”. Any other element of $N$ represents the interaction of a pair of genes within the GS, where 1 represents activation and −1 represents inhibition. For the construction of $N$ the database KEGG can be used. Thus the network-based kernel incorporates external knowledge of the biological mechanisms within the GS of interest into the GSA. However, the matrices to build $K(z_i, z_j)$ can get very bulky for large gene-sets which comprise hundreds of markers, in particular for a GWAS with a large sample size.

As alternative approach, Chen et al. proposed a “gene-set ridge regression in association studies (GRASS)” algorithm. $^{86}$ In order to reduce the amount of data finally used for modelling, the genetic variation of all SNPs belonging to a single gene is decomposed by applying principal component analysis. Thereby, orthogonal $eigenSNPs$ per gene are generated, but
only those that explain 95% of the genetic variation of a gene will further be included in the
statistical model. Hence the variable $z_i$ denotes here the vector of eigenSNPs of the $i$th individual. In contrast to the method above, GRASS utilises a logistic fully-parametrical regression model:

$$
\text{logit}(p(y_i = 1)) = x_i^T \beta_1 + z_i^T \beta_2
$$

but the estimated parameters $\beta_2$ are constrained under the usually penalty function of a ridge regression:

$$
\|\beta_2\|_\gamma = \left(\sum_{j=1}^{p} |\beta_j|^\gamma\right)^{\frac{1}{\gamma}}
$$

with $0 < \gamma < 1$ and $p$ the number of parameters $\beta_j$.

The second methods, as all others testing $\beta_2$ in the manner of Q1, may be invalid if e.g. the assumptions of the models are not fulfilled. This might be the case if variables (e.g. markers) present a multi-collinear structure (strong LD). They can also lead to false positive findings, e.g. if not sufficiently adjusted population stratification causes spurious association.

Other methods to test Q1 are based on the weighted Fisher’s inverse $\chi^2$-method to pool statistical evidence in combining several p-values (further denoted as SPP: simple p-value pooling):

$$
M = -2 \sum_{m=1}^{n_m} w_m \ln(p_m)
$$

with $m$ an index for a marker, $n_m$ the number of markers belonging to GS and $w_m$ a weight.

Luo et al. proposed first to generate the correlation matrix $R_m$ of the normal-quantiles of the $p_m$-values $Z_m = \Phi^{-1}(1 - p_m)$ for all markers belonging to a gene $g$. Then a linear combination in the form

$$
T_g = \frac{e^{e^2 Z_m}}{\sqrt{e^{e R_m e}}} \quad \text{with} \quad e = (1, 1, \ldots, 1)^T
$$

is calculated and a gene-wide $p_g$-value, assuming $T_g$ to be standard-normally distributed, is derived. Finally, all $p_g$-values of genes belonging to the GS are used to calculate $M$ (instead of $p_m$) and the corresponding GS-wide $p_{GS}$-value, setting $w_g = 1$ (instead of $w_m$) for all genes.

De la Cruz et al. proposed the method SLAT (Set Level Association Testing), which differs from the above mentioned approach by truncating $p_m$ lower than a pre-specified threshold (=restriction of accounted markers) and assigning weights $w_m$ according the number of markers in LD per gene.
2.2 Methods for the competitive null-hypothesis

Alternative to such “one-model” approaches, methods testing the competitive null-hypothesis (Q2) have been developed. These usually compare single marker association results (e.g. p-values or odds ratios) of markers/genes belonging to GS with those of all remaining genes in the genome (remaining GS’). Thus, pathways enriched with moderate association signals at several markers may be discovered which would be missed by a single marker approach.

As before, comprehensive comparisons of these methods outline the pros and cons, but without pointing to one approach as the overall best.\textsuperscript{68,69,71,79} Two methods are exemplarily introduced here, to demonstrate the methodical challenges in the way to define such a common model:

The first and most popular approach briefly described here is GSEA (Gene-Set Enrichment Analysis), proposed by Subramanian et al.\textsuperscript{90,91} A given list of markers is first ranked by $p_m$-values. To evaluate the degree of “enrichment” the GSEA method calculates an Enrichment Score (ES) by walking down this ordered list. At the beginning, a cumulative sum $C$ is set to zero and will be increased when a marker belongs to GS and decreased otherwise. The increment of increase and decrease is chosen in the way that $C$ is equal to zero at the end of the list again. If a GS is enriched with markers of low $p_m$-values, than $C$ will get very large quickly. If the GS is not associated to the trait of interest, $C$ will follow a random walk around zero. The maximum departure of $C$ from zero is taken as ES and can be interpreted as a weighted Kolmogorov-Smirnov statistic. Hence GSEA tests, whether the distribution of $p_m$-values follows a uniform distribution. GSEA normalizes the ES for each GS to account for the variation in set sizes, yielding a normalized enrichment score (NES). To avoid dependencies of markers in LD, one may choose the smallest $p_m$-value of markers belonging to the same gene.

The second approach briefly described here is ALIGATOR (Association List Go AnnoTatOR)\textsuperscript{92}. It uses of all $p_m$-values per marker resulting from a previous GWAS, regardless whether a marker belongs to GS or GS’. The first step consists of counting the numbers of significant markers $m_{\text{sig,GS}}$ belonging to a GS and $m_{\text{sig,GS}'}$ belonging to the remaining GS’. These counts are then compared like in Fisher’s exact test, however the null-distribution of the test statistic is generated by a Monte-Carlo-Permutation procedure. It is desirable to correct the achieved GS-specific p-values for the number of GSs being tested. Because the GSs are not generally independent, standard methods, such as the Bonferroni and Sidak corrections, are inappropriate; a bootstrap approach is applied instead. This method was used to accomplish GSA for this dissertation.

The input datasets for this method are small, because the thousands of genotypes of each individual are no more required; but the accomplishment is time-intensive. However, because of the massive use of permutation and boot-strapping almost no model assumption needs to be fulfilled.
The diversity of the upper mentioned methods illustrates that GSA itself is a generic term describing an analytical strategy rather than a single method. Approaches differ for example by the way to combine results of several makers to a single quantity for a gene, by the way to assign genes to pathways or to statistically contrast genes belonging to the GS of interest with all remaining genes (providing a \( p_{GS} \)-value for a GS).

The performance of some strategies has been compared \(^{25,31,32}\); e.g. by G. Fehringer et al.\(^{33}\) who compared GSA applied to two independent GWASs formed by several ILCCO/TRICL data sets. They found “a highly plausible association for the acetylcholine receptor activity pathway”, but concluded: “Difficulty in replicating associations (between the independent GWAS datasets) hindered our comparison ...”. They used a heuristic approach by declaring those pathways as highly plausibly associated to lung cancer which were highly ranked (low gene-set \( p \)-values) by at least two GSA methods in both GWAS data sets, without taking additional information into account.

From a scientific point of view such an ad-hoc approach is not satisfying, because significance and consistency need to be addressed in a systematic quantitative manner. GSA usually provides \( p \)-values for GSs without giving any kind of effect/association estimate. Hence, Fisher’s inverse \( \chi^2 \) method (here also denoted as SPP) - a well-established method to pool significance by a meta-analysis - might be applied. However, even if a GS of interest is found significantly “enriched” within all of several independent GWAS data sets, it is not guaranteed that the underlying single-marker association results are consistent regarding the direction of the association (as risk or protection factor for the disease).
3 Summaries

3.1 META-GSA: A novel method to synthesise statistical evidence of several gene-set analyses

I have proposed the quantitative approach META-GSA to combine results of GSAs, respectively $p_{GS}$-values of GS, by incorporating concordance of single-marker association patterns between studies, relevant for the GS of interest.\(^{11}\)

These $p_{GS}$-values usually result from one-sided statistical tests (e.g. the larger ES the lower $p_{GS}$). Hence, low $p_{GS}$-values result from low single marker $p$-values $p_m$ of genes in GS. Nevertheless, low $p_{GS}$-values can theoretically arise through accumulated marker significance in which the minor alleles of all markers are observed for example as protective factors in one study, while being seen at the same time as risk factors in another study. Thus, significance for GS can appear simultaneously in several studies without concordance of the patterns of observed associations of markers, respectively genes, contained in GS (briefly denoted as association pattern). Consequently, concordance of the direction of the test (of $p_{GS}$) is not given a priori.

The main steps of META-GSA are first to determine the concordance of association patterns; second, to use these to derive a weight for each study; and third, to apply a weighted version of Fisher's inverse $\chi^2$-method\(^ {93}\) to summarize significance of GSAs in a single meta-analytical $p$-value. Thus significance of GSA-results and concordance of single-marker association are combined. META-GSA can be further considered as an approach testing GS-significance conditional to, or in the presence of concordance of association patterns.

META-GSA was intended to use as little input data as possible. This makes it feasible even if the access to original study data is restricted, e.g. if data owners are reluctant to share detailed information due to legal or ethical causes. Because META-GSA is designed to rely on $p$-values as quantity of significance, it was necessary to define a measure that reflects also the direction of the observed association for a marker (indicating a protective or risk factor). Thus, a directed reversed $p$-value (PDR) was defined as:

$$p_{m,s}' = d_{m,s} \cdot (1 - p_{m,s})$$

where $m$ is an index for a marker, $s$ is an index for a study and $d_{m,s} \in \{-1, +1\}$ is the direction of the observed association.

Next, the correlation of all $p_{m,s}'$-values of all markers belonging to GS is calculated. Because it is not assumed that each study used the same GSA approach, a rank-correlation, which is independent of the scale of $p_{m,s}'$, is used. Once the correlation matrix is found, principal component analysis (PCA) is used to determine the load of each study on a common but unknown general PDR-profile of the GS. We assume that these loads can be represented by the
first principal component \((PC_1_s)\). Study weights \(w_s\) are then calculated as the product of the normalized load and the proportion of effective studies:

\[
w_s = \frac{PC_1_s}{\sum_{j=1}^{n_s} PC_1_j} \cdot \frac{EV1}{n_s}
\]

where \(n_s\) denotes the number of studies in the meta-analysis and \(EV1\) denotes the first eigenvalue. \(EV1/\sum_{j=1}^{n_s} EV1_j\) is the fraction of explained variance, since \(\sum_{j=1}^{n_s} EV1_j = n_s\) for a correlation matrix, \(EV1/n_s\) can be considered as the proportion of effective studies.

Finally, the test statistic

\[
M_{n_s} = -2 \sum_{s=1}^{n_s} w_s \ln(p_{GS,s})
\]

is calculated and a corresponding \(p_{META-GSA,GS}\)-value is derived. Since the mathematical conditions of the weighted version of Fisher’s inverse \(\chi^2\)-method to assume \(M_{n_s}\) as \(\chi^2\)-distributed are not fulfilled, the application of a CPU-intensive permutation procedure is required.

Accordingly, the core test-statistic \(M_{n_s}\) of META-GSA depends on necessary aspects being considered: GS-significance by \(p_{GS,s}\)-values and concordance of association patterns by the weights \(w_s\).

Furthermore, I have investigated the benefits and the effort of META-GSA in comparison with SPP, which is fast but does not address concordance of association patterns in any way. Both methods keep the type 1 error at the specified level. However, under \(H_0\) the results of META-GSA and SPP were found to be almost uncorrelated. False-positive gene sets found by META-GSA and SPP only partially overlap. We assume that this may result at least in part from including the concordance of association patterns, for the evaluation.

Moreover, META-GSA was found to be more powerful than SPP. The greater the number of studies combined, the larger the advantages in power became.

We also compared META-GSA to a pooledGWAS-GSA approach. For the latter, the combining of studies is switched to the level of markers, followed by a single GSA performed on the pooled marker-specific associations. In general, we found META-GSA to outperform pooledGWAS-GSA.

The method META-GSA is described in detail in Rosenberger et al. *PLoS One* 2015. All programs were implemented in SAS 9.3 (SAS Institute, NC, USA) and are provided as supplemental material.
3.2 Gene-set analysis with respect to lung cancer: A meta-analysis

I have further performed a meta-analysis of seven GSAs for lung cancer, applying the method META-GSA. Overall, information taken from 11,365 cases and 22,505 controls from within the TRICL/ILCCO consortia was used to investigate a total of 234 pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

In the original GWASs, a log-additive mode of inheritance was fitted for each marker, adjusting for age, sex, smoking status, study centre (if applicable), and the first three principal components to account for hidden genomic structure. There was agreement within TRICL/ILCCO on the model to be used for single-marker association analysis. The calculations had been performed by the study centres themselves and shared within the consortium. The results of marker-by-marker association tests were used as input information for the GSAs. The marker-to-gene annotation (MtG) for humans from the ENSEMBL database was used. For some data set GSA results were already available. If necessary, GSA was performed using the program ALIGATOR.

This revealed the systemic lupus erythematosus KEGG pathway hsa05322, driven by the gene region 6p21-22, but no other investigated KEGG pathway, as implicated in lung cancer (p=0.0306, corrected for multiple testing). This gene region is known to be associated with squamous cell lung carcinoma. The most important genes driving the significance of this pathway belong to the genomic areas HIST1-H4L, -1BN, -2BN, -H2AK, -H4K and C2/C4A/C4B. Within these areas, the markers most significantly associated with LC are rs13194781 (located within HIST12BN) and rs1270942 (located between C2 and C4A).

4 Discussion

About 50,000 individuals are diagnosed with lung cancer in Germany every year. They face a poor prognosis, since only 21% of women and 16% of men survive five years after being diagnosed. Most cases can be attributed to tobacco smoking, but there is also scientific evidence that genetic factors contribute to the susceptibility of Lung cancer. In the recent two decades genome-wide association studies (GWASs) have been carried out to identify genetic variants that are associated to lung cancer or to a specific histological subtype. Genomic risk loci in European population were identified at chromosomes e.g. 5p15.33, 6p21-22 and 15q25, after pooling data of several GWASs within the International Lung Cancer Consortium (ILCCO), overlapping with the Transdisciplinary Research in Cancer of the Lung (TRICL). It has also been increasingly recognized that GSA can extend GWAS approaches by incorporating existing knowledge of biological processes, with the aim of identifying disease-related pathways. GSA has gained great popularity and several approaches have been proposed. Although the pros and cons have been discussed and points to improve have been formulated, it has only been mentioned that there is a need to replicate pathway association findings to avoid false positive results. According to our knowledge, there is no formal method to combine the results of several GSAs. The basic criticism on applying simple p-pooling (SPP) is a lack of interpretability of results when single markers differ in their attributed role as risk or predictive factor. I have proposed the quantitative approach META-GSA to combine such results, respectively, GS-significance, by incorporating concordance of single-marker association patterns between studies, relevant for the GS of interest. This method was then applied to the data shared within ILCCO/TRICL, to further investigate genetic risk factor for lung cancer.

META-GSA was found to outperform SPP and a pooledGWAS-GSA approach. One advantage of META-GSA is the fact that heterogeneity in the strength of association for single markers or genes, respectively, between studies does not necessarily cause lower power, as long as other genes belonging to GS compensate such deficiency. “Between-study heterogeneity … can offer valuables insight for further clarification of gene-disease associations”.

Furthermore, META-GSA is applicable to any GSA method selected, even those using individual participants’ genotype data, which may prove to be more suitable and more powerful than methods based on GWAS summary results (pooledGWAS-GSA). Resting the GSA on common effect estimates can become critical in the case of strong study heterogeneity for few or many markers, since the existence of a common marker-specific association in such a situation is doubtful. For META-GSA, study heterogeneity results in low concordance of the patterns of study-specific association estimates and subsequently in low study weights, which simply reduces the power, however without violating such a critical assumption.

All the same, META-GSA has some critical points that need to be mentioned. First, it is necessary to estimate LD between neighboring markers to be able to calculate PDRs (the core
quantity of META-GSA) of differing markers across studies. It is known that LD-patterns in human subpopulations are different.\textsuperscript{97} This can be problematic if one aims to combine results across different ethnicities or if the source population of a study is known to be admixed.

Secondly, the implemented way to combine marker-level estimates to a gene-level statistic is only one of several published alternatives.\textsuperscript{98–101}

Moreover, to perform GWASs and GSAs one needs to annotate marker to genes (MtG) which are annotated to pathways (GtP) in a static way, referring to public databases. In contrast, genes act dynamically and may have for instance several transcripts which can be active or passive in a certain pathway. Thus, a static annotation maybe doesn’t match well to a dynamic biological process. META-GSA is concerned to this drawback to the same extend as the GSA approaches aimed to be combined.

Next, META-GSA is applicable to any GSA method selected for the analysis of a single GWAS, irrespectively if the \textit{self-contained null-hypothesis} (Q1) or the \textit{competitive null-hypothesis} (Q2) was tested. Combining \( p_{GS,s} \)-values for Q1 and Q2 is calculative feasible but induces a lack of interpretability.

In addition, the use of a permutation procedure is time and CPU-intensive.

The biggest disadvantage is perhaps that p-values for each tested gene-set are the only results. META-GSA is designed to rely on p-values as quantity of significance. A p-value can be used to justify the existences of an association; however it is not solely determined by the strength of the observed effect, but also by factors like sample size, the used statistical model and the applied test procedure. Hence META-GSA is unable to estimate the fraction of the risk for an investigated disease that can be attributed to the identified driving genes or the whole gene-set. As with most GSA approaches META-GSA does not deliver any effect estimation.

All in all, I could demonstrate that META-GSA may be a powerful add-on tool in the research of the genetic architecture of complex traits or diseases. I have applied this new tool to discover an accumulation of genomic association with lung cancer in the KEGG pathway hsa05322, which comprises genes related to \textit{systemic lupus erythematosus} (SLE). This suggests some cross-phenotype association with lung cancer and SLE.

Regarding the application, all \( p_{GS} \)-driving genes identified in the reported meta-analysis are located within or next to the major histocompatibility complex (MHC) on chromosome 6p21-22, albeit in two separate areas, about 3000 kb apart. The first area comprises the genes of histone cluster I: \textit{HIST1-H4L}, \textit{-1BN}, \textit{-2BN}, \textit{-H2AK}, \textit{-H4K}. The second area comprises the genes \textit{C2}, \textit{C4A}, and \textit{C4B}. Neither of both areas were before identified as associated to lung cancer. However, the identification of disease-relevant genes in the MHC region (6p21–6p22) and far beyond is complicated owing to the strong and extensive LD across both common and
rare haplotypes.\textsuperscript{12} Hence any observed cross-phenotype association will probably tag plenty of genes.

In conclusion, the gene region 6p21-22 appears to be more extensively associated with lung cancer than previously assumed. Given wide-stretched linkage disequilibrium to the area \textit{APOM/BAG6/MSH5}, there is currently simply not enough information or evidence to conclude whether the potential pleiotropy of lung cancer and SLE is spurious, biological, or mediated.
References


77. 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,


B Curriculum Vitae

Albert Rosenberger  born in Vienna, 16th July 1968
Nationality: Austrian

Professional Career

<table>
<thead>
<tr>
<th>Date</th>
<th>Institution</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>since 11/2001</td>
<td>University Medical Centre Göttingen</td>
<td>scientific employee</td>
</tr>
<tr>
<td>08/1997 – 11/2001</td>
<td>University Hospital Tübingen</td>
<td>scientific employee</td>
</tr>
<tr>
<td>01/1994 – 07/1997</td>
<td>University of Vienna</td>
<td>research associate</td>
</tr>
<tr>
<td>08/1996 – 03/1997</td>
<td>publishing house OSTAG / Vienna</td>
<td>technical co-worker and assistant of the management</td>
</tr>
<tr>
<td>11/1996 – 03/1997</td>
<td>AKH – General Hospital Vienna</td>
<td>freelancer and research associate</td>
</tr>
<tr>
<td>11/1994 – 06/1995</td>
<td>AKH – General Hospital Vienna</td>
<td>freelancer and research associate</td>
</tr>
<tr>
<td>01/1990 – 09/1990</td>
<td>SIEMENS Austria</td>
<td>technical co-worker</td>
</tr>
</tbody>
</table>

Education

<table>
<thead>
<tr>
<th>Year</th>
<th>Institution</th>
<th>Field of Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995 - 1996</td>
<td>University College Dublin:</td>
<td>1 year academic courses in statistics and sociology</td>
</tr>
<tr>
<td>1990 - 1996</td>
<td>University of Vienna:</td>
<td>regular studies of statistics and sociology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>graduation: Mag. rer. soc. oec.</td>
</tr>
<tr>
<td>1988 - 1989</td>
<td>obligatory military service</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>graduation: Matura</td>
</tr>
<tr>
<td>1974 - 1983</td>
<td>elementary and secondary school in Vienna</td>
<td></td>
</tr>
</tbody>
</table>
Publications

2017

- Pathway-analysis of published genome-wide association studies of lung cancer: A potential role for the CYP4F3 locus

- A Novel Genetic Variant in Long Non-coding RNA Gene NEXN-AS is Associated with Risk of Lung Cancer

- Genetic Risk Can Be Decreased: Quitting Smoking Decreases and Delays Lung Cancer for Smokers With High and Low CHRNA5 Risk Genotypes - A Meta-Analysis

- Polymorphisms of the centrosomal gene (FGFR1OP) and lung cancer risk: a meta-analysis of 14,463 cases and 44,188 controls
  Carcinogenesis. 2016 Mar;37(3):280-9

- Association of Patient Risk Factors and Frequency of Nevus-Associated Cutaneous Melanomas
  Haenssle H, Mograby N, Ngassa A, Buhl T, Emmert St, Schöm M, Rosenberger A, Bertsch HP.
  zitiert in: Nväusassozierte Melanome: Risiken quantifiziert, um Früherkennung zu optimieren; Dtsch Arztebl 2016; 113(25): A-1217 / B-1022 / C-1006

2015

- Inzidenz der Candidädie nach gastrointestinal-chirurgischen Eingriffen in deutschen Akutkrankenhäusern [Incidence of Candidaemia Following Abdominal Surgery in German Hospitals].
  Kuhns M, Rosenberger A, Bader O, Reichard U, Gross U, Weig M.
  Zentralbl Chir 2015; 140(6): 617-623

- Identification of lung cancer histology-specific variants applying Bayesian framework variant prioritization approaches within the TRICL and ILCCO consortia
  Carcinogenesis. 2015 Nov;36(11):1314-26
o META-GSA: Combining findings from gene-set analyses across several genome-wide association studies

PloS One. 2015 Oct 26;10(10):e0140179

o Identification of shared and unique susceptibility pathways among cancers of the lung, breast, and prostate from genome-wide association studies and tissue-specific protein interactions.


o CHRNAS Risk Variant Predicts Earlier Lung Cancer Diagnosis and Delayed Smoking Cessation - A Meta-analysis


o Informed genome-wide association analysis with family history as a secondary phenotype identifies novel loci of lung cancer


o Assessment of melanoma histotypes and associated patient related factors: Basis for a predictive statistical model

Haenssle H, Hoffmann S, Buhl T, Emmert St, Schön M; Bertsch HP, Rosenberger A

o Melanoma thickness: The role of patients’ characteristics, risk indicators and patterns of diagnosis


2014

o Permutationstest bei komplexen Teststrategien

Rosenberger A

o Re: Time to Smoke First Morning Cigarette (TTFC) and Lung cancer in a Case-Control Study

Rosenberger A

o Kernel score statistic for dependent data

Malzahn D, Friedrichs St, Rosenberger A, Bickeböller H

o Genetic factors of individual radiation sensitivity in lung cancer families

o Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer
Nat Genet. 2014 Jul;46(7):736-41. doi: 10.1038/ng.3002. Epub 2014 Jun 1

2013

o Identification of lung cancer histology-specific variants applying Bayesian framework variant prioritization approaches within the TRICL and ILCCO consortia
Brenner D.R., Amos C.I. et al. Rosenberger A, Hung RJ

o Mild overexpression of MeCP2 in mice causes a higher susceptibility toward seizures

o Limited diagnostic value of Wells-score and D-dimer testing in hospitalized dermatologic patients with symptoms of deep vein thrombosis
Haenssle H, Ayad N, Buhl T, Rosenberger A, Thoms KM, Schön MP, Zutt M

o Inzidenz der Candidämie nach gastrointestinal-chirurgischen Eingriffen in deutschen Akutkrankenhäusern (Incidence of Candidemia following abdominal surgery in German hospitals)
Kuhns M*, Rosenberger A*, Bader O, Reichard U, Groß U and Weig M (* first authors)

o Influence of Common Genetic Variation on Lung Cancer Risk: Meta-Analysis of 14,900 Cases and 29,485 Controls

o Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls

2012

o PI3K inhibition enhances doxorubicin-induced apoptosis in sarcoma cells
Informed conditioning on clinical covariates increases power in case-control association studies

Previous Lung Diseases and Lung Cancer Risk: A Pooled Analysis from the International Lung Cancer Consortium

Increased genetic vulnerability to smoking at CHRNA5 in early-onset smokers

Controlled-rate freezzer cryopreservation of highly concentrated peripheral blood mononuclear cells results in higher cell yields and superior autologous T-cell stimulation for dendritic cell-based immunotherapy
Buhl T, Legler TJ, Rosenberger A, Schardt A, Schön MP, Haensslie HA

Heritability of radiation induced DNA damage response in lung cancer families

Asthma and Lung Cancer Risk Study: A Systematic Investigation by the International Lung Cancer Consortium
o Calcitriol inhibits Hedgehog signaling and induces vitamin D receptor-signaling and differentiation in the Patched mouse model of embryonal rhabdomyosarcoma

o Integrating static and dynamic features of melanoma: The DynaMel algorithm

o Validation of three BRCA1/2 mutation-carrier probability models MYRIAD, BRCAPRO and BOADICEA in a population-based series of 183 German families
Schneegans SM*, Rosenberger A*, Engel U, Sander M, Emons G, Shoukier M
* both first authors

o A novel method for testing the veridicality of dental colour assessments
Knösel M, Reus M, Rosenberger A, Ziebolz D

o mentioned in the acknowledgements
Novel Roles of Caenorhabditis elegans Heterochromatin Protein HP and Linker Histone in the Regulation of Innate Immune Gene Expression

2011

o Interaction of human multidrug and toxin extrusion 1 (MATE1) transporter with antineoplastic agents
Grottker J, Rosenberger A, Burckhardt G, Hagos Y

o Antitumoral Effects of Calcitriol in Basal Cell Carcinomas Involve Inhibition of Hedgehog Signaling and Induction of Vitamin D Receptor Signaling and Differentiation

o Thrombophilia in patients with chronic venous leg ulcers-a study on patients with or without post-thrombotic syndrome
Zutt M, Krüger U, Rosenberger A, Schön M.P., Neumann Ch., von Ahsen N., Kretschmer L

o Aspirin and NSAID use and risk of lung cancer: International Lung Cancer Consortium pooled analysis

o Increased Lipoprotein (a) concentrations in patients with chronic venous leg ulcers: a study on patients with or without postthrombotic syndrome
o Durability of bleaching results achieved with 15% carbamide peroxide and 38% hydrogen peroxide in vitro
Knösel M, Reus M, Rosenberger A, Attin T, Ziebolz D
Eur J Esthet Dent. 2011 Autumn;6(3):342-56

INAUGURAL – DISSERTATION of Reus M. unter dem Titel „Vergleich der Farbstabilität milder und hochkonzentrierter Bleichmittel - Eine In-vitro-Untersuchung” (2010, Georg-August-Universität zu Göttingen

o A Phenotype-Based Genetic Association Study Reveals the Contribution of Neuregulin1 Gene Variants to Age of Onset and Positive Symptom Severity in Schizophrenia

o Validation of a fully automated COMET assay: 1.75 million single cells measured over a year period

2010

o GPC rs2352028 variant and risk of lung cancer in never smokers

o Activation of the hedgehog pathway confers a poor prognosis in embryonal and fusion gene-negative alveolar rhabdomyosarcoma

o Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium

o Inverse relationship between cerebrovascular lesions and severity of Lewy body pathology in patients with Lewy body diseases
Ghebremedhin E, Rosenberger A, Rüb U, Vuksic M, Berhe T, Bickeboller H, de Vos, RAL, Thal DR, Delter Th

o Tumor stroma-derived Wnta induces differentiation of basal cell carcinoma of Ptc1 mutant mice via CaMKII
Cancer Res. 2010 Apr 1;70(7):2739-48. doi: 10.1158/0008-5472.CAN-09-3743. Epub 2010 Mar 16

o Selection of patients for long-term digital dermoscopy by assessment of melanoma risk factors
Seven-point checklist for dermoscopy: Performance during years of prospective surveillance of patients at increased melanoma risk

Genetic polymorphism of MPO, GSTT, GSTM, GSTP, EPHX and NQO as risk factors of early onset lung cancer

2009

Integration of a priori gene set information into genome-wide association studies
Soehn M, Rosenberger A, Bickeböller H

Effect of DNA repair host factors on temozolomide or dacarbazine melanoma treatment in Caucasians

A genome-wide association study of lung cancer identifies a region of chromosome 6p associated with risk for adenocarcinoma

Intracellular delivery of MHC class I-binding epitopes: DCs loaded and matured with cationic peptide/polyl(C) complexes efficiently activate T cells
Haenssle HA, Riedl P, Buhl T, Schardt A, Rosenberger A, Schön MP, Schirmbeck R

Associations between the oxytocin receptor gene (OXTR) and affect, loneliness and intelligence in normal subjects
Lucht MJ, Barnow S, Sonnenfeld Ch, Rosenberge A, Grabe HJ, Schroeder W, Voelzke H, Freyberger HJ, Herrmann FH, Kroemer h Roskopf D

CYP polymorphisms as risk factors for early onset lung cancer: gender specific differences

TLR Gene Polymorphisms and Liver Disease Manifestations in Chronic Hepatitis C

A multiplex real-time PCR method for detection of GSTM and GSTT
Timofeeva M, Jäger B, Rosenberger A, Sauter W, Wichmann H-E, KORA Study Group, Bickeböller H, Risch A
Antitumor effects of a combined -aza'-deoxycytidine and valproic acid treatment on rhabdomyosarcoma and medulloblastoma in Ptch mutant mice

Impact of Endothelin- LysAsn polymorphism for coronary artery disease and endorgan damage in hypertensives

Matrix metalloproteinase (MMP) is associated with early-onset lung cancer


Detection of ATM Gene Mutations in Young Lung Cancer Patients: A Population-based Control Study


CD ligation during dendritic cell maturation reduces cell death and prevents interleukin--induced re expression to macrophage-like monocytes

Cyclosporine treatment of full-blown Hh/Pch-associated RMS results in interference with Hh/Pch signalling, but not in inhibition of tumor growth

Spontaneous elimination of hepatitis C virus infection: A retrospective study on demographic, clinical, and serological correlates

Sex-dependent genetic markers of CYPA expression and activity in human liver microsomes

Meta-analysis of whole-genome linkage scans for Parkinson's disease
2006

- Significant association of a MV independent polymorphism in the 5' UTR of the PRNP gene with sporadic Creutzfeldt-Jakob disease in a large German case-control study
  Vollmert C, Windl O, Xiang W, Rosenberger A, Zerr I, Wichmann H-E, Bickeboller H, Illig T, the KORA group, Kretzschmar H A
  Journal of Medical Genetics 2006 Oct;43(10):e53

- A functional polymorphism within plasminogen activator urokinase (PLAU) is associated with Alzheimer's disease

- Results from an observational trial: digital epiluminescence microscopy follow-up of atypical nevi increases the sensitivity and the chance of success of conventional dermoscopy in detecting melanoma
  J Invest Dermatol. 2006 May;126(5):980-5

The working group of Dr. Hänßle had been awarded with the Förderpreis der Niedersächsischen Krebsgesellschaft

- Relationship of apolipoprotein E and age at onset to Parkinson disease neuropathology
  Ghebremedhin E, Del Tredici K, Vuksic M, Rub U, Thal DR, Burbach GJ, Rosenberger A, Bickeboller H, Deller T, de Vos RA, Jansen Steur EN, Braak H

2005

- Detection of the CCR-Delta HIV resistance gene in Bronze Age skeletons
  Hummel S, Schmidt D, Kremeyer B, Herrmann B, Oppermann M. (Rosenberger A acknowledged for detailed statistical analysis)

- Surrogate phenotype definition for alcohol use disorders: a genome-wide search for linkage and association
  Rosenberger A, Janicke N, Köhler K, Korb K, Kulle B, Bickeboller H
  BMC Genet. 2005 Dec 30;6 Suppl 1:555

- Dissection of heterogeneous phenotypes for quantitative trait mapping
  Genet Epidemiol. 2005;29 Suppl 1:S41-7

- NAD(P)H oxidase and multidrug resistance protein genetic polymorphisms are associated with doxorubicin-induced cardiotoxicity

- Effects of socio-demographic variables on health-related quality of life determined by the quality of life index - German version
  Hum Psychopharmacol. 2005 Jul;20(5):359-65

- (mentioned in the acknowledgment)
  Detection of the CCR-Delta HIV resistance gene in Bronze Age skeletons
  Hummel S, Schmidt D, Kremeyer B, Herrmann B, Oppermann M
2004

- Results of a surveillance programme for patients at high risk of malignant melanoma using digital and conventional dermoscopy
  Eur J Cancer Prev 2004 Apr;13(2):133-8

2003

- Nonparametric longitudinal allele-sharing model
  Kulle B, Köhler K, Rosenberger A, Loesgen S, Bickeböller H
  BMC Genet 2003 Dec 31;4 Suppl 1:S85

- Modeling and dissection of longitudinal blood pressure and hypertension phenotypes in genetic epidemiological studies
  Bickeböller H, Barrett JH, Jacobs KB, Rosenberger A
  Genet Epidemiol. 2003;25 Suppl 1:S72-7

- Retrospective study of malignant melanoma patients treated with mistletoe extracts
  Stumpf C, Rosenberger A, Rieger S, Tröger W, Schietzel M, Stein GM

2002

- Sleeping habits in the Austrian population
  Sleep Med. 2002 Jan;3(1):21-8

- Methods in evidence-based medicine consensus findings - a case study in trauma surgery
  Rosenberger A, Selbmann HK
  Z Arztl Fortbild Qualitatssich 2003 Jul;97(4-5):311-9

- Childhood diseases, infectious diseases, and fever as potential risk factors for cancer?
  Hoffmann C, Rosenberger A, Tröger W, Bühring M
  Forsch Komplementarmed Klass Naturheilkld. 2002 Dec;9(6):324-30

- Validation of questionnaires from several medical fields regarding the constitution of patients
  Hoffmann C, Rosenberger A, Tröger W, Stange R, Bühring M
  Forsch Komplementarmed Klass Naturheilkld. 2002 Feb;9(1):37-44

2001

- N-CPAP rejecters--a specific group of noncompliant patients with obstructive sleep apnea syndrome
  Wien Klin Wochenschr. 2001 Apr 17;113(7-8):245-8

2000

- Sleep and quality of life in the Austrian population

- Mistletoe extracts in the therapy of malignant, hematological and lymphatic diseases - a monocentric, retrospective analysis over years
  Stumpf C, Rosenberger A, Rieger S, Tröger W, Schietzel M
  Forsch Komplementarmed Klass Naturheilkld. 2000 Jun;7(3):139-46

1999

- Development of lymphocyte subsets in tumor patients after subcutaneous administration of mistletoe extracts
  Büssing A, Rosenberger A, Stumpf C, Schietzel M
  Forsch Komplementarmed. 1999 Aug;6(4):196-204
1998

- Self-reported prevalence of venous diseases in the general population of Austria--results of the SERMO (Self-Reported Morbidity) Study
  Schmeiser-Rieder A, Kunze U, Mitsche N, Rosenberger A, Kunze M

- Technical report: public health monitor: SERMO Study--significance for immunization programs
  Schmeiser-Reider A, Kunze U, Mitsche N, Rosenberger A, Kunze M

- Impact of nasal continuous positive airway pressure treatment on quality of life in patients with obstructive sleep apnoea
  Eur Respir J. 1998 Apr;11(4):890-4