Dartmouth College
Dartmouth Digital Commons

Open Dartmouth: Faculty Open Access Articles

3-8-2017

Gene-Set Meta-Analysis of Lung Cancer Identifies Pathway Related to Systemic Lupus Erythematosus

Albert Rosenberger
Georg-August-University Gottingen

Melanie Sohns
Georg-August-University Gottingen

Stefanie Friedrichs
Georg-August-University Gottingen

Rayjean J. Hung
Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto

Gord Fehringer
Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto

See next page for additional authors

Follow this and additional works at: https://digitalcommons.dartmouth.edu/facoa
Part of the Medicine and Health Sciences Commons

Recommended Citation

Authors

Albert Rosenberger, Melanie Sohns, Stefanie Friedrichs, Rayjean J. Hung, Gord Fehringer, John McLaughlin, Christopher I. Amos, Paul Brennan, Angela Risch, Irene Bruske, Neil E. Caporaso, Maria Teresa Landi, David C. Christiani, Yongyue Wei, and Heike Bickeboller
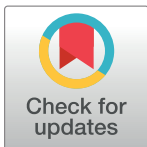
RESEARCH ARTICLE

Gene-set meta-analysis of lung cancer identifies pathway related to systemic lupus erythematosus

Albert Rosenberger[1]*, Melanie Sohns[1], Stefanie Friedrichs[1], Rayjean J. Hung[2,3], Gord Fehringer[2], John McLaughlin[4], Christopher I. Amos[5], Paul Brennan[6], Angela Risch[7], Irene Brüske[8], Neil E. Caporaso[9], Maria Teresa Landi[9], David C. Christiani[10], Yongyue Wei[10], Heike Bickeböller[1]

1 Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, Göttingen, Germany, **2** Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Canada, **3** Dalla Lana School of Public Health, University of Toronto, Toronto, Canada, **4** Public Health Ontario, Toronto, Canada, **5** Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, United States of America, **6** International Agency for Research on Cancer, Lyon, France, **7** Division of Molecular Biology, University Salzburg, Salzburg, Austria, **8** Institute of Epidemiology I, Helmholtz Center Munich, Munich, Germany, **9** Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America, **10** Harvard University School of Public Health, Boston, Massachusetts, United States of America

* arosenb@gwdg.de

Abstract

Introduction

Gene-set analysis (GSA) is an approach using the results of single-marker genome-wide association studies when investigating pathways as a whole with respect to the genetic basis of a disease.

Methods

We performed a meta-analysis of seven GSAs for lung cancer, applying the method META-GSA. Overall, the 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.

Results

META-GSA reveals the systemic lupus erythematosus KEGG pathway *hsa05322*, driven by the gene region 6p21-22, as also implicated in lung cancer (p = 0.0306). 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*).

Conclusions

We have discovered a pathway currently marked as specific to systemic lupus erythematosus 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. Given wide-stretched linkage disequilibrium to the area *APOM/BAG6/MSH5*, there is currently simply not enough information or evidence to conclude whether the potential pleiotropy of lung cancer and systemic lupus erythematosus is spurious, biological, or mediated. Further research into this pathway and gene region will be necessary.

Introduction

Since the beginning of the 20th century, lung cancer (LC) occurrence has been increasing rapidly and has become the most common cancer in males. It is the main cause of cancer-related death worldwide [1] and tobacco smoke is its major risk factor. The risk of developing LC in current smokers is 7.6 to 9.3 times higher compared to that of never smokers [2]. However, around every fourth LC case is not attributable to smoking [3]. A five-fold increased risk of developing early-onset LC in the presence of a family history of early-onset LC in any first-degree relatives has also been observed [4, 5]. This and other evidence has led to the general acceptance that a genetic component in early-onset LC development exists. However, an increased risk of developing LC has also been observed in patients with other disease, such as COPD, pneumonia, tuberculosis, or the autoimmune disorder *systemic lupus erythematosus* (SLE) [6, 7]. In the case of patients with SLE, an increased relative risk (RR) of developing LC was observed as being 1.68 (95%-CI: 1-33-2.13) [6]. In spite of multiform clinical manifestations and outcomes, it is generally accepted that genetics plays a role in SLE [8]. In light of the results of this investigation, we will discuss a shared genetic susceptibility as a possible connection between SLE and LC.

Genome-wide association studies (GWASs) have revealed that genomic variations at e.g. 5p15.33, 6p21-22 and 15q25 influence LC risk in European populations [9–16]. Further weakly associated single markers in at least 12 genes have been found given their known role within certain molecular mechanisms [17–21]. Since associated genes are elements of respective pathways, one may assume that nicotine dependency [14], inflammation [16, 22], or DNA repair [23], among others, play a role in an individual's susceptibility to developing LC.

The usual approach to identify such molecular mechanisms with GWAS is primarily to investigate single-marker-association and then allocate these markers to genes and finally the genes to pathways. Doing so, either the marginal effect of a single marker and/or the sample size needs to be large, because a low genome-wide level of significance of 1×10^{-7} or smaller is needed owing to multiple testing. Gene-set analysis (GSA) strategies were proposed as complementary approaches in the investigation of the genetic basis of a disease using GWAS results [24–26], by seeking to identify sets of genes (GS) with sufficient enrichment of marker-specific significance for an association with a phenotype.

GSA approaches provide no effect estimates of the association, but only p-values (p_{GS}). To pool the p_{GS}-values of several GSAs, it is important to take into account the concordance across studies of all single-marker-association point estimates related to every gene in a considered gene set [27]. However, one only needs to correct for multiple testing using the lower number of GSs being investigated instead of the larger number of genotyped markers. Once a GS has been found to be significantly associated, a search may be conducted for the genes that drive its significance and for the hosted markers which are concordant across studies based on their observed associations.

Here we aimed to identify pathways taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [28] as being associated with LC. KEGG provides a collection of

manually drawn pathway maps representing an up-to-date knowledge on the molecular interaction and reaction networks. This includes pathways for metabolisms (e.g. nicotinate and nicotinamide metabolism), for genetic information processing (e.g. DNA repair), for environmental information processing (e.g. Wnt signaling), for cellular processes (e.g. cell cycle), for organismal systems (e.g. circadian rhythm) and last but not least for human diseases (e.g. LC or SLE) [29]. We refrained from restricting the KEEG collection, because pathways that are potentially involved in the etiology of LC (examples are given above in brackets) are contained in every upper mentioned category.

Our subsequent goal was to determine the driving genes in the pathways identified in the first step. To this end, we combined the results of seven LC GWASs from the Transdisciplinary Research in Cancer of the Lung / International Lung Cancer Consortium (TRICL / ILCCO) in a meta-analysis.

Materials and methods

Description of studies

The meta-analysis was based on summary data from seven previously reported LC GWASs form TRICL / ILCCO (Fig 1). We included 11,365 LC cases and 22,505 controls of European descent in the analysis. An overview as well as study name abbreviations are given in Table 1. Details and references are provided Supplement S1 File.

Strategy and methods

In the original GWASs, a log-additive mode of inheritance was fitted for each marker, adjusting for age, sex, smoking status, study center (if applicable), and the first three principal components to account for hidden genomic structure. The results of marker-by-marker association testing were used as input information for the GSAs.

For this meta-analysis, we set up a two-phase seamless design consisting of a screening phase and a replication phase. In the screening phase, the results of MDACC, TORONTO, GLC, and CE were combined, because GSA of these studies was performed for 234 KEGG pathways previously [30, 31]. In the replication phase, the results of the remaining studies NCI, deCODE, and HARVARD were combined to investigate only those pathways whose findings in the screening phase proved promising. If necessary, GSA was performed using the program ALIGATOR [32]. The method META-GSA [27] was performed to pool GSA results (p-values $p_{GS,s}$) at each stage. The aim of META-GSA is to increase statistical evidence by pooling the p-values $p_{GS,s}$ of GSAs, taking also into account the concordance of the signs of single-marker-association point estimates and related p-values of all markers ($p_{m,s}$) assigned to genes contained in the GS [27]. The core element of this approach is a directed p-value (PDR), combining significance and direction of single markers and LD to other markers. Necessary estimates of LD were based on the genotype data of GLC, with imputation of missing markers based on the 1000-Genome Project [33], the 1000-GenomePilot 1-Panel or the HapMap3-Panel as available using the SNAP online tool [34].

The SNP-to-gene annotation (StG) for humans of the ENSEMBL database [35] was used. Markers with LD of at least $r^2 \geq 0.8$ to any marker inside a gene were additionally assigned to that gene [36]. All genes were then annotated to 234 gene sets from the KEGG database (gene-to-pathway annotation (GtP)).

Both phases can be considered as the first and the second stage of a seamless, adaptive study with interim selection of gene sets ("drop-loser design" [37]). The investigation of every KEGG pathway with a pooled $p_{scr.} < \beta_1 = 1/234$ in the screening phase was stopped early for futility. The significance, combining screening and replication phase, was assessed according
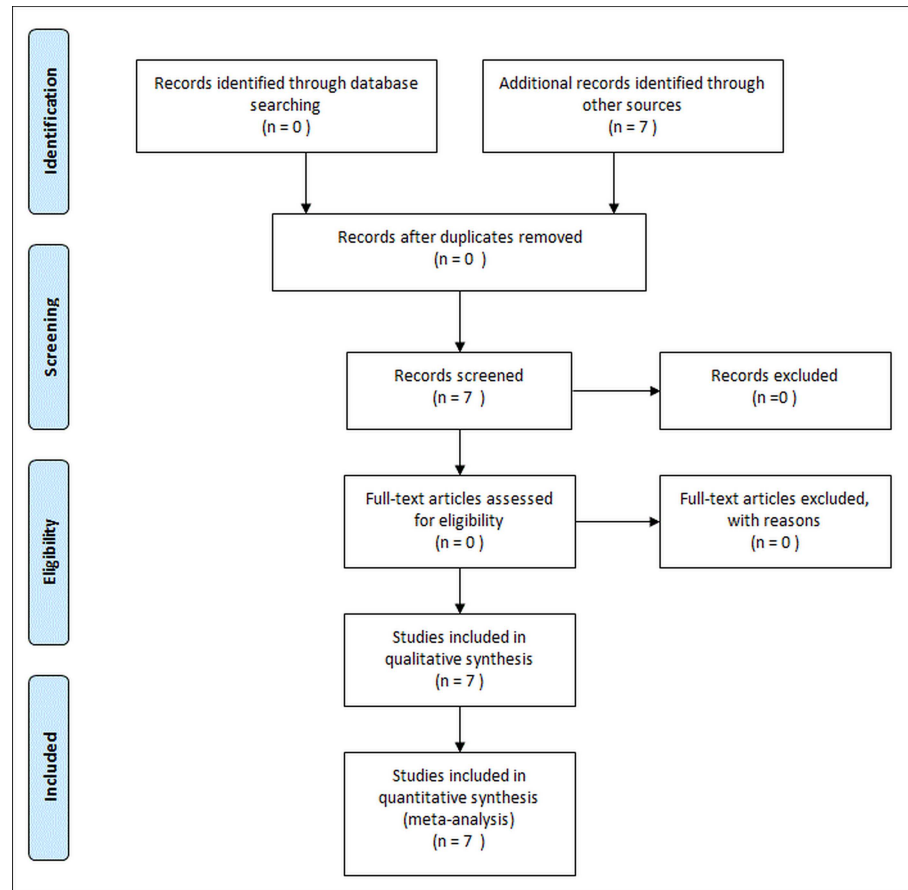
Fig 1. Study selection flow cart.

to the "method based on the sum of p-values" (MSP) [37, 38]. The p-value was then calculated by the equation $p_{GS} = \beta_1 (p_{scr.} + p_{rep.}) - 0.5\beta_1^2$. This p_{GS} needs to be corrected for multiple testing by taking into account the total number of 234 pathways. Due to pathway overlap we estimated the number of independent tests t_{eff} according to the lowest slope method (LSM) [39] considering all $p_{scr.}$-values of the screening phase. Applying a Bonferroni-like correction then yields the final p-value $p_{GS,corr.} = min(1, t_{eff} \cdot p_{GS})$. Furthermore, META-GSA was also applied to all seven studies and all pathways surviving the screening phase to take into account the concordance of single-marker-association point estimates across <u>all</u> considered studies at the same time.

The next step was to identify the main genes driving the significance of gene sets (denoted as p_{GS}–driving genes). Thus we contrasted the mean of PDRs across studies for each gene ($\overline{PDR}_g$ as a measure of concordance) with pooled p-values regarding the gene-level statistics (p_{gene} as measure of significance, calculated according to Fisher's χ^2-method). To judge these findings adequately, we also calculated $\overline{PDR}_g$ for the known LC-related genes *CLTM1L*, *TERT*, *CHRNB4*, *CHRNA3*, *CHRNA5*, *MSH5*, *BAG6*, *RAD52* and *CDKN2B*. Within these genes we looked markers with a large mean of PDRs across studies ($\overline{PDR}_m$).

Finally, we performed a sub-group meta-analysis for the one identified KEGG pathway according to histological subtype (AdenoLC, SqCLC, SCLC and LCLC), sex, age (older or younger than 50 years), and smoking behavior (current, former, ever and never smokers).

Table 1. Characteristics of lung cancer GWASs of the International Lung Cancer Consortium (ILCCO).

Study	Cases	Controls	Location	Study design	Illumina genotyping platform	Number of SNPs
Scanning phase						
MDACC[a]	1 150	1 134	Texas, USA	Hospital-based case–control	317K	312 829
TORONTO[b]	331	499	Toronto, CA	Hospital-based case–control	317K	314 285
CE (IARC[c])	1 854	2 453	Romania, Hungary, Slovakia, Poland, Russia, Czech Republic	Multicenter hospital-based case–control	317K, 370Duo	
GLC[d]	487	480	Germany	Population-based case–control (<50 years)	HumanHap550K	503 381
Replication phase						
DeCODE Genetics	830	11 228	Iceland	Population-based case–control	317K, 370Duo	290 386
HARVARD	984	970	Massachusetts, USA	Hospital-based case–control	610Quad	543 697
NCI GWAS						506 062
EAGLE[e]	1 920	1 979	Italy	Population-based case–control	HumanHap550v3_B, 610Quad	
ATBC[f]	1 732	1 271	Finland	Cohort	HumanHap550K, HumanHap610	
PLCO[g]	1 380	1 817	10 US Centers	Cohort-Cancer Prevention Trial	317K / 240S, HumanHap550v3_B, HumanHap610	
CPS-II[h]	697	674	all US states	Cohort	HumanHap550K, 610Quad	
Overall	**11 365**	**22 505**				

[a] MD Anderson Cancer Center.

[b] Toronto study by Lunenfeld-Tanenbaum Research Institute.

[c] Central Europe Study of the International Agency for Research on Cancer.

[d] German Lung Cancer Study.

[e] Environment And Genetics in Lung cancer Etiology study.

[f] Alpha-Tocopherol, Beta-Carotene Cancer Prevention study.

[g] Prostate, Lung, Colon, Ovary screening trial.

[h] Cancer Prevention Study II nutrition cohort.

doi:10.1371/journal.pone.0173339.t001

During this investigation the region 6p21-22 became of interest. Respective correlation of marker genotypes and gene expression (eQTL) was previously measured in non-neoplastic pulmonary parenchymal samples taken some distance from the primary tumor in LC patients [40]. We used the estimated correlation between every SNP located between 31.6MB and 32.2 MB (all within 6p21-22) and the expression of the genes *APOM*, *BAG6*, *MSH5* (reported as relevant in LC), *C2*, *C4B*, *SKIV2L*, *STK19* (closely located to genes driving the significance in this META-GSA application) and *TNXB* (reported as relevant for SLE), in total 5,572 estimated correlations. Estimating t_{eff} = 5309 independent tests (by LSM) yields a global threshold for significance of 1×10^{-7}.

Results

Association of pathways: Screening and replication phase

Only three of the 234 pathways investigated revealed a p-value lower than the futility threshold and were selected for the replication phase: *hsa05322*: *systemic lupus erythematosus (SLE)*, *hsa00790*: *folate biosynthesis* and *hsa04940*: *type I diabetes mellitus* (Table 2). Only for the SLE

Table 2. Significant results of META-GSA.

KEGG pathways	number of genes	screening		replication	MSP combination		all
		4 studies		3 studies			7 studies
	n_{genes}	$p_{scr.}$	$p_{scr.corr.}$ $^\$$	$p_{rep.}$	p_{GS}	$p_{GS,corr.}$ $^\$$	p_{GS}
hsa05322 (SLE)	128	***0.0003	*0.0457	0.0857	***0.0004	0.0615	*0.0306
hsa00790 (folate bio.)	13	***0.0003	0.0543	0.9122	***0.0046	0.6672	0.3154
hsa04940 (T1DM)	42	***0.0011	0.1940	0.4890	***0.0024	0.3570	0.3952
231 other gene sets		>0.0043	futility	stopping			

SLE—systemic lupus erythematosus; *folate bio* folate biosynthesis; *T1DM*—type I diabetes mellitus, *MSP*—combined p-values according to the method based on the sum of p-values (adaptive designed approach for early futility stopping); $p_{scr.}$—p-value of the screening phase; $p_{scr.corr.}$—p-value of the screening phase corrected for multiple testing; $p_{rep.}$—p-value of the replication phase, p_{GS}—p-value of the gene set (combining p_{scr} and p_{rep}); $p_{GS,corr.}$—p-value of the gene set corrected for multiple testing; effective number of independent gene sets according the lowest slope method (LSM).

$^\$$: t_{eff} = 171.5.

* P ≤ 0.05.

** P ≤ 0.01.

*** P ≤ 0.001.

doi:10.1371/journal.pone.0173339.t002

pathway we were able to achieve a low p-value when combining screening and replication phase and correcting for multiple testing ($p_{GS,corr}$ = 0.0615). Combining all seven studies in a single META-GSA, in order to take the concordance of single-marker-association point estimates of all studies into account adequately, yielded a p_{GS}-value of 0.0306 for this SLE pathway. This indicates sufficient enrichment and satisfactory concordance of marker-specific significance for an association with LC.

Genes driving significance

Four genes of the SLE pathway (*HIST1-H4L,-1BN, -H2AK, -H4K*) and their close neighbor *HIST1H2BN* strike out by concordance of marker-specific association ($abs(\overline{PDR}_g) \sim 0.8$) across studies and a gene-level p_{gene} _value lower than_ 0.01 (Table 3). All five genes belong to the histone cluster 1 and are closely located within 41 kb of each other on 6p22.1. Weaker concordance was observed for further two less significant genes (p_{gene} _value_ < 0.05): C4A ($\overline{PDR}_g$ = -0.41) and *C2* ($\overline{PDR}_g$ = 0.33).

Markers driving significance

The markers rs13194781, rs1270942 and rs389884 are those with the largest $\overline{PDR}_m$-values (all >0.7) and the strongest associations with LC (in terms of OR). For rs13194781, which is located within *HIST1H2BN* (ENSEMBL definition), an OR of 1.23 (p = 0.0032) was estimated. The markers rs1270942 and rs389884 are perfect proxies for each other according to the 1000-Genome Pilot 1-panel [33]. They are closely located upstream of C2 and downstream of C4A, respectively. There is no LD with the first marker rs13194781 (Table 4).

Subgroup meta-analysis

We revealed more evidence for an association of the SLE pathway with AdenoLC (p_{GS} = 0.0030) than for any other histotype. We also found the association to be significant in women (p_{GS} = 0.0112) but not in men (p_{GS} = 0.1453) and in older cases (p_{GS} = 0.0002) but not in younger (p_{GS} = 0.0588). No significant association was observed when stratifying according to smoking behavior (Table 5). Significance within the considered subgroups is driven by same p_{GS}-driving genes of

Table 3. Significance and concordance of selected genes of interest.

gene	location	number of studies with	concordance	significance
		$p_{gene,study}$ < 5%	$\overline{PDR}_g$	p_{gene}
significant genes belonging to the significant gene set hsa05322 (SLE)				
HIST1H4K	6p22.1	2	-0.84	0.0056
HIST1H2BN	6p22.1	2	-0.80	0.0091
HIST1H2AK	6p22.1	2	-0.80	0.0091
HIST1H1B	6p22.1	2	+0.75	0.0093
HIST1H2AL	6p22.1	2	+0.75	0.0093
C2	6p21.3	2	+0.33	0.0109
C4A	6p21.3	1	-0.41	0.0319
genes known to be associated with LC (for comparison only)				
CLPTM1L	5q15.33	4	-0.53	< .0001
TERT	5q15.33	4	+0.49	0.0013
CHRNB4	15q24	3	-0.63	< .0001
CHRNA3	15q24	4	-0.58	< .0001
CHRNA5	15q24	3	-0.45	0.0009
MSH5	6p21.3	3	+0.67	< .0001
BAG6	6p21.3	--	+0.39	0.1425
RAD52	12p13.33	1	+0.23	0.3143
CDKN2B	9p21.3	--	-0.13	0.6729

$p_{gene,study}$ is the study specific p-value for gene; $\overline{PDR}_g$ is the mean of study specific PDRs for a gene (95% random interval derived from all 16.000 assigned genes: [±0.306]); pooled p_{gene}—$p_{gene,study}$-values combined by Fisher's inverse χ^2-method.

doi:10.1371/journal.pone.0173339.t003

the region 6p22.1–22.2 as in the total sample (C2 and the genes of the histone 1 cluster). Also, most of the more moderate concordant genes that drive significance of *hsa05322* in at least one of the considered subgroups are histone-coding genes.

SNP × eQTL correlation

Both aforementioned SNPs belonging to *C2/C4A*, rs1270942 and rs389884, are significant correlated with the expression of the gene *APOM* (p<10⁻¹³), which is located about 500 kb away (Fig 2). However, the expression pattern is this region is puzzling, since other markers within *C2* (rs537160, rs622871, rs630379) are also correlated with the gene expression in non-neoplastic samples of LC patients of the neighboring gene *C4B* (not part of the investigated KEGG pathway, although related to SLE). It is also remarkably that the correlation of SNPs belonging to *C2/C4A* with the expression of *C2* is less significant (p ~10⁻³) than with the expression of SKIV2L (p ~10⁻⁵), which is not related to SLE.

Discussion

We could demonstrate an accumulation of genomic association with LC in the KEGG pathway *hsa05322*, which comprises genes related to SLE. This suggests some cross-phenotype (CP) association with LC and SLE. The significance was higher in the subgroup of AdenoLC patients than within other histological subtypes and in women compared to men. This fits our expectations in view of women, who predominantly develop AdenoLC, are more often affected with SLE than men [41], who predominantly develop smoking-related SqCLC [1, 42].

All p_{GS}–driving genes identified in this meta-analysis are located within or next to the major histocompatibility complex (MHC) on chromosome 6p21-22 (Fig 2), albeit in two

Table 4. Markers with <0.5 in genes of interest on 6p21-22.

SNP	allocated to	Position	MAF	r² to		D' to		$\overline{PDR}_m$	LC		SqCLC	
				(A)	(B)	(A)	(B)		OR	p-value	OR	p-value
rs200991	HIST1++	27847716	0.12[a]	0.646[b]		1[a]		0.598	1.14	0.0021	1.16	3.1×10⁻⁵
rs13194781 (A)	HIST1++	27847861	0,08	1		1		0.719	1.23	0.0032	1.22	9.7×10⁻⁶
rs9262143	MDC1	30685004	0.16[§]					0.769	1.25	0.0027	1.25	1.3×10⁻⁷
rs3094127	MDC1	30729670	0.18					0.664	0.84	0.0029	1.10	4.0×10⁻²
rs3128982	HCP5	31449414	0.30					0.578	1.07	0.0032	1.12	1.1×10⁻³
rs3117582	BAG6	31652743	0.09[b]		0.881[b]		1[b]	0.485	1.27	0.0049	1.30	4.5×10⁻¹⁰
rs3131379	MSH5	31753256	0.09[b]		0.881[b]		1[b]	0.461	1.20	0.0074	1.28	3.8×10⁻⁷
rs652888	C2	31883457	0.17		0.336[b]		1[b]	0.538	1.14	0.0013	1.18	1.3×10⁻⁴
rs535586	C2	31892560	0.35		0.131[b]		1[b]	0.606	1.09	0.0001	1.11	1.2×10⁻³
rs659445	C2	31896527	0.35		0.131[b]		1[b]	0.711	1.09	3.7×10⁻⁶	1.10	3.1×10⁻³
rs1270942 (B)	C2	31951083	0.09[b]		1		1[b]	0.728	1.27	0.0090	1.29	5.8×10⁻⁶
rs438999	C2	31960529	0.06		0.005[b]		1[b]	-.517	0.91	0.0027	0.85	1.0×10⁻²
rs454212	C4A	31966595	0.08					-.556	0.95	0.0034	0.84	1.7×10⁻²
rs389884	C4A	31973120	0.09[b]		1[$]		1[b]	0.724	1.27	0.0080	1.28	7.2×10⁻⁶

Odds ratios (OR), corresponding p-values from a random effects meta-analysis model; single study ORs were adjusted for age, sex, smoking and genetic background; r² and D' were calculated according to the HapMap3-panel.

([a]) or the 1000 Genome Pilot 1-panel.

([b]) using SNAP Version 2.2; HIST1++ denotes the gene cluster HIST1-H4L/H2BN/H2AK/H2BN/H4K; LC—lung cancer (all histological subtypes), SqCLC – squamous-cell lung cancer; markers with largest $\overline{PDR}_m$ with genes driving the significance of the SLE gene set (HIST1++, C2 and C4A) are printed in bold. Position of SNPs is given according to NCBI Build 37. MAF . . . minor allele frequencies in controlls.

separate areas, about 3000 kb apart. The first area comprises the genes of histone cluster I: *HIST1-H4L, -1BN, -2BN, -H2AK, -H4K* (the strongest associated marker is rs13194781; OR = 1.23, p = 0.0032). It is well known that a variety of histone related modifications are either related to cancer or to SLE, or to both [8, 43]. They play a role e.g. in DNA repair, cell cycle or gene expression [8, 44], which by themselves are associated to LC or SLE, respectively [23, 45]. Interestingly enough, we detected associations to LC of the DNA signature of histone coding genes, rather than with respect to some kind of epigenetic outcome.

The second area comprises the genes *C2, C4A,* and *C4B* (the strongest associated markers are rs1270942 and rs389884; OR = 1.27, p = 0.009). It is well established, that reduced gene expression of C2 and C4A can predispose to SLE [46]. This two genes, and perhaps also *C4B*, are involved in the clearance of apoptotic bodies [8]. This is in turn crucially important for controlling inflammation, which plays a role in the development of LC [3].

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 [47]. Hence any observed CP association will probably tag plenty of genes. An association of the gene area APOM/*BAG6/MSH5* in the MHC region with LC has previously been reported, which is strongest for SqCLC and AdenoLC [9, 13]. The strongest associations with SqCLC in this area was previously reported for the markers rs3117582 (located within *BAG6 and APOM;* OR = 1.3, p = 4.5×10⁻¹⁰), which was found associated also with SLE (OR = 2.2, p = 4.2×10⁻²¹) [48]. This marker is about 220 kB apart but in strong LD with the newly identified markers rs1270942 and rs389884 (located close to C2; Table 4 and Fig 2). More important, a highly significant correlation between markers of the area *C2/C4A/C4B* with the expression of the gene *APOM* in non-neoplastic samples taken from LC patients was also recently reported [40] (Fig 2). APOM is involved in lipid transport and is linked with high-density lipoprotein

Table 5. Subgroup analysis for *hsa05322*: histological subtypes, sex, age, smoking.

hsa05322: SLE	META-GSA	Gene	Location	concordance	significance
	p_{GS}			$\overline{PDR_g}$	p_{gene}
AdenoLC	**0.0030**	HIST2-1q21.2	1q21.2	-0.6	0.1666
SqCLC	**0.0376**	H2AFV	7p13	0.5	0.7209
SCLC	0.0626	C1QA	1p36.12	-0.5	0.7101
		HIST2-1q21.2	1q21.2	-0.5	0.0577
		ELANE	19p13.3	0.5	0.4864
		HIST1-6p22.2	6p22.2	0.5	0.2177
		HIST1-6p22.2	6p22.2	0.5	0.2177
LCLC	0.2056	--			
male	0.1453	HIST1H3C	6p22.2	-0.5	0.3726
female	**0.0112**	HIST1H2AL	6p22.1	0.5	0.1229
old (>50)	**0.0002**	HIST1-6p22.1a	6p22.1	-0.7	0.0054
		HIST1-6p22.1b	6p22.1	0.5	0.1578
		C2	6p21.3	0.5	0.0013
		H2AFV	7p1	0.6	0.4005
young ($\leq$50)	0.0588	--			
current smokers	0.3563	HIST1-6p22.1a	6p22.1	0.4	0.1720
		H3F3C	12p11.21	0.4	0.5821
		HIST3H3	1q42	0.4	0.6468
		HIST1-6p22.1b	6p22.1	0.4	0.2028
		HIST1-6p22.1c	6p22.1	0.4	0.3375
former smokers	0.4691	--			
ever smokers	0.5132	HIST1-6p22.1a	6p22.1	0.5	0.0462
		HIST1-6p22.1c	6p22.1	0.5	0.1587
never smokers	0.5429	FCGR3A	1q23	-0.5	0.2300
		CTSG	14q11.2	-0.5	0.3403

Listed are genes, respectively regions containing genes with $abs(\overline{PDR_g}) \geq 0.5$.

HIST2-1q21.2: HIST2H2AA3 / HIST2H2AA4 / HIST2H3C / HIST2H4B.

HIST3-1q42: HIST3H2A / HIST3H2BB / HIST3H3.

HIST1-6p22.1a: HIST1H4K / HIST1H2AK / HIST1H2AL / HIST1H2BM / HIST1H2BN / HIST1H3I / HIST1H4L / HIST1H3J / HIST1H4J (27.800K).

HIST1-6p22.1b: HIST1H2AG / HIST1H2BK (27.150 K).

HIST1-6p22.1c: HIST1H2BI / HIST1H3G / HIST1H4H (26.280 K).

HIST1-6p22.2: HIST1H3E / HIST1H2AE / HIST1H2BG / HIST1H4E (26.200 K).

The numbers in brackets are the approximate locations according to dbGENE.

doi:10.1371/journal.pone.0173339.t005

cholesterol in the pathogenesis of emphysema, which is on the other hand considered as associated with LC [49, 50]. But other explanations of the observed associations have been given, too; for instant a connection to embryonic lethality with defects in the development of the lung (related to the function of *BAG6*) or deficits in mismatch excision repair (related to the function of *MSH5*) [13]. Moreover, the association of MSH5 with SLE was reported as not shared with other autoimmune/inflammatory diseases [51].

Apart from all this, some remarks about the applied method need to be made. The whole approach is an intensive investigation of p-values, which—in the context of this project—are indicators of evidence for or against the rejection of a null-hypothesis of no genetic association. We used the program ALIGATOR to perform GSA, which circumvents bias due to uneven counts of markers per gene as well as genes per gene set [32]. Choosing another algorithm
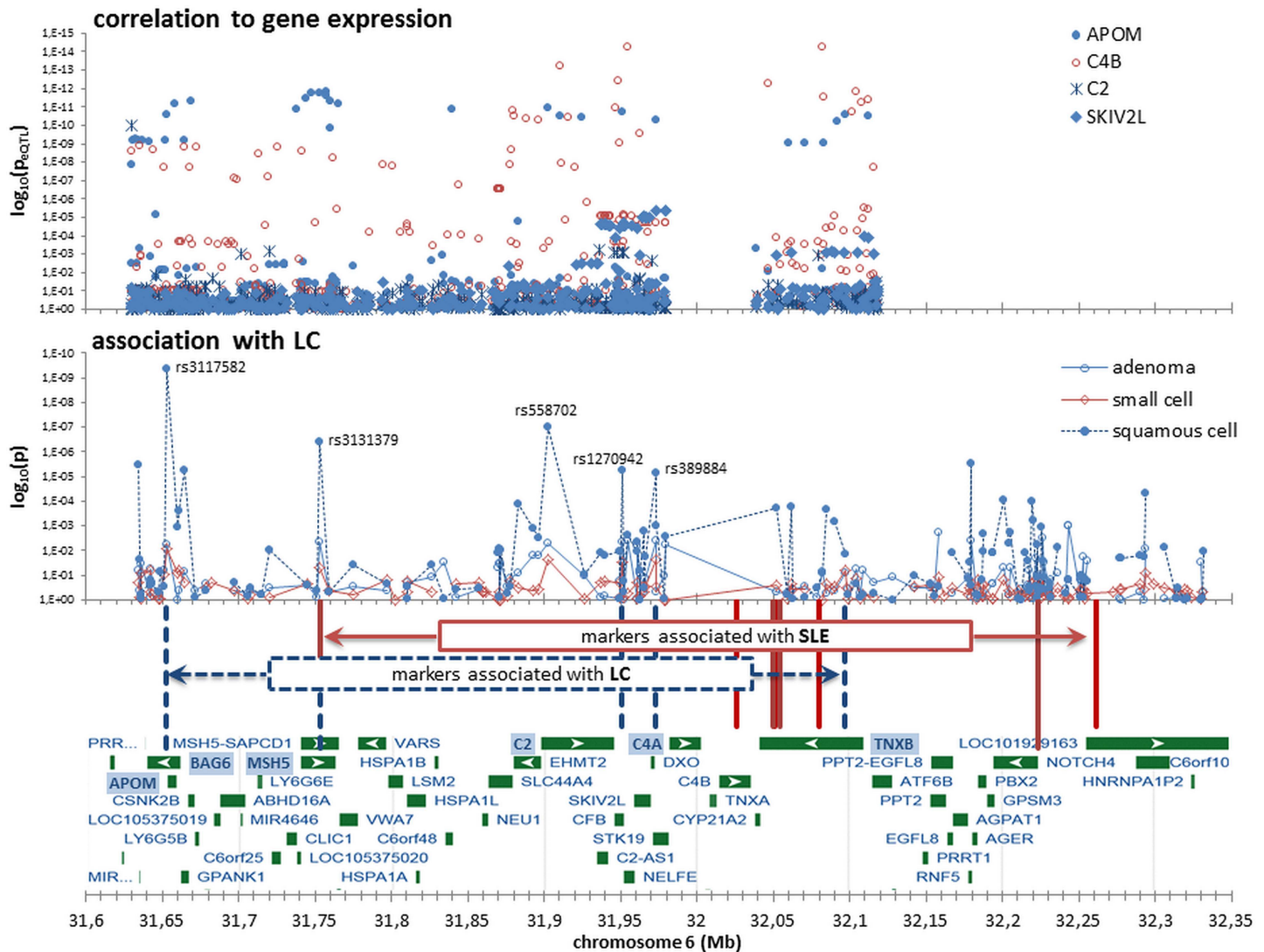
Fig 2. Association and correlation with gene expression in the chromosome 6p21-22 region. *LC*—lung cancer, *SLE*—systemic lupus erythematosus; correlation to gene expression: pooled p-values as reported by Nguyen et al., 2014 [40]; association with LC: pooled p-values as reported by Timofeeva et al. 2012 [13].

doi:10.1371/journal.pone.0173339.g002

would probably lead to different results [31]. In addition, 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 we can present significance of our findings but are unable to estimate the part of LC risk that can be attributed to the identified genes or gene sets.

Conclusion

We were able to identify CP risk factors by first pooling results of gene set analyses and looking afterwards for those genes driving the significance of discovered gene sets. In doing so, we have discovered a pathway that is currently marked as specific to SLE as being significantly implicated in LC. The gene region 6p21-22 in this pathway appears to be more extensively associated with lung cancer than previously assumed. Given wide-stretched linkage

disequilibrium to the area *APOM/BAG6/MSH5*, there is currently simply not enough information or evidence to conclude whether the potential pleiotropy of LC and SLE is spurious, biological, or mediated. Further research into this pathway and gene region will be necessary.

Supporting information

S1 File. Detailed study description.
(DOCX)

S2 File. PRISMA Checklist.
(DOCX)

S3 File. Meta-analysis on Genetic Association Studies Checklist | PLOS ONE.
(DOCX)

Acknowledgments

Author Contributions

Conceptualization: A. Rosenberger.

Formal analysis: A. Rosenberger SF MS GF.

Funding acquisition: RH JM CA PB A. Risch IB NC ML DC HB.

Investigation: RH JM CA PB A. Risch IB NC ML DC HB.

Methodology: A. Rosenberger.

Project administration: CA RH HB.

Resources: RH JM CA PB A. Risch IB NC ML DC YW HB.

Writing – original draft: A. Rosenberger HB RH NC.

Writing – review & editing: A. Rosenberger.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011; 61(2):69–90. doi: 10.3322/caac.20107 PMID: 21296855

2. Lee PN, Forey BA, Coombs KJ. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. BMC Cancer. 2012; 12:385. PubMed Central PMCID: PMC3505152. doi: 10.1186/1471-2407-12-385 PMID: 22943444

3. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. Nature reviews Cancer. 2007; 7(10):778–90. doi: 10.1038/nrc2190 PMID: 17882278

4. Cassidy A, Myles JP, Duffy SW, Liloglou T, Field JK. Family history and risk of lung cancer: age-at-diagnosis in cases and first-degree relatives. British journal of cancer. 2006; 95(9):1288–90. PubMed Central PMCID: PMC2360569. doi: 10.1038/sj.bjc.6603386 PMID: 17003779

5. Kreuzer M, Kreienbrock L, Gerken M, Heinrich J, Bruske-Hohlfeld I, Muller KM, et al. Risk factors for lung cancer in young adults. Am J Epidemiol. 1998; 147(11):1028–37. PMID: 9620046

6.   Ni J, Qiu LJ, Hu LF, Cen H, Zhang M, Wen PF, et al. Lung, liver, prostate, bladder malignancies risk in systemic lupus erythematosus: evidence from a meta-analysis. Lupus. 2014; 23(3):284–92. doi: 10.1177/0961203313520060 PMID: 24429300

7.   Brenner DR, Boffetta P, Duell EJ, Bickeboller H, Rosenberger A, McCormack V, et al. Previous lung diseases and lung cancer risk: a pooled analysis from the International Lung Cancer Consortium. Am J Epidemiol. 2012; 176(7):573–85. PubMed Central PMCID: PMCPMC3530374. doi: 10.1093/aje/kws151 PMID: 22986146

8.   Costa-Reis P, Sullivan KE. Genetics and epigenetics of systemic lupus erythematosus. Current rheumatology reports. 2013; 15(9):369. doi: 10.1007/s11926-013-0369-4 PMID: 23943494

9.   Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. NatGenet. 2008; 40(12):1407–9.

10. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008; 452 (7187):633–7. doi: 10.1038/nature06885 PMID: 18385738

11. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008; 40(5):616–22. doi: 10.1038/ng.109 PMID: 18385676

12. Truong T, Hung RJ, Amos CI, Wu X, Bickeboller H, Rosenberger A, et al. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. J NatlCancer Inst. 2010; 102(13):959–71.

13. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeboller H, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Human molecular genetics. 2012; 21(22):4980–95. Epub 2012/08/18. PubMed Central PMCID: PMCPMC3607485. doi: 10.1093/hmg/dds334 PMID: 22899653

14. Brennan P, Hainaut P, Boffetta P. Genetics of lung-cancer susceptibility. The lancet oncology. 2011; 12 (4):399–408. Epub 2010/10/19. doi: 10.1016/S1470-2045(10)70126-1 PMID: 20951091

15. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014; 46(7):736–41. PubMed Central PMCID: PMC4074058. doi: 10.1038/ng.3002 PMID: 24880342

16. Fehringer G, Liu G, Pintilie M, Sykes J, Cheng D, Liu N, et al. Association of the 15q25 and 5p15 lung cancer susceptibility regions with gene expression in lung tumor tissue. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012; 21(7):1097–104. Epub 2012/04/28.

17. Timofeeva M, Kropp S, Sauter W, Beckmann L, Rosenberger A, Illig T, et al. Genetic polymorphisms of MPO, GSTT1, GSTM1, GSTP1, EPHX1 and NQO1 as risk factors of early-onset lung cancer. IntJ Cancer. 2010; 127(7):1547–61.

18. Leng S, Picchi MA, Liu Y, Thomas CL, Willis DG, Bernauer AM, et al. 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. PubMed Central PMCID: PMC3643420. doi: 10.1093/carcin/bgt024 PMID: 23354305

19. Hung RJ, Christiani DC, Risch A, Popanda O, Haugen A, Zienolddiny S, et al. International Lung Cancer Consortium: pooled analysis of sequence variants in DNA repair and cell cycle pathways. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008; 17(11):3081–9. PubMed Central PMCID: PMC2756735.

20. Manuguerra M, Saletta F, Karagas MR, Berwick M, Veglia F, Vineis P, et al. XRCC3 and XPD/ERCC2 single nucleotide polymorphisms and the risk of cancer: a HuGE review. Am J Epidemiol. 2006; 164 (4):297–302. doi: 10.1093/aje/kwj189 PMID: 16707649

21. Brenner DR, Brennan P, Boffetta P, Amos CI, Spitz MR, Chen C, et al. Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls. Hum Genet. 2013; 132(5):579–89. Epub 2013/02/02. PubMed Central PMCID: PMCPMC3628758. doi: 10.1007/s00439-013-1270-y PMID: 23370545

22. Gomes M, Teixeira AL, Coelho A, Araujo A, Medeiros R. The role of inflammation in lung cancer. Advances in experimental medicine and biology. 2014; 816:1–23. doi: 10.1007/978-3-0348-0837-8_1 PMID: 24818717

23. Kiyohara C, Takayama K, Nakanishi Y. Lung cancer risk and genetic polymorphisms in DNA repair pathways: a meta-analysis. Journal of nucleic acids. 2010; 2010:701760. PubMed Central PMCID: PMC2958337. doi: 10.4061/2010/701760 PMID: 20981350

24. Sohns M, Rosenberger A, Bickeboller H. Integration of a priori gene set information into genome-wide association studies. BMC proceedings. 2009; 3 Suppl 7:S95. PubMed Central PMCID: PMC2795999.

25. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. European journal of human genetics: EJHG. 2010; 18(1):111–7. doi: 10.1038/ejhg.2009.115 PMID: 19584899

26. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M. Genome-wide gene and pathway analysis. European journal of human genetics: EJHG. 2010; 18(9):1045–53. doi: 10.1038/ejhg.2010.62 PMID: 20442747

27. Rosenberger A, Friedrichs S, Amos CI, Brennan P, Fehringer G, Heinrich J, et al. META-GSA: Combining Findings from Gene-Set Analyses across Several Genome-Wide Association Studies. PLoS One. 2015; 10(10):e0140179. PubMed Central PMCID: PMCPMC4621033. doi: 10.1371/journal.pone. 0140179 PMID: 26501144

28. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012; 40(Database issue):D109–D14. doi: 10.1093/nar/ gkr988 PMID: 22080510

29. KEEG Pathway Database [Internet]. Kanehisa Laboratories. 1995–2017. Available from: http://www. genome.jp/kegg/pathway.html.

30. Tintle N, Lantieri F, Lebrec J, Sohns M, Ballard D, Bickeboller H. Inclusion of a priori information in genome-wide association analysis. Genetic epidemiology. 2009; 33 Suppl 1:S74–80. PubMed Central PMCID: PMC2922922.

31. Fehringer G, Liu G, Briollais L, Brennan P, Amos CI, Spitz MR, et al. Comparison of pathway analysis approaches using lung cancer GWAS data sets. PLoS One. 2012; 7(2):e31816. Epub 2012/03/01. PubMed Central PMCID: PMCPMC3283683. doi: 10.1371/journal.pone.0031816 PMID: 22363742

32. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet. 2009; 85 (1):13–24. doi: 10.1016/j.ajhg.2009.05.011 PMID: 19539887

33. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. PubMed Central PMCID: PMC3498066. doi: 10.1038/nature11632 PMID: 23128226

34. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008; 24(24):2938–9. PubMed Central PMCID: PMC2720775. doi: 10.1093/bioinformatics/btn564 PMID: 18974171

35. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2014.

36. Malzahn D, Friedrichs S, Bickeböller H. Comparing strategies for combined testing of rare and common variants in whole sequence and genome-wide genotype data. BMC proceedings. 2015:accepted for publication.

37. Chang M. Adaptive design theory and implementation using SAS and R: Boca Raton: Chapman & Hall/CRC; 2008.

38. Chang M. Adaptive design method based on sum of p-values. Statistics in medicine. 2007; 26 (14):2772–84. doi: 10.1002/sim.2755 PMID: 17133651

39. Hsueh HM, Chen JJ, Kodell RL. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. Journal of biopharmaceutical statistics. 2003; 13(4):675–89. doi: 10.1081/ BIP-120024202 PMID: 14584715

40. Nguyen JD, Lamontagne M, Couture C, Conti M, Pare PD, Sin DD, et al. Susceptibility loci for lung cancer are associated with mRNA levels of nearby genes in the lung. Carcinogenesis. 2014; 35(12):2653–9. PubMed Central PMCID: PMC4247514. doi: 10.1093/carcin/bgu184 PMID: 25187487

41. Brinks R, Fischer-Betz R, Sander O, Richter JG, Chehab G, Schneider M. Age-specific prevalence of diagnosed systemic lupus erythematosus in Germany 2002 and projection to 2030. Lupus. 2014; 23 (13):1407–11. doi: 10.1177/0961203314540352 PMID: 24928831

42. Devesa SS, Bray F, Vizcaino AP, Parkin DM. International lung cancer trends by histologic type: male: female differences diminishing and adenocarcinoma rates rising. International journal of cancer Journal international du cancer. 2005; 117(2):294–9. doi: 10.1002/ijc.21183 PMID: 15900604

43. Chervona Y, Costa M. Histone modifications and cancer: biomarkers of prognosis? American journal of cancer research. 2012; 2(5):589–97. PubMed Central PMCID: PMC3433108. PMID: 22957310

44. House NC, Koch MR, Freudenreich CH. Chromatin modifications and DNA repair: beyond double-strand breaks. Frontiers in genetics. 2014; 5:296. PubMed Central PMCID: PMC4155812. doi: 10. 3389/fgene.2014.00296 PMID: 25250043

45. Kazma R, Babron MC, Gaborieau V, Genin E, Brennan P, Hung RJ, et al. Lung cancer and DNA repair genes: multilevel association analysis from the International Lung Cancer Consortium. Carcinogenesis.

2012; 33(5):1059–64. Epub 2012/03/03. PubMed Central PMCID: PMCPMC3334518. doi: 10.1093/carcin/bgs116 PMID: 22382497

46. Leffler J, Bengtsson AA, Blom AM. The complement system in systemic lupus erythematosus: an update. Annals of the rheumatic diseases. 2014; 73(9):1601–6. doi: 10.1136/annrheumdis-2014-205287 PMID: 24845390

47. Ahmad T, Neville M, Marshall SE, Armuzzi A, Mulcahy-Hawes K, Crawshaw J, et al. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. Human molecular genetics. 2003; 12(6):647–56. PMID: 12620970

48. Alonso MD, Martinez-Vazquez F, Riancho-Zarrabeitia L, Diaz de Teran T, Miranda-Filloy JA, Blanco R, et al. Sex differences in patients with systemic lupus erythematosus from Northwest Spain. Rheumatology international. 2014; 34(1):11–24. doi: 10.1007/s00296-013-2798-9 PMID: 23812032

49. Burkart KM, Manichaikul A, Wilk JB, Ahmed FS, Burke GL, Enright P, et al. APOM and high-density lipoprotein cholesterol are associated with lung function and per cent emphysema. The European respiratory journal. 2014; 43(4):1003–17. PubMed Central PMCID: PMC4041087. doi: 10.1183/09031936.00147612 PMID: 23900982

50. Brenner DR, Hung RJ, Tsao MS, Shepherd FA, Johnston MR, Narod S, et al. Lung cancer risk in never-smokers: a population-based case-control study of epidemiologic risk factors. BMC Cancer. 2010; 10:285. doi: 10.1186/1471-2407-10-285 PMID: 20546590

51. Fernando MM, Freudenberg J, Lee A, Morris DL, Boteva L, Rhodes B, et al. Transancestral mapping of the MHC region in systemic lupus erythematosus identifies new independent and interacting loci at MSH5, HLA-DPB1 and HLA-G. Annals of the rheumatic diseases. 2012; 71(5):777–84. PubMed Central PMCID: PMC3329227. doi: 10.1136/annrheumdis-2011-200808 PMID: 22233601