Identifizierung genetischer Suszeptibilitätsloci für granulomatöse Lungenkrankheiten

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Dipl.-Biol. Benjamin Schmid

Kiel, 2014

Erste Gutachterin: Prof. Dr. Manuela Dittmar

Zweite Gutachterin: Prof. Dr. Almut Nebel

Tag der mündlichen Prüfung:.....30.01.2014

Zum Druck genehmigt: 25.04.2014

_____

Dekan Prof. Dr. Wolfgang J. Duschl

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AIDS	erworbenes Immundefektsyndrom (engl. *acquired immunodeficiency syndrome*)
ANXA11	Annexin A11
ATP	Adenosintriphosphat
BAL	bronchoalveoläre Lavagen
bp	Basenpaare
BTNL2	*butyrophilin-like 2 protein*
CCL2	*chemokine (C-C Motif) ligand 2*
CCR2	*C-C chemokine receptor type 2*
CD4+	CD4-Rezeptor
CD8+	CD8-Rezeptor
cDNA	komplementäre DNA (engl. *complemetary DNA*)
CEU	Bewohner Utahs mit nord- und westeuropäischen Vorfahren
ChIP	Chromatin-Immunpräzipitation
ChIP-Seq	Chromatin-Immunpräzipitation mit DNA-Sequenzierung
CI	Konfidenzintervall (engl. *confidence interval*)
cM	Centimorgan
CpG-Ort	CpG-Dinukleotid
CR	Genotypisierungsrate (engl. *call rate*)
ddNTP	Didesoxyribonukleosid-Triphosphate
DEPC	Diethylpyrocarbonat
DNA	Desoxyribonukleinsäure (engl. *deoxyribonucleic acid*)
dNTP	Desoxyribonukleosid-Triphosphate
DRB1	*HLA class II histocompatibility antigen, DRB1-9 beta chain*
EDTA	Ethylendiamintetraacetat
eQTL	Quantitativer merkmalsbeeinflussender Expressions-Locus (engl. *expression quantitative trait loci*)
ER	Endoplasmatisches Retikulum
EREG	Epiregulin
FAM-Farbstoff	6-FAM-phosphoramidit Fluoreszenzfarbstoff
FK	Feinkartierungsstichprobe
GAPDH	Glycerinaldehyd-3-phosphat-Dehydrogenase
GLM	Generalisierte Lineare Modelle
Gm12878-Zelllinie	lymphoblastoide Zelllinie

GWAS	genomweite Assoziationsstudie
H_0	Nullhypothese
H3K27Ac	Acetylierung des Lysins an Position 27 am Histon 3
H3K4Me1	Monomethylierung des Lysins an Position 4 des Histon 3
H3K4Me3	Trimethylierung des Lysins an Position 4 des Histon 3
Häm	Fe-Protopophyrin IX (Häm *b*)
HIV	Humanes Immundefizienz-Virus (engl. *human immunodeficiency virus*)
HLA	humane Leukozytenantigene (engl. *human leukocyte antigen)*
Hsp70	Familie der Hitzeschockproteine 70
htSNP	haplotypidentifizierender SNP (engl. *haplotype tag SNP*)
HWE	Hardy-Weinberg-Gleichgewicht (engl. *Hardy-Weinberg equilibrium*)
IBS	Identität-durch-Status (engl. *identity-by-state*)
IFNG	Gamma-Interferon I (engl. *interferon gamma*)
IFN-γ	Gamma-Interferon I (engl. *interferon gamma*)
IKMB	Institut für Klinische Molekularbiologie der Christian-Albrechts-Universität zu Kiel
IL-10	Interleukin-10
IQR	Quartilsabstand (engl. *interquartile range*)
IRE1	Inositol-abhängiges Protein 1(engl. *Inositol-requiring protein 1*)
JNK	c-Jun N-terminale Kinase (engl. *c-Jun N-terminal kinase*)
kb	Kilobasenpaar
KCNQ1	spannungsgesteuertes Kalium-Kanal Protein (engl. *potassium voltage-gated channel, KQT-like subfamily, member 1*)
KORA	Kooperative Gesundheitsforschung in der Region Augsburg
LD	Kopplungsungleichgewicht (engl. *linkage disequilibrium*)
lncRNA	lange nicht-codierende RNA (engl. *long non-coding RNA*)
LPS	Lipopolysaccharide
LWK	Bevölkerung der Luhya in Webuye, Kenya
m/v	Masse/Volumen
MAF	Allelfrequenz des seltenen Allels (engl. *minor allele frequency*)
MALDI-TOF	Matrix-unterstützte Laser-Desorption/Ionisation und Massenspektrometrie mit Flugzeitanalysator (engl. *matrix-assisted laser desorption/ionization time-of-flight mass spectrometer*)
ManLAM	Glycolipid Lipoarabinomannan
Mb	Megabasenpaar
MDA	*multiple displacement amplification*
MEX	Bewohner mit mexikanischer Abstammung aus Los Angeles, California

MHC2TA	MHC Klasse II Transaktivator (engl. *MHC class II transactivator*)
miRNA	microRNA, kurze nicht-codierende RNA
mKatG	Katalase-Peroxidase
MKK	Bevölkerung der Maasai in Kinyawa, Kenia
MMLV	Murines Leukämievirus (engl. *moloney-murine leukemia virus*)
mRNA	Boten-RNA (engl. *messenger RNA*)
NHLF-Zelllinie	normale humane Lungenfibroblasten-Zelllinie
NIH	National Institute of Health
NOTCH4	*neurogenic locus notch homolog 4*
NOTCH-Protein	Familie von Transmembranproteinen mit einer extrazellulären Domäne
Nsp I	Restriktionsendonuklease Nsp I
nsSNP	nicht-synonymer SNP (engl. *non-synonymous SNP*)
OR	Quotenverhältnis (engl. *odds ratio*)
OS9	*osteosarcoma amplified 9*
P2X7	*P2X purinoceptor 7*
PBS	Phosphatgepufferte Salzlösung (engl. *phosphate buffered saline*)
PCA	Hauptkomponentenanalyse (engl. *principal component analysis*)
PCR	Polymerase-Kettenreaktion (engl. *polymerase chain reaction*)
POPGEN Biobank	Biobank des populationsgenetischen Forschungsprojekts des Nationalen Genomforschungsnetzes
PRDM9-Protein	PR Domäne Zinkfinger Protein (engl. *PR domain zinc finger protein 9*)
QC	Qualitätskontrolle (engl. *quality control*)
r^2	Masseinheit des Kopplungsungleichgewichts
RAB23	*Ras-related protein Rab-23*
RNA	Ribonukleinsäure (engl. *ribonucleic acid*)
RNAi	RNA-Interferenz
RR	relatives Risiko
RT	Raumtemperatur
RT-PCR	Reverse Transkriptase Polymerase-Kettenreaktion (engl. *reverse transcription polymerase chain reaction*)
SA	Sarkoidose
SD	Standardabweichung (engl. *standard deviation*)
SLC11A1 (NRAMP1)	*natural resistance-associated macrophage protein 1*
SNP	Einzelbasen-Polymorphismen (engl. *single nucleotide polymorphism*)
spaSNP	Spleißstellen SNP
Sty I	Restriktionsendonuklease Sty I

tagging SNP	haplotypidentifizierender SNP
TB	Tuberkulose
TDT	Test auf Transmissions-Ungleichgewicht (engl. *transmission disequilibrium test*)
TFBS	Transkriptionsfaktorbindestellen
TNFA	Tumornekrosefaktor-α Gen
TNF-α	Tumornekrosefaktor-α
T_{Reg}	Regulatorische T-Zellen
TSI	Bewohner der Toskana, Italien
TST	Tuberkulin Hauttest (engl. *tuberculin skin test*)
T-Zellen	T-Lymphozyten
UV	ultraviolettes Licht
VIC-Farbstoff	VIC-Fluoreszenzfarbstoff
WGA	Gesamt Genomamplifikation (engl. *whole genome amplification*)
WHO	Weltgesundheitsorganisation (engl. *world health organisation)*
WTCCC	*Wellcome Trust Case Control Consortium*
YRI	Bevölkerung der Yoruba in Ibadan, Nigeria
χ^2	Chi-Quadrat

1 Einleitung

1.1 Genetische Faktoren bei Krankheiten

1.1.1 Komplexe Krankheiten

Krankheiten, die auf genetischen Faktoren beruhen, können aus einer einzelnen genetischen Mutation (monogenetische Erkrankung) entstehen, bei der die - meist vererbte - Mutation eines Gens alleinverantwortlich ein charakteristisches Krankheitsbild hervorruft. Es gibt aber auch Fälle, in denen erst eine Kombination von verschiedenen Faktoren wie Umwelteinflüssen, Lebensgewohnheiten und mehreren veränderten Genen zu einem Krankheitsausbruch führt. Krankheiten, die auf einer solchen Vielzahl von Faktoren basieren, werden daher als multifaktorielle oder komplexe Krankheit bezeichnet (Dempfle et al. 2008; N. J. Risch 2000). Da komplexe Krankheiten, neben den anderen möglichen Faktoren, in der Regel auf einer größeren Anzahl von verschiedenen Genen beruhen, haben die einzelnen Gene häufig nur einen moderaten Einfluss auf die Krankheitsentstehung. Einzelne Risikoallele können auch in der gesunden Normalbevölkerung zu finden sein, d.h. erst eine spezifische Kombination mehrerer Risikoallele und deren Wechselwirkung mit den spezifischen Umweltfaktoren führen zur Krankheit (Page et al. 2003). Die Wahrscheinlichkeit, mit der ein bestimmter Genotyp zur Ausprägung des dazugehörigen Phänotyps führt, wird als Penetranz bezeichnet. In komplexen Krankheiten liegt stets eine unvollständige Penetranz vor, was die Analyse der Krankheitsursachen erschwert. Ein weiteres Problem bei der Erforschung von komplexen Krankheiten besteht darin, dass Risikoallele häufig nur in bestimmten Bevölkerungsgruppen (wie z.B. bei Afro-Amerikanern, Kaukasiern oder Japanern) eine Rolle spielen. Ergebnisse aus Assoziationsstudien lassen sich daher nicht generell auf andere Populationen übertragen und müssen erst in den jeweiligen Populationen validiert werden. Zu der genetischen Heterogenität kommt bei komplexen Krankheiten häufig auch eine phänotypische Heterogenität hinzu, und es können unterschiedliche Ausprägungen der Krankheit vorkommen (häufig dann als Subphänotyp der Krankheit bezeichnet). Bei der krankheitsbezogenen genetischen Forschung ist daher die sorgfältige Klassifizierung der Patienten von großer Bedeutung, um eine möglichst homogene Studienpopulation zu erhalten.

1.1.2 Infektionskrankheiten

Die Analyse der genetischen Basis für die Anfälligkeit gegenüber Infektionskrankheiten ist möglicherweise eine der komplexesten Herausforderungen in der Genetik von Krankheiten. Infektionskrankheiten sind wahrscheinlich nicht nur polygene Krankheiten mit einer starken

genetischen Komponente, sondern zusätzlich benötigt es in allen Fällen einen essentiellen Umweltfaktor (Pathogen), welcher fast immer ein eigenes Genom besitzt. Nichtsdestotrotz gibt es einen stetigen Fortschritt in der Erforschung der komplexen Wechselwirkungen von Wirtsgenen und Mikroorganismen. Seit den 1930er Jahren konnten verschiedene Zwillingsstudien zeigen, dass die Wirtsgenetik eine wesentliche Rolle in der unterschiedlichen Anfälligkeit für Tuberkulose spielt (Comstock 1978; Diehl et al. 1936; Kallmann and Reisner 1943). Solche Unterschiede konnten auch bei weiteren Infektionskrankheiten (z.B. Poliomyelitis und Hepatitis B) beobachtet werden (Herndon and Jennings 1951; Lin et al. 1989). Die Beobachtung, dass einige Infektionskrankheiten ein vererbbares Element haben, ist nicht neu, genauso wenig wie die Beobachtung, dass Individuen unterschiedlich auf bestimmte Infektionen reagieren. Im Jahr 1930 wurden in Lübeck unbeabsichtigterweise Säuglinge in einem Krankenhaus mit dem *Mycobacterium tuberculosis* infiziert (American Journal of Public Health and the Nation's Health 1931). In den meisten Fällen führte dies zur Erkrankung an Tuberkulose, doch in einigen Fällen zeigten die Säuglinge keine Symptome und blieben gesund. Solche Beobachtungen stützen die These einer genetischen Komponente bei Infektionskrankheiten. Die Rolle des Immunsystems bei der Bekämpfung von Infektionen ist hinlänglich bekannt und es ist nachvollziehbar, dass Mutationen in Genen, die für Bestandteile des Immunsystems codieren, zu einer Anfälligkeit gegenüber Infektionskrankheiten führen können. So konnte schon gezeigt werden, dass Mutationen in dem humanen Leukozyten-Antigen-System (engl. *human leukocyte antigen,* HLA) mit der Suszeptibilität gegenüber verschiedenen Infektionskrankheiten (Tuberkulose, Lepra, Hepatitis B, AIDS (engl. *acquired immunodeficiency syndrome*)) in Verbindung gebracht werden (Almarri and Batchelor 1994; Brahmajothi et al. 1991; Carrington et al. 1999; Visentainer et al. 1997). Aber auch außerhalb des Bereichs des HLA-Systems konnten schon Gene identifiziert werden, die die Anfälligkeit gegenüber Infektionen beeinflussen können, wie z.B. Polymorphismen im Promoterbereich des *TNFA*-Gens (Tumornekrosefaktor-α), die Auswirkungen auf die Anfälligkeit für Malaria haben können (McGuire et al. 1999). Über die Identifizierung von Genen außerhalb des HLA-Systems können somit Signaltransduktionswege aufgedeckt werden, die in der Pathogenese der untersuchten Infektionskrankheit eine Rolle spielen. Mit diesem durch die genetische Forschung gewonnenem Wissen können dann möglicherweise neue Therapien entwickelt werden.

1.1.3 Einzelbasen-Polymorphismen

Eine der wesentlichen Aufgaben in der genetischen Forschung ist es DNA-Sequenzvariationen mit vererbbaren Phänotypen zu assoziieren, um zugrunde liegende genetische Faktoren zu identifizieren. Die Kosten für die Sequenzierung eines vollständigen humanen Genoms sind noch sehr hoch, und

wenn umfangreiche Stichproben untersucht werden sollen, beschränkt man sich heutzutage üblicherweise auf Bereiche des Genoms, die einen Großteil der vorkommenden Sequenzvariationen abdecken.

Um im großen Umfang Faktoren, die zum Ausbruch einer Krankheit beitragen, auf genetischer Ebene zu untersuchen, werden genetische Marker genutzt. Genetische Marker sind DNA-Loci mit bekannter Position im Genom, welche einen gewissen Grad an Variabilität zwischen Individuen zeigen. Für eine Analyse des humanen Genoms braucht es eine große Anzahl von genetischen Markern, möglichst gleichmäßig verteilt im Genom. Des Weiteren sollten diese Marker möglichst Polymorphismen sein, die schnell, akkurat und kostengünstig bestimmt werden können. Einzelbasen-Polymorphismen (engl. *single nucleotide polymorphism*, SNP) sind solche genetischen Marker. Mit ihnen kann das Genom auf krankheitsbeeinflussende Faktoren - also Basenaustausche, die zu einer Veränderung des Phänotyps führen - untersucht werden.

Häufig wird der Begriff SNP allerdings auch für Mutationen verwendet, bei denen an einer Stelle im Genom Insertionen oder Deletionen (genannt Indel) auftauchen. Die daraus resultierenden Polymorphismen werden wie die klassischen Einzelbasen-Polymorphismen als genetische Marker benutzt (Abb. 1-1). Klassischerweise beziehen sich SNP-Marker aber auf eine einzige Position im Genom, bei der eine Base durch eine andere, z.B. durch eine Punktmutation, substituiert wurde (Abb. 1-1).

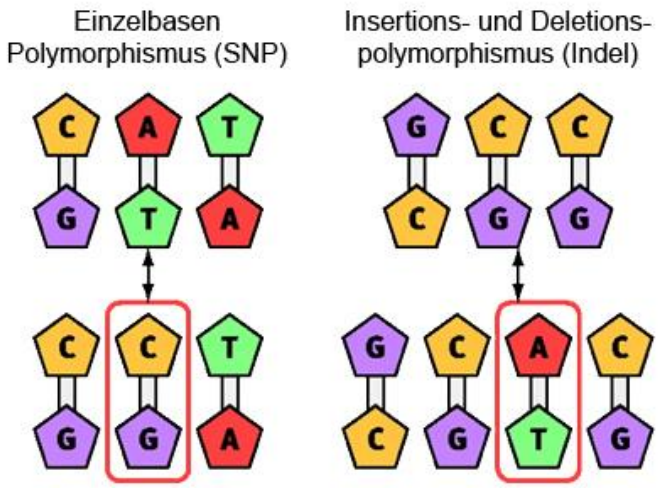

Abb. 1-1: Schematische Darstellung des Einzelbasen Polymorphismus (SNP) und des Insertions- und Deletionspolymorphismus (Indel). Diese Polymorphismen werden als genetische Marker in genomweiten Assoziationsstudien (GWAS) genutzt. Modifiziert nach (Hingorani et al. 2010).

Obwohl bei einem SNP nach dem Austausch des Nukleotids alle vier Allele möglich wären, sind beim Menschen die meisten SNPs biallelisch, wobei ungefähr zwei Drittel aller SNPs die Allele Cytosin und Thymin (C und T) oder Guanin und Adenin (G und A) tragen. Die hohe Anzahl dieser Cytosin/Thymin-SNPs ist wahrscheinlich zum Teil durch die häufig stattfindende 5-Methylcytosin Desaminierungs-

Reaktion, besonders an CpG-Orten, zu erklären (Holliday and Grigg 1993). Die ursprüngliche Definition von SNPs verlangte, dass sie mindestens eine Frequenz von 1% des seltenen Allels (engl. *minor allele frequency*, MAF) in der untersuchten Population aufweisen. Heutzutage, mit dem verbesserten Verständnis der evolutionären Hintergründe vererbter Mutationen, definiert eine MAF > 1% mittlerweile einen SNP als „häufigen" (engl. *common*) SNP, bei einer MAF < 1% spricht man von einem „seltenen" (engl. *rare*) SNP (http://hapmap.ncbi.nlm.nih.gov/).

SNPs treten im Schnitt alle 100 bis 300 Basen auf und stellen mit ca. 90% die häufigste Variation im menschlichen Genom dar. Sie zeigen genomweit eine heterogene Verteilung; Regionen mit einer hohen Dichte an SNPs (z.B. die sog. HLA-Region) wechseln sich mit SNP-armen Regionen ab (Guillaudeux et al. 1998; Horton et al. 1998; Koboldt et al. 2006; Nachman et al. 1998; Sainudiin et al. 2007; Varela and Amos 2010). Ungefähr 90% der heterozygoten SNPs eines Individuums sind häufige Varianten, die es mit anderen Individuen der gleichen Population teilt (D. Altshuler et al. 2008). Mittlerweile haben das 'Human Genome Project' (Lander et al. 2001), das 'SNP Consortium' (Sachidanandam et al. 2001) und das 'International HapMap Project' (International HapMap Consortium 2005) zusammen über zehn Millionen häufige DNA-Varianten, hauptsächlich SNPs, in einer Auswahl von DNA-Proben wichtiger Bevölkerungsgruppen beschrieben.

1.1.4 Kopplungsungleichgewicht

SNPs sind zwar über das gesamte Genom verteilt, werden aber nicht notwendigerweise vollständig unabhängig voneinander vererbt. Treten gewisse Kombinationen von Allelen benachbarter SNPs in bestimmten chromosomalen Bereichen gehäuft auf, liegt zwischen ihnen eine Kopplung vor. Das Kopplungsungleichgewicht (engl. *linkage disequilibrium*, LD) beschreibt die nicht-zufällige Assoziation von Allelen an zwei Loci auf einem Chromosom in einer sich natürlich vermehrenden Population (zusammengefasst in Slatkin 2008). Dieser chromosomale Bereich wird als LD-Block und eine bestimmte Kombination von SNP-Allelen innerhalb des LD-Blocks als Haplotyp bezeichnet. Wenn nun diese chromosomale Region in den nächsten Generationen repliziert wird, wird der Haplotyp sehr wahrscheinlich intakt bleiben. In diesem Fall wird ein vollständiges LD zwischen den benachbarten SNP-Allelen herrschen.

Der Grad des LDs wird durch eine Reihe von Faktoren beeinflusst, wie Genkopplung, Selektion, Rekombinationsrate, Mutationsrate, Gendrift, nicht-zufällige Paarung und der Populationsstruktur (zusammengefasst in A. Collins 2009). Diese Faktoren wirken aber nicht gleichmäßig an jedem Bereich des humanen Genoms, sondern es wechseln sich Regionen, in denen sich das LD über weite Abstände erstreckt, mit Regionen ab, in denen LD kaum detektierbar ist (S. B. Gabriel et al. 2002). Die Regionen mit einem sehr geringen LD werden als Rekombinations-Hotspots bezeichnet. Hier findet,

im Gegensatz zu Regionen mit hohem LD, eine Vielzahl von Rekombinationsereignissen statt. Rekombinations-Hotspots zeichnen sich durch DNA-Sequenzmotive aus, die von dem PRDM9-Protein erkannt werden, welches die Initiation der meiotischen Rekombination vermittelt (Baudat et al. 2010; Billings et al. 2013; Jeffreys et al. 2013). Zur Quantifizierung für paarweises LD zwischen biallelischen SNPs wird am häufigsten die Maßeinheit r^2 (quadrierter Korrelationskoeffizient als Maßeinheit des LDs zwischen zwei Loci) benutzt. Diese beschreibt den Anteil der beobachteten Fälle, in denen zwei SNP-Marker gemeinsam auftreten. Die möglichen Werte von r^2 bewegen sich zwischen 0 und 1. Wenn zwei Marker den Wert $r^2 = 1$ haben, dann werden die Allele stets gemeinsam vererbt. Hat r^2 den Wert 0, dann sind die in Frage stehenden Marker nicht gekoppelt und verhalten sich statistisch unabhängig voneinander (Raychaudhuri 2011). Mit dem Wissen über das Vorkommen eines SNP-Allels kann mit Hilfe des LDs auf die Wahrscheinlichkeit des Vorkommens von anderen Allelen geschlossen werden. Somit kann das LD dazu genutzt werden, mit Hilfe von relativ wenigen genotypisierten SNPs eine Assoziation einer bestimmten genomischen Region nachzuweisen (z.B. mit einer Krankheit in Assoziationsanalysen). Dies gilt, solange die genotypisierten SNPs repräsentativ für die jeweiligen LD-Blöcke sind und in einem hohen LD mit den anderen SNPs des LD-Blocks stehen (engl. *tagging-SNP*). Der ursprünglich untersuchte Marker muss daher als solcher keine funktionellen Effekte aufzeigen, sondern nur mit der kausativen Variante in LD stehen um ein Assoziationssignal zu zeigen (zusammengefasst in Kwok and Gu 1999). Die Suche nach den kausativen Varianten kann dann auf die Regionen konzentriert werden, in denen SNPs eine Assoziation mit der Krankheit zeigen (F. S. Collins et al. 1997). Mit der Verfügbarkeit von genomweiten LD-Daten wird es möglich, die Technik der genomweiten Assoziationsanalyse (engl. *genome-wide association study*, GWAS) umzusetzen, bei der, durch die Genotypisierung einer verhältnismäßig geringen Anzahl von *tagging*-SNPs, mit Hilfe des LD eine genomweite Abdeckung erfolgen kann.

1.1.5 Genomweite Assoziationsstudien

Die in den Vereinigten Staaten von Amerika ansässige Behörde „National Institute of Health“ (NIH) definiert eine genomweite Assoziationsstudie (GWAS) als eine Analyse von häufigen genetischen Varianten (SNPs) des humanen Genoms mit dem Ziel genetische Faktoren zu identifizieren, die einen Einfluss auf die Gesundheit haben. In dieser Hinsicht gleichen GWAS den klassischen epidemiologischen Ansätzen, wobei die dort untersuchten Risikofaktoren in einer GWAS das Allel eines untersuchten genetischen Markers darstellen. Der Vorteil einer GWAS ist jedoch, dass sie eine hypothesenfreie genomweite Analyse darstellt. Heutzutage sind GWAS bei der Erforschung humangenetischer Erkrankungen nicht mehr fortzudenken. Sie führten zur Entdeckung einer großen Anzahl von häufigen Varianten, deren Auftreten mit verschiedenen häufigen Krankheiten wie z.B.

Herzerkrankungen, Autoimmunkrankheiten, entzündlichen Erkrankungen und psychiatrischen Krankheiten korreliert (Visscher et al. 2012).

In einer GWAS wird eine möglichst große Anzahl häufiger genetischer Marker (SNPs; hauptsächlich *tagging*-SNPs) in einer großen Anzahl von Individuen auf eine Assoziation mit einem bestimmten Merkmal untersucht. Mit Hilfe von Fall-Kontroll-Studien können SNPs entdeckt werden, die einen Unterschied in der Allelfrequenz zwischen betroffenen Individuen und nicht verwandten gesunden Kontrollpersonen aufweisen.

1.1.5.1 Genomweite Assoziationsanalysen komplexer Krankheiten

Der Wahrscheinlichkeitswert, mit dem in einer genomweiten Assoziationsstudie ein echter genetischer Effekt nachgewiesen werden kann, wird mittels der statistischen Teststärke (engl. *statistical power*) ausgedrückt. Bei Fall-Kontroll-Assoziationsstudien hängt die Teststärke direkt proportional mit der Stichprobengröße, der Größe des genetischen Effekts, der Frequenz des untersuchten Allels und der Signifikanzgrenze (i.d.R. $\alpha = 0{,}05$) zusammen. Um bei komplexen genetischen Erkrankungen Risikofaktoren zu identifizieren, ist es wichtig, eine möglichst große Anzahl von Individuen (Stichproben) zu untersuchen, da in komplexen Krankheiten erst viele verschiedene Gene mit jeweils nur einem moderaten Einfluss zum Krankheitsausbruch führen und einzelne dieser Risikoallele auch in den gesunden Kontrollindividuen zu finden sind (Page et al. 2003) (Abb. 1-2).

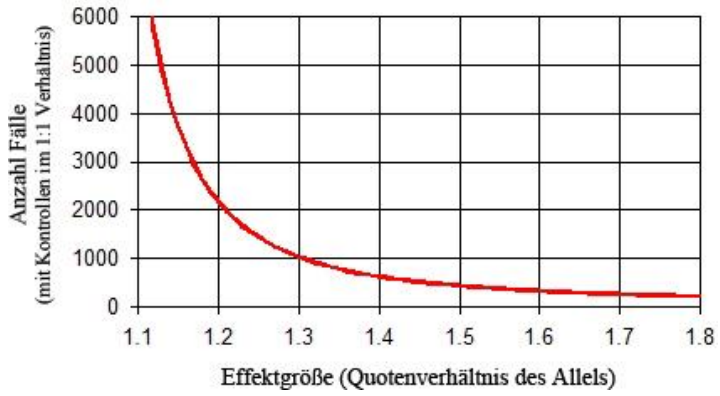

Abb. 1-2: Zusammenhang zwischen dem Quotenverhältnis des Allels und der Stichprobengröße. Der Anzahl der Fälle entspricht eine identische Anzahl von gesunden Kontrollen. Für die Berechnung wurde von einer Teststärke von 80%, einer Signifikanzgrenze von $\alpha = 0{,}05$ und einer Risikoallelfrequenz von 30% ausgegangen. Darstellung mit dem „PS Power and Sample Size Program“ (Dupont and Plummer 1997).

Mit der ersten erfolgreichen GWAS 2005 (Klein et al. 2005), welche die altersbedingte Makuladegeneration untersuchte, und besonders mit der ersten großen, sorgfältig konzipierten GWAS des WTCCC (engl. *Wellcome Trust Case Control Consortium*) (Wellcome Trust Case Control Consortium 2007) für komplexe Krankheiten begann die Ära der GWAS in der genetischen Forschung.

Seitdem wurden über 2.000 Loci signifikant mit einem oder mehreren komplexen Merkmalen assoziiert (Abb. 1-3).

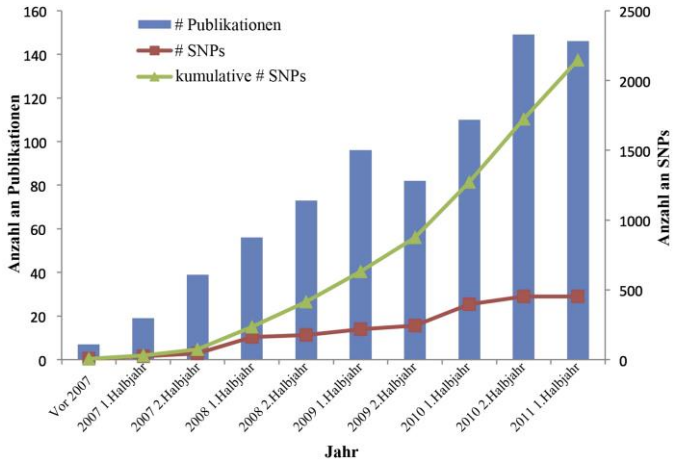

Abb. 1-3: GWAS Publikationen aus dem GWAS Katalog des ‚National Human Genome Research Institute' (NHGRI) über einen Zeitraum von 5 Jahren. Es sind nur SNPs dargestellt, welche eine Assoziiation von $p < 5 \times 10^{-8}$ mit einem Locus zeigen. Um doppelte Zählungen zu vermeiden, wurden die für ein gleiches Merkmal identifizierten SNPs mit einem LD von $r^2 > 0.8$ in den gesamten HapMap Proben ausgeschlossen. Nach (Visscher et al. 2012).

Die Idee der genomweiten Assoziationsstudien basiert auf der Beobachtung, dass häufige Varianten zu der Entstehung von häufigen Krankheiten beitragen, und aus dieser Beobachtung entwickelte sich die Hypothese '*common disease – common variant*' (häufige Krankheit – häufige Variante) (Chakravarti 1999; Lander 1996; D. E. Reich and Lander 2001; N. Risch and Merikangas 1996). Die Hypothese besagt, dass, obwohl häufige Krankheiten durch eine Vielzahl von Loci beeinflusst werden können, ein Großteil des genetischen Risikos seine Ursache in einer oder einer kleinen Anzahl von häufigen Varianten hat. Die Definition von häufigen Varianten verlangt, dass diese eine Frequenz des selteneren Allels (engl. *minor allele frequency*, MAF) von > 1% haben. Die Hypothese impliziert, dass alle kausalen Mutationen eines Locus häufige Varianten sind, doch ist es in vielen Fällen so, dass häufige Varianten nur dazu dienen Loci für detailliertere Untersuchungen auszuwählen, in denen dann die kausalen seltenen Varianten identifiziert werden können. Aus dieser Beobachtung entstand die Hypothese '*common disease – rare variant*' (häufige Krankheit – seltene Variante) (Bodmer and Bonilla 2008; J. K. Pritchard 2001; Schork et al. 2009). Diese Hypothese besagt, dass bei einer häufigen Krankheit (mit einer Prävalenz höher als 1-5%), die auf genetischen Ursachen beruht, die Ursachen nicht zwingend auch häufige Varianten in der gleichen Population darstellen müssen. Eher ist die Krankheit ein Ergebnis von mehreren in der Population eher seltenen Varianten in dem gleichen Gen (allelische Heterogenität) oder in mehreren Genen (Locus-Heterogenität). Dabei stellt jede Variante nur eine moderate, aber feststellbare Erhöhung des Erkrankungsrisikos dar und erst eine Aufsummierung der Effekte von einer Reihe von niedrigfrequenten dominanten und unabhängig wirkenden Varianten führt zur Krankheit. Wegen ihrer geringen Frequenz und ihres individuell

geringen Beitrags zu der vererbten Anfälligkeit für die Krankheit können seltene Varianten selbst mit sehr großen GWAS nicht (oder nur sehr schwer) identifiziert werden. Doch es scheint möglich, dass in einer GWAS detektierte Signale den Effekt von mehreren seltenen Varianten reflektieren (Dickson et al. 2010) und somit in solchen Studien seltene Varianten indirekt identifiziert werden können.

Mit den üblichen GWAS können seltene Varianten nicht (oder nur sehr schwer) nachgewiesen werden, da in Fall-Kontroll-Studien die Allelfrequenz der Marker direkt die statistische Teststärke beeinflusst (Abb. 1-4). Diesen Zusammenhang zeigt McCarthy (2008) im Rahmen einer Diabetes-Studie. Das Gen *KCNQ1* wurde in zwei GWAS (jeweils weniger als 200 Fälle) in ostasiatischen Populationen mit Typ2-Diabetes-Risiko assoziiert. Diese Assoziation konnte in Europäischen Populationen erst in einer umfangreichen GWAS repliziert werden, da die Allelfrequenz des Risikoallels bei Europäern nur 5% - im Gegensatz zu ca. 40% bei den Ostasiaten - beträgt (Abb. 1-4).

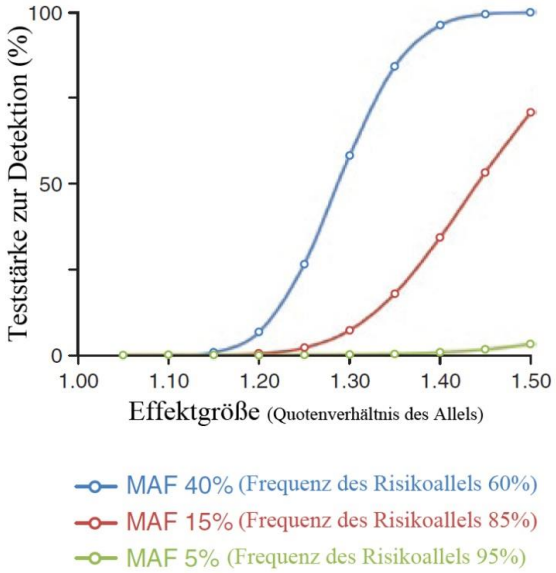

Abb. 1-4: Teststärke, mit der Varianten mit gleicher Effektgröße, aber unterschiedlicher Risikoallelfrequenz detektiert werden können. Bei einer Effektgröße von 1,35 wird eine Teststärke von über 80% erreicht, wenn die MAF des protektiven Allels 40% beträgt. Wenn die MAF 5% beträgt, fällt die Teststärke unter 1%. Die Berechnungen wurden in 2.000 Fällen mit 2.000 Kontrollen und einer Signifikanzschwelle von $p = 5 \times 10^{-8}$ durchgeführt. Nach (McCarthy 2008).

In genomweiten Ansätzen werden daher alternative statistische Methoden zur Entdeckung von seltenen Varianten benötigt, oder es müssen sehr umfangreiche Stichproben untersucht werden. Durch die Entwicklung der Methode der Genotypimputation können mittlerweile verschiedene GWAS zusammengefasst und gemeinsam analysiert werden (de Bakker et al. 2008). Damit erhöht sich die Stichprobengröße und die Identifizierung von SNP-Markern mit geringerer Allelfrequenz ist möglich (Abb. 1-2 und Abb. 1-5). Die eigentliche Aufgabe der Genotypimputation ist es, die Anzahl der Marker in einer GWAS unter der Nutzung von Referenzdaten zu erhöhen. Mit den kommerziell erhältlichen SNP-Arrays (ein DNA-Mikroarray mit dem Einzelbasen-Polymorphismen detektiert werden) können jeweils nur ca. eine Million SNPs genotypisiert werden, doch mittels spezialisierter

Hochdurchsatz-Genotypisierungsverfahren wurden im humanen Genom mittlerweile schon mehr als zehn Millionen SNPs (mit einer MAF ≥ 0,05) identifiziert und verifiziert (Abecasis et al. 2010). Konsortien wie das „*HapMap Project*“ (International HapMap Consortium 2003; Pemberton et al. 2010) oder das „*1000 Genomes Project*“ (1000 Genomes Project Consortium 2010) haben für verschiedene charakteristische humane Populationen Haplotypdaten durch spezialisierte Hochdurchsatz-Genotypisierungsverfahren gewonnen. Auf Basis dieser Referenzdaten ist es möglich, nicht-typisierte Marker des SNP-Arrays zu schätzen (imputieren): Durch Nutzung der Haplotyp- und LD-Daten können anhand der auf dem Array genotypisierten SNPs sehr akkurate Schätzungen über das Vorhandensein weiterer nicht-typisierter Marker gemacht werden (S. R. Browning and Browning 2007; B. L. Browning and Browning 2009; Li et al. 2009). Damit wird zunächst einmal die Anzahl der SNPs in einer GWAS und somit die genomweite Abdeckung an Markern erhöht. Durch das Imputieren von GWAS wird es jedoch auch möglich, unterschiedliche Datensätze in einer Metaanalyse zu kombinieren, da durch die Imputation Unterschiede in der Markerdichte der Datensätze aufgehoben werden (de Bakker et al. 2008; Zeggini and Ioannidis 2009). Beispiel siehe Abb. 1-5.

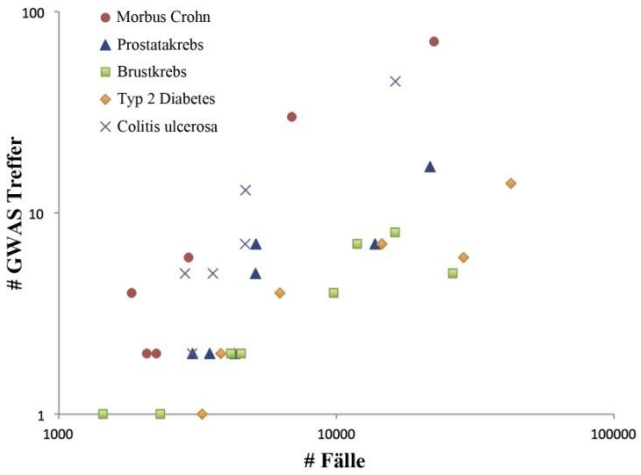

Abb. 1-5: Anzahl von Identifizierungen von mit einer Krankheit assoziierbaren SNP-Loci als Funktion des Stichprobenumfangs. Die Koordinaten sind logarithmisch aufgetragen. Nach (Visscher et al. 2012).

1.1.5.2 Populationsspezifische genetische Strukturen

GWAS können in jeder Population durchgeführt werden, doch bei der Betrachtung der mit den Studien gewonnenen Daten müssen in einigen Fällen populationsspezifische Besonderheiten beachtet werden: Die LD-Strukturen im Genom können zwischen humanen Populationen stark variieren. Die höchste genetische Variation wird in afrikanischen Populationen beobachtet. Sie ist in nicht-afrikanischen Populationen deutlich geringer. Afrikanische Populationen weisen ein geringeres Ausmaß an LD auf und die Bereiche, über die sich starkes LD erstreckt, sind kürzer. Der moderne

Mensch (*Homo sapiens*) entwickelte sich vor rund 200.000 Jahren in Afrika und die „*Out of Africa*"-Theorie besagt, dass der moderne Mensch sich erst innerhalb der letzten ca. 100.000 Jahre über die restliche Erde ausgebreitet hat (Campbell and Tishkoff 2008; Tishkoff and Verrelli 2003). Mit der Migration aus Afrika kam es zu einem genetischen „Flaschenhalseffekt" (oder Gründereffekt; engl. *founder effect*). Gründungspopulationen außerhalb Afrikas sind genetisch deutlich weniger variabel, wozu kommt, dass diese Populationen auch ein verstärktes LD und sehr ähnliche Muster in ihren LD-Strukturen aufweisen (Abb. 1-6).

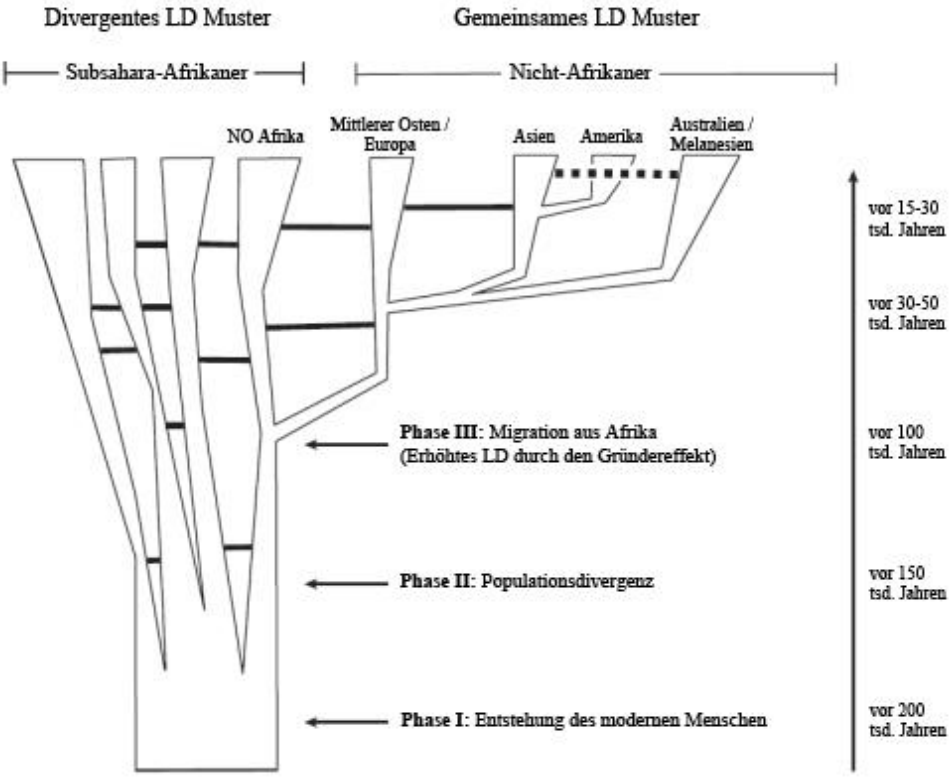

Abb. 1-6: Übersicht der Evolutionsgeschichte humaner Kopplungsungleichgewichte (LD). Die afrikanischen Ursprungspopulationen haben große, in sich unterteilte Populationsstrukturen behalten. In der afrikanischen Evolutionsgeschichte finden sich in diesen Populationen komplexe Muster von Populationsexpansionen, -reduktionen, Migrationen und Vermischungen statt. Der genetische Flaschenhals, der bei der Gründung der nicht-afrikanischen Populationen (vor ca. 50-100 Tsd. Jahren) entstand, sorgte für ein geringeres Level an genetischer Diversität, ein höheres LD und auch für ähnlichere LD-Muster. Die horizontalen schwarzen Linien symbolisieren den Genfluss zwischen Populationen und die gestrichelte horizontale Linie zeigt den jüngsten Genfluss von Asien nach Australien/Melanesien. Nach (Campbell and Tishkoff 2008).

Unter evolutionären Gesichtspunkten hat in Afrika, im Vergleich mit jeder anderen geographischen Region, am längsten eine relativ große effektive Populationsgröße existiert. Daher haben auf diese Population über einen sehr viel längeren Zeitraum populationsgenetische Ereignisse wie natürliche Selektion, Gendrift, Mutationen und Genfluss gewirkt. Auch die vielen sprachlich bedingten Sub-Populationen, die sich in Afrika ausgeprägt haben, und die damit einher gehenden

Migrationsereignisse, Vermischungen und Fluktuationen in der Population trugen zu der hohen genetischen Variabilität bei, die in europäischen Populationen z.B. nicht so stark ausgeprägt ist.

1.2 Krankheiten

1.2.1 Sarkoidose

Sarkoidose, früher auch als Morbus Boeck bezeichnet, wurde erstmalig von Ernest Besnier und Cæsar Peter Møller Boeck in den Jahren 1889 und 1899 beschrieben (Besnier 1889; Boeck 1899). Besnier und Boeck beobachteten damals eine entzündliche Veränderung der Haut, die in 25-50% der Fälle bei einer Sarkoidoseerkrankung auftreten kann. Sarkoidose ist jedoch eine systemische entzündliche Erkrankung, die hauptsächlich die Lunge und die Lymphknoten betrifft, wobei gleichwohl auch jedes andere Organ betroffen sein kann.

Die Krankheit betrifft hauptsächlich junge Erwachsene, häufiger Frauen, zwischen dem 20. und 40. Lebensjahr und hat eine geschätzte weltweite Prävalenz von 2 bis 64/100.000 Einwohnern (Deutschland: 40/100.000) (Haimovic et al. 2012). Insgesamt weist die Krankheit eine starke Heterogenität in Bezug auf die betroffenen Organe und den Krankheitsbeginn und -verlauf auf. Der Phänotyp der Sarkoidose variiert bei den verschiedenen Geschlechtern und Ethnien (L. S. Newman et al. 1997). Je nach Krankheitsverlauf werden zwei Subphänotypen differenziert:

I) Bei dem *akuten* Subphänotyp der Sarkoidose tritt eine abrupt ausbrechende starke Entzündungsreaktion auf, die jedoch einen spontanen Heilungsverlauf innerhalb von zwei Jahren nimmt. Eine besondere Form der *akuten* Sarkoidose stellt das Löfgren-Syndrom dar (Lofgren 1953). Charakteristisch für das Löfgren-Syndrom ist eine Vergrößerung der Lymphknoten, des Weiteren treten akute Entzündungen der Subkutis, meist an den Unterschenkeln, Füßen und Händen, mit einem Durchmesser bis 10 cm auf.

II) Bei der *chronischen* Sarkoidose zeigen sich zunächst nur schwache Entzündungsreaktionen, die sich mit dem Fortschreiten der Krankheit verschlimmern und zu Funktionsverlust und permanenter Schädigung der betroffenen Organe führen können. Generell ist der *chronische* Subphänotyp definiert durch länger als zwei Jahre andauernde Symptome (American Thoracic Society 1999) und führt in einigen schweren Fällen durch Schädigung der Atemwege, des Herzens oder des Zentralnervensystems schließlich zum Tod (Thomas and Hunninghake 2003).

Eines der universell charakteristischen Merkmale der Sarkoidose ist die Granulombildung innerhalb des Bindegewebes betroffener Organe (Zissel et al. 2010), die in allen Formen der Sarkoidose beobachtet wird. Trotz der heterogenen klinischen Erscheinungsform der Krankheit deutet die Granulombildung darauf hin, dass alle Subgruppen der Krankheit gemeinsame Faktoren haben

müssen. Die Granulome der Sarkoidose besitzen normalerweise keinen nekrotisierenden Kern, es wurden allerdings in einigen Fällen Granulome mit einem nekrotisierenden Kern beobachtet (Kunitake et al. 1999; Rosen 2007; Yeboah et al. 2012) (Abb. 1-7).

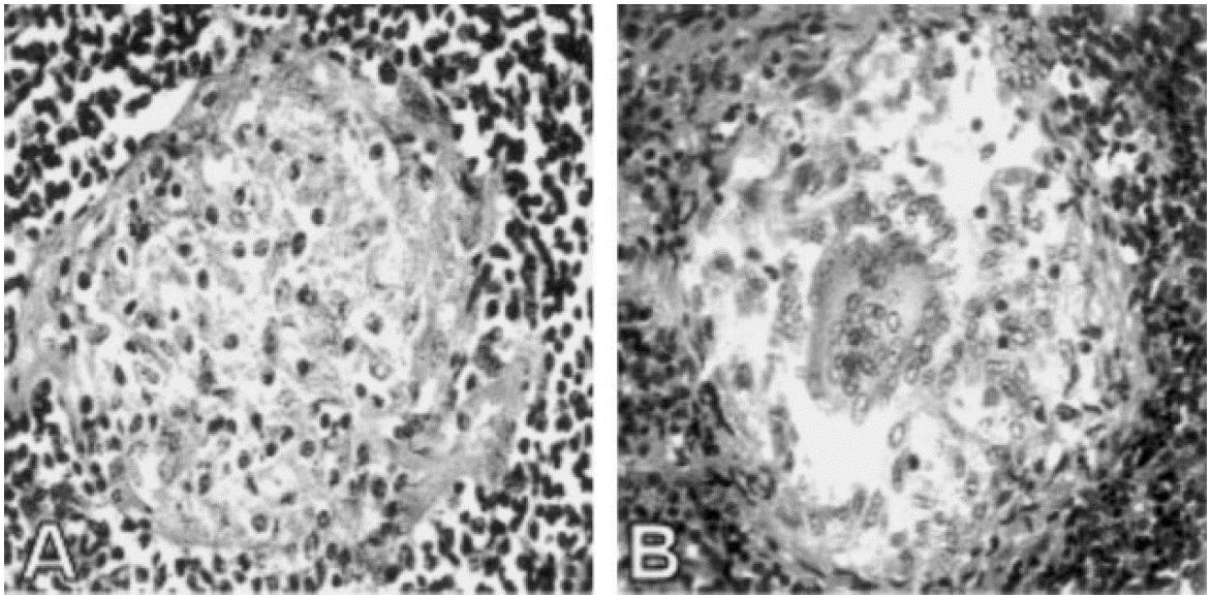

Abb. 1-7: (A) Nicht-nekrotisierendes Sakoidosegranulom. **(B)** Nicht-nekrotisierendes Sarkoidosegranulom mit multinukleären Riesenzellen. Nach (Rosen 2007).

Die Granulombildung wird häufig als Hinweis auf eine zugrunde liegende mikrobielle Infektion oder als Hinweis auf ein anderes fremdes Antigen als Krankheitsauslöser herangezogen (Chen et al. 2008; Gupta et al. 2007; Z. Song et al. 2005). Der Tuberkuloseerreger *Mycobacterium tuberculosis* (Oswald-Richter et al. 2010; Oswald-Richter et al. 2012; Richmond and Drake 2010; Svendsen et al. 2011) wird schon lange auch als Krankheitserreger der Sarkoidose diskutiert, da er z.B. von letzterer klinisch und pathologisch nicht unterscheidbare Hautläsionen verursachen kann. Bei der Suche nach auslösenden Faktoren konnte auch gezeigt werden, dass Sarkoidosepatienten Antikörper gegen mikrobielle Hitzeschock-Proteine besitzen (Dubaniewicz et al. 2012). Zudem wurden T-Zellen, die die Katalase-Peroxidase (mKatG) aus *M. tuberculosis* erkennen und darauf reagieren, bei Patienten mit aktiver Sarkoidose, nicht aber bei Patienten mit inaktiver Sarkoidose nachgewiesen (Chen et al. 2008). Es werden im Zusammenhang mit der Krankheit aber auch Bakterien wie *Rikettsia helvetica, Propionibacterium acnes* und *Borrelia burgdorferi* sowie verschiedene Viren als Auslöser diskutiert (Ezzie and Crouser 2007), jedoch konnte bis heute kein krankheitsauslösendes Antigen identifiziert werden.

1.2.1.1 Sarkoidose: Eine komplexe Krankheit

Nicht nur herkömmliche Krankheitserreger werden als mögliche Ursache für die Krankheit angesehen, auch verschiedene Umwelteinflüsse, genetische Prädisposition und vor allem die Wechselwirkung zwischen diesen Faktoren stehen im Verdacht zur Entstehung von Sarkoidose beizutragen oder sie gar zu verursachen (Abb. 1-8). Eine höhere Erkrankungsrate bestimmter Berufsgruppen, wie z.B. bei Feuerwehrleuten, Angehörigen des US-Militärs, Berufen in der Landwirtschaft und im Gesundheitswesen, deuten auf einen Einfluss der Umweltfaktoren auf die Krankheitsentstehung hin (L. S. Newman et al. 2004). Zudem gibt es Hinweise, dass eine Reihe von organischen und anorganischen Substanzen wie z.B. solche aus Nadelbäumen und Tonböden sowie Zirkonium, Aluminium und Talk eine krankheitsinduzierende Wirkung haben können. Eine der Studien konnte zum Beispiel zeigen, dass die Arbeiter, die nach dem World Trade Center Unglück bei den Aufräumarbeiten halfen, ein erhöhtes Risiko haben an Sarkoidose zu erkranken (Jordan et al. 2011). Die Beobachtung dieser zeitlichen und räumlichen Häufungen sowie die Übertragbarkeit der Krankheit durch Organspenden (Burke et al. 1990) unterstützen die Theorie, dass Sarkoidose durch ein Antigen ausgelöst wird.

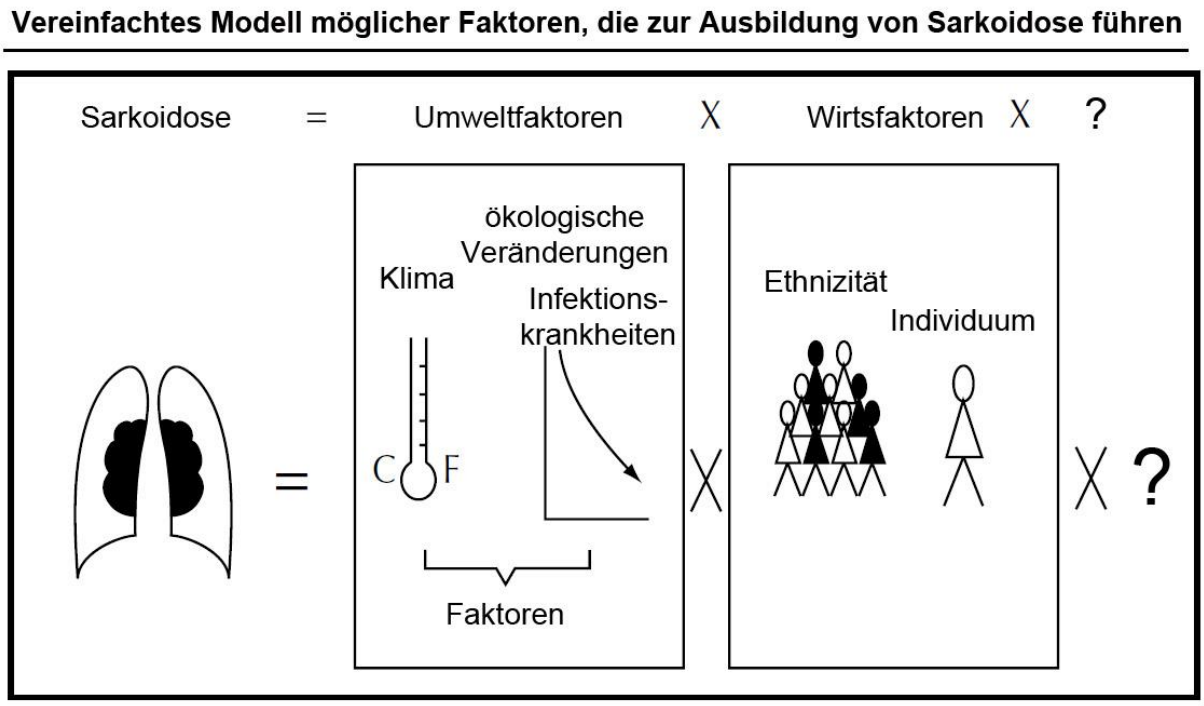

Abb. 1-8: Ältere schematische Darstellung der Faktoren, die der Sarkoidose zugrunde liegen. Nach (Y. Hosoda et al. 2002).

Es gibt aber auch umfangreiche Hinweise auf eine genetische Prädisposition für Sarkoidose. Bei 5% - 16% der Patienten wird eine familiäre Häufung der Krankheit beobachtet (Rybicki et al. 2001) und Zwillingsstudien mit monozygoten und dizygoten Zwillingen zeigten, dass monozygote Zwillinge mit einem erkrankten Geschwisterpaar ein 80-fach höheres Risiko haben, ebenfalls an Sarkoidose zu

erkranken, als der Durchschnittsbürger. Das Risiko für dizygote Zwillinge ist im Vergleich deutlich geringer, aber immerhin noch 7-fach höher als bei Durchschnittsbürgern (British Thoracic and Tuberculosis Association 1973; Sverrild et al. 2008). Sarkoidose tritt in verschiedenen Bevölkerungsgruppen mit unterschiedlicher Inzidenz auf (Yutaka Hosoda et al. 1997), so erkranken z.B. Afro-Amerikaner im Vergleich zu Amerikanern mit weißer Hautfarbe besonders häufig an Sarkoidose (35/100.000 zu 11/100.000) (Rybicki et al. 1997).

Mittlerweile konnten mit Kopplungsstudien und GWAS eine Reihe von Loci als Risikogene für Sarkoidose identifiziert werden: Das Immunsystem spielt in der Pathogenese der Krankheit eine wichtige Rolle und Studien konnten zeigen, dass Varianten von Klasse-II-Antigenen des humanen Leukozytenantigen-Systems (engl. *human leukocyte antigen,* HLA) populationsübergreifend zur Krankheitsanfälligkeit beitragen. Die stärksten Assoziationen wurden bisher mit den *DRB1*-Allelen festgestellt und es scheint ein allgemeines HLA-Klasse-II Assoziationsmuster zu geben, wobei *HLA-DRB1*01* und *DRB1*04* schützend wirken und *DRB1*03*, *DRB1*11*, *DRB1*12*, *DRB1*14* und *DRB1*15* Risikofaktoren für Sarkoidose darstellen (Grutters et al. 2003). Die Gene des HLA-Systems codieren für Antigen-präsentierende Proteine, welche für die Immunerkennung wichtig sind. Neben den Genen, die für die sechs HLA-Hauptproteine codieren (HLA-A, B, C, DR, DP und DQ), existiert in dem HLA-Komplex eine große Anzahl weiterer Gene, welche größtenteils in der Immunabwehr involviert sind. Mutationen in diesen Genen können zur Anfälligkeit gegenüber bestimmten Krankheiten beitragen oder aber auch zu Autoimmunreaktionen führen. Über eine Kopplungsstudie konnte das Gen *BTNL2*, welches ebenfalls im HLA-Komplex liegt, als Risikofaktor für Sarkoidose identifiziert werden (Rybicki et al. 2005; Valentonyte et al. 2005). Es konnte gezeigt werden, dass das von *BTNL2* codierte Protein eine Rolle in der Aktivierung von T-Zellen spielt (Nguyen et al. 2006).

Doch auch außerhalb des HLA-Komplexes liegende Gene tragen zur Anfälligkeit gegenüber Sarkoidose bei, z.B. konnte mit einer Kandidatengenstudie das Gen *IL23R* als ein weiterer Risikofaktor identifiziert werden (Fischer et al. 2011). Das Gen codiert für den Interleukin-23-Rezeptor, welcher sich auf verschiedenen Zellen des Immunsystems befindet und an der Erkennung von körperfremden Substanzen beteiligt ist. Weitere davor nicht mit der Krankheitsentstehung in Verbindung gebrachte Risikoloci konnten durch GWAS entdeckt werden: *ANXA11* (Hofmann et al. 2008), *RAB23* (Hofmann et al. 2011), *OS9* (Hofmann et al. 2013) und *NOTCH4* (Adrianto et al. 2012).

Alle diese Gene konnten mit für die Sarkoidose relevanten Funktionen in Verbindung gebracht werden, doch scheinen diese jeweils verschiedene Funktionen und Signaltransduktionswege zu betreffen. Während das von *ANXA11* codierte Protein Annexin A11 eine Rolle im Signalübertragungsweg der Apoptose von aktivierten Entzündungszellen zu spielen scheint (zusammengefasst in Gerke and Moss 2002; Moss and Morgan 2004), wirkt das RAB23-Protein

anscheinend als Antagonist für den Hedgehog-Signalweg (Evans et al. 2005), welcher eine Rolle in der CD4+ T-Zellen-Aktivierung spielt (Stewart et al. 2003). Das Protein OS9 interagiert mit dem Protein DC-STAMP, welches wiederum eine Funktion in der Bildung von multinukleären Riesenzellen zu haben scheint (Jansen et al. 2009). NOTCH-Proteine spielen hingegen eine Rolle in der Steuerung der Zelldifferenzierung und es gibt Anzeichen dafür, dass NOTCH-Proteine an der Differenzierung und Aktivität von T-Zellen beteiligt sind (zusammengefasst in Maillard et al. 2003).

Viele verschiedene Signalübertragungswege können also in der Pathogenese der Sarkoidose eine Rolle spielen und erklären vielleicht teilweise die beobachtet Heterogenität der Krankheit.

Den klinischen Unterschieden in den beiden Subphänotypen der Sarkoidose entsprechen zum Teil auch Unterschiede auf der genetischen Ebene. Die Loci *HLA-DRB1*03* (Grunewald and Eklund 2009), *DRB1*0301* (Sato et al. 2010), *DQB1*0201* (Sato et al. 2010), *CCR2* (Spagnolo et al. 2003), *MHC2TA* (Grunewald et al. 2010) und *OS9* (Hofmann et al. 2013) scheinen eher zu dem *akuten* Subphänotyp beizutragen, während *BTNL2* (Li et al. 2006; Rybicki et al. 2005) und *IL23R* (Fischer et al. 2010) mit dem *chronischen* Subphänotyp assoziiert sind. Individuelle bedingte genetische Risikofaktoren beeinflussen demnach also die Wahrscheinlichkeit einen bestimmten Sarkoidose-Subphänotyp zu entwickeln.

1.2.2 Tuberkulose

Tuberkulose ist eine systemische entzündliche bakterielle Infektionskrankheit, deren charakteristisches Merkmal nekrotisierende Granulome sind. Früher auch unter Schwindsucht oder Morbus Koch bekannt, ist Tuberkulose heutzutage eine der tödlichsten Infektionskrankheiten überhaupt. Im Jahr 2011 wurden weltweit 8,7 Millionen Neuerkrankungen und 1,4 Millionen Todesfälle im Zusammenhang mit Tuberkulose durch die Weltgesundheitsorganisation (engl. *world health organisation*, WHO) registriert (WHO 2012). Tuberkulose betrifft hauptsächlich junge Erwachsene, aber auch alle anderen Altersgruppen und wird durch verschiedene Stämme von Mykobakterien (in erster Linie durch das *Mycobacterium tuberculosis*, aber auch durch *M. bovis*, *M. africanum*, *M. canetti* und *M. microti*) ausgelöst (Russell 2011; van Soolingen et al. 1997). Die Infektion manifestiert sich hauptsächlich in der Lunge (pulmonale Tuberkulose), da das Antigen durch Tröpfcheninfektion übertragen wird, allerdings kann sich eine extrapulmonale Tuberkulose auch im gesamten Körper ausbreiten und jedes andere Organ befallen (Ramesh et al. 1987).

Tuberkulose ist evolutionsgeschichtlich gesehen eine sehr alte Krankheit, es wurden Anzeichen einer Infektion bei etwa 500.000 Jahre alten Fossilien eines *Homo erectus* gefunden (Kappelman et al. 2008) und Spuren von *M. tuberculosis* konnten in ca. 9.000 Jahren alten menschlichen Knochen nachgewiesen werden (Hershkovitz et al. 2008). Epidemiologen vermuten, dass aktuell ungefähr ein

Drittel der Weltbevölkerung mit *M. tuberculosis* infiziert ist, doch sind die meisten Infektionen latent und nur aus ungefähr einer von zehn Infektionen entwickelt sich eine aktive Tuberkulose (Enarson and Murray 1996; WHO 2012). Die weltweite Inzidenz variiert stark, während sie in Deutschland bei 5,7/100.000 Einwohnern liegt, geht man in Afrika von einer Inzidenz von 351/100.000 aus (WHO 2012). Die Behandlung der Krankheit ist schwierig und benötigt die Anwendung verschiedener Antibiotika über einen langen Zeitraum, daher sind hohe Tuberkuloseraten besonders in Ländern mit schlechter Gesundheitsversorgung zu beobachten (z.B. in Teilen Asiens und Afrikas) (WHO 2012).

Ist *M. tuberculosis* in die Lunge gelangt, wird es dort von Makrophagen und dendritischen Zellen phagozytiert. Im Phagosom der Makrophagen wird das Bakterium jedoch nicht abgebaut, da die Zellwand nicht aufgebrochen werden kann. Bestandteile der Zellwand (das Glycolipid Lipoarabinomannan (ManLAM)) des *M. tuberculosis* können die Fusion des Phagosoms mit dem Lysosom verhindern und außerdem eine Apoptose der Zellen unterbinden. Obendrein wird die Aktivierung der infizierten Makrophagen über den Interferon-Gamma (IFN-γ) Signalweg gestört (Briken et al. 2004). Die infizierten Makrophagen wandern über das Alveolarepithel in das Lungengewebe. Dort wird dann ein Granulom gebildet (Dannenberg 1993). Die erfolgreiche Einkapselung des Pathogens im Granulom führt in über 90% der Fälle zu einer latenten Infektion, indem das Immunsystem diesen kontrollierten Zustand aufrechterhalten kann und eine aktive Tuberkulose verhindert (Marks et al. 2000; Ulrichs and Kaufmann 2002; Y. Zhang 2004) (Abb. 1-9 und Abb. 1-10).

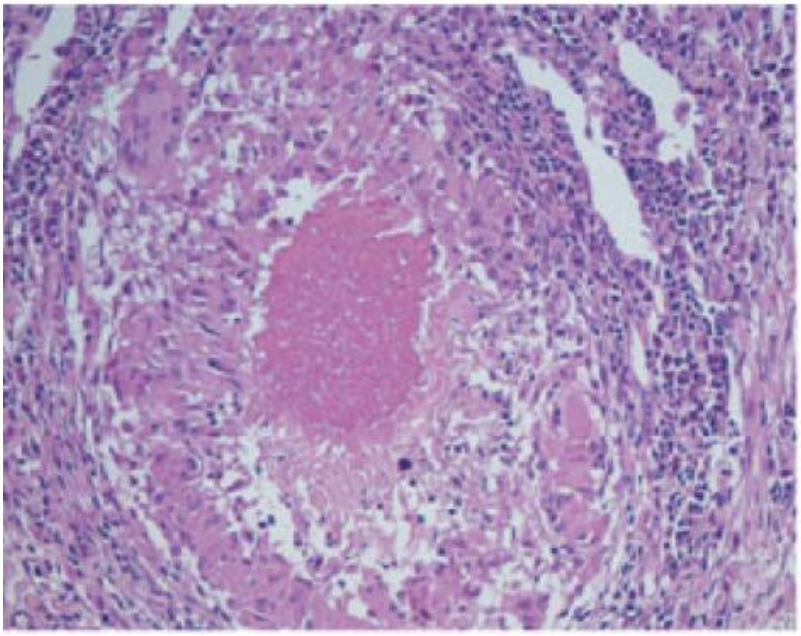

Abb. 1-9: Humanes nekrotisierendes Tuberkulosegranulom in der Lunge (x100). Nach (M. J. Kim et al. 2010).

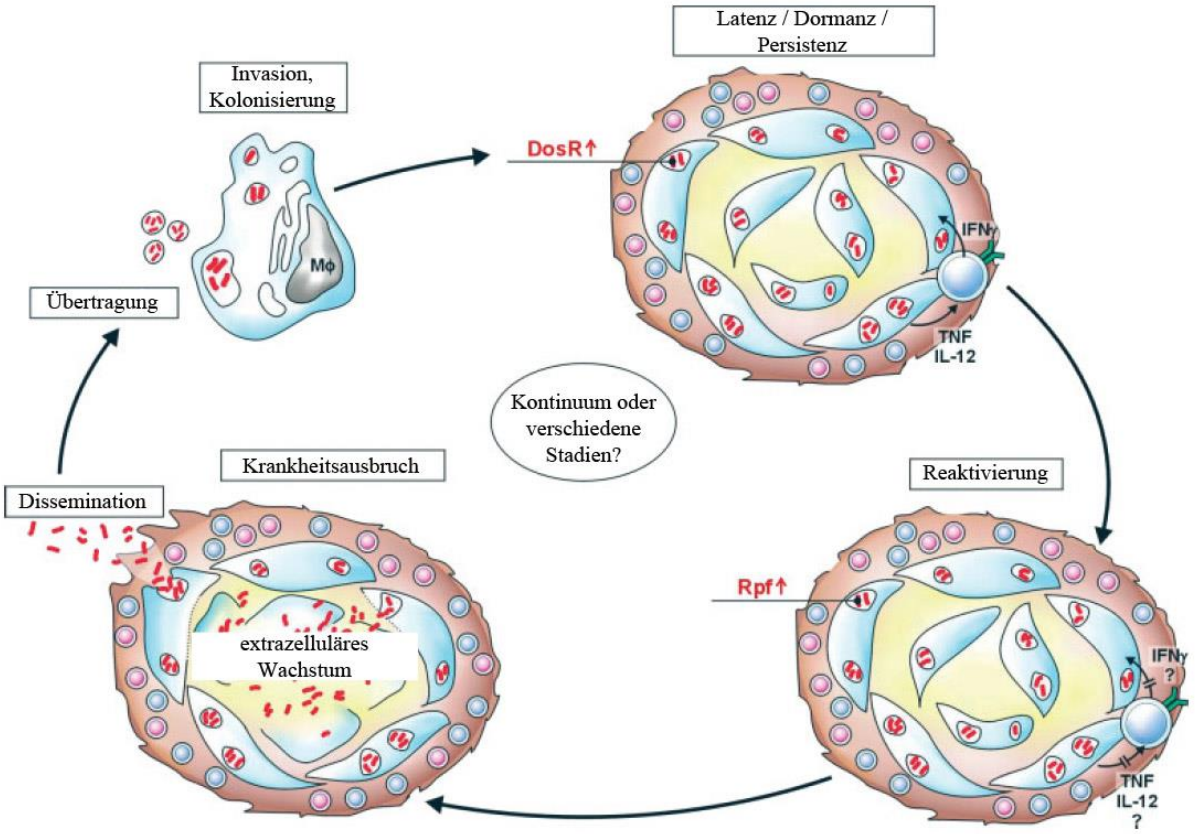

Abb. 1-10: Ablauf einer Tuberkuloseinfektion mit Lantenzphase. Nach der Invasion von Makrophagen, überlebt M. tuberculosis intrazellulär in den Granulomen. Der Transkriptionsfaktor DosR fördert die Produktion von Sensor-Kinasen, welche die Wirtsantworten auf Hypoxie und Stickoxidstress erkennen und regulieren können. Das DosR-System ist ein entscheidender Faktor für die mykobakterielle Dormanz. Durch Veränderungen der Bedingungen (z.B. ein geschwächtes Immunsystem) kommt es zur Produktion von Rpf-Proteinen (engl. *resuscitation-promoting factor*) und eine Reaktivierung der Mykobakterien tritt ein. Die mit M. tuberculosis infizierten Makrophagen sterben ab und extrazelluläres Wachstum der Mykobakterien tritt ein. Kann das Immunsystem die Struktur des Granuloms nicht mehr aufrechterhalten, breitet sich M. tuberculosis im Körper aus und infiziert weitere Makrophagen. Nach (Ulrichs and Kaufmann 2006).

1.2.2.1 Infektionskrankheit oder Wirtsprädisposition?

M. tuberculosis und die anderen Stämme des *Mycobacterium tuberculosis complex* sind als Auslöser von Tuberkulose anerkannt. Nach einer Infektion entwickelt sich aber nur in ungefähr 10% der Fälle eine aktive Tuberkulose (Marks et al. 2000; Ulrichs and Kaufmann 2002; Y. Zhang 2004). Dies zeigt, dass weitere Faktoren eine Rolle bei der Pathogenese spielen müssen. Das Risiko einer Infektion nach dem Kontakt mit dem Antigen wird hauptsächlich von exogenen Faktoren beeinflusst wie z.B. von der Infektiosität des Überträgers, von der Nähe des Kontakts, vom Zugang zu medizinischer Versorgung und von Umweltfaktoren (z.B. Rauchen, Alkoholkonsum und Luftverschmutzung in Innenräumen). Endogene (wirtsspezifische) Faktoren wiederum beeinflussen das Fortschreiten von der Infektion zur aktiven Krankheit, hierzu zählen z.B. Alter, Geschlecht, Fehl- oder Unterernährung, Diabetes und Immunstatus des betroffenen Individuums. Die Bedeutung des Immunsystems im Zusammenhang mit der Erkrankung wird dadurch unterstrichen, dass bei HIV-positiven (engl. *human*

immunodeficiency virus, HIV) und anderen immunsupprimierten Individuen das Risiko, eine aktive Tuberkulose zu entwickeln, um ungefähr 10% pro Jahr erhöht ist (Corbett et al. 2003; Girardi et al. 2000). Mittlerweile konnten durch Kopplungsstudien, Zwillingsstudien, Kandidatengenstudien und einige wenige GWAS schon etliche Genpolymorphismen mit der Prädisposition für Tuberkulose in Verbindung gebracht werden. Neben Polymorphismen in der HLA-Region (Singh et al. 1983) konnten bisher die Gene *SLC11A1* (*NRAMP1*) (Cervino et al. 2000; Greenwood et al. 2000; Meilang et al. 2012), *IFNG* (Pacheco et al. 2008; Vallinoto et al. 2010), *EREG* (Thuong et al. 2012), *P2X7* (Singla et al. 2012; Xiao et al. 2010), *BTNL2* (Lian et al. 2010), *IL-10* (Ates et al. 2008; J. Zhang et al. 2011b) und *CCL2* (Feng et al. 2012) als mit der Tuberkulosesuszeptibilität zusammenhängend identifiziert werden. Dies zeigt deutlich, dass neben einer Infektion mit *M. tuberculosis* die genetische Prädisposition eine wesentliche Rolle bei der Krankheitsentstehung zu spielen scheint. Auch Zwillingsstudien untermauern einen genetischen Aspekt der Krankheit, da bei monozygoten Zwillingen eine signifikant höhere Konkordanz als bei dizygoten Zwillingen beobachtet wurde (Comstock 1978). Auf Grund verschiedener Zwillingsstudien wird bei der Tuberkulose von einer starken Erblichkeit (36% bis 80%) der Krankheit ausgegangen (Comstock 1978; Jepson et al. 2001; Kallmann and Reisner 1943; Kimman et al. 2006; Newport et al. 2004).

1.3 Gemeinsame pathologische Kennzeichen von Tuberkulose und Sarkoidose

Sowohl bei Sarkoidose (SA) als auch bei Tuberkulose (TB) ist die Granulombildung das charakteristische Merkmal der Krankheit. Generell ist ein Granulom eine Reaktion des Immunsystems auf ein langlebiges und schwer abbaubares Antigen, welches eine lokale von Typ1-T-Helferzellen vermittelte Immunantwort auslöst. Alveolare Phagozyten (Makrophagen und dendritischen Zellen) nehmen über Phagozytose die Mykobakterien (TB) oder das bisher unbekannte Antigen (SA) auf. Einige dieser Zellen wandern dann in die sekundären lymphatischen Organe (Lymphknoten) und aktivieren T-Lymphozyten (Bhatt et al. 2004; Demangel et al. 2002; Khader et al. 2006). In den lymphatischen Organen differenzieren sich nach der Stimulation hauptsächlich CD4+ T-Zellen, doch auch CD8+ T-Zellen werden gebildet. Die aktivierten T-Zellen wandern zum Ort der Entzündung, akkumulieren um die Makrophagen herum und sekretieren dort Zytokine (z.B. IFN-γ, TNF-α, IL-2). Zytokine spielen bei Differenzierungs- und Rekrutierungsprozessen der verschiedenen Zelltypen der adaptiven Immunantwort eine wichtige Rolle. Durch die Zytokine (wie IFN-γ und TNF-α) differenzieren sich einige der Makrophagen in Epitheloidzellen und multinukleäre Riesenzellen. Das Zentrum des Granuloms besteht aus nekrotisierenden (TB) oder nicht-nekrotisierenden (SA) Makrophagen. Nur in seltenen Fällen konnten bei Sarkoidose nekrotisierende Granulome beobachtet

werden (zusammengefasst in Yeboah et al. 2012). Um diesen Kern formieren sich konzentrische Lagen von Makrophagen, Epitheloidzellen, multinukleäre Riesenzellen, Schaumzellen, T-Lymphozyten und B-Lymphozyten (Abb. 1-11 und Abb. 1-12) (Aaron et al. 2004; Agostini et al. 2000; Boros 1978; Ducati et al. 2006; Flynn and Chan 2001; Gal and Koss 2002; Grunewald and Eklund 2007; Mariano 1995; J. M. Reich 2012; Rosen 2007; Thomas and Hunninghake 2003; Ulrichs and Kaufmann 2006). Im Zentrum dieses Gebildes befinden sich vornehmlich CD4+ T-Zellen, während CD8+ T-Zellen hauptsächlich in der Peripherie des Granuloms akkumulieren.

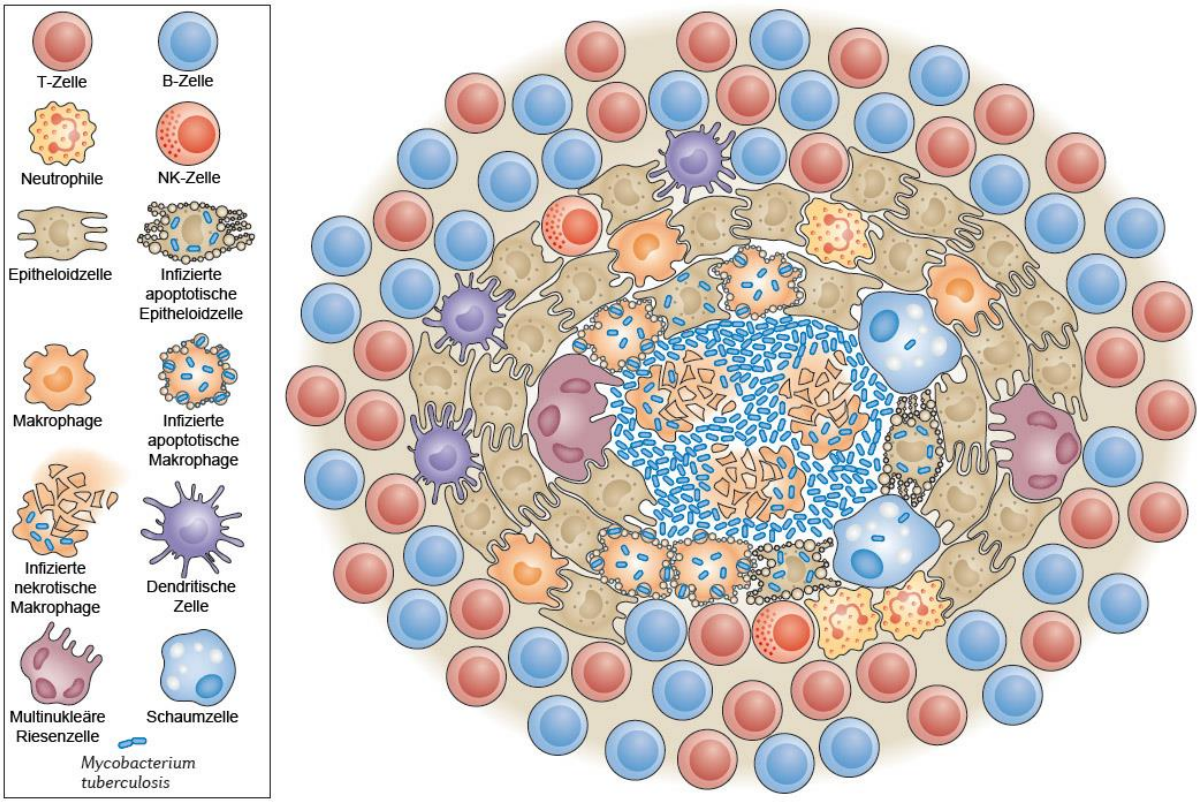

Abb. 1-11: Struktur und zelluläre Bestandteile eines Tuberkulosegranuloms. In seiner einfachsten Form kann man ein Granulom als ein kompaktes, organisiertes Aggregat von Epitheloidzellen (differenzierte Makrophagen, welche durch ineinander greifende Zellmembranen mit angrenzenden Zellen verbunden sind) bezeichnen. Epitheloidzellen können eine hohe phagozytische Aktivität aufweisen, in einigen Fällen aber enthalten sie überhaupt gar keine Bakterien. Die Makrophagen im Granulom können auch zu multinukleären Riesenzellen fusionieren oder sich zu Schaumzellen differenzieren. Schaumzellen lokalisieren sich am häufigsten an der Außenzone des nekrotischen Zentrums des voll ausgebildeten Granuloms. Die Gründe für diese Veränderungen sind nicht bekannt und allgemein wurden bisher in Schaumzellen und multinukleären Riesenzellen selten phagozytierte Bakterien beobachtet. Bakterien befinden sich hauptsächlich in dem nekrotischen Zentrum, in welchem sich tote und absterbende Makrophagen befinden. Des Weiteren sind viele andere Zelltypen in einem Granulom zu finden: Neutrophile, dendritische Zellen, B- und T-Zellen, natürliche Killerzellen (NK Zellen), Fibroblasten und Zellen, die extrazelluläre Matrixkomponenten sekretieren. Mittlerweile wird vermutet, dass auch die das Granulom umgebenden Epithelzellen (nicht gezeigt) an der Formation beteiligt sind. Nach (Ramakrishnan 2012).

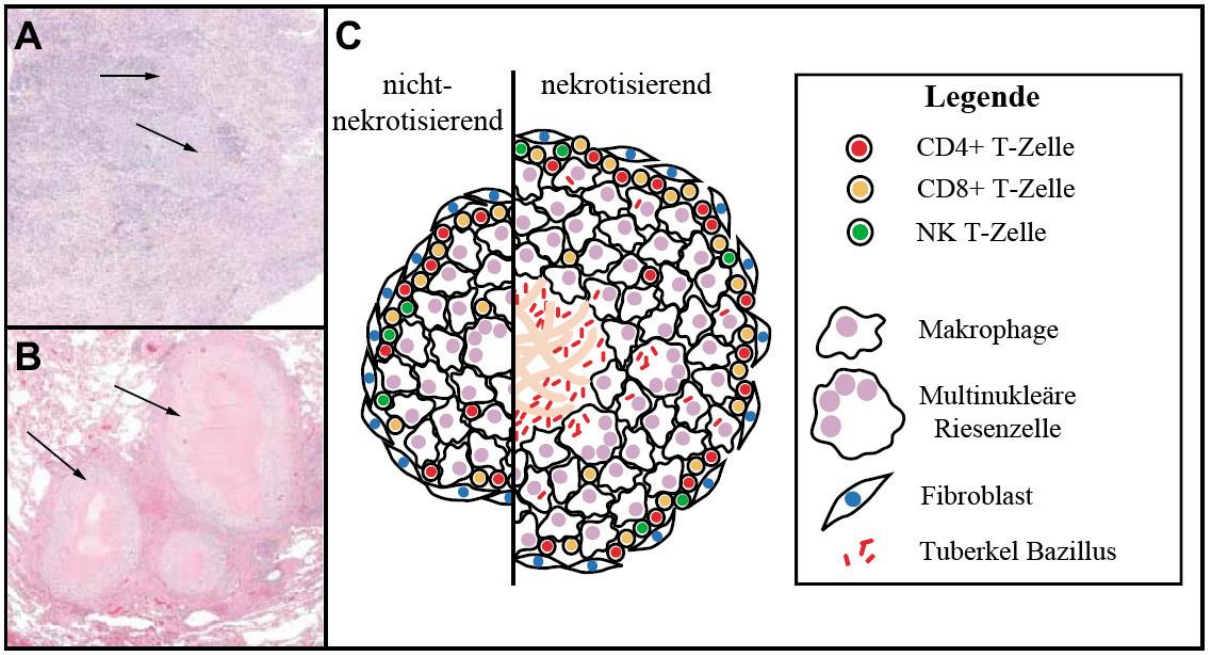

Abb. 1-12: Struktur humaner Granulome. (A) Nicht-nekrotisierende Sarkoidosegranulome (Pfeile). **(B)** Nekrotisierende Tuberkulosegranulome (Pfeile). **(C)** Schematische Darstellung von nicht-nekrotisierenden und nekrotisierenden Granulomen. Abkürzungen: Natürliche Killer T-Zelle (NK T Zelle). Nach (Co et al. 2004).

Der fast identische Aufbau der Granulome zeigt, dass beiden Krankheiten anscheinend die gleichen Mechanismen für die Granulombildung zugrunde liegen. Die Granulombildung ist ein sehr komplexer Vorgang, bei dem viele Faktoren eine Rolle spielen. Eine Fehlregulation in diesem komplexen Netzwerk an beteiligten Faktoren könnte somit zur Pathogenese der Krankheiten beitragen: So konnte bei Sarkoidosepatienten im Vergleich zu gesunden Kontrollpersonen ein Ungleichgewicht innerhalb der T-Zellpopulationen an den Orten der Entzündung beobachtet werden (Grunewald and Eklund 2007; Ho et al. 2005; Rossi et al. 1984).

Analysen des Genexpressionsprofils von Blutzellen bei Tuberkulose- und Sarkoidosepatienten im Vergleich mit gesunden Kontrollpersonen zeigten eine hohe Übereinstimmung im Muster der Genregulation in beiden Patientengruppen (Koth et al. 2011; Maertzdorf et al. 2012). Maertzdorf *et al.* (2012) konnten insgesamt nur vier Gene identifizieren, die in ihrer Regulation einen signifikanten Unterschied zwischen Tuberkulose- und Sarkoidosepatienten zeigen. Auch wenn bei den beiden Krankheiten noch weitere Faktoren eine Rolle spielen, machen diese Daten wahrscheinlich, dass bei beiden Krankheiten gleiche oder zumindest sehr ähnliche Immunprozesse ablaufen und somit auf gemeinsame zugrunde liegende pathologische Prozesse hindeuten.

Des Weiteren befallen beide Krankheiten identische Organe. Am häufigsten wird die Lunge befallen (TB: ~80%; SA: ~90%), aber auch Lymphsystem, Zentralnervensystem, Skelett, Haut, Augen und alle weiteren Organe können bei beiden Krankheiten betroffen sein (American Thoracic Society 1999, 2000; Costabel 2001; Golden and Vikram 2005). Die Gemeinsamkeiten zwischen den Krankheiten

sind offensichtlich und in einigen früheren Veröffentlichungen wurde Sarkoidose auch als eine Unterform von Tuberkulose angesehen (zusammengefasst in L. Newman 2005). Diese weitreichenden Überschneidungen legen den Schluss nahe, dass auch auf genetischer Ebene Überschneidungen vorhanden sind und gemeinsame genetische Faktoren für Sarkoidose und Tuberkulose existieren.

1.4 Ziele der Arbeit

1.4.1 Untersuchung von genetischen Faktoren für entzündliche granulomatöse Lungenkrankheiten

1.4.1.1 Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen

Die Krankheitsentstehung von Sarkoidose ist sehr komplex und bisher weitestgehend unbekannt. Neben genetischen Faktoren scheinen auch Umweltfaktoren in der Ätiologie beteiligt zu sein, doch obwohl verschiedene Antigene in diesem Zusammenhang diskutiert werden, konnte bisher keines als Ursache für die Krankheit identifiziert werden. Aufgrund dieser Komplexität stellt die Identifizierung der der Sarkoidose zugrundeliegenden Mechanismen eine große Herausforderung dar. Die von Sverrild *et al.* (2008) in einer Zwillingsstudie berechnete Erblichkeit von Sarkoidose beträgt 66% und lässt somit auf eine starke genetische Komponente bei der Ausprägung der Krankheit schließen. Die bisher entdeckten Risikoloci können diese Erblichkeit der Krankheit jedoch nur zu einem kleinen Teil erklären, daher liegt die Vermutung nahe, dass weitere Risikoloci für Sarkoidose existieren. Das Ziel dieser Arbeit ist es weitere existierende Genvarianten, welche zur Krankheitsentstehung beitragen, zu identifizieren, genau zu lokalisieren und zu untersuchen. Dafür werden in einem hypothesenfreien Ansatz die Genotypdaten von 1.294.967 SNPs eines imputierten GWAS-Datensatzes analysiert.

Ein interessantes Merkmal der Sarkoidose ist das Auftreten der zwei im Krankheitsverlauf sehr unterschiedlichen Subphänotypen (siehe Kapitel 1.2.1). Einige genetische Polymorphismen, die mit Sarkoidose im Zusammenhang stehen, konnten schon distinkt mit einem der beiden Subphänotypen assoziiert werden (Fischer et al. 2010; Grunewald and Eklund 2009; Grunewald et al. 2010; Hofmann et al. 2013; Rybicki et al. 2005; Sato et al. 2010; Spagnolo et al. 2003). Genetische Faktoren haben also eine Auswirkung auf das beim Patienten auftretende Krankheitsbild, und da bisher andere Sarkoidose auslösende Faktoren nicht zweifelsfrei identifiziert werden konnten, spielen offensichtlich genetische Faktoren für die Ausbildung der Krankheit und ihrer Unterformen eine große Rolle. Genetische Polymorphismen könnten also helfen, die der Krankheit zugrunde liegenden Mechanismen weiter zu entziffern und zu verstehen. Mit den vorliegenden Daten sollen daher darüber hinaus die genetischen Unterschiede zwischen dem *akuten* und *chronischen* Sarkoidose-

Subphänotyp untersucht werden, mit dem Ziel genetische Faktoren zu identifizieren, welche spezifisch die Ausprägung eines der beiden Subphänotypen bedingen. Dafür stehen für die Analyse zwei Stichproben (GWAS und Validierungsstichprobe) mit ausführlichen Phänotypinformationen zur Verfügung.

1.4.1.2 Identifizierung gemeinsamer genetischer Faktoren der Krankheiten Tuberkulose und Sarkoidose

Die Beziehung zwischen Tuberkulose und Sarkoidose ist bis heute Gegenstand der Forschung. Schon vor über 110 Jahren vermutete Caesar Boeck in einer Sarkoidose-Fall-Beschreibung eine Verwandtschaft mit Tuberkulose (Boeck 1899). In beiden Krankheiten werden körperliche Symptome wie Fieber, Unwohlsein, Ermüdungserscheinungen, Anorexie und Gewichtsverlust beobachtet und beide sind in ihrem Phänotyp schwer zu unterscheiden. Bei Sarkoidosepatienten wurde zudem gehäuft ein vorausgehender Kontakt mit Tuberkulosepatienten beobachtet, was auf eine mögliche epidemiologische Verbindung zwischen den Krankheiten hindeutet (Bunn and Johnston 1972; Parsons 1960) und in ethnischen Minderheiten mit einer hohen Inzidenz für Tuberkulose wurden auch vermehrt Sarkoidosefälle beobachtet (Brett 1965). Das *Mycobacterium tuberculosis* wurde daher seitdem immer wieder als kausatives Antigen der Sarkoidose vermutet, doch konnte diese Annahme bisher nicht einwandfrei bestätigt werden (Brownell et al. 2011). Wie schon oben für die Sarkoidose beschrieben, weisen auch bei der Tuberkulose Zwillingsstudien auf eine hohe Erblichkeit (36% bis 80%) der Krankheit und damit auf eine starke genetische Komponente hin (Comstock 1978; Jepson et al. 2001; Kallmann and Reisner 1943; Kimman et al. 2006; Newport et al. 2004).

Das Granulom stellt bei beiden Krankheiten das charakteristische histologische Merkmal dar und zeigt, wie eingangs beschrieben, einen nahezu identischen Aufbau. Die Granulombildung wird in beiden Krankheiten von einer Immunantwort auslöst, die von Typ1-T-Helferzellen vermittelt wird. Bei einer so komplexen Gewebeneubildung wie bei der eines Granuloms muss eine Vielzahl von Genen aktiviert oder inhibiert werden und eine Vielzahl von verschiedenen Zelltypen ist an diesem Vorgang beteiligt und muss reguliert werden. Das Genexpressionsprofil der Blutzellen von Tuberkulose- und Sarkoidosepatienten gibt zudem einen weiteren Hinweis darauf, dass in den beiden Krankheiten identische immunologische Reaktionen ablaufen (Koth et al. 2011; Maertzdorf et al. 2012). Mutationen in den an diesen Vorgängen beteiligten Genen könnten daher zu einer generellen Anfälligkeit gegenüber solchen entzündlichen granulomatösen Lungenkrankheiten beitragen. Studien konnten Mutationen des *BTNL2*-Gens mit Sarkoidose und Tuberkulose in Verbindung bringen (Lian et al. 2010; Rybicki et al. 2005; Valentonyte et al. 2005), wobei der SNP rs2076530 zur Sarkoidose- und der SNP rs9268492 zur Tuberkulose-Suszeptibilität beizutragen zu scheint (Lian et al. 2010).

Bei anderen entzündlichen Krankheiten wurde schon nachgewiesen, dass sie auf gemeinsamen Gendefekten beruhen (z.B. Sarkoidose und Morbus Crohn (Franke et al. 2008) oder Morbus Crohn und Psoriasis (D. Ellinghaus et al. 2012)).

Ein Ziel dieser Arbeit ist es, in einer kombinierten Analyse der Krankheiten Sarkoidose und Tuberkulose genetische Faktoren aufzudecken, die in den Mechanismen eine Rolle spielen, die den beiden Krankheiten zugrunde liegenden, und somit vielleicht allgemeine genetische Faktoren entzündlicher granulomatöser Lungenkrankheiten darstellen. Diese Fragestellung soll mit Hilfe einer gemeinsamen Analyse einer imputierten Sarkoidose-GWAS und einer imputierten Tuberkulose-GWAS in einem hypothesenfreien Ansatz bearbeitet und mit zwei unabhängigen Replikationsphasen bestätigt werden.

2 Material und Methoden

2.1 Material

2.1.1 Patienten- und Kontrollstichproben

Die folgende Tabelle 2-1 zeigt einen Überblick über die in dieser Arbeit verwendeten Patienten- und Kontrollpopulationen. Diese Populationen setzen sich - bis auf die Familientriostichprobe - ausnahmslos aus nichtverwandten Individuen zusammen, zeigen untereinander keine Überschneidungen und stellen somit unabhängige Stichproben dar.

Tab. 2-1: Übersicht der in dieser Arbeit verwendeten Studienpopulationen. Die Zahl der Individuen bezieht sich auf die Stichprobengröße vor der Qualitätsfilterung. *Anzahl der Sarkoidose-Familientrios. **Diese Stichprobe wurde zur zusätzlichen Nachverfolgung der Kandidaten-SNPs verwendet

Krankheit	Population	Nationalität	Patienten	Kontrollpersonen
Sarkoidose	Stichprobe A (GWAS)	Deutschland	646	1.770
	Stichprobe B (Validierung)	Deutschland	1.530	2.204
	Stichprobe C-I (Replikation)	Deutschland	307	285
	Stichprobe C-II (Replikation)	Tschechien	267	330
	Stichprobe C-III (Replikation)	Schweden	1.066	940
	Stichprobe C-IV (Familientrios)	Deutschland	342*	-
Tuberkulose	Stichprobe D (GWAS)	Ghana	1.368	1.910
	Stichprobe E-I (Validierung)	Ghana	795	1.714
	Stichprobe E-II** (Validierung)	Ghana	-	1.012
	Stichprobe F (Replikation)	Südafrika	471	491

2.1.2 GWAS-Datensätze

In dieser Arbeit wurden die Originaldaten zweier bereits publizierter GWAS-Datensätze verwendet:

2.1.2.1 Stichprobe A (Sarkoidose-GWAS):

Von dem Institut für Klinische Molekularbiologie (IKMB) der Christian-Albrechts-Universität zu Kiel wurden die Daten eines Sarkoidose-GWAS zur Verfügung gestellt. In zwei Publikationen (Fischer et al. 2012; Hofmann et al. 2013) wurden diese Daten dieses Affymetrix Genome-Wide Human SNP-Arrays 6.0 bereits verwendet und im Rahmen dieser Arbeit erneut analysiert.

Die deutschen Sarkoidosepatienten für diese GWAS wurden mit Hilfe der Deutschen Sarkoidose-Vereinigung e.V., der Krankenkassen und spezialisierten Krankenhäuser und Ärzte rekrutiert. Die Diagnose wurde, entsprechend den internationalen Standards (American Thoracic Society 1999), bei allen teilnehmenden Patienten durch eine histologische Untersuchung verifiziert, Patienten mit einer unsicheren Diagnose wurden ausgeschlossen. Alle Patienten gaben in einem Fragebogen Auskunft über den Krankheitsverlauf, durch den auch eine unterschiedliche Klassifizierung von *akuter* und *chronischer* Sarkoidose möglich war. Von den insgesamt 646 Patienten der Stichprobe A waren Subphänotypinformationen für 608 Patienten vorhanden (*akuter* Subphänotyp: n = 194; *chronischer* Subphänotyp: n = 414).

Die verwendeten deutschen Kontrollpersonen wurden durch die KORA Studie Augsburg (Gesundheitsforschungsprogramm in der Region Augsburg) und die POPGEN Biobank (Krawczak et al. 2006) rekrutiert. Alle Teilnehmer der Studienpopulationen gaben schriftlich ihre Einwilligung zur Teilnahme. Die Probenkollektion, die experimentelle Vorgehensweise und die Einhaltung des Datenschutzes wurden von der Ethikkommission des Universitätsklinikums Schleswig-Holstein und dem lokalen Datenschutzbeauftragten gemäß den datenschutzrechtlichen Bestimmungen geprüft und beaufsichtigt.

Insgesamt umfasst der Affymetrix SNP Array 6.0 Datensatz die Genotypdaten von 646 Sarkoidose-Patienten und 1.770 gesunden Kontrollpersonen. Nach der Anwendung konservativer und allgemein etablierter Genotypisierungsqualitätskontroll-Verfahren (Franke et al. 2010b) wurden diese Individuen erfolgreich für 703.104 SNPs genotypisiert. Diese SNP-Genotypen wurden im Rahmen dieser Arbeit mit dem Programm BEAGLE, unter Verwendung der HapMap3 Referenz-Genotypen der CEU, TSI und MEX Stichproben, imputiert (siehe Kapitel 2.2.4). Es wurden nur Schätzungen von autosomalen Genotypen mit einer Imputations-Aussagewahrscheinlichkeit (der *INFO score*) $r^2 > 0{,}3$ und einer Frequenz des selteneren Alleles (MAF) > 1% (in Fällen oder Kontrollen) verwendet. Zur Analyse stand somit ein Datensatz von 1.294.967 SNPs zur Verfügung.

2.1.2.2 Stichprobe D (Tuberkulose-GWAS)

Von dem Bernhard-Nocht-Institut für Tropenmedizin in Hamburg wurden die Daten eines Tuberkulose-GWAS zur Verfügung gestellt. In zwei Publikationen (Thye et al. 2010; Thye et al. 2012) wurde ein Großteil der Daten des Affymetrix Genome-Wide Human SNP-Arrays 6.0 bereits verwendet und im Rahmen dieser Arbeit erneut analysiert.

Die für die GWAS verwendeten ghanaischen Tuberkulosepatienten wurden in Ghana in den Krankenhäusern ‚Korle Bu Teaching Hospital', ‚Komfo Anokye Teaching Hospital' und weiteren 15 Krankenhäusern und Polikliniken in Accra und Kumasi rekrutiert. Die Diagnose wurde durch

umfassende Dokumentation der Krankengeschichte, ärztliche Untersuchungen, HIV-1/2 Tests, Brust posterior-anterior Röntgenaufnahmen, Ziehl-Neelsen Färbung von zwei unabhängigen Sputumabstrichen, sowie durch Kultivierung von Isolaten des *M. tuberculosis*-Komplexes auf Loewenstein-Jensen Medium verifiziert. Die Hauptsymptome wurden zusätzlich über einen von den Patienten auszufüllenden Fragebogen ermittelt. Alle Patienten dieser Studienpopulation waren HIV-negativ und hatten bei der Röntgenuntersuchung die für Lungentuberkulose charakteristischen Läsionen in der Lunge. Alle Patienten wurden im Rahmen des DOTS Programms (Directly Observed Treatment Short-Course strategy), des ghanaischen nationalen Tuberkuloseprogramms behandelt.
Die Kontrollpersonen wurden im Rahmen einer Assoziationsstudie zu Malaria und im Rahmen einer ghanaischen Tuberkulosestudie rekrutiert. Alle Kontrollpersonen wurden nach einer ärztlichen Untersuchung als gesund eingestuft. Die Probenkollektion, die experimentelle Vorgehensweise und der Datenschutz wurden durch das ‚Committee on Human Research' der ‚Kwame Nkrumah University of Science and Technology', Kumasi, Ghana und das ‚Ethics Committee of the Ghana Health Service', Accra, Ghana geprüft und genehmigt. Für die ghanaische Tuberkulosestudie wurde der Gesundheitsstatus der Teilnehmer durch klinische Untersuchungen, Krankenakten und durch posterior-anterior Röntgenaufnahmen der Brust überprüft. Alle Teilnehmer der Studie willigten in diese schriftlich oder im Fall von Analphabetismus mit ihrem Daumenabdruck ein.
Insgesamt umfasste der Affymetrix SNP-Array 6.0 Tuberkulosedatensatz die Genotypdaten von 1.368 Tuberkulose-Patienten und 1.910 gesunden Kontrollpersonen. Nach der Anwendung konservativer und allgemein etablierter Genotypisierungsqualitätskontroll-Verfahren (Thye et al. 2009) wurden diese Individuen erfolgreich für 811.349 SNPs genotypisiert. Diese SNP-Genotypen wurden im Zuge dieser Arbeit ebenfalls mit dem Programm BEAGLE, unter Verwendung der HapMap3 Referenz-Genotypen YRI, LWK und MKK Stichproben, imputiert (siehe Kapitel 2.2.4). Es wurden nur Schätzungen von autosomalen Genotypen mit einer Imputations-Aussagewahrscheinlichkeit (der *INFO score*) $r^2 > 0{,}3$ und einer Frequenz des selteneren Alleles (MAF) > 1% (in Fällen oder Kontrollen) verwendet. Damit stand für die Analyse ein Datensatz mit 1.358.606 SNPs zur Verfügung.

2.1.3 Kollektiv der Validierungsstudie

2.1.3.1 Stichprobe B

Von dem Institut für Klinische Molekularbiologie (IKMB) der Christian-Albrechts-Universität zu Kiel wurde die DNA einer deutschen Stichprobe von Sarkoidosepatienten und Kontrollpersonen zur Verfügung gestellt.

Die Patienten der Stichprobe B wurden nach den schon für Stichprobe A beschriebenen Standards rekrutiert und ebenfalls durch die oben genannten Institutionen, ebenso gaben alle Patienten in einem Fragebogen Auskunft über den Krankheitsverlauf, wodurch wieder eine unterscheidende Klassifizierung von *akuter* und *chronischer* Sarkoidose möglich war. Somit waren Subphänotypinformationen für 1.393 Patienten (*akuter* Subphänotyp: n = 502; *chronischer* Subphänotyp: n = 891) von den insgesamt 1.530 Patienten der Stichprobe B vorhanden.

Die verwendeten deutschen Kontrollpersonen wurden durch die POPGEN Biobank (Krawczak et al. 2006) rekrutiert. Alle Teilnehmer der Studienpopulationen gaben schriftlich ihre Einwilligung. Die Probenkollektion, die experimentelle Vorgehensweise und die Einhaltung des Datenschutzes wurden von der Ethikkommission des Universitätsklinikums Schleswig-Holstein und dem lokalen Datenschutzbeauftragten gemäß den datenschutzrechtlichen Bestimmungen geprüft und beaufsichtigt.

Insgesamt umfasste die Stichprobe B DNA-Proben von 1.530 Patienten und 2.204 gesunden Kontrollpersonen.

2.1.3.2 Stichprobe E-I

Von dem Bernhard-Nocht-Institut für Tropenmedizin in Hamburg wurde die DNA einer ghanaischen Stichprobe von Tuberkulosepatienten und Kontrollpersonen zur Verfügung gestellt.

Die Rekrutierung der Patienten der Stichprobe E-I erfolgte durch die schon für Stichprobe D beschriebenen Institutionen und nach denselben Standards.

Die ghanaischen Kontrollpersonen dieser Stichprobe wurden im Rahmen einer ghanaischen Tuberkulosestudie rekrutiert. Alle Teilnehmer der Studienpopulation willigten in diese schriftlich oder im Fall von Analphabetismus mit ihrem Daumenabdruck ein. Die Probenkollektion, die experimentelle Vorgehensweise und der Datenschutz wurden durch das ‚Committee on Human Research' der ‚Kwame Nkrumah University of Science and Technology', Kumasi, Ghana und das ‚Ethics Committee oft the Ghana Health Service', Accra, Ghana geprüft und genehmigt.

Insgesamt umfasste die Stichprobe E-I DNA-Proben von 795 Patienten und 1.714 gesunden Kontrollpersonen.

2.1.3.3 Stichprobe E-II

Von dem Bernhard-Nocht-Institut für Tropenmedizin in Hamburg wurde die DNA einer ghanaischen Stichprobe von Kontrollpersonen zur Verfügung gestellt.

Die Rekrutierung der gesunden Kontrollpersonen der Stichprobe E-II erfolgte durch die schon für Stichprobe E-I beschriebenen Institutionen und nach denselben Standards.
Die Stichprobe E-II umfasste DNA-Proben von 1.012 gesunden Kontrollpersonen.

2.1.4 Kollektiv der Replikationsstudie

2.1.4.1 Stichprobe C-I

Von der Medizinischen Klinik II des Universitätsklinikums Bonn wurde die DNA einer deutschen Stichprobe von Sarkoidosepatienten und Kontrollpersonen zur Verfügung gestellt.
Die Rekrutierung der Individuen der Stichprobe C-I erfolgte nach den schon für die Stichprobe A beschriebenen Standards und überschneidet sich komplett mit der in der Veröffentlichung von Pabst *et al.* (2011) beschriebenen Stichprobe. Subphänotypinformationen waren in dieser Stichprobe nur für einen Teil der verwendeten Proben vorhanden (*akuter* Subphänotyp: n = 40; *chronischer* Subphänotyp: n = 61).
Insgesamt umfasste die Stichprobe C-I DNA-Proben von 307 Patienten und 285 gesunden Kontrollpersonen.

2.1.4.2 Stichprobe C-II

Von der ‚Faculty of Medicine and Dentistry' der Palacký Universität in Olmütz, Tschechien wurde die DNA einer tschechischen Stichprobe von Sarkoidosepatienten und Kontrollpersonen zur Verfügung gestellt.
Die tschechischen Sarkoidosepatienten wurden durch die ‚Medical Faculty of Palacký University and University Hospital' in Olmütz, Tschechien rekrutiert. Die Diagnose wurde histologisch verifiziert und Patienten mit einer unsicheren Diagnose wurden aus der Studie ausgeschlossen. Die in dieser Arbeit verwendete Stichprobe C-II überschneidet sich im Wesentlichen mit der Stichprobe, beschrieben in einer Veröffentlichung von Mrazek *et al.* (2011). Für diese Stichprobe standen keine Subphänotypinformationen zur Verfügung.
Insgesamt umfasste die Stichprobe C-II DNA-Proben von 267 Patienten und 330 gesunden Kontrollpersonen.

2.1.4.3 Stichprobe C-III

Von dem ‚Karolinska University Hospital' der Gemeinde Solna, Schweden wurde die DNA einer schwedischen Stichprobe von Sarkoidosepatienten und Kontrollpersonen zur Verfügung gestellt.

Die schwedischen Sarkoidosepatienten wurden in der pulmonären Abteilung der Ambulanz der Karolinska Universitätsklinik in Solna, Schweden rekrutiert. Die verwendeten Kontrollpersonen wurden über die ‚Epidemiological Investigation of Rheumatoid Arthritis' (EIRA)-Studie rekrutiert und überschneiden sich mit der in der Veröffentlichung von Klareskog *et al.* (2006) beschriebenen Kontrollpersonenstichprobe. Subphänotypinformationen waren in dieser Stichprobe nur für einen Teil der verwendeten Proben vorhanden (Löfgren-Syndrom: n = 311).

Insgesamt umfasste die Stichprobe C-III DNA-Proben von 1.066 Patienten und 940 gesunden Kontrollpersonen.

2.1.4.4 Stichprobe C-IV

Von dem Institut für Klinische Molekularbiologie (IKMB) der Christian-Albrechts-Universität zu Kiel wurde die DNA einer deutschen Familientriostichprobe zur Verfügung gestellt.

Die deutsche Familientriostichprobe C-IV wurde nach den bei Stichprobe A beschriebenen Standards rekrutiert, wobei hier zusätzlich Eltern und weitere Familienmitglieder rekrutiert wurden. Bei Familien mit zwei oder mehr Betroffenen wurden ausführliche Informationen über die Krankheitsgeschichte und Familienstruktur eingeholt. Diese Stichprobe überschneidet sich vollständig mit der Stichprobe in der Veröffentlichung von Fischer *et al.* (2011). Subphänotypinformationen standen für einen Großteil dieser Stichprobe zur Verfügung (*akuter* Subphänotyp: n = 94; *chronischer* Subphänotyp: n = 212).

Insgesamt umfasste die Stichprobe C-IV DNA-Proben von 342 Trios.

2.1.4.5 Stichprobe F

Von der ‚Faculty of Health Sciences' der Stellenbosch Universität in Stellenbosch, Südafrika wurde die DNA einer südafrikanischen Stichprobe von Tuberkulosepatienten und Kontrollpersonen zur Verfügung gestellt.

Die südafrikanischen Tuberkulosepatienten wurden alle aus der Metropolregion von Kapstadt in der Provinz Westkap in Südafrika rekrutiert. Diese Region weist eine hohe Inzidenz an Tuberkuloseerkrankungen und eine einheitliche ethnische Zugehörigkeit, einen einheitlichen sozioökonomischem Status und eine niedrige Prävalenz an humanen Immundefizienz-Virus-Erkrankungen (HIV) auf. Die Tuberkulosepatienten wurden mittels eines bakteriologischen Nachweises (durch Sputumabstrich und/oder Auswertung von Kulturen) identifiziert. Gesunde Kontrollpersonen wurden in derselben Region rekrutiert und lebten unter den gleichen Bedingungen wie die Patienten. Tuberkulin-Hauttest (engl. *tuberculin skin test*, TST)-Untersuchungen der gesunden

Kontrollpersonen ergaben eine hohe Positiv-Rate (Gallant et al. 2010), welche ein Hinweis auf eine latente *M. tuberculosis* Infektion ist. Ein Großteil der verwendeten Kontrollpersonen war TST-positiv, es wird von einer Positiv-Rate von ~80% ausgegangen. Die gesunden Kontrollpersonen hatten in ihrer Vorgeschichte keine Tuberkulose oder Tuberkulosebehandlung und zeigten keine Verwandtschaft mit anderen Individuen der Stichprobe. Alle HIV-positiven Individuen wurden aus der Studie ausgeschlossen. Die in dieser Arbeit verwendete Stichprobe F überschneidet sich im Wesentlichen mit der in einer Veröffentlichung von Adams *et al.* (2011) beschrieben Stichprobe.

Die Probenkollektion, die experimentelle Vorgehensweise und der Datenschutz wurden durch das Ethikkomitee der ‚Faculty of Health Sciences' der Stellenbosch Universität (Projektnummer 95/072) geprüft und genehmigt. Alle Individuen der Studienpopulationen gaben schriftlich ihre Einwilligung zur Teilnahme.

Insgesamt umfasste die Stichprobe F DNA-Proben von 471 Patienten und 491 gesunden Kontrollpersonen.

2.1.5 Feinkartierungsstichprobe

Die Feinkartierung wurde mit einer Stichprobe (im Folgenden Feinkartierungsstichprobe (FK) genannt), die sich aus der kompletten Stichprobe B und Teilen der Patienten aus der Stichprobe A zusammensetzte, durchgeführt. Insgesamt bestand diese Stichprobe aus 1.815 Sarkoidosepatienten und 2.204 Kontrollpersonen. Für die Stichprobe waren für einen Großteil der Patienten Subphänotypinformationen vorhanden (*akuter* Subphänotyp: n = 597; *chronischer* Subphänotyp: n = 1.055).

2.1.6 Übersicht der verwendeten Stichproben

2.1.6.1 Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen

Die Abbildung 2-1 gibt nochmals eine grafische Übersicht über die verschiedenen in diesem Teilprojekt der Arbeit genutzten Sarkoidosestichproben.

Stich-probe	Patienten (vor / nach QC)	Kontrollpersonen (vor / nach QC)
A	646/564	1770/1575
B	1530/1486	2204/2137
C-I	307/303	285/281
C-II	267/264	330/325
C-III	1066/1027	940/916
C-IV	342/301*	-
FK	1815/1810	2204/2182

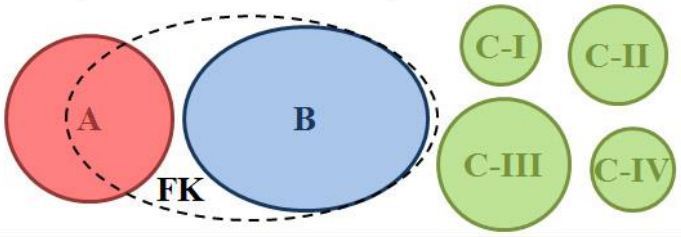

Abb. 2-1: Übersicht der Stichproben des Sarkoidoseprojekts. Die Sarkoidosestichproben sind mit ihren in der Arbeit verwendeten Abkürzungen aufgeführt, wobei die Abkürzung ‚FK' für Feinkartierungsstichprobe steht. Die Individuenanzahlen der Stichproben beziehen sich auf den Stichprobenumfang vor und nach der Qualitätsfilterung (QC). *Die Stichprobe der Familientrios beinhaltete keine unabhängigen Kontrollpersonen.

2.1.6.2 Identifizierung gemeinsamer genetischer Faktoren der Krankheiten Tuberkulose und Sarkoidose

In der Abbildung 2-2 ist nochmals eine grafische Übersicht über die verschiedenen in diesem Projekt genutzten Stichproben zu sehen.

Stich-probe	Patienten (vor / nach QC)	Kontrollpersonen (vor / nach QC)
A	646/564	1770/1575
D	1368/1288	1910/1790
B	1530/1486	2204/2137
E-I	795/780	1714/1674
E-II**	-	1012/925
C-I	307/303	285/281
C-II	267/264	330/325
F	471/386	491/372

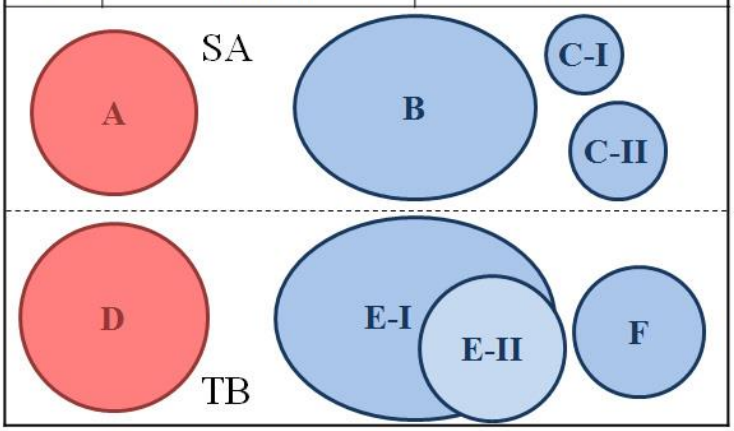

Abb. 2-2: Übersicht der Stichproben des Tuberkulose- und Sarkoidoseprojekts. Die verschiedenen Stichproben sind mit ihren in der Arbeit verwendeten Abkürzungen aufgeführt. Die Individuenanzahlen der Stichproben beziehen sich auf den Stichprobenumfang vor und nach der Qualitätsfilterung (QC). Die Abkürzungen ‚SA' und ‚TB' stehen für Sarkoidose und Tuberkulose. **Diese Stichprobe wurde zur zusätzlichen Nachverfolgung der Kandidaten-SNPs verwendet.

2.1.7 Gebrauchsmittel und Reagenzien

Material	Hersteller
0,5-30,0 µL 96-Kanal Spitzen	Sequenom; San Diego, USA
100 bp DNA-Leiter	Invitrogen; Karlsruhe, Deutschland
384er Deep Well Lagerplatte	Abgene; Epsom, UK
Advantage RT-for-PCR Kit	Clontech; BD Biosciences, Heidelberg, Deutschland
Agarose	Eurogentec; Köln, Deutschland
AmpliTaq DNA Polymerase	Applied Biosystems; Weiterstadt, Deutschland
AmpliTaq Gold DNA Polymerase	Applied Biosystems; Weiterstadt, Deutschland

BigDye® v1.1 Terminator Cycle Sequenzing Kit	Applied Biosystems; Weiterstadt, Deutschland
BigDye® Xterminator™ Purification Kit	Applied Biosystems; Weiterstadt, Deutschland
Bromphenolblau	Sigma; München, Deutschland
DMEM Medium	PAA Laboratories; Pasching, Österreich
dNTP Set (100 mM)	PEQLAB Biotechnology GmbH; Erlangen, Deutschland
Easy peel Folie	Abgene; Epsom, UK
EDTA	Sigma; München, Deutschland
EDTA Blutröhrchen 9 ml	Sarstedt; Nümbrecht, Deutschland
Ethanol	Merck; Darmstadt, Deutschland
Ethidiumbromid (10mg/ml)	Invitrogen; Karlsruhe, Deutschland
Exonuklease1	Fisher Scientific; Schwerte, Deutschland
GeneAmp PCR-Puffer	Applied Biosystems; Weiterstadt, Deutschland
Glyzerol	Sigma; München, Deutschland
HCl	Sigma; München, Deutschland
Human Multiple Tissue cDNA Panel	Clontech; BD Biosciences, Heidelberg, Deutschland
Invisorb Blood Giga Kit	Invitek; Berlin, Deutschland
iPLEX Gold Reagenzien und Chip-Kit	Sequenom; San Diego, USA
iPLEX Gold Reagenzienkit	Sequenom; San Diego, USA
Isopropanol	Merck; Darmstadt, Deutschland
Komplettes Genotypisierungs-Reagenzienset	Sequenom; San Diego, USA
$MgCl_2$	Merck; Darmstadt, Deutschland
MicroAmp optical 96-Well-Platte	Applied Biosystems; Weiterstadt, Deutschland
Microtiter 384-Well-Platte	Applied Biosystems; Weiterstadt, Deutschland
Microtiter 384-Well-Platte	Greiner Bio-One GmbH; Frickenhausen, Deutschland
Microtiter 96-Well-Platte	Costar Corning Incorporated; Cambridge, MA, USA
Microtiter 96-Well-Platte	Sarstedt; Nürnberg, Deutschland
NaCl	Sigma; München, Deutschland
PCR Zubehör- und Enzym-Set	Sequenom; San Diego, USA
Penicillin/Streptomycin	Biochrom; Berlin, Deutschland
PEQLAB DNA Isolationssystem	PEQLAB Biotechnology GmbH; Erlangen, D.
PicoGreen	Molecular Probes Europe BV; Leiden, Niederlande
Pipettenfilterspitzen (10 / 200 / 1000 µl)	Sarstedt; Nürnberg, Deutschland
Pipettenspitzen (steril 5 / 10 / 25 ml)	Sarstedt; Nümbrecht, Deutschland
Proteinase K	Invitek; Berlin, Deutschland

Reagenzien-Reservoirs	Sequenom; San Diego, USA
RNAeasy Kit	Qiagen; Hilden, Deutschland
Röhrchen (0,5 / 1,5 / 2,0 ml)	Eppendorf; Köln, Deutschland
Röhrchen (60 ml)	Sarstedt; Nümbrecht, Deutschland
Röhrchen, steril (15 ml)	Sarstedt; Nümbrecht, Deutschland
Röhrchen, steril (50 ml)	BD Biosciences; Heidelberg, Deutschland
RPMI1640 Medium	PAA Laboratories; Pasching, Österreich
SAP Shrimp Alkaline Phosphatase	Fisher Scientific; Schwerte, Deutschland
Selbsthaftende Sealingfolie	Marsh Biomedical Products, Inc.; Rochester, USA
SmartLadder DNA marker	Eurogentec; Köln, Deutschland
SpectroCHIP Arrays und Clean-Resin-Kit	Sequenom; San Diego, USA
TAE Puffer 25x	Amresco; Solon, OH, USA
Taqman Gen-Expressions-Assays	Applied Biosystems; Weiterstadt, Deutschland
TaqMan Universal PCR Master Mix	Applied Biosystems; Weiterstadt, Deutschland
Taqman-Assays	Applied Biosystems; Weiterstadt, Deutschland
TBE Puffer 10x	Amresco; Solon, OH, USA
Tris(hydroxymethyl)-aminomethan	Merck; Darmstadt, Deutschland
Triton-X	Sigma; München, Deutschland
Trypsin / EDTA (0,25 % / 1 mM)	Invitrogen/Gibco; Karlsruhe, Deutschland
Zellkultur-Flaschen (250 ml)	BD Biosciences; Heidelberg, Deutschland
Zelllinien	DSMZ; Braunschweig, Deutschland

2.1.8 Geräte

Gerät	**Hersteller**
3700 DNA Analyzer	Applied Biosystems Inc.; Foster City, CA, USA
3730xl DNA Analyzer	Applied Biosystems Inc.; Foster City, CA, USA
7900HT Fast Real-time PCR-Systems	Applied Biosystems Inc.; Foster City, CA, USA
ABI Prism 7700 Sequence Detector	Applied Biosystems Inc.; Foster City, CA, USA
BioDoc Analyzer	Biometra; Göttingen, Deutschland
Biometra T Gradient	Whatman Biometra GmbH; Göttingen, Deutschland
Biometra T1 Thermocycler	Whatman Biometra GmbH; Göttingen, Deutschland
Gel Doc XR	Bio-Rad; München, Deutschland
GeneAmp PCR System 9700	Applied Biosystems Inc.; Foster City, CA, USA
Heraeus 3 incubator	Kendro; Hanau, Deutschland

Heraeus Kelvitron t	Kendro; Hanau, Deutschland
Heraeus Biofuge fresco	Kendro; Hanau, Deutschland
Heraeus Biofuge pico	Kendro; Hanau, Deutschland
Heraeus Labofuge 400	Kendro; Hanau, Deutschland
Heraeus Multifuge 3S-R	Kendro; Hanau, Deutschland
Heraeus Varifuge 3.2RS	Kendro; Hanau, Deutschland
High Performance UV Transilluminator	VWR; Hamburg, Deutschland
Horizontal Electrophoresis Apparatus	Bio-Rad; München, Deutschland
Hydra 384 Robbins Scientific	Dunn Labortechnik; Asbach, Deutschland
Hydra 96 Robbins Scientific	Dunn Labortechnik; Asbach, Deutschland
Micro Zentrifuge	Roth; Karlsruhe, Deutschland
Microwave R-2V18	Sharp Electronics; Hamburg, Deutschland
Mini Vortexer VM-3000	VWR; Darmstadt, Deutschland
Plattensealer ALPS-300	Abgene; Epsom, UK
Power Pac 300 Electrophoresis Power Supply	Bio-Rad; München, Deutschland
Tecan Freedom Evo 150	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Tecan Freedom Evo 200	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Tecan Genesis RSP 150	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Tecan Genesis Workstation 150	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Tecan Genesis Workstation 200	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Tecan Spectrafluor Plus	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Te-MO	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Te-MO mit Cooling Rack	Tecan, Deutschland GmbH; Crailsheim, Deutschland
Thermomixer 5437	Eppendorf; Köln, Deutschland
Ultrospec 3100pro	Amersham Biosciences; Freiburg, Deutschland
UVB light source	Philips; Hamburg, Deutschland
Vortex-Genie 2 G-560E	Scientific Industries; Bohemia, NY, USA
Sequenom MassARRAY Analyzer 4	Sequenom; San Diego, USA

2.1.9 Software und Datenbanken

Name	**URL**
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/

Datenbank in der Einzelbasen-Polymorphismen (SNPs) verzeichnet sind.

EIGENSOFT 5.0.1 http://www.hsph.harvard.edu/alkes-price/software/

Programm mit dem populationsgenetische Methoden angewandt werden können.

eQTL resources http://eqtl.uchicago.edu/Home.html

Datenbank in der Informationen über SNPs gesammelt werden, die die Genregulation beeinflussen können.

GOLD v1.1.0 http://www.sph.umich.edu/csg/abecasis/GOLD

Programm zur grafischen Darstellung des zwischen Markern herrschenden LDs.

GraphPad Prism 6.0 http://www.graphpad.com/scientific-software/prism/

Programm zur Analyse und grafischer Darstellung von experimentellen Daten. Rohdaten aus Laborexperimenten können mit dem Programm analysiert werden und die Ergebnisse grafisch dargestellt werden.

Haploview v4.1 http://www.broad.mit.edu/mpg/haploview

Programm mit dem verschiedene Haplotypanalysen von SNP-Datensätzen durchgeführt werden können, wie z.B. LD- und Haplotyp-Block Analyse, Schätzungen der Haplotypfrequenz in Populationen, Assoziationstests von SNPs und Haplotypen

HapMap http://www.hapmap.org

Datenbank mit Haplotypkarten des humanen Genoms, in der DNA-Sequenzvariationen verzeichnet sind. Die Haplotypkarten sind für eine Auswahl an verschiedenen Populationen verfügbar.

miRDB http://mirdb.org

Datenbank in der miRNA-Zielsequenzen und funktionellen Annotationen verzeichnet sind.

NIEHS SNPinfo http://www.niehs.nih.gov/research/resources/index.cfm

SNP-Datenbank in der funktionelle Charakteristiken von SNPs gesammelt werden.

PLINK v1.07 http://pngu.mgh.harvard.edu/~purcell/plink

Programm zur Analyse von genomweiten Genotypdaten, ausgelegt grundlegende Analysen von großen Datensätzen.

PS v3.0.43 http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize

Mit diesem Programm können verschiedenen Teststärke- und Stichprobengröße-Berechnungen durchgeführt werden.

R 2.10.1 http://www.r-project.org

Programmiersprache, mit der statistische Berechnungen durchgeführt und grafische Darstellungen erstellt werden können.

RefSeq http://www.ncbi.nlm.nih.gov/refseq/

Referenzsequenzen-Datenbank, in der Genomische-, Transkript- und Proteinsequenzen annotiert sind.

SNAP Pairwise LD http://www.broadinstitute.org/mpg/snap/ldsearchpw.php

Datenbank in der Informationen für das paarweise LD zwischen SNPs gesammelt werden.

UCSC genome browser http://genome.ucsc.edu/cgi-bin/hgGateway

Datenbank für Referenzsequenzen verschiedener Genome. In dieser Datenbank sind u.a. die ENCODE- und Neandertal-Datenbank implementiert.

2.1.9.1 IBDbase

Mit der Sequenom®- oder TaqMan®-Technologie gewonnenen Daten und eine Vielzahl von weiteren Informationen, wie z.B. Krankheitsstatus der Studienteilnehmer, wurden in einer im Institut für Klinische Molekularbiologie des Universitätsklinikums Schleswig-Holstein in Kiel entwickelten Datenbank (IBDbase) gespeichert (Hampe et al. 2001). Mit der ebenfalls in diesem Institut entwickelten Software ‚XPTools' wurde zudem ein Programm für die vereinfachte Datenbankabfrage entwickelt (Teuber et al. 2009). Die Genotypinformation der untersuchten Proben kann mit dieser Software z.B. für die weitere Analyse im gebräuchlichen Linkageformat exportiert werden.

2.1.9.2 PLINK Software

Mit der frei verfügbaren Software PLINK v1.07 stand für diese Arbeit eine potente Software zur Analyse von Genotyp-Daten im Linkageformat zur Verfügung (Purcell et al. 2007). Mit PLINK können auch sehr umfangreiche GWAS-Datensätze schnell und mit den gängigen Analysemethoden

analysiert werden. PLINK wurde im Rahmen dieser Arbeit hauptsächlich für Assoziationsanalysen und für die Überprüfung auf Populationsstratifikation genutzt.

2.2 Methoden

2.2.1 Studiendesign

Für die Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen und für die Identifizierung gemeinsamer genetischer Faktoren von Tuberkulose und Sarkoidose wurde jeweils ein zweistufiges Studiendesign angewendet (Satagopan et al. 2002). Obwohl in beiden Analysen ein zweistufiges Studiendesign verwendet wurde, unterschied sich das jeweilige Studiendesign in einigen Details:

2.2.1.1 Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen

Für die Analyse von Sarkoidose und ihrer Subphänotypen wurde ein „klassisches“ zweistufiges Studiendesign angewendet (Satagopan et al. 2002) (Abb. 2-3). In der ersten Stufe werden dabei hypothesenfrei genomweit alle Marker (der in der Stufe analysierten Individuen) auf Assoziationen mit der zu untersuchenden Krankheit überprüft. Diese erste Stufe wird in dieser Arbeit mit dem GWAS-Datensatz (Stichprobe A) durchgeführt. Die in der ersten Stufe auf Basis eines signifikanten Unterschieds in der Genotypfrequenz zwischen Patienten und Kontrollpersonen identifizierten Marker werden dann in der zweiten Stufe an Individuen einer weiteren unabhängigen Stichprobe getestet (validiert). Da normalerweise in der ersten Stufe eine Vielzahl von assoziierten Markern detektiert wird, werden in der zweiten Stufe häufig nur die vielversprechendsten Marker evaluiert. Um aus dieser Vielzahl der detektierten Marker eine Auswahl zu treffen, werden Schwellenwerte für die Marker definiert und zum Teil weitere Analysemethoden für die Markerauswahl genutzt (Kapitel 2.2.6). Mit dieser Auswahl soll erreicht werden, dass nur „echte“ Assoziationen nachverfolgt werden und „falsch-positive“ Assoziationen damit weitestgehend ausgeschlossen werden. Diese zweite Stufe (Validierung) wurde mit der Stichprobe B durchgeführt. Dabei werden die in der zweiten Stufe detektierten Assoziationen mit der Bonferroni-Methode auf multiples Testen korrigiert (Kapitel 2.2.7.3). Nur wenn Marker nach dieser Korrektur eine Assoziation von $p < 0{,}05$ aufweisen, gilt die Assoziation des Markers als validiert. Assoziationen von Markern, die sich in der zweiten Stufe bestätigt hatten, wurden in weiteren unabhängigen Stichproben getestet, um festzustellen, ob die Marker definitiv und populationsunabhängig mit der Krankheit assoziiert sind (Replikation).

Abb. 2-3: Schematische Darstellung einer zweistufigen Assoziationsstudie. In der GWAS werden genomweit alle SNPs der Individuen der Stichprobe auf eine Assoziation mit der Krankheit getestet. In der Validierung werden die vielversprechendsten SNP aus der ersten Stufe in Individuen einer weiteren Stichprobe erneut auf Assoziation getestet. Bestätigte Assoziationen werden dann in weiteren unabhängigen Stichproben repliziert.

2.2.1.2 Identifizierung gemeinsamer genetischer Faktoren der Krankheiten Tuberkulose und Sarkoidose

Für die Identifizierung gemeinsamer genetischer Faktoren von Tuberkulose und Sarkoidose wurde eine Abwandlung des zweistufigen Studiendesigns angewendet. Dabei wird die erste Stufe wie schon zuvor als „Entdeckungsstichprobe“ (engl. *discovery sample*) genutzt, um hypothesenfrei genomweit alle Marker in einer Stichprobe zu untersuchen. In dieser Arbeit werden dafür ein imputierter Tuberkulose-GWAS und ein imputierter Sarkoidose-GWAS verwendet. Hierbei werden beide GWAS-Datensätze zusammen analysiert und es werden Marker identifiziert, die eine Assoziation mit beiden Phänotypen zeigen. Anhand der in der ersten Stufe gewonnenen Assoziationsdaten werden wiederum die vielversprechendsten SNPs für die weiterführenden Analysen ausgewählt. Nun werden in der zweiten Stufe die Assoziationen der ausgewählten Marker in „Replikationsstichproben“ (engl. *replication sample*) getestet. Dies kann, wie in dieser Arbeit, in mehreren „Replikationsstichproben“ geschehen, um einen größeren Stichprobenumfang zu erreichen (siehe Abb. 2-2). Für diese Arbeit wurde die Replikation in zwei aufeinanderfolgenden Phasen durchgeführt. Hierbei werden jedoch die in der zweiten Stufe erzielten Assoziationsergebnisse nicht auf multiples Testen korrigiert. Es werden stattdessen, nach Durchführung aller Genotypisierungen, die Daten der Marker (aus den verschiedenen Stichproben) in einer Metaanalyse zusammengefasst und anhand der genomweiten Signifikanz ($p = 5 \times 10^{-8}$) bewertet. Zeigen SNPs p-Werte unterhalb dieses Schwellenwerts, sind diese als echte Assoziationen anzusehen (Thompson et al. 2011).

2.2.2 Probenpräparation

2.2.2.1 DNA-Isolierung aus Vollblut und Lagerung

Die Extraktion von genomischer DNA aus Blutproben erfolgte mit dem Invisorb® Blood Giga Kit (Invitek) nach folgendem Protokoll: Bei -80°C eingelagerte Blutproben wurden zunächst in einem kühlen Wasserbad langsam aufgetaut. 10 ml EDTA-Vollblut wurden mit 10 ml Puffer I versetzt und

für 10 min bei Raumtemperatur inkubiert, um die Erythrozyten zur Lyse zu bringen. Die Probe wurde dann 3 min bei 3000 U/min zentrifugiert und der Überstand verworfen. Die Aufreinigung wurde im Folgenden durch Resuspendierung in 20 ml Puffer I und erneute Zentrifugation so lange wiederholt, bis das verbleibende Pellet aus Zellkernen der Leukozyten farblos und somit frei vom Häm der Erythrozyten war. Das Pellet wurde danach mit 3 ml Puffer II und 50µl Proteinase K resuspendiert und für zwei Stunden bei 60°C in einem Wasserbad unter Schütteln inkubiert. Dann wurde die Probe mit 1,8 ml Puffer III versetzt und nach kurzem Schütteln für 5 min auf Eis gelagert, um die Proteine von der DNA zu trennen. Danach wurde bei 5000 U/min für 15 min zentrifugiert und der Überstand in ein neues 15 ml Zentrifugationsröhrchen überführt. Mit der Methode der Alkoholfällung wurde die DNA daraufhin präzipitiert: Zu dem Überstand wurden 10 ml Ethanol (96%) hinzugefügt und das Gemisch vorsichtig geschwenkt. Die präzipitierte DNA wurde vorsichtig abpipettiert und in ein 2 ml Eppendorf-Röhrchen überführt. Zur Entfernung der Salzreste wurde 1 ml Ethanol (70%) hinzugefügt, gründlich gemischt und bei 13000 U/min für 2 min zentrifugiert. Der Ethanolüberstand wurde abpipettiert und das übriggebliebene DNA-Pellet für 10 min bei Raumtemperatur getrocknet. Zum Abschluss wurde das gereinigte DNA-Pellet in 500 ml Tris-EDTA-Puffer resuspendiert und bei -20°C eingelagert.

2.2.2.2 Probenvorbereitung zur Sequenom®- und TaqMan®-Genotypisierung

Um die vorhandene genomische DNA der Proben möglichst effizient zu nutzen, wurde diese mit der Methode der Gesamt-Genom-Amplifizierung (engl. *whole genome amplification*; WGA) vermehrt. Bei der Sequenom®- oder TaqMan®-Technik kann sowohl genomische DNA als auch WGA-DNA für die Genotypisierung eingesetzt werden.

Mit Hilfe der aus dem Bakteriophagen Phi29 stammenden DNA-Polymerase und mit randomisierten Hexamer-Primern wird die DNA in der sogenannten ‚*multiple displacement amplification*' (MDA) vervielfältigt. Die Polymerase besteht nur aus einer Untereinheit, ist sehr stabil und hat eine Korrekturlesefunktion (Alsmadi et al. 2009; Blanco et al. 1989; Garmendia et al. 1992). Da die Primer gleichzeitig an vielen Stellen in der denaturierten genomischen DNA binden können, ist es möglich eine Vielzahl von Replikationen parallel zu initiieren. Während der Replikation wird der von der Polymerase synthetisierte DNA-Rückwärtsstrang von seinem Gegenstrang abgelöst. Somit bildet sich ein neuer Einzelstrang, der wiederum als Matrize für weitere Polymerasen dient. Mit dieser Methode kann die vorhandene Menge an doppelsträngiger DNA ca. 100-fach vervielfältigt werden - die erhaltende DNA besitzt eine Fragmentlänge von 10 bis 100kb - wobei durch diese Methode Unterschiede in der DNA-Konzentration zwischen verschiedenen Proben tendenziell ausgeglichen werden (Dean et al. 2002).

Für die WGA-Reaktion wurde das *GenomiPhi* Kit (GE Healthcare, UK) eingesetzt und je Probe 1 µl genomische DNA mit einer Konzentration von 60 ng/µl verwendet. Alle Schritte wurden entsprechend dem Herstellerprotokoll durchgeführt. Eine Konzentrationsbestimmung der prozessierten Proben erfolgte mit einem PicoGreen-Messprotokoll (Gallagher and Desjardins 1989).

2.2.3 Genotypisierung

2.2.3.1 Affymetrix Genome-Wide Human SNP Array 6.0®

Mit dem *Genome-Wide Human SNP Array 6.0* wurde die gleichzeitige Genotypisierung von über 906.600 Einzelbasen Polymorphismen (SNPs) realisiert. Die Methode, mit der diese große Anzahl von SNPs gleichzeitig genotypisiert werden kann, beruht auf einer Weiterentwicklung der 2004 von Matsuzaki *et al.* vorgestellten Genotypisierungs-Technik (Matsuzaki et al. 2004). Mit dem *Genome-Wide Human SNP Array 6.0* können in einer DNA-Probe genomweit Proben (SNPs) gleichzeitig, ohne Benutzung von locus-spezifischen Primern, analysiert werden (Kennedy et al. 2003).

Genomische DNA wird hierbei in zwei unabhängigen Ansätzen mit den Nsp I- und Sty I-Restriktionsenzymen verdaut (Abb. 2-4). Die Restriktionsfragmente werden an Adapter-DNA-Moleküle ligiert, die die klebrigen 4 Nukleotid langen Enden (engl. *sticky ends*) binden. Alle Fragmente des Restriktionsverdaus, unabhängig von ihrer Größe, sind Substrate für die Adapter-Ligation. Mit einem allgemeinen Primer, der die Adapter-Sequenz erkennt, werden die mit den Adapter-Molekülen verbundenen DNA-Fragmente mit einer Größe von 200 bis 1.100 Nukleotiden amplifiziert. Die Produkte der Amplifikationen der zwei Ansätze werden dann zusammengeführt und mit Polystyren-Kügelchen aufgereinigt. Die DNA wird dann fragmentiert, markiert und mit dem SNP-Array 6.0 hybridisiert. Die DNA-Fragmente binden spezifisch die Sonden, die sich auf dem Chip befinden, und die Menge der gebundenen DNA-Fragmente wird mittels Fluoreszenz ermittelt (Abb. 2-4). Die Fluoreszenzintensitäten werden mit dem Affymetrix Scanner gemessen und die gemessenen Pixel-Intensitäten in CEL-Dateien abgespeichert. Die Genotypen werden mit der ‚Affymetrix Power Tools'-Software, welche den *Birdseed*-Algorithmus (Hong et al. 2010; Korn et al. 2008) nutzt, durch die Auswertung der Intensitäten bestimmt.

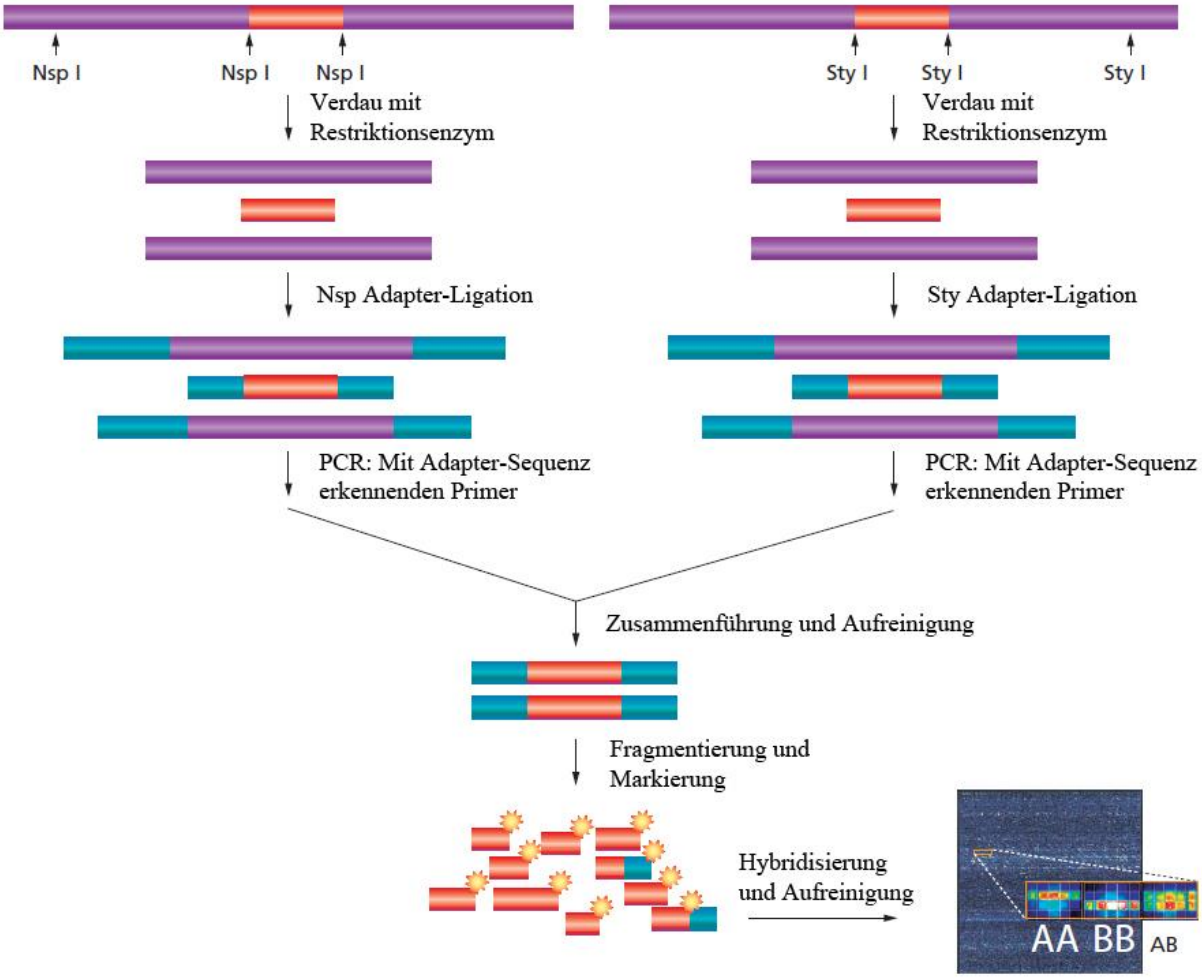

Abb. 2-4: Übersicht der DNA-Verarbeitung mit dem „Genome-Wide Human SNP Nsp/Sty Assay Kit 5.0/6.0". Nach (Affymetrix 2009).

2.2.3.2 Sequenom® Genotypisierung

Mit der Sequenom®-Technologie (im Folgenden: Sequenom) können bis zu 40 SNPs auf einmal mit einer PCR-Reaktion und mittels Massenspektrometrie genotypisiert werden. Mit Sequenom können 384 Platten mit WGA-amplifizierten DNA-Proben auf das Vorhandensein des entsprechenden Genotyps von bis zu 40 SNPs untersucht werden. Um mit Sequenom SNPs in genomischer DNA zu typisieren, wird der den SNP beinhaltende DNA-Abschnitt in einer PCR vervielfältigt. Dabei werden Primer benutzt, die eine zehn Nukleotid umspannende Erkennungssequenz an ihrem 5'-Ende besitzen (Abb. 2-5). Das so gewonnene Amplikon hat eine Größe von 80 bis 100 Basenpaaren zuzüglich der Erkennungssequenz. Die zehn Nukleotide lange Sequenz dient in der späteren Reaktion als Erkennungssequenz für die sogenannten ‚*Extension Primer*'. Diese ‚*Extension Primer*' haben eine Länge von 15 bis 30 Nukleotiden, deren atomare Masse (Dalton; Da) später in einer Massenspektroskopie gemessen wird. Zur eindeutigen Unterscheidung der verschiedenen ‚*Extension Primer*' bei der späteren Auswertung ist es wichtig, dass sie sich in ihrer Masse um mindestens 30 Da

unterscheiden. Die ‚*Extension Primer*' werden so designt, dass ihr 3'-Ende direkt an den SNP angrenzt, damit dieser dann in einer folgenden Reaktion mit einem Didesoxyribonukleosid-Triphosphat (ddNTP) an seinem 3'-Ende verlängert wird. Solche ddNTPs gleichen den normalen Desoxyribonukleotid-Triphosphaten (dNTPs), jedoch ist die Ribose an Position 2' und 3' desoxydiert. Durch die fehlende Hydroxylgruppe an Position 3' bricht die Polymerisation ab, da das nächste Nukleotid nicht mehr angehängt werden kann. Um in der späteren Massenspektrometrie die verschiedenen dNTPs an der Position des SNPs zu erkennen, wurde die Masse der artifiziellen ddNTPs modifiziert. Von den so verarbeiteten Proben wurden dann ungefähr 25 nl auf Matrixpunkte, die sich auf einem Silica-Chip befinden, aufgetragen und eine MALDI-TOF (Matrix-unterstützte Laser-Desorption/Ionisation (MALDI) und Massenspektrometrie mit Flugzeitanalysator (TOF)) durchgeführt. Diese Schritte wurden mit dem ‚MassARRAY® Analyzer 4'-Gerät (Sequenom, San Diego) von Sequenom durchgeführt. Anhand der gemessenen Massen der einzelnen ‚*Extension Primer*' mit dem dazugehörigen ddNTP kann die vorhandene Base an der SNP-Position in den einzelnen Proben identifiziert werden. Bei der MALDI-TOF Massenspektrometrie werden mit einem UV-Laser auf eine Matrix aufgebrachte Biomoleküle verdampft, dabei werden die Proben ionisiert und in einem elektrischen Feld in Richtung eines Ionendetektors beschleunigt. Je nach Ladung werden die Ionen unterschiedlich stark beschleunigt und ihre Flugzeit von der Matrix bis zum Ionendetektor gemessen. Letztlich erhält man ein Spektrum der elektrischen Signale, welches mittels der MassARRAY Typer® Software von Sequenom® analysiert und ausgewertet wird. Hierbei werden die Spektren vektorgraphisch in ein zweidimensionales Koordinatensystem übertragen, so dass sich identische Genotypen an derselben Stelle in dem Koordinatensystem konzentrieren (engl. *cluster*). Die automatische Zuordnung der Genotypen durch die Software wurde stets manuell kontrolliert und gegebenenfalls korrigiert.

Die Verarbeitung der Proben erfolgte nach dem Sequenom® Standardprotokoll (S. Gabriel et al. 2009).

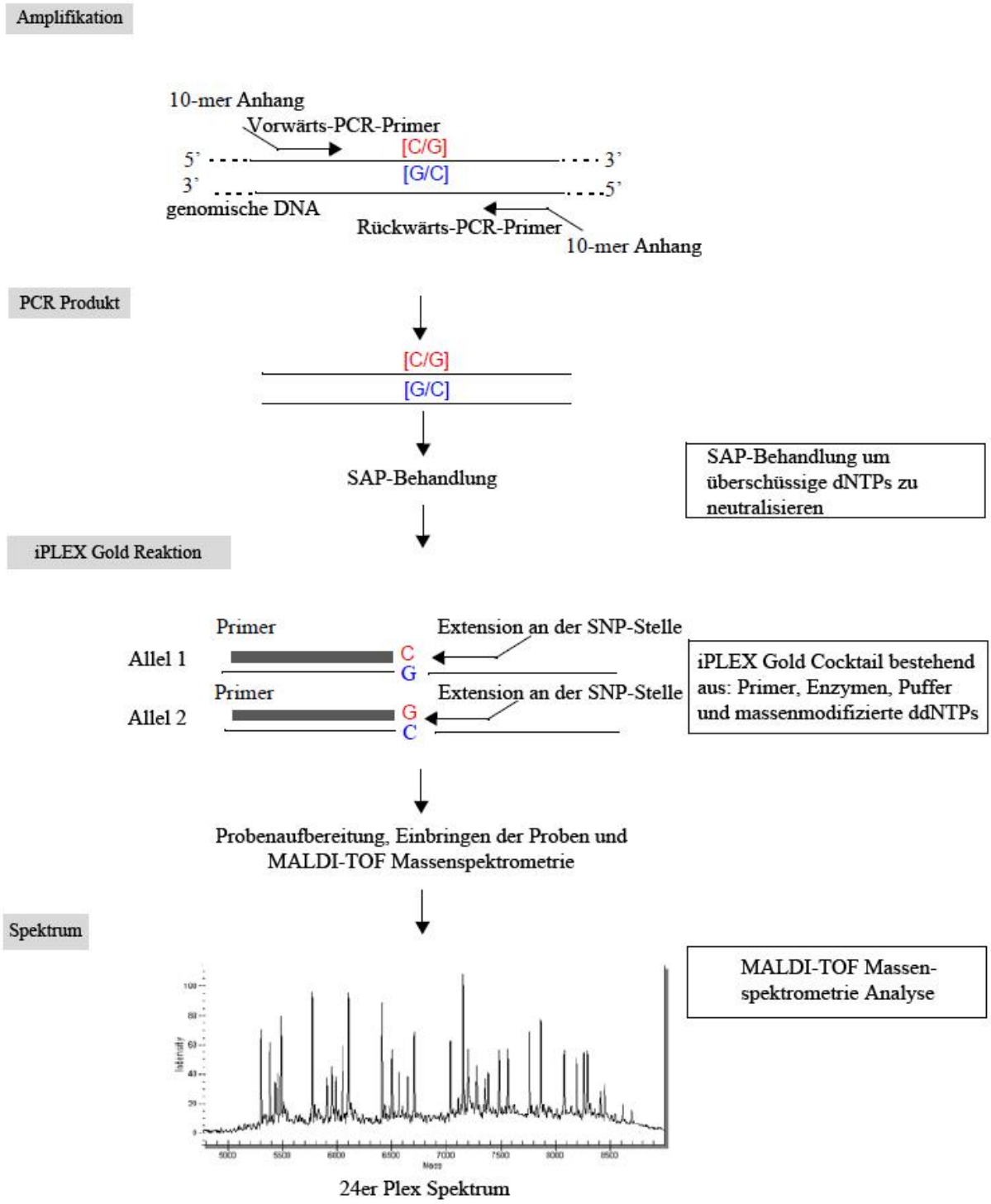

Abb. 2-5: Schematische Darstellung des iPLEX Assays. Das Schema stellt einen einzelnen Assay dar. Nach (Sequenom 2006).

2.2.3.3 TaqMan® Genotypisierung

Um einige wenige SNPs zu genotypisieren, wurde die kostengünstigere und robuste TaqMan®-Technologie verwendet. TaqMan® basiert auf dem PCR-Verfahren (Mullis et al. 1986), bei dem die 5'-Exonukleaseaktivität der DNA-Polymerase ausgenutzt wird (De la Vega et al. 2005; Livak 1999).

Neben den Primern werden in der TaqMan®-PCR zwei Oligonukleotide, welche jeweils mit einem Fluoreszenzfarbstoff (meist VIC und FAM) an ihrem 5‘-Ende markiert sind, eingesetzt (Abb. 2-6). Gleichzeitig ist an diese sogenannten TaqMan®-Sonden an ihrem 3‘-Ende ein Quencher gebunden. Dieser Quencher verhindert durch seine räumliche Nähe innerhalb der Sonde die Fluoreszenz des an der Sonde gebundenen Farbstoffes. Die Oligonukleotidsequenz der Sonden wird so designt, dass sie allelspezifisch die den SNP umgebende DNA-Sequenz bindet. Die Primer werden so designt, dass sie ein ca. 200 Nukleotid langes, den SNP umschließendes Fragment amplifizieren. Während der PCR binden die Sonden allelspezifisch die SNP-Region, die zwischen den Primer-Bindestellen liegt. Mit ihrer 5‘->3‘ Exonukleaseaktivität beginnt die Polymerase die gebundene Sonde vom 5‘-Ende her zu spalten, sobald diese erreicht wird. Durch den Abbau der Sonde werden das Fluorophor und der Quencher voneinander getrennt, so dass der Quencher keine Wirkung mehr auf den Fluoreszenzfarbstoff hat und der Farbstoff sichtbar wird. Mit jeder Amplifizierung der DNA-Fragmente durch die PCR nimmt auch die detektierte Fluoreszenz zu. Bei einem homozygoten Genotyp wird nach Ende der PCR mittels Laser-Scan-Technologie ein einziges Fluoreszenzsignal detektiert, je nach SNP-Allel entweder der VIC- oder FAM-Farbstoff. Bei einem heterozygoten Genotyp können beide Signale detektiert werden. Dies erlaubt eine direkte Zuordnung der Genotypen in den jeweiligen Proben.

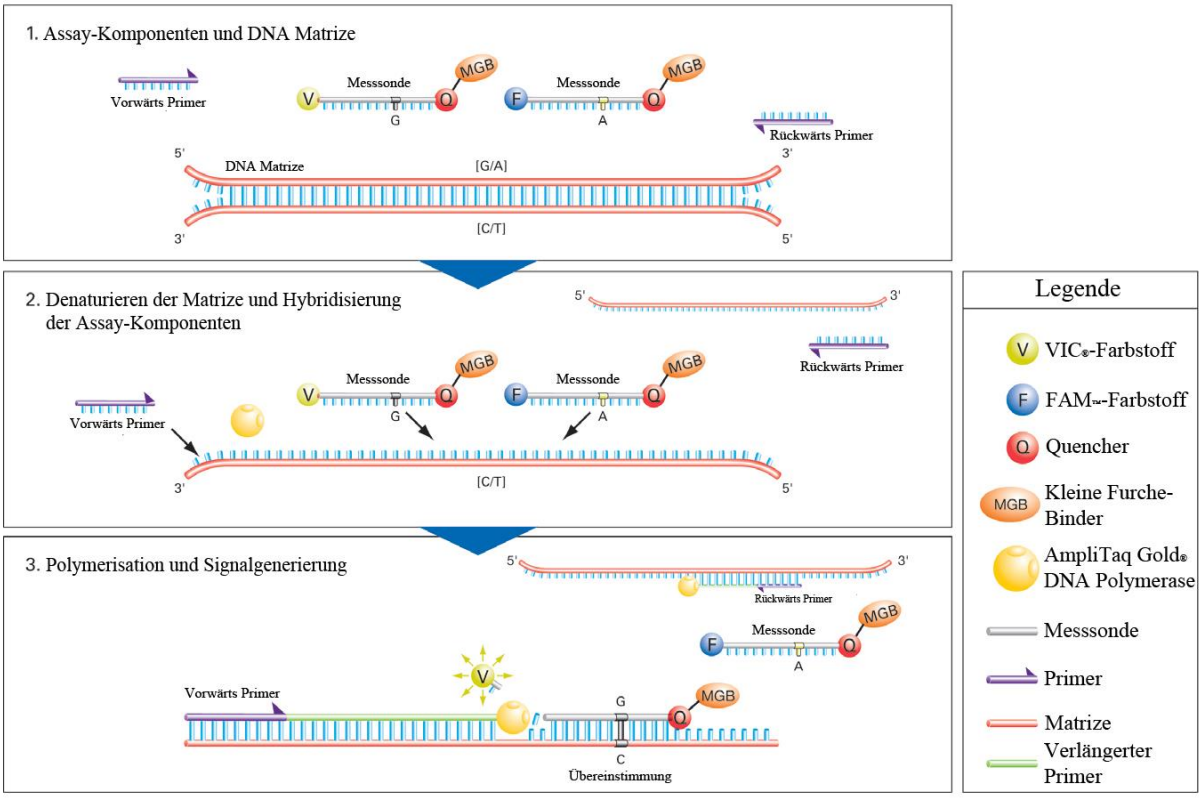

Abb. 2-6: Die allelische Diskriminierung (SNP Genotypisierung) wird durch selektive Hybridisierung der TaqMan®-MGB Messsonden erreicht. Nach (Applied Biosystems 2011).

Die TaqMan® Genotypisierung wurde mit 2 ng getrockneter, WGA-amplifizierter DNA nach einem einheitlichen PCR-Protokoll durchgeführt. Um für jede Probe die optimalen Ergebnisse zu erzielen, wurde gegebenenfalls die primäre 60°C-Anlagerungstemperatur auf 58°C oder 62°C angepasst. Die PCR wurde mit dem GeneAmp PCR System 9700 (Applied Biosystems, Forster City) durchgeführt. Für die Detektion der Fluoreszenz wurde das Fluoreszenz-Messgerät ABI Prism 7900HT (Applied Biosystems, Forster City) verwendet. Die Messergebnisse wurden mit der ‚Sequence Detection System'-Software 2.0® als Koordinaten in einem zweidimensionalem Koordinatensystem eingetragen. Wie bei der Auswertung der Sequenom-Daten konzentrieren sich identische Genotypen in sogenannten Clustern. Die automatische Zuordnung der Genotypen wurde auch hier manuell kontrolliert und gegebenenfalls korrigiert.

2.2.4 Genotyp-Imputation

Mit dem Affymetrix SNP-Array 6.0 können bis zu 906.600 SNPs genotypisiert werden. Verglichen mit der erwarteten Anzahl von mehr als zehn Millionen SNPs (mit einer MAF von ≥ 0,05) im menschlichen Genom kann mit dem Affymetrix SNP-Array 6.0 nur ein Bruchteil aller häufigen Varianten untersucht werden (Frazer et al. 2007). Die Genotyp-Imputation ist eine statistische Methode, mit der die in einer Probe fehlenden Marker, unter Zuhilfenahme von Referenz-Datensätzen, geschätzt (imputiert) werden. Mit den durch das „*HapMap*"-Konsortium (International HapMap Consortium 2003; Pemberton et al. 2010) und durch das „*1000 Genomes Project*" (1000 Genomes Project Consortium 2010) mittels spezialisierten Hochdurchsatz-Genotypisierungsverfahren gewonnenen Haplotyp-Daten als Grundlage war es möglich die Methode der Genotyp-Imputation zu entwickeln. Mit Hilfe dieser Referenzdaten können mit dem Array nicht-typisierte Marker imputiert werden. Die Genotyp-Imputation kann somit die Teststärke eines bestehenden GWAS durch das Hinzufügen von Markern, die sich nicht auf dem ursprünglichen Array befinden, erhöhen (Guan and Stephens 2008; Li et al. 2009; Marchini et al. 2007). Ein weiterer interessanter Aspekt ist, dass durch Genotyp-Imputation Metaanalysen von Datensätzen mit unterschiedlicher Markerdichte möglich wurden (de Bakker et al. 2008; Zeggini and Ioannidis 2009), da die Ungleichheit in der Marker-Abdeckung in den verschiedenen Datensätzen durch die Imputation verringert wird. Bei der Imputation werden die genotypisierten Marker einer Probe mit den Markern in den Referenz-Datensätzen abgeglichen. Dabei werden Regionen identifiziert, die sich die Probe mit der Referenz teilt. Die zusätzlichen Marker der Referenz-Datensätze in diesen Regionen werden nun in die Proben-Datensätze integriert (Abb. 2-7). Die Imputation führt so normalerweise zu einem Anstieg in der statistischen Teststärke um ca. 10% im Vergleich zu den ursprünglichen GWAS-Daten (Spencer et al. 2009).

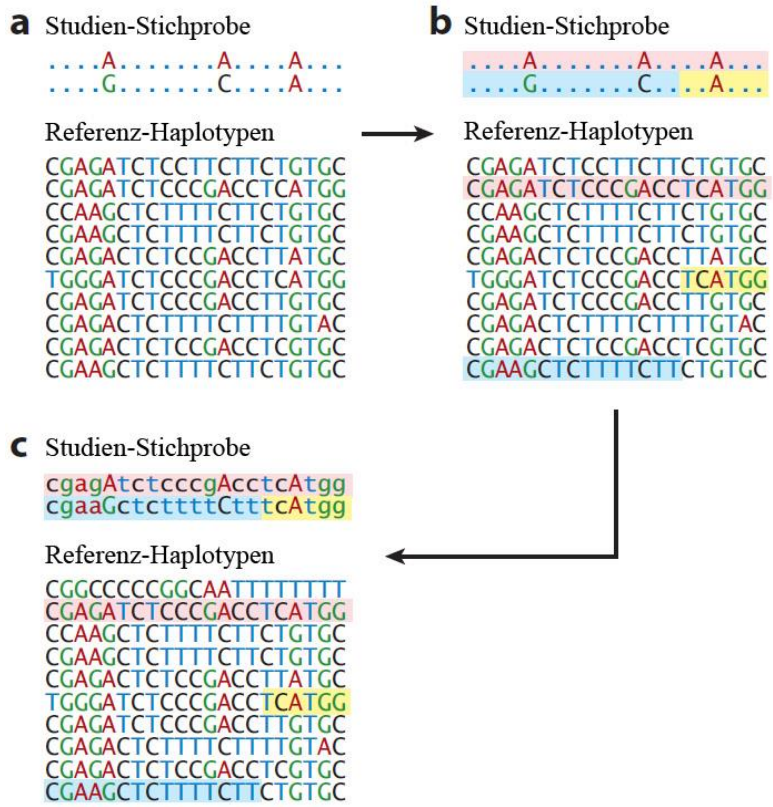

Abb. 2-7: Genotyp-Imputation in einer Stichprobe unverwandter Individuen. (a) Die in der Studie erhaltenen Daten (Studien-Stichprobe s.o. Bild) zeigen nur eine geringe Anzahl von genetischen Markern in den einzelnen untersuchten Individuen. Detaillierte Haplotypinformationen sind für eine Referenzpopulation verfügbar. **(b)** Es werden Chromosomregionen identifiziert, die sich die untersuchten Individuen mit der Referenzpopulation teilen. In europäischen Populationen werden normalerweise Abschnitte von >100 kb identifiziert. **(c)** Die Studiendaten und die als geteilt identifizierten Haplotypinformationen werden kombiniert um die fehlenden Genotypen in den Studiendaten vorherzusagen. Nach (Li et al. 2009).

Die in dieser Arbeit verwendeten GWAS-Datensätze wurden von Dr. David Ellinghaus mit dem Programm BEAGLE (S. R. Browning 2006; S. R. Browning and Browning 2007) unter Verwendung entsprechender Referenzdatensäten von ‚*HapMap Phase 3*' (D. M. Altshuler et al. 2010) imputiert. Dabei wurden für die Stichprobe A die Referenzdatensätze CEU, TSI und MEX verwendet, während für die Stichprobe D die Referenzdatensätze YRI, LWK und MKK verwendet wurden. Das Programm BEAGLE wurde mit den Standardeinstellungen ausgeführt und die Berechnungen mit dem Markov-Modell wurden zehnmal wiederholt (java -Xmx27377m -jar beagle.jar phased=phased.input.bgl unphased=unphased.input.bgl markers=marker.ids missing=0 gprobs=true niterations=10 out=out_file). Die allel-basierte Assoziationsanalyse zwischen dem Phänotyp und den prognostizierten Allel-Anzahlen wurde mit dem Programm PLINK (Purcell et al. 2007) unter der Nutzung des logistischen Regressionsmodells für Dosis-Daten (engl. *dosage data*) durchgeführt, um möglichen Ungenauigkeiten bei der Imputation der Genotypdaten Rechnung zu tragen. Um die Daten für potentiell verzerrende Faktoren der Populationsstratifikation zu korrigieren, wurden diese

mit dem Programm EIGENSTRAT (Price et al. 2006) für die 6 Eigenvektoren mit den höchsten Eigenwerten bereinigt.

2.2.5 Auswertung und Qualitätskontrolle

2.2.5.1 Datenverwaltung und Qualitätskontrolle

In einer GWAS wird eine große Anzahl von Marker-Loci (SNPs) auf Assoziation getestet. Bei einer Zahl von ca. 1 Million Markern führen selbst geringe Fehlerraten oder Messabweichungen dazu, dass sich die Teststärke, mit der Assoziationen festgestellt werden können, verringert und sich die Möglichkeit von falsch-positiven Assoziationen erhöht. Die Datenqualität einer GWAS ist also essentiell, um sicherzugehen, dass man echte genetische Assoziationen nachweist. Um falsch-positive Assoziationen zu vermeiden, war es also nötig verschiedene Schritte der Qualitätskontrolle durchzuführen und dabei Marker oder Individuen mit hohen Genotypisierungs-Fehlerraten zu entfernen. In gleicher Weise war es nötig außer den GWAS-Stichproben auch alle anderen in dieser Arbeit verwendeten Stichproben einer Qualitätskontrolle zu unterziehen.

2.2.5.2 Qualitätskontrolle auf Markerbasis

Fehler in der Genotypisierung führen häufig dazu, dass SNP-Genotypen fälschlicherweise als homozygot erkannt werden und es somit zu einem Fehlen von heterozygoten SNP-Genotypen in der Stichprobe kommen kann. In einem solchen Fall kommt es zu einer Abweichung von dem Hardy-Weinberg-Gleichgewicht (engl. *Hardy-Weinberg equilibrium*; HWE). Das nach dem Mathematiker G. H. Hardy und dem Arzt Wilhelm Weinberg benannte Modell beschreibt das Verhältnis zwischen Allel- und Genotypfrequenz in einer großen, sich zufällig paarenden Population ohne die Faktoren Selektion, Mutation und Migration (G. H. Hardy 1908; Weinberg 1908). Bei einem Locus mit den zwei Allelen α und *A* und den dazugehörigen Allelfrequenzen $p_\alpha = p$ und $p_A = q$ entsprechen nach dem HWE die Wahrscheinlichkeiten der drei möglichen Genotypen αα, α*A*, *AA* den Genotypfrequenzen p^2, $2pq$, q^2. Für eine große, auf zufälliger Paarung basierende Population hat dies folgende Auswirkungen: Die Allelfrequenzen sind über Generationen hinweg konstant. Des Weiteren können die Genotypfrequenzen bei den Nachkommen mithilfe der Allelfrequenzen der Elterngeneration vorhergesagt werden. Abweichungen von diesem Gleichgewichtszustand werden als Hardy-Weinberg-Ungleichgewicht bezeichnet und weisen entweder auf Genotypisierungsfehler oder das Auftreten von Selektion, Mutation und Migration hin. In Stichproben mit kranken Individuen kann eine solche Abweichung allerdings auch ein Hinweis auf eine mögliche Krankheitsassoziation sein. In Assoziationsstudien wird daher auf Abweichungen vom HWE in den Kontroll-Individuen getestet, um

Genotypisierungsfehler aufzudecken. Abweichungen vom HWE werden heutzutage mit einem von Wigginton *et al.* (2005) beschriebenen exakten Test überprüft; dieser Test ist in dem Softwarepaket PLINK (Purcell et al. 2007) implementiert.

GWAS-Daten werden, abgesehen davon, dass jeder SNP auf signifikante Abweichungen vom HWE ($p < 0{,}001$) getestet wird, routinemäßig weiteren Qualitätskontrollen unterzogen (Anderson et al. 2010):

1) SNPs mit einer zu hohen Rate an fehlenden Genotypdaten (> 5%) werden ausgeschlossen.

2) Marker mit einer zu geringen Frequenz des seltenen Allels (MAF < 1%) werden entfernt.

3) SNPs mit einer signifikant unterschiedlichen Rate an fehlenden Genotypdaten zwischen Fällen und Kontrollen werden ausgeschlossen.

All diese Schritte wurden bei der Qualitätskontrolle für sämtliche in dieser Arbeit verwendeten Stichproben durchgeführt. Die Qualitätskontrolle der GWAS-Datensätze wurde von Dr. David Ellinghaus durchgeführt.

2.2.5.3 Qualitätskontrolle einzelner Individuen

Die Qualitätskontrolle der in den GWAS Daten enthaltenen Individuen sollte mindestens fünf Schritte umfassen (Anderson et al. 2010):

1) Den Ausschluss von Individuen mit diskordanten Geschlechtsinformationen.

2) Ausschluss von Individuen mit einer zu hohen Rate an fehlenden Genotypdaten (> 5%).

3) Individuen mit außerhalb des Medians der Stichprobe liegenden Heterozygositätsraten (engl. *outlying heterozygosity rates*) werden entfernt.

4) Ausschluss von doppelten oder nahverwandten Individuen.

5) Individuen mit abweichender Herkunft werden entfernt.

Um Diskrepanzen zwischen Genotypinfomationen und festgestelltem Geschlecht eines Individuums zu ermitteln, werden die Homozygositätsraten für jedes Individuum für alle X-chromosomalen SNPs berechnet und mit den erwarteten Werten verglichen. Normalerweise erwartet man bei Männern eine Homozygositätsrate nahe 1 (wegen Genotypisierungsfehlern kann diese leicht variieren) und bei Frauen eine Homozygositätsrate von < 0,2 (Anderson et al. 2010). Individuen mit nicht übereinstimmenden Geschlechtsinformationen werden für die weitere Analyse entfernt.

Der Anteil an fehlenden Genotypen ist ein nützlicher Indikator für schlechte Genotypisierungsqualität der Proben. Ein großer Anteil an fehlenden Genotypen weist normalerweise auf Probleme mit der Hybridisierung hin, die durch fehlerhafte SNP-Arrays oder schlechte DNA-Qualität entstehen können

(Ziegler et al. 2010). Proben, die Genotypen für weniger als 95% der SNPs besitzen, werden normalerweise von der Analyse ausgeschlossen.

Die Heterozygositätsrate beschreibt die Proportion der heterozygoten Genotypen für ein Individuum und kann ein Gradmesser für die DNA-Qualität sein. Wenn man die Verteilung der mittleren Heterozygosität (ausgenommen der Geschlechts-Chromosomen) über alle Individuen betrachtet, können Individuen mit einem überhöhten oder verringerten Maß an heterozygoten Genotypen (mit einer Standardabweichung +/- 5 von der mittleren Heterozygosität) erkannt und ausgeschlossen werden.

Duplikate und Verwandte ersten oder zweiten Grades können einen systematischen Fehler in populationsbasierten Fall-Kontroll-Assoziationsstudien einführen, da bestimmte Genotypen in Familien überrepräsentiert sind. Um solche Duplikate und verwandte Individuen zu identifizieren, werden Identität-durch-Status-Werte (engl. *identity-by-state*; IBS) für jedes in der Studie vorkommende Individuen-Paar, basierend auf dem durchschnittlichen Anteil der gemeinsamen Allele an (unkorrelierten) genotypisierten SNPs (ausgenommen der Geschlechts-Chromosomen), berechnet. Verwandte Individuen teilen sich mehr Allele, als man nach einer Zufallsverteilung erwarten würde, und haben einen höheren IBS-Wert (dabei zeigt ein IBS-Wert von 1 eine maximale Verwandtschaft (Duplikat) an), d.h. mit dem Grad an geteilten Allelen steigt auch der Grad der Verwandtschaft und der IBS-Wert nähert sich dem Wert 1 an. (Anderson et al. 2010). Bei Individuen mit hohen IBS-Werten wurde jeweils das Individuum mit der schlechteren Genotypisierungsrate von der Analyse ausgeschlossen.

All diese Schritte wurden bei der Qualitätskontrolle für sämtliche in dieser Arbeit verwendeten Stichproben durchgeführt. Die Qualitätskontrolle der GWAS-Datensätze wurde von Dr. David Ellinghaus durchgeführt.

2.2.5.4 Populationsstratifikation

Um Populationsstratifikationen (die systematische Differenz in Allelfrequenzen zwischen Subpopulationen einer Population) zu erfassen und zu korrigieren, wird in genomweiten Ansätzen das Verfahren der Hauptkomponentenanalyse (engl. *principal component analysis*; PCA) verwendet (Price et al. 2006). Die PCA ist eine Methode der multivarianten Statistik, die dazu benutzt wird, mehrere unkorrelierte Variablen (die Hauptkomponenten, engl. *principal components*) aus einer Datenmatrix, die Observationen einer Reihe von potentiell miteinander korrelierten Variablen enthält, zu berechnen. Die Hauptkomponenten werden so konstruiert, dass die erste Hauptkomponente für den größten Teil der Variation in den Originaldaten verantwortlich ist. Dieser folgen dann die zweite Hauptkomponente und die weiteren in absteigender Bedeutung.

Unterschiede in der Herkunft der Proben können auch geographisch interpretiert werden, wobei schon die ersten beiden Hauptkomponenten ausreichen, damit sich Individuen aus verschiedenen europäischen Ländern getrennt gruppieren (Heath et al. 2008). Die PCA-Methode zur Detektion der Herkunft ist in dem Softwarepaket EIGENSOFT implementiert (Price et al. 2006). Mit dieser Methode wurden die verwendeten GWAS-Datensätze von Dr. David Ellinghaus auf Populationsstratifikation überprüft. Mit dem genomischen Inflationsfaktor λ_{GC} kann der Grad der Populationsstratifikation dargestellt werden. Je stärker der Inflationsfaktor von $\lambda_{GC} = 1$ abweicht, desto größer ist der Grad an Populationsstratifikation in der untersuchten Probe. Dabei wird die Verteilung der χ^2-Statistik in einem sog. Quantil-Quantil-Diagramm (engl. *quantile-quantile-plot, Q-Q-Plot*) aufgetragen (Beispiel in Abb. 2-8).

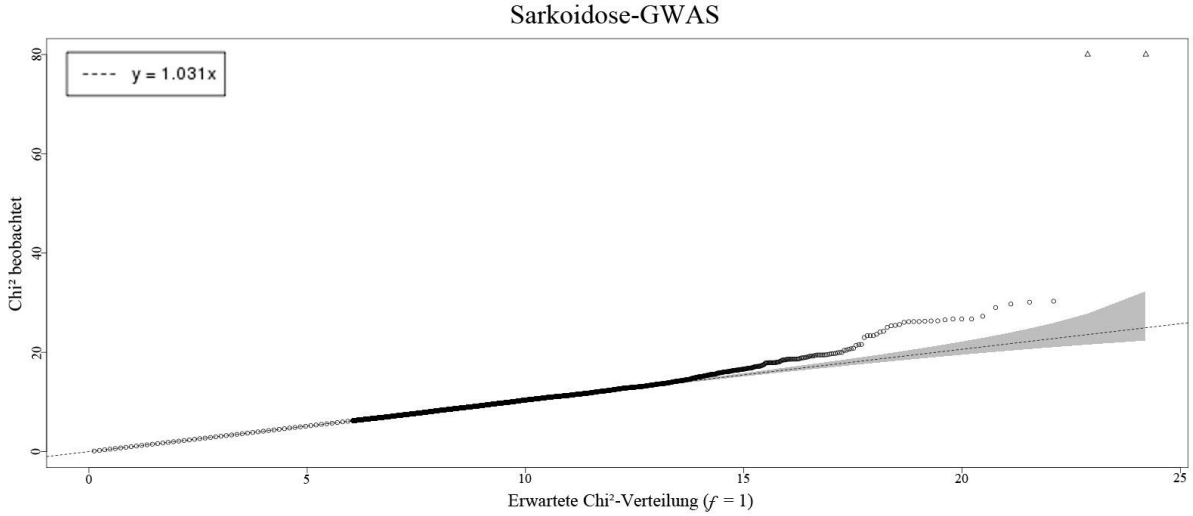

Abb. 2-8: Quantil-Quantil-Diagramm (Q-Q-Plot) für die Sarkoidosestichprobe A. Die theoretischen Quantile einer Verteilung (Erwartete X^2-Verteilung) werden gegen die empirischen Quantile von beobachteten Merkmalswerten (hier X^2 beobachtet) aufgetragen. Wenn die Merkmalswerte aus der Vergleichsverteilung stammen, stimmen die empirischen und die theoretischen Quantile gut überein und liegen nahe der gestrichelten Linie. Der grau unterlegte Bereich entspricht einer Konfidenz von 95 % für jeden Vergleich zwischen dem empirischem und dem theoretischen Quantil. Die Abbildung wurde mit der qq.chisq-Funktion des R-package snpMatrix v1.6.1 (http://www-gene.cimr.cam.ac.uk/clayton/software/) erzeugt.

2.2.6 Methoden zur Analyse von genomweiten Assoziationsdaten

2.2.6.1 Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen

Die Analysen des Sarkoidose-GWAS (564 Sarkoidosefälle, davon 176 *akut* und 354 *chronisch*, und 1.575 Kontrollpersonen) wurden in drei verschiedenen Fall-Kontroll-Zusammensetzungen durchgeführt, damit neben dem „normalen“ Phänotyp Sarkoidose auch dessen Subphänotypen untersucht werden konnten. Für die Analysen wurden nur SNPs verwendet, die mindestens einen „Aggregation“-SNP (engl. *clumping SNP*) ($p \leq 0{,}05$) im starken LD ($r^2 \geq 0{,}8$) besitzen, damit mögliche Genotypisierungs- und Imputationsartefakte in dem Datensatz weitestgehend ausgeschlossen

werden konnten. Für die SNP-Auswahl wurden *p*-Schwellenwerte definiert, die stringent genug sind, um die Akkumulation von falsch-positiven Assoziationen zu vermeiden, dabei aber immer noch die Detektion vielversprechender Assoziationssignale zulassen:

1) In der ersten Analyse wurden alle Sarkoidosefälle zusammen mit den Kontrollpersonen untersucht (im Folgenden als „Allgemeine Analyse“ bezeichnet). Die SNPs wurden nach ihrem *p*-Wert im allelbasierten χ^2-Test (ƒ = 1) in aufsteigender Reihenfolge geordnet und einer visuellen Überprüfung in Form einer graphischen Darstellung des SNPs und seiner Umgebung (engl. *regional plot*, Beispiel in Abb. 2-9) unterzogen. Alle SNPs mit einem *p*-Wert < 10^{-5} wurden für die weiterführenden Analysen ausgewählt.

2) In der zweiten Analyse wurden nur die *akuten* Sarkoidosefälle zusammen mit den Kontrollpersonen untersucht (als „Analyse akut“ bezeichnet). Die SNPs wurden ebenfalls nach ihrem *p*-Wert geordnet und visuell überprüft. Alle SNPs mit einem *p*-Wert von $p < 10^{-4}$ wurden für die weiteren Untersuchungen ausgewählt.

3) In einer dritten Analyse wurden nur *chronische* Sarkoidosefälle zusammen mit den Kontrollpersonen untersucht (als „Analyse chronisch“ bezeichnet). Die SNP-Auswahl erfolgte entsprechend den Kriterien der „Analyse akut“.

Die statistischen Analysen der Genotypdaten wurden mit dem Programm PLINK (Purcell et al. 2007) durchgeführt.

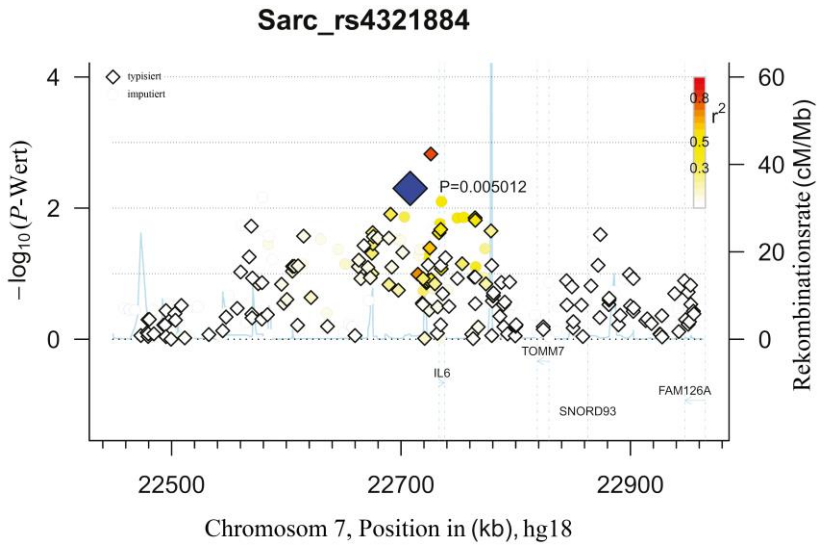

Abb. 2-9: Regional Plot des SNPs rs4321884 (GWAS Daten). Chromosomale Positionen beziehen sich auf das „human reference assembly GRCh36“. Die Form der SNPs bezeichnet ihren Genotypisierungsstatus: Typisierte SNPs sind in Diamantenform dargestellt und imputierte SNPs als Kreisform. Der untersuchte SNP ist vergrößert und in blau dargestellt. Der Farbcode kennzeichnet das LD zwischen dem SNP rs4321884 und den umgebenden Markern. Die Rekombinationsrate ist als blaue Linie dargestellt und beruht auf den HapMap-Daten. Die Genpositionen und Gennomenklatur basiert auf der RefSeq-Datenbank.

2.2.6.2 Identifizierung gemeinsamer genetischer Faktoren der Krankheiten Tuberkulose und Sarkoidose

Mit den GWAS-Datensätzen wurden zunächst für jede Krankheit einzeln Assoziationsanalysen durchgeführt. Danach wurden die Assoziationsdaten der GWAS-Datensätze gefiltert, um mögliche Genotypisierungs- und Imputationsartefakte auszuschließen. Für die weiteren Analysen wurden nur SNPs verwendet, die mindestens einen *clumping* SNP ($p \leq 0{,}05$) im starken LD ($r^2 \geq 0{,}8$) besitzen. Die gefilterten Daten der Assoziationsanalysen der beiden GWAS wurden nun gemeinsam analysiert. Da die verwendeten Datensätze aus zwei verschiedenen Regionen stammen (Ghana und Deutschland), zwischen deren Populationen eine starke genetische Varianz besteht, wurden die Genotypdaten der imputierten GWAS mit unterschiedlichen Methoden analysiert. Insgesamt wurden vier unterschiedliche Analysestrategien angewendet, um SNPs für die weiteren Untersuchungen aus den beiden GWAS-Datensätzen auszuwählen. Dabei wurden für die SNP-Auswahl die *p*-Schwellenwerte so definiert, dass diese stringent genug sind, um die Akkumulation von falsch-positiven Assoziationen zu vermeiden, dabei aber immer noch die Detektion gemeinsamer Assoziationssignale der beiden Krankheiten zulassen:

1) Im ersten Ansatz (im Folgenden „*meta-analysis fixed effects*“ genannt) wurden die GWAS-Datensätze in einer Metaanalyse mit einem festen Effekt-Modell (engl. *meta-analysis fixed effects*) analysiert. Dieser Analyse liegt die Annahme zugrunde, dass sich beide Krankheiten einen Effekt mit einer gleicher Effektrichtung (Risiko erhöhend oder protektiv) teilen.

2) Ein zweiter Ansatz (im Folgenden „*meta-analysis opposite effects*“ genannt) bestand aus einer Metaanalyse mit einem entgegengesetzten Effekt-Modell (engl. *meta-analysis opposite effects*). In dieser Analyse wurden die ORs aller SNPs der Stichprobe D (Tuberkulose-GWAS) umgedreht, um einen möglichen entgegengesetzten Effekt an einem Marker zu berücksichtigen. Für die beiden Metaanalysen wurden nur SNPs mit einem *p*-Wert $< 1 \times 10^{-2}$ in den einzelnen Datensätzen berücksichtigt. Nach Durchführung der Metaanalysen wurden die SNPs in beiden Ansätzen nach ihrem *p*-Wert im allelbasierten χ^2-Test ($f = 1$) in aufsteigender Reihenfolge geordnet und einer visuellen Überprüfung durch *regional plots* unterzogen. In beiden Ansätzen wurden alle SNPs mit einem *p*-Wert $< 1 \times 10^{-4}$ für die weiterführenden Analysen ausgewählt. Die Metaanalysen wurden mit dem Programm PLINK durchgeführt (Purcell et al. 2007).

3) In einem dritten Ansatz (im Folgenden „*LD cluster ranking*“ genannt) wurde eine auf dem LD-Muster basierende SNP-Ranking-Analyse durchgeführt. Studien haben gezeigt, dass sich das LD auf Bereiche von über 100kb im humanen Genom erstrecken kann (Abecasis et al. 2001; A. Collins et al. 1999; D. E. Reich et al. 2001; Taillon-Miller et al. 2000). Da in dieser Arbeit zwei in ihrer genomischen Struktur stark unterschiedliche Populationen untersucht werden (Ghanaer und Deutsche), sollte eine

mögliche Überschneidung von LD-Blocks berücksichtigt werden. Um möglichst alle vorhandenen LD-Block-Überschneidungen zu detektieren, wurde in diesem Ansatz ein sehr großzügiger Bereich (250kb) auf LD in den einzelnen Populationen überprüft.

Die SNPs der jeweiligen GWAS wurden dafür in LD-Gruppen zusammengefasst und mit diesen LD-Gruppen wurde dann eine Ranking-Analyse durchgeführt. Für die Bildung der LD-Gruppen wurde aus jedem assoziierten Locus der SNP mit dem stärksten Assoziationssignal ausgewählt (im Folgenden ‚lead-SNP' genannt; $p_{LEAD} < 5 \times 10^{-3}$), diese ‚lead-SNPs' fungierten nun bei der Ranglistenbildung als Repräsentant der jeweiligen LD-Gruppen. Alle SNPs ($p_{LD} < 0,05$), die in einem starken LD ($r^2 \geq 0,8$) mit dem ‚lead-SNP' standen und nicht weiter als 250kb vom ‚lead-SNP' entfernt lagen, wurden den jeweiligen LD-Gruppen zugeordnet (Abb. 2-10). Hierfür wurden die Datensätze mit dem „*LD-based result clumping*"-Verfahren des Programms PLINK (Purcell et al. 2007) gefiltert ($p_{LEAD} < 5x\ 10^{-3}$, $p_{LD} < 0,05$, $r^2 \geq 0,8$, kb = 250). Die ‚lead-SNPs' wurden nun benutzt, um die LD-Gruppen in krankheitsspezifische Ranglisten nach *p*-Wert in aufsteigender Reihenfolge anzuordnen. Diese beiden Ranglisten wurden dann auf SNP-Überschneidungen zwischen den beiden Krankheiten überprüft. Bei Überschneidungen von SNPs zwischen den krankheitsspezifischen Ranglisten wurden die Ränge der ‚lead-SNPs' der jeweiligen LD-Gruppen addiert und in eine neue Rangliste übertragen. Diese Rangliste wurde in aufsteigender Reihenfolge der Ränge geordnet und anhand derer SNPs für die weiteren Untersuchungen ausgewählt. In dem Fall, dass unterschiedliche ‚lead-SNPs' hochrangige LD-Gruppen repräsentierten, wurden beide SNPs für die Nachverfolgung ausgewählt. Alle ausgewählten SNPs wurden ebenfalls visuell anhand von *regional plots* überprüft.

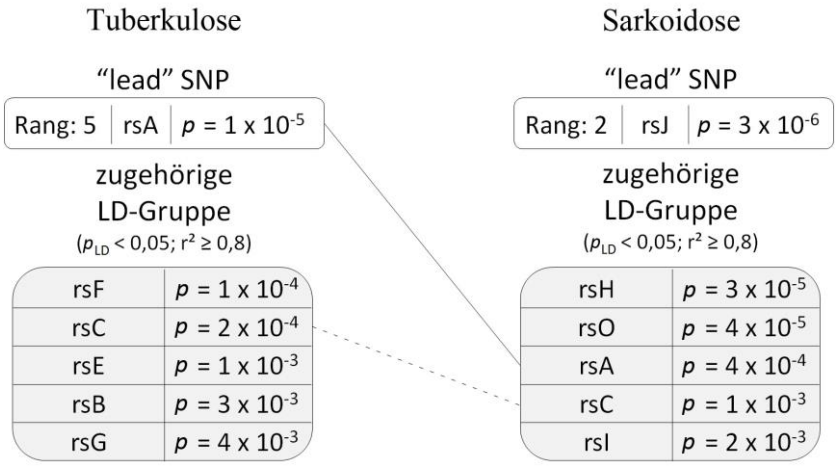

Abb. 2-10: Schematische Darstellung des *‚LD cluster rankings'*. Für jede Krankheit ist eine LD-Gruppe mit ihrem dazugehörigen lead-SNP dargestellt. Gibt es eine Überschneidung von SNPs zwischen den Krankheiten, werden die Ränge der lead-SNPs addiert und in eine „kombinierte" Rangliste übertragen. Dabei ist es unerheblich, ob die zwei SNPs aus der LD-Gruppe sind (rsC, gestrichelte Linie) oder ob einer der SNPs ein lead-SNP ist (rsA, durchgezogene Linie). In beiden genannten Fällen würden in der neuen Rangliste zwei SNPs (rsA, rsJ) den Rang 7 repräsentieren, da die lead-SNPs nicht identisch sind.

4) In dem letzten Ansatz (im Folgenden „*gene ranking*“ genannt) wurde eine Rankingmethode entwickelt, die auf einer SNP-zu-Gen Zuweisung beruht. Hierfür wurden für jede Krankheit alle SNPs ($p < 10^{-3}$), die innerhalb der transkribierten Regionen (+/- 1kb für potentielle regulatorische Bereiche; basierend auf der Ensembl-Datenbank (Flicek et al. 2010)) eines Gens liegen, diesem zugeordnet. In jedem Datensatz wurde den Genen (wie bei dem *LD cluster ranking*) ein ‚lead-SNP' (der SNP mit dem niedrigsten *p*-Wert in der definierten Region; min. $p < 10^{-3}$) zugewiesen. Mithilfe des ‚lead-SNPs' wurden in den jeweiligen Datensätzen Ranglisten für die Gene erstellt. Bei Überschneidungen wurden die Ränge der Gene addiert und in eine neue Rangliste übertragen. Von allen Genen dieser Rangliste wurden die ‚lead-SNPs' für die weiteren Analysen ausgewählt, wobei diese zuvor visuell anhand der *regional plots* überprüft wurden. Die *gene ranking*-Analyse wurde mit dem Programm R (R Core Team 2013) durchgeführt.

2.2.7 Statistische Methoden

2.2.7.1 Genetische Assoziation

In einer populationsbasierten genetischen Fall-Kontroll-Assoziationsstudie werden die Frequenzen von Allelen oder Genotypen an Marker-Loci von Individuen einer gegebenen Population (mit und ohne einem gegebenen Krankheitsmerkmal) verglichen, um zu bestimmen, ob eine statistische Assoziation zwischen einem Marker und dem Krankheitsmerkmal besteht. Ein biallelischer SNP (mit den zwei Allelen α und *A*) hat die möglichen Genotypen αα, α*A*, *AA*. Stellt eines der Allele (z.B. *A*) des SNPs einen Risikofaktor für die Krankheit dar, dann können die Allele α und *A* des SNPs der n Individuen einer Fall-Kontroll-Gruppe in einer 2x2 Kontingenztafel aufgetragen werden (Tab. 2-2) (Clarke et al. 2011).

Tab. 2-2: 2 x 2 Kontingenztafel die den Krankheitsstatus beschreibt.

Allel	α	*A*	Summe
Fälle	m_{11}	m_{12}	$m_{1.}$
Kontrollen	m_{21}	m_{22}	$m_{2.}$
Summe	$m_{.1}$	$m_{.2}$	$2n$

Die Penetranz der Krankheit, die mit einem gegebenen Genotyp assoziiert ist, ist die konditionelle Wahrscheinlichkeit $P(\chi|g)$ für das Ereignis zu erkranken χ, wenn ein spezifischer Genotyp *g* vorhanden ist (Ziegler et al. 2010). Die Standardmodelle für die Penetranz einer Krankheit sind: Multiplikative, additive, rezessive und dominante Modelle. Bei einem additiven Modell mit dem genetischen Penetranz-Parameter γ (γ > 1) ist das Krankheitsrisiko bei dem Genotyp α*A* um den Faktor γ und bei dem Genotyp *AA* um 2 γ erhöht. Die Stärke einer Assoziation kann auch mit dem

relativen Risiko (RR) ausgedrückt werden. Das relative genotypische Risiko stellt das erhöhte Risiko eines Individuums mit dem spezifischen Genotyp *g* gegenüber einem Individuum ohne krankheitsauslösende Allele dar (Ziegler et al. 2010). Auf Penetranzen basierte Schätzungen über das relative Risiko können nur direkt von prospektiven Kohorten-Studien durchgeführt werden. In diesen Studien werden Individuen einer gleichen Population, die einem Faktor ausgesetzt oder nicht ausgesetzt sind, über einen längeren Zeitraum verfolgt, um herauszufinden, bei welchen Individuen sich eine Krankheit entwickelt (Clarke et al. 2011). In Fall-Kontroll-Studien ist es allerdings nicht möglich direkte Schätzungen über die Penetranz einer Krankheit zu machen, so dass diese daher nicht mit dem relativen Risiko dargestellt werden kann. In Fall-Kontroll-Studien wird die Stärke einer Assoziation durch das Quotenverhältnis (engl. *odds ratio*; OR) ausgedrückt. Das OR beschreibt die Wahrscheinlichkeit, ob eine Krankheit vorhanden ist, gegenüber der Wahrscheinlichkeit, dass die Krankheit nicht vorhanden ist, in einer Gruppe von Individuen mit Risikofaktor im Vergleich mit einer Gruppe ohne Risikofaktor. Das allelische OR beschreibt die Assoziation zwischen Krankheit und Allel, durch den Vergleich der Chance auf eine Erkrankung für ein Individuum mit dem Allel *A* mit der Chance auf eine Erkrankung für ein Individuum mit dem Allel α (Clarke et al. 2011):

$$OR_A = \frac{m_{12}m_{21}}{m_{11}m_{22}}$$

Wobei erst das dazugehörige 95 %-Konfidenzintervall (KI) (engl. *confidence interval*; CI):

$$KI = \ln(OR) \pm \left[1{,}96 \sqrt{\frac{1}{m_{12}} + \frac{1}{m_{11}} + \frac{1}{m_{22}} + \frac{1}{m_{21}}} \right]$$

eine Beurteilung des OR ermöglicht. Nur wenn das KI den Wert 1 nicht einschließt, ist ein Ergebnis signifikant.

Bei einer geringen Penetranz einer Krankheit besteht günstigenfalls kaum ein Unterschied zwischen RRs und ORs (d.h. RR ≈ OR). Des Weiteren kann das OR auch bei multivariaten Analysemethoden angewandt werden, mit denen weitere SNPs, Risikofaktoren und andere klinische Variablen berücksichtigt werden können (Clarke et al. 2011). Solche Analysemethoden schließen auch die logistische Regression und weitere Log-lineare Modelle ein (Bishop et al. 1975).

Tests auf genetische Assoziationen werden normalerweise für jeden SNP einzeln durchgeführt. Unter der Annahme der Nullhypothese (H_0), welche keine Assoziation mit der Krankheit zugrunde legt, erwartet man keinen Unterschied in den relativen Allel-Frequenzen zwischen der Fall- und Kontroll-Gruppe. Ein Test auf allelische Assoziation kann mit einem einfachen χ^2 (Chi-Quadrat)-Test durch Überprüfen auf Unabhängigkeit der Zeilen und Spalten der Kontingenztafel durchgeführt werden (Clarke et al. 2011):

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(m_{ij} - E[m_{ij}])^2}{E[m_{ij}]} \text{ wenn } E[m_{ij}] = \frac{m_{i\bullet}m_{\bullet j}}{2n}$$

χ^2 hat unter der Annahme der Nullhypothese eine χ^2 Verteilung mit einem Freiheitsgrad. Die Nullhypothese, die von keiner Assoziation mit der Krankheit ausgeht, wird bei hohen Werten der χ^2-Teststatistik widerlegt. Der Test auf allelische Assoziation mit einem Freiheitsgrad hat eine bessere Teststärke als jeder genotypische Assoziationstest mit zwei Freiheitsgraden, wenn die Penetranz des Genotyps zwischen der Penetranz von zwei homozygoten Genotypen liegt (Clarke et al. 2011). Berechnungen der allelischen Assoziationen wurden in dieser Arbeit mit dem Programm PLINK (Purcell et al. 2007) durchgeführt.

2.2.7.2 Exakter Chi-Quadrat-Test (Exakter Fisher-Test)

Wenn die Anzahl der statistischen Beobachtungen sehr selten ist, wie z.B. bei seltenen Allelen (Varianten), dann kann der χ^2-Test von Karl Pearson irreführende Ergebnisse liefern. In diesem Fall ist der exakte Fisher-Test eine Form der Analyse, die bei seltenen Beobachtungen exakte Ergebnisse liefert (wenn auf Signifikanz in einer 2 x 2 Kontingenztafel getestet wird). Der Test nutzt, anstatt der Annäherung an die χ^2-Verteilung, die exakte hypergeometrische Verteilung, um auf Unabhängigkeit in der Kontingenztafel zu testen. In dieser Arbeit wurde der exakte Fisher-Test genutzt, wenn die Allelanzahl in einem der Felder der Kontingenztafel weniger als 5 betrug oder wenn die Allelanzahlen sehr ungleich waren. Bei normalen 2 x 2 Kontingenztafeln wurde der exakte Fisher-Test nicht genutzt, da für normale Verteilungen dieser Test ein über-konservativer Test ist. Analysen mit dem exakten Fisher-Test wurde mit dem Programm PLINK (Purcell et al. 2007) durchgeführt.

2.2.7.3 Mehrfaches Testen (Alphafehler-Kumulierung)

Wird ein SNP auf die Assoziation mit einer Krankheit getestet, dann sollte die Wahrscheinlichkeit *p*, mit der die Nullhypothese H_0 widerlegt werden kann, die in der Biologie etablierte Signifikanzgrenze α (Fehler 1. Art) von α = 0,05 nicht übersteigen. Da häufig mehrere Nullhypothesen in einem Datensatz getestet werden, erhöht sich die Wahrscheinlichkeit, dass eine von ihnen fälschlicherweise widerlegt wird (multiples Testproblem). Dieser Kumulierung oder Inflation von α-Fehlern kann mit verschiedenen statistischen Methoden entgegengewirkt werden. In der Biologie hat sich die Bonferroni-Methode als einfachste und konservativste Form für die Korrektur bei α-Fehler-Inflation durchgesetzt. Dabei wird die Signifikanzgrenze (α) auf die Anzahl der durchgeführten Tests (*n*) hin korrigiert (α/n).

Heutzutage beinhalten GWAS zwischen 10^5 und 10^6 SNPs. Geht man z.B. von 100.000 durchgeführten Tests aus, dann wird erwartet, dass 5% der SNPs (5.000 Tests) durch Zufall einen *p*-Wert < 0,05 haben (wenn die H_0 für alle Tests gilt). Simulationen haben ergeben, dass eine genomweite Untersuchung von häufigen Varianten (SNPs) dem Äquivalent von ~1 Million unabhängiger Hypothesen entspricht (Pe'er et al. 2008). Um falsch-positive Ergebnisse, die durch das mehrfache Testen auftreten können, zu vermeiden, ist eine stringentere Signifikanzgrenze für solche Datensätze nötig. Verschiedene Methoden wurden diskutiert, um die genomweite Signifikanzgrenze für GWAS zu bestimmen (Dudbridge and Gusnanto 2008; Duggal et al. 2008). Letztlich hat sich die sehr konservative Bonferroni-Korrektur als Standardmethode für mehrfaches Testen durchgesetzt, die auch in dieser Arbeit verwendet wurde. Mit der Bonferroni-Korrektur ergibt sich für häufige Varianten (MAF > 1%) eine Signifikanzgrenze von $\alpha_p = \alpha/10^6 = 5 \times 10^{-8}$. Diese genomweite Signifikanzgrenze ist überkonservativ, da die einzelnen statistischen Tests auf einem Chromosom wegen des vorherrschenden LDs zwischen den Markern nicht unabhängig voneinander sind. Um jedoch ein Übermaß an falsch-positiven Assoziationen in GWAS zu vermeiden, ist diese Signifikanzgrenze seit 2008 als allgemeine Norm akzeptiert (Pe'er et al. 2008).

2.2.7.4 Multivariate logistische Regression

Die logistische Regression gehört zu den statistischen Modellen der Generalisierten Linearen Modelle (GLM) und kann für genetische Analysen genutzt werden. Ist die abhängige Variable dichotomer Natur (z.B. ‚gesund' oder ‚krank'), wird die logistische Regression genutzt, um Kovariablen zu bestimmen, welche die abhängige Variable am besten erklären.

In dieser Arbeit wurde die logistische Regression benutzt, um SNP-SNP-Interaktionen zu bestimmen. Dabei wird untersucht, ob eine gefundene Mutation (SNP) alleine zur Vorhersage des Krankheitsstatus geeignet ist oder ob weitere SNPs das Vorhersagemodell verbessern können und somit den Effekt der gefundenen Mutation teilweise oder ganz erklären. D.h. also, ob der Effekt des SNPs auf den Krankheitsstatus abhängig oder unabhängig von weiteren SNPs ist.

Die genotypbasierte logistische Regressionsanalyse wurde mit dem Programm PLINK (Purcell et al. 2007) durchgeführt. Analysen zur Überprüfung der OR-Homogenität (Breslow-Day Test) wurden ebenfalls mit dem Programm PLINK berechnet.

2.2.7.5 Likelihood-Quotienten-Test

Der Likelihood-Quotienten-Test ist ein statistischer Test, der die Eignung von zwei theoretischen Hypothesen überprüft, mit welcher Wahrscheinlichkeit diese zuzutreffen. Dabei wird, auf Basis des

Wahrscheinlichkeitsquotienten, die Nullmodellwahrscheinlichkeit mit der Alternativmodellwahrscheinlichkeit verglichen. Der Wahrscheinlichkeitsquotient gibt also an, um wie viel wahrscheinlicher ein Ergebnis unter dem einen Modell als unter dem Alternativmodell ist. Mit dem Wahrscheinlichkeitsquotienten kann dann ein *p*-Wert berechnet werden, der die Wahrscheinlichkeit eines der Modelle ausdrückt. D.h. beide konkurrierenden Modelle, das Nullmodell und das Alternativmodell, werden unabhängig voneinander den Daten angepasst und der Wahrscheinlichkeitsquotient ergibt sich aus dem Vergleich der beiden Modelle.

Mit dem Programm PLINK (Purcell et al. 2007) wurde der Likelihood-Quotienten-Test verwendet, um zu testen, ob eine allelische Assoziation zwischen zwei Markern (SNPs) besteht. D.h. es wurde getestet, ob spezielle Paare von Allelen an zwei Loci häufiger oder weniger häufig, als der Zufall es erwarten lässt, zusammen in einem Individuum auftreten. Dabei wurde die Testmethode genau wie von Lyles *et al.* beschrieben verwendet (Lyles et al. 2007).

2.2.7.6 Epistase

Mit dem Begriff Epistase wird eine Interaktion zwischen verschiedenen Genen bezeichnet. Allgemein bezieht sich die Definition auf die Interaktion von Genen zwischen Allelen an unterschiedlichen Genloci. Der Begriff „Epistase“ wurde das erste Mal 1909 von Bateson verwendet, er beschrieb damit einen maskierenden Effekt, bei dem eine Variante oder ein Allel an einem Locus eine andere Variante an einem anderen Locus in ihrem Effekt unterdrückt (Bateson 1909). In der Humangenetik treten allerdings einige Probleme mit dieser Definition auf, wenn diese auf binäre Merkmale angewendet wird. Der untersuchte Phänotyp ist häufig qualitativ und normalerweise dichotom in Hinblick auf das Vorhandensein oder Fehlen einer Krankheit. Betrachtet man nun zwei Loci (A und B) mit jeweils zwei möglichen Allelen (*A*, *a* und *B*, *b*): Wenn die Annahme gilt, ein prädisponierendes Allel (*A*, *B*) an beiden Loci ist nötig, um das Merkmal zu beobachten, kann der Effekt von Allel *A* nur beobachtet werden, wenn auch das Allel *B* vorhanden ist. Locus B ist also epistatisch zu Locus A, denn wenn der Genotyp *b*/*b* am Locus B vorhanden ist, kann der Effekt von Locus A nicht beobachtet werden. Allerdings stimmt auch die Behauptung, Locus A ist epistatisch zu Locus B, da bei dem vorhandenem Genotyp *a*/*a* am Locus A kein Effekt von Locus B beobachtet werden kann. Man kann also nicht, wie in der ursprünglichen Definition von Bateson, behaupten, einer der Loci ist epistatisch gegenüber dem anderen. Wegen der Symmetrie zwischen den Effekten von Locus A und B ist kein Locus epistatisch über dem anderen (Tab. 2-3).

Tab. 2-3: Penetranztabelle für zwei Loci. Beispiel einer Penetranztabelle für epistatisch interagierende Loci (A und B) unter allgemeinen Bedingungen. Wobei 0 keinem beobachten Effekt entspricht und bei 1 ein Effekt beobachtet wird.

	Genotyp am Locus B		
Genotyp am Locus A	*b/b*	*b/B*	*B/B*
a/a	0	0	0
a/A	0	1	1
A/A	0	1	1

So eine Penetranztabelle stellt eine allgemeinere Form und Definition der Epistase zwischen unterschiedlichen Loci (z.B. bei komplexen Krankheiten) dar (Neuman and Rice 1992).

Standardmodelle der logistischen Regression werden für die Analyse auf Epistase bei Fall-Kontroll-Daten verwendet, welche für zwei diallelische Kandidaten-Loci genotypisiert wurden. Zunächst wird folgendes Modell angepasst:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa}x_1x_2 + i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2$$

und wird dann mit der Nullhypothese

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2$$

verglichen.

Wobei x_i und z_i Dummy-Variablen für den zugrunde liegenden Phänotyp am Locus i sind; die Allele sind mit 1 und 2 bezeichnet und die Koeffizienten $\mu, a_1, d_1, a_2 \text{ und } d_2$ repräsentieren genetische Parameter, welche entsprechend dem durchschnittlichen, additiven oder dominanten Effekt an den zwei Loci geschätzt werden; $i_{aa}, i_{ad}, i_{da} \text{ und } i_{dd}$ entsprechen den epistatischen Interaktionseffekten. p_{ij} definiert die Penetranz für den Genotyp *i* am Locus 1 und Genotyp *j* am Locus 2.

Um auf Epistase zwischen Markern zu testen wurden in der Arbeit die betroffenen Marker auf ihren krankheitsbeeinflussenden Effekt mit dem Program PLINK (Purcell et al. 2007) analysiert.

2.2.7.7 Transmissions-Ungleichgewichts-Test

Wie eine Assoziationsstudie vergleicht der Test auf Transmissions-Ungleichgewicht (engl. *transmission disequilibrium test*, TDT), ob ein spezifisches Allel überproportional häufig bei betroffenen Individuen auftritt (Spielman et al. 1993). Bei dem TDT werden familiäre Kontrollen statt unabhängiger Kontrollen aus der gleichen Population genutzt, um die Effekte von Populationsstratifikation zu vermeiden, die z.B. entstehen können, wenn die ethnische Herkunft der Kontrollindividuen nicht komplett mit den erkrankten Individuen übereinstimmt. Die Idee hinter dem TDT ist, dass, wenn eine Assoziation zwischen einem genetischen Marker und der Anfälligkeit

gegenüber einer Krankheit besteht, die Wahrscheinlichkeit der Übertragung dieses Markers von den Eltern auf das betroffene Kind entweder signifikant erhöht (engl. *over-transmission*) oder reduziert (engl. *under-transmission*) ist und sich von dem bei der mendelschen Regel der Vererbung prognostiziertem Wert 0,5 (50% Übertragung wird bei einer nicht vorhandenden Kopplung und nicht vorhandenem LD erwartet) unterscheidet.

Hierfür werden bei dem TDT Studiendesign Kernfamilien (engl. *nuclear families*) verwendet, d.h. es werden üblicherweise die gesunden Eltern („interne Kontrollindividuen“) und ein betroffenes Kind in das Testverfahren eingeschlossen (Falk and Rubinstein 1987). Dabei werden die Allele der Eltern, die an die betroffenen Kinder transmittiert werden, als Fälle betrachtet (Kopplungsinformation), während die nicht-transmittierten Allele als fiktive interne Kontrollallele (Assoziationsinformation) dienen (Abb. 2-11).

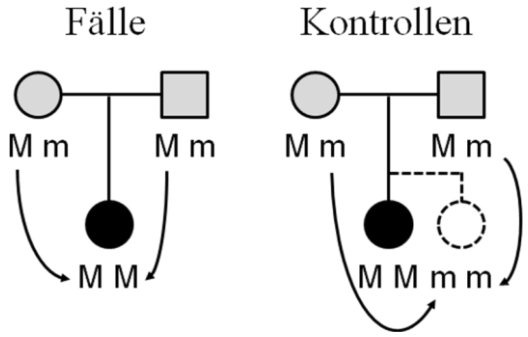

Abb. 2-11: Schematische Darstellung einer TDT-Analyse. Die Kreise repräsentieren weibliche und die Quadrate männliche Individuen. Die elterliche Generation ist hier in grau dargestellt, kranke Individuen sind in schwarz und gesunde Individuen in weiß dargestellt. Die an das erkrankte Kind transmittierten Allele MM dienen als Fall und die nicht an das Kind transmittierten elterlichen Allele mm als Allele eines fiktiven Kontrollindividuums. Nur von heterozygoten Eltern können Segregationsinformationen gewonnen werden, Transmissionen von homozygoten Eltern liefern keine Segregationsinformationen und werden daher vom TDT ausgeschlossen.

Es ist nicht relevant, ob der jeweilige Elternteil betroffen ist oder nicht, jedoch liefern nur heterozygote Eltern Segregationsinformationen. Daher gehen homozygote Eltern nicht in die Berechnungen ein und werden von dem TDT ausgeschlossen. Bei dem TDT wird die Anzahl der transmittierten Allele heterozygoter Eltern an betroffene Kinder mit der Anzahl der nicht-transmittierten Allele verglichen (Tab. 2-4).

Tab. 2-4: 2 x 2 Kontingenztafel für den TDT.

		nicht-transmittiert	
		Allel M	Allel m
transmittiert	Allel M	*a*	*b*
	Allel m	*c*	*d*

Zur Berechnung eines Transmissionsungleichgewichtes wurde die Teststatistik des McNemar-Test genutzt, welche einer asymptotischen Chi-Quadrat-Verteilung mit einem Freiheitsgrad entspricht (Agresti and Coull 1996):

$$\chi^2 = \frac{(b-c)^2}{(b+c)}$$

Hinsichtlich der Effizienz ist das TDT-Studiendesign dem Fall-Kontroll-Studiendesign unterlegen. In dem TDT liefert jedes Trio Informationen für zwei Transmissionen und zwei Nicht-Transmissionen, während in Fall-Kontroll-Studien drei Individuen sechs informative Allele in die Studie einbringen (Morton and Collins 1998). Daher wurde der TDT in der vorliegenden Arbeit nur zur Bestätigung von positiven Assoziationsbefunden durchgeführt. TDT-Analysen wurden mit dem Programm PLINK durchgeführt (Purcell et al. 2007).

2.2.7.8 Multidimensionale Skalierung

Mit der multidimensionalen Skalierung (engl. *multidimensional scaling*) kann der Grad der Ähnlichkeit zwischen einzelnen Objekten eines Datensatzes visuell dargestellt werden. Dabei werden die Objekte (z.B. Individuen einer Stichprobe) räumlich so angeordnet, dass die Abstände zwischen den Objekten im Raum möglichst exakt den Unterschieden zwischen den einzelnen Objekten entsprechen. In dieser Arbeit wurde der Ansatz der metrischen multidimensionalen Skalierung nach Kruskal (1964) verwendet.

Mit der metrischen multidimensionalen Skalierung können Objekte (I) mit dem Abstand d_{ij} im hoch dimensionalen Raum in einem kleineren r-dimensionalen Raum so angeordnet werden, dass die euklidischen Distanzen in diesem Raum möglichst genau den Distanzen d_{ij} gleichen. Werden als Werte Ähnlichkeitsmaße c_{ij} anstatt Distanzen zwischen Objekten betrachtet, dann lassen sich diese durch die Transformation:

$$d_{ij} = \sqrt{c_{ii} + c_{jj} - 2c_{ij}}$$

in Distanzen überführen.

Wenn

$$\text{Matrix}\, A \;=\; (a_{ij}) \,\text{mit}\, (a_{ij}) \;=\; -\frac{1}{2} d_{ij}^2$$

und

$$\text{Matrix}\, B \;=\; (b_{ij}) \,\text{mit}\, (b_{ij}) \;=\; a_{ij} - \left(\frac{1}{n}\sum_{j=1}^{n} a_{ij}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} a_{ij}\right) + \left(\frac{1}{n^2}\sum_{i=1}^{n}\;\sum_{j=1}^{n} a_{ij}\right)$$

Dann können die Eigenwerte λ_i und die zugehörigen Eigenvektoren $\gamma_i = (\gamma_{ij})$ der Matrix B mit der Eigenschaft

$$\sum_{j=1}^{n} \gamma_{ij}^2 = \lambda_i$$

bestimmt werden.

Die Koordinaten der skalierten Datenpunkte im r dimensionalen Raum ergeben sich dann aus den Eigenvektoren zu den r größten Eigenwerten:

$$\chi_i = \gamma_i \sqrt{\lambda_i}$$

Die multidimensionalen Skalierungsanalysen wurden mit dem Programm EIGENSOFT (Price et al. 2006) durchgeführt.

2.2.7.9 Wilcoxon-Mann-Whitney-Test

Der Wilcoxon-Mann-Whitney-Test ist ein nichtparametrischer statistischer Test für unverbundene Stichproben. Der Test wurde von Henry Mann und Donald Whitney (1947) sowie Frank Wilcoxon (1945) entwickelt (Mann and Whitney 1947; Wilcoxon 1945).

Gegeben sind zwei identisch verteilte Teilstichproben $A = X_1, \ldots, X_m$ und $B = Y_1, \ldots, Y_n$ die unabhängig voneinander erhoben wurden. Folgende Nullhypothese soll untersucht werden: H_0 = Beide Teilstichproben entstammen der gleichen Grundgesamtheit. Beide Teilstichproben werden nun zusammengefasst und jedem Wert wird ein Rang zugeordnet (kleinste Beobachtung: Rang 1; größte Beobachtung: Rang $m + n$). Von jeder Teilstichprobe werden die Rangsummen R_x und R_y berechnet, und folgende Größen bestimmt:

$$U_x = mn + \frac{m(m+1)}{2} - R_x \text{ und } U_y = mn + \frac{n(n+1)}{2} - R_y$$

Für die Teststatistik gilt dann:

$$U := \min\{U_x, U_y\}$$

Die Testgröße U wird nun mit dem kritischen Wert U_{krit} verglichen:

$$U_{krit} = A + B$$

Abgelehnt wird H_0 wenn:

$$U < U_{krit}$$

Der Wilcoxon-Mann-Whitney-Test wurde mit dem Programm GraphPad Prism 6 berechnet (GraphPad Software).

2.2.7.10 Kruskal-Wallis-Test

Der Kruskal-Wallis-Test ist ebenfalls ein nichtparametrischer statistischer Test für unverbundene Stichproben. Mit dem Test können aber, auf Basis von Rangplatzsummen, Vergleiche von mehr als zwei Teilstichproben durchgeführt werden.

Folgende Nullhypothese H_0 wird untersucht: Zwischen den Teilstichproben besteht kein Unterschied. Als Prüfgröße wird der sogenannte H -Wert berechnet:

Der Rang R_i für *n* Beobachtungen in der Vereinigung der Teilstichproben wird bestimmt. Daraus ergeben sich dann die Rangsummen S_h für die einzelnen Teilstichproben und folgende Teststatistik

$$H = \frac{12}{n(n+1)} \sum_h \frac{s_h^2}{n_h} - 3(n+1)$$

Die Teststatistik folgt unter H_0 einer χ^2-Verteilung. Die Freiheitsgrade (Df) berechnen sich nach $Df = k - 1$, wobei *k* die Anzahl der Teilstichproben ist. Die berechnete Prüfgröße H wird mit der theoretischen Größe ($H(\chi^2)$) der χ^2-Verteilung (unter dem erwünschten Signifikanzniveau) verglichen. Die Nullhypothese H_0 wird abgelehnt, wenn

$$H \geq H(\chi^2)$$

Der Kruskal-Wallis-Test wurde mit dem Programm GraphPad Prism 6 berechnet (GraphPad Software).

2.3 Molekularbiologische Methoden

2.3.1 Agarose-Gelelektrophorese

Die Agarose-Gelelektrophorese wurde verwendet, um lineare DNA-Fragmente ab einer Größe von 200 bp zu trennen und zu identifizieren. In Abhängigkeit von der Agarosekonzentration im Gel [0,5 bis 2,0% (w/v)] werden die DNA-Fragmente von 200-50.000 bp entsprechend ihrer Größe und Gestalt aufgetrennt (Gelsiebeffekt). Die DNA-Proben wurden mit zehnfachem Auftragspuffer versetzt, zusammen mit einem Marker in die Taschen des ethidiumbromidhaltigen Gels aufgetragen und bei konstanter Spannung von 90 Volt 20-60 min aufgetrennt. Ethidiumbromid interkaliert in die DNA und fluoresziert im UV-Licht bei 312 nm. Die DNA wurde für qualitative und quantitative Analysen mittels einer UV-Geldokumentationsanlage (BioDoc Analyze, Biometra, Göttingen) dokumentiert und ausgewertet.

2.3.2 Isolierung von Gesamt-RNA

Die RNA-Isolierung aus Zellen wurde mit dem RNeasy-Kit von Qiagen (Hilden) durchgeführt. Die Zellen wurden mit PBS gewaschen, in 350 µl RLT Puffer (supplementiert mit 10 µl/ml 2-Mercaptoethanol) lysiert und durch Zentrifugation über eine QiaShredder®-Säule (13.000x *g*, 2 min, RT) aufgeschlossen. Das Eluat wurde mit gleichem Volumen 70%igem Ethanol versetzt und in die RNeasy-Säule überführt. Durch Zentrifugation (15 s; 8000 U/min) wurde die RNA an die Membran gebunden und das Eluat verworfen. Nach einem Waschschritt mit 350 µl Puffer RLT (15 s; 8000 U/min) wurde die nun auf der Säule gebundene RNA durch zwei 30-minütige Inkubationen mit DNase-Lösung (10 µl DNase I in 70 µl RDD Puffer) von verbliebener genomischer DNA gereinigt. Die Säule wurde dreimal mit 700 µl RW1 Puffer und zweimal mit 500 µl RPE Puffer gewaschen, in ein unbenutztes Auffangröhrchen überführt und zum Trocknen für 2 min bei 13.000x *g* zentrifugiert. Im Anschluss wurde die Säulenmatrix mit 30–50 µl RNase-freiem Wasser (Qiagen) versetzt und für 5 min bei RT inkubiert. Die Elution der Gesamt-RNA erfolgte durch einen letzten Zentrifugationsschritt (2 min, 13.000x *g*) in ein frisches RNA-Röhrchen. Die Lagerung der RNA erfolgte bei -80 °C. Alle weiteren Arbeitsschritte wurden auf Eis durchgeführt.

2.3.3 Qualitätskontrolle der RNA

Die Qualität der RNA wurde durch Agarose-Gelelektrophorese überprüft. Während bei intakten RNA-Proben lediglich die 18S- und 28S-rRNA-Banden zu erkennen sind, deuten zusätzliche Banden oder ein Verschmieren der rRNA-Banden auf Kontamination oder Degradation der RNA hin. Um Kontaminationen der RNA-Proben mit genomischer DNA auszuschließen, wurde eine Test-PCR mit isolierter RNA als Matrize durchgeführt. Nur RNA-Proben, bei denen keine GAPDH-Bande amplifiziert wurde, wurden für die reverse Transkription verwendet.

2.3.4 Polymerase-Kettenreaktion (PCR)

PCRs wurden in einem Thermocycler (Biometra, Göttingen) durchgeführt. Für Reverse-Transkriptase-Polymerase-Kettenreaktionen (RT-PCR) wurde die DNA-Polymerase GoTaq™ (Promeaga, Madison, USA) verwendet. In Tabelle 2-5 ist der verwendete PCR-Ansatz gezeigt.

Tab. 2-5: Reaktionsansatz und Programm einer PCR. *Die verwendeten Oligonukleotide sind in Tabelle 2-7 aufgelistet.

PCR-Komponenten	Volumen [µl]
Matrizen-DNA (10pg – 200ng)	X
5x GoTaq™-Reaktionspuffer	5
Oligonukleotide (je 10 µM)*	1
dNTPs (10 mM)	1
GoTaq™ DNA-Polymerase	0,4
A. bidest	ad 20

Temperatur [°C]	Zeit [min]	Zyklen
95	5:00	1
95	0:30	25-35
T_{Hybrid}*	0:30	
72	1:00	
72	5:00	1
4	Pause	1

2.3.5 cDNA-Synthese

Mit Hilfe des Enzyms Reverse-Transkriptase kann aus einem RNA-Template die dazu komplementäre DNA (engl. *complemetary DNA*; cDNA) hergestellt werden. Für die vorliegende Arbeit wurde eine RNA-abhängige DNA-Polymerase (= Reverse-Transkriptase) aus dem Retrovirus *MMLV* (*moloney-murine leukemia virus*) verwendet. Diese spezifische RNA-abhängige DNA-Polymerase benötigt einen Primer zur Synthese, der an die RNA bindet. Meist wird ein Oligo-dT-Abschnitt verwendet (10-15 Thyminbasen), der komplementär zum Poly-A-Schwanz der eukaryotischen mRNA ist. Alternativ werden sog. *random* Hexamer-Primer (Oligonukleotide, die aus sechs zufällig zusammengesetzten Nukleinbasen bestehen) eingesetzt. Für die cDNA-Synthesen wurde das Advantage® RT-for-PCR Kit (Clontech, Palo Alto, CA) verwendet, welches neben einer retroviralen *MMLV*-Transkriptase alle weiteren benötigten Komponenten enthielt. Die Zusammensetzung und der Ablauf einer cDNA-Synthese sind in Tabelle 2-6 dargestellt.

Die cDNA wurde durch Zugabe von 80 µl DEPC-behandeltem Wasser (Ambion/*Applied Biosystems*, Darmstadt) 1:5 verdünnt und bis zur weiteren Nutzung bei -80 °C gelagert.

Tab. 2-6: Reaktionsansatz für cDNA-Synthese.

Komponente	Volumen [µl]	Inkubation
Gesamt-RNA	500 ng	2 min bei 70 °C
Oligo(dT)18-Primer	1,0 µl	
Aqua bidest. (RNase-frei)	Ad 13,5 µl	
5-fach Reaktionspuffer	4,0 µl	1 h bei 42 °C + 5 min bei 95 °C
dNTP-Mix (jeweils 10 mM)	1,0 µl	
Rekombinanter RNase-Inhibitor (40 U/µl)	0,5 µl	
MMLV-Reverse-Transkriptase (200 U/µl)	1,0 µl	
Gesamtvolumen	20,0 µl	

2.3.6 Reverse-Transkriptase-Polymerase-Kettenreaktion (RT-PCR)

Die gewebespezifische Expression von Genen der Chromosom *11q13.1* Region auf mRNA-Ebene wurde mit Hilfe genspezifischer Primer analysiert. Basierend auf der Sequenzinformation der jeweiligen mRNA wurden die Primerpaare so gewählt, dass die Länge des resultierenden Amplikons

zwischen 200 und 1000 bp betrug. Die Synthese der Oligonukleotide erfolgte durch die Firma Eurogentec (Köln). Amplifiziert wurden die DNA-Fragmente durch die Polymerase-Kettenreaktion (PCR) unter Verwendung der DNA-Polymerase GoTaq™ (Promeaga, Madison, USA) in einem Gradienten *Thermocycler* (Biometra, Göttingen). Eine Auflistung der verwendeten Oligonukleotide findet sich in der Tabelle 2-7.

Tab. 2-7: Verwendete Oligonukleotide für die Real Time-PCR.

Gen	Identifikationsnummer	Sequenz Vorwärts-Primer (5'→3')	Sequenz Rückwärts-Primer (5'→3')
PLCB3*	NM_000932	ccagaacagacaggtgcaga	atgcttgtccctcatcttgg
BAD*	NM_032989.2	ccgagtgagcaggaagactc	ggtaggagctgtggcgact
GPR137*	NM_001170881.1	tgcttctgtatgggcacaag	acaccacctgggcaaagtag
C11orf20*	NM_001039496.1	atctgggggagaaggacact	ctgccgatccctaactggta
KCNK4*	NM_033310.2	gtgccaccggagctagtaag	cacggtggtaagcgtcacta
ESRRA*	NM_004451.3	tcgctgtctgaccagatgtc	aagggccaaggcctttagta
PRDX5*	NM_012094.4	atgggactagctggcgtgt	agcggctctgctgaaactg
CCDC88B*	NM_032251.5	acctgattcagacccacagg	ctgtctctcctcccagcaac
CCDC88B†		agtacctggaccagcttaatgcc	gcatcagcctcttcaccttgtct

*Primer die für die qRT-PCR des Gewebs-Panel und des BAL-Panel I verwendet wurden
† Primer die für die qRT-PCR des BAL-Panel II verwendet wurden

3 Ergebnisse

3.1 Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen

Um die Krankheit auf Assoziationen im Hinblick auf den Phänotyp Sarkoidose allgemein, aber auch spezifisch auf einen der Subphänotypen zu überprüfen, stand ein GWAS-Datensatz zur Verfügung (Kapitel 2.1.2.1). Für diese Analyse wurde ein zweistufiges Studiendesign verwendet (Kapitel 2.2.1.1), bei dem in der ersten Stufe (GWAS) alle Marker jedes Individuums auf Assoziation mit der Krankheit getestet wurden. In der zweiten Stufe wurden nur die vielversprechendsten Marker der ersten Stufe in weiteren Individuen (Stichprobe B) auf Assoziation getestet. Die in der zweiten Stufe bestätigten Marker wurden in zusätzlichen unabhängigen Stichproben auf Assoziation getestet, um die detektierte Assoziation zu replizieren (Siehe Kapitel 2.2.1.1). Die Ergebnisse dieses Teils der Arbeit wurden bereits 2012 in dem Journal „*American Journal of Respiratory and Critical Care Medicine*“ veröffentlicht (Fischer et al. 2012).

3.1.1 Sarkoidose-GWAS

Nach Qualitätskontrolle (siehe Kapitel 2.2.5) umfasste der Sarkoidose GWAS-Datensatz Proben von 2.139 Individuen (564 Sarkoidosepatienten, davon 176 Patienten mit *akuter* Sarkoidose und 354 Patienten mit *chronischer* Sarkoidose, und 1.575 gesunde Kontrollpersonen). Dabei galt:

- Individuen mit mehr als 5% fehlender Daten wurden von der Analyse ausgeschlossen (n = 195);
- von unerwarteten Duplikaten oder Verwandten wurde jeweils das Individuum mit der schlechteren Genotypisierungsrate ausgeschlossen (n = 63);
- Individuen mit einer Heterozygositätsrate mit mehr als +/- 5 SD Abweichung vom Mittelwert der Heterozygosität (32,33%) wurden ausgeschlossen (n = 5);
- nach Überprüfung auf Populationsstratifikation wurden 14 Individuen ausgeschlossen.

Nach Genotypimputation (Kapitel 2.2.4) umfasste der Datensatz 1.294.967 SNPs für weiterführende Analysen. Die genetische Heterogenität dieses Datensatzes war mit einem *„genomic inflation factor“* von λ_{GC} = 1,031 gering, und die multidimensionale Skalierungsanalyse bestätigte die erwartete europäische Abstammung aller Individuen, was auch in der dazugehörigen Quantil-Quantil-Abbildung zu erkennen ist (Abb. 3-1).

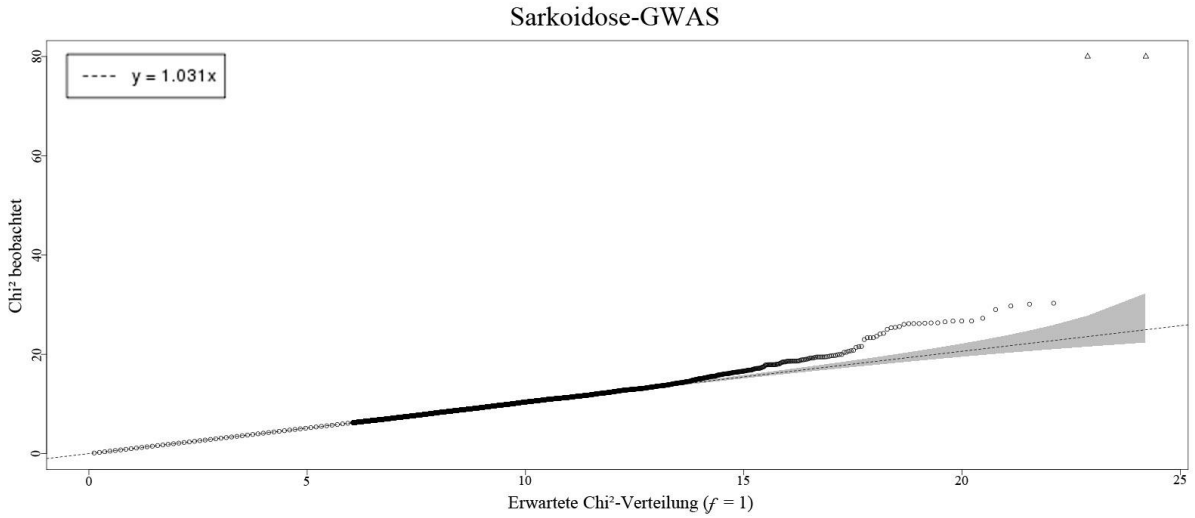

Abb. 3-1: Quantil-Quantil-Diagramm der χ^2-Teststatistik für den Sarkoidose-GWAS. Q-Q-Abbildung der χ^2-Teststatistik mit 95%-Konfidenzbereich (grau unterlegt), für alle SNPs in dem Sarkoidose-GWAS, die den Qualitätskriterien entsprachen.

3.1.2 Identifizierung von Risikoloci der Sarkoidose und ihre Subphänotypen

Die Genotypen der GWAS wurden in drei verschiedenen Zusammensetzungen in Fall-Kontroll-Assoziationsanalysen untersucht (siehe Kapitel 2.2.6.1):

1) In der „Allgemeine Analyse“ wurden 13 SNPs mit einem *p*-Wert < 10^{-5} für die weiterführenden Analysen ausgewählt.

2) In der „Analyse akut“ wurden sechs SNPs mit einem *p*-Wert von $p < 10^{-4}$ für die weiteren Untersuchungen ausgewählt.

3) In der „Analyse chronisch“ wurden 11 SNPs mit einem *p*-Wert von $p < 10^{-4}$ ausgewählt.

Unter Berücksichtigung der jeweiligen Selektionskriterien (s.o.) wurden somit insgesamt 30 SNPs für die Validierung ausgewählt (Tab. 3-1).

Tab. 3-1: Ergebnisse für die 30 SNPs mit den höchsten Rängen aus den drei verschiedenen GWAS-Analysen. Die SNPs sind nach den verschiedenen Analysen gruppiert (Akut = Analyse akut; Allg. = Allgemeine Analyse; Chronisch = Analyse chronisch). A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; Chr. = Chromosom; CI = Konfidenzintervall (engl. *confidence interval*); ds = stromabwärts (engl. *downstream*); e = exonisch; i = intronisch; Kont. = Kontrollen; OR = Quotenverhältnis (engl. *odds ratio*); us = stromaufwärts (engl. *upstream*).

Chr.	Position (bp)	dbSNP ID	Analysetyp	Gen	A1	AF Fälle GWAS	AF Kont. GWAS	p_{CCA} GWAS	OR [95% CI] GWAS
1	9,776,339	rs3934934	Allg.	*PIK3CD, i*	G	0.14	0.11	5,06E-06	1,82 [1,41-2,35]
1	53,574,076	rs1679938	Allg.	*SLC1A7, i*	G	0,28	0,23	2,74E-05	1,50 [1,24-1,81]
1	151,827,813	rs4845624	Allg.	*THEM5, us*	G	0,37	0,32	4,71E-05	1,43 [1,20-1,70]
2	27,083,489	rs6756245	Allg.	*DPYSL5, i*	C	0,39	0,44	9,85E-06	0,69 [0,59-0,82]
2	203,266,610	rs16839127	Allg.	*BMPR2, i*	A	0,09	0,06	2,92E-05	1,91 [1,41-2,59]
11	5,146,869	rs12577357	Allg.	intergenisch	T	0,22	0,28	5,54E-05	0,67 [0,56-0,82]
11	64,107,477	rs479777	Allg.	*CCDC88B, us*	C	0,29	0,36	4,92E-06	0,67 [0,56-0,79]
11	82,065,238	rs7930387	Allg.	intergenisch	G	0,23	0,19	4,20E-05	1,49 [1,23-1,80]
11	118,172,520	rs894678	Allg.	intergenisch	A	0,31	0,38	3,11E-05	0,70 [0,59-0,83]
14	21,230,576	rs12895983	Allg.	intergenisch	T	0,06	0,10	3,53E-05	0,53 [0,39-0,72]
15	79,296,648	rs3825929	Allg.	*RASGRF, i*	A	0,49	0,43	3,37E-06	1,46 [1,25-1,72]
17	27,871,025	rs17766675	Allg.	*TAOK1, e*	G	0,24	0,19	5,35E-05	1,48 [1,22-1,72]
22	39,529,133	rs4820369	Allg.	*CBX7, e*	A	0,18	0,13	4,09E-05	1,57 [1,26-1,94]
2	109,742,307	rs6542797	Akut	*SH3RF3, us*	C	0,33	0,41	1,86E-04	0,60 [0,46-0,78]
4	79,019,831	rs13106755	Akut	*FRAS1, i*	C	0,34	0,25	2,90E-05	1,81 [1,37-2,39]
5	6,710,771	rs274667	Akut	*POLS, us*	A	0,28	0,36	1,91E-04	0,59 [0,45-0,78]
7	84,294,717	rs10954758	Akut	intergenisch	C	0,49	0,40	1,63E-04	1,65 [1,27-2,14]
7	146,344,283	rs10262146	Akut	*CNTNAP2, i*	T	0,33	0,22	1,23E-05	1,84 [1,40-2,41]
16	54,963,926	rs7203657	Akut	*PNAS-108, ds*	T	0,30	0,21	4,78E-06	1,94 [1,46-2,57]
1	162,139,692	rs7546353	Chronisch	*NOS1AP, i*	C	0,23	0,18	2,03E-04	1,53 [1,22-1,91]
2	36,260,998	rs1387287	Chronisch	intergenisch	T	0,11	0,07	4,36E-05	1,96 [1,42-2,72]
2	224,183,485	rs1517634	Chronisch	intergenisch	G	0,30	0,23	2,06E-05	1,57 [1,27-1,93]
4	175,503,012	rs2250175	Chronisch	intergenisch	A	0,52	0,46	7,61E-05	1,46 [1,21-1,75]
5	60,993,594	rs2030888	Chronisch	*FLJ37543, i*	T	0,26	0,20	3,89E-05	1,59 [1,27-1,98]
8	33,853,868	rs7465297	Chronisch	intergenisch	G	0,28	0,21	7,02E-05	1,54 [1,24-1,90]
10	46,146,746	rs17157941	Chronisch	*ANUBL1, i*	C	0,06	0,03	3,50E-05	2,46 [1,61-3,78]
11	5,168,952	rs6578556	Chronisch	*OR52A1, us*	C	0,41	0,49	3,29E-05	0,67 [0,56-0,81]
11	107,420,192	rs638234	Chronisch	*ABH8, i*	G	0,19	0,15	1,91E-04	1,58 [1,24-2,02]
15	58,877,599	rs1869133	Chronisch	*ADAM10, us*	C	0,24	0,19	8,34E-05	1,55 [1,25-1,93]
19	17,182,017	rs7259348	Chronisch	*HAUS8, i*	T	0,21	0,29	1,70E-04	0,66 [0,53-0,82]

3.1.3 Bekannte Sarkoidose-Risikogene in dem vorliegenden GWAS-Datensatz

Einige frühere Veröffentlichungen haben die Assoziation von Risikogenen mit Sarkoidose nachgewiesen (Adrianto et al. 2012; Fischer et al. 2011; Grutters et al. 2003; Hofmann et al. 2008; Hofmann et al. 2011; Hofmann et al. 2013; Rybicki et al. 2005; Valentonyte et al. 2005). In dem für diese Arbeit vorliegendem GWAS-Datensatz (Kapitel 2.1.2.1) wurden die Assoziationen von folgenden in den Veröffentlichungen identifizierten Markern verifiziert: Die Marker der bekannten Risikogene *BTNL2* (repräsentiert durch rs2076530) und *ANXA11* (rs1953600) zeigten in dem GWAS-Datensatz in der Einzelmarker-Fall-Kontroll-Analyse eine Assoziation sowohl mit dem *akuten* als auch mit dem *chronischen* Subphänotyp (Akut: $p = 1{,}39 \times 10^{-4}$ und $p = 6{,}07 \times 10^{-5}$; chronisch: $p = 2{,}75 \times 10^{-11}$ und $p = 9{,}55 \times 10^{-7}$). Auch die *RAB23*-Region (rs3957366) und der gemeinsame Suszeptibilitätslocus für Sarkoidose und Morbus Crohn auf Chromosom 10p12.2 (rs1398024) zeigten in dem GWAS-Datensatz ähnliche Assoziationsmesswerte wie in den damaligen Studien ($p = 1{,}36 \times 10^{-3}$ und $p = 0{,}04$).

3.1.4 Validierung der Kandidaten-SNPs

Für die Validierung der mit der GWAS-Analyse identifizierten Kandidaten-SNPs stand die unabhängige Stichprobe B (Kapitel 2.1.3.1) zur Verfügung. Nach der Qualitätskontrolle (Kapitel 2.2.5) umfasste die Stichprobe B 3.623 Individuen. Die Gesamtmenge der 1.486 Sarkoidosepatienten beinhaltete 488 *akute* Sarkoidosefälle und 865 *chronische* Sarkoidosefälle für die subphänotypspezifische Analyse. Die 30 Kandidaten-SNPs wurden mit der Sequenom®-Technology in der Stichprobe B genotypisiert (Tab. 3-2). Fünf SNPs zeigten in den Fall-Kontroll-Analysen eine nominell signifikante Assoziation: Drei SNPs aus der Allgemeinen Analyse (rs4845624, *p*-Wert = $1{,}61 \times 10^{-2}$; rs479777, *p*-Wert = $3{,}26 \times 10^{-7}$; rs12895983, *p*-Wert = $8{,}3 \times 10^{-3}$) und jeweils ein SNP aus der Analyse akut (rs7203657, *p*-Wert = $3{,}13 \times 10^{-2}$) und Analyse chronisch (rs17157941, *p*-Wert = $5{,}25 \times 10^{-3}$). Nach der Korrektur für multiples Testen (Bonferroni-Korrektur) zeigte aber nur ein SNP aus der allgemeinen Analyse eine signifikante Assoziation (rs479777, *p*-Wert = $9{,}79 \times 10^{-6}$; OR = 0,77; [95% CI] 0,70-0,85) (Tab. 3-2). Die *odds ratios* des SNPs bei dem *akuten* (OR = 0,79) und bei dem *chronischen* Subphänotyp (OR = 0,78) in der Stichprobe B zeigten keinen signifikanten Unterschied zwischen den Subphänotypen. In einer Metaanalyse der Stichproben A und B zeigte der SNP rs479777 eine Assoziation mit Sarkoidose mit einer genomweiten Signifikanz ($p_{META} = 2{,}05 \times 10^{-12}$; OR = 0,75).

Tab. 3-2: Ergebnisse der Validierungsstufe für die 30 am stärksten assoziierten SNPs aus den GWAS Analysen. Die SNPs sind nach den verschiedenen Analysen gruppiert (Akut = akute Sarkoidosefälle versus Kontrollpersonen; Allg. = Allgemeine Analyse; Chronisch = chronische Sarkoidosefälle versus Kontrollpersonen). A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Die kursiv dargestellten SNPs zeigten eine nominelle Assoziation vor der Bonferroni-Korrektur, der fett gedruckte SNP zeigte auch nach der Bonferroni-Korrektur eine signifikante Assoziation. Abkürzungen: AF = Allelfrequenz; Chr. = Chromosom; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; p_{Korr} = Nach Bonferroni korrigierte *p*-Werte; OR = Quotenverhältnis (engl. *odds ratio*); Val. = Validierung.

Chr.	dbSNP ID	Analysetyp	A1	AF Fälle GWAS	AF Kont. GWAS	*p*-Wert GWAS	OR [95% CI] GWAS	AF Fälle Val.	AF Kont. Val.	p_{Korr}-Wert Val.	OR [95% CI] Val.
1	rs3934934	Allg.	G	0,14	0,11	$5,06 \times 10^{-06}$	1,82 [1,41-2,35]	0,10	0,10	1	1,03 [0,89-1,20]
1	rs1679938	Allg.	G	0,28	0,23	$2,74 \times 10^{-05}$	1,50 [1,24-1,81]	0,25	0,25	1	1,04 [0,93-1,16]
1	*rs4845624*	*Allg.*	*G*	*0,37*	*0,32*	*$4,71 \times 10^{-05}$*	*1,43 [1,20-1,70]*	*0,33*	*0,30*	*$4,82 \times 10^{-01}$*	*1,13 [1,02-1,25]*
2	rs6756245	Allg.	C	0,39	0,44	$9,85 \times 10^{-06}$	0,69 [0,59-0,82]	0,42	0,42	1	1,00 [0,91-1,10]
2	rs16839127	Allg.	A	0,09	0,06	$2,92 \times 10^{-05}$	1,91 [1,41-2,59]	0,07	0,07	1	1,13 [0,94-1,35]
11	rs12577357	Allg.	T	0,22	0,28	$5,54 \times 10^{-05}$	0,67 [0,56-0,82]	0,26	0,26	1	1,01 [0,91-1,13]
11	**rs479777**	**Allg.**	**C**	**0,29**	**0,36**	**$4,92 \times 10^{-06}$**	**0,67 [0,56-0,79]**	**0,30**	**0,36**	**$9,79 \times 10^{-06}$**	**0,77 [0,70-0,85]**
11	rs7930387	Allg.	G	0,23	0,19	$4,20 \times 10^{-05}$	1,49 [1,23-1,80]	0,19	0,19	1	0,99 [0,88-1,11]
11	rs894678	Allg.	A	0,31	0,38	$3,11 \times 10^{-05}$	0,70 [0,59-0,83]	0,35	0,37	1	0,93 [0,85-1,03]
14	*rs12895983*	*Allg.*	*T*	*0,06*	*0,10*	*$3,53 \times 10^{-05}$*	*0,53 [0,39-0,72]*	*0,10*	*0,08*	*$2,49 \times 10^{-01}$*	*1,24 [1,06-1,46]*
15	rs3825929	Allg.	A	0,49	0,43	$3,37 \times 10^{-06}$	1,46 [1,25-1,72]	0,45	0,43	1	1,07 [0,98-1,18]
17	rs17766675	Allg.	G	0,24	0,19	$5,35 \times 10^{-05}$	1,48 [1,22-1,72]	0,21	0,20	1	1,04 [0,92-1,16]
22	rs4820369	Allg.	A	0,18	0,13	$4,09 \times 10^{-05}$	1,57 [1,26-1,94]	0,15	0,16	1	0,92 [0,81-1,05]
2	rs6542797	Akut	C	0,33	0,41	$1,86 \times 10^{-04}$	0,60 [0,46-0,78]	0,39	0,39	1	0,98 [0,85-1,13]
4	rs13106755	Akut	C	0,34	0,25	$2,90 \times 10^{-05}$	1,81 [1,37-2,39]	0,27	0,28	1	0,95 [0,82-1,12]
5	rs274667	Akut	A	0,28	0,36	$1,91 \times 10^{-04}$	0,59 [0,45-0,78]	0,36	0,37	1	0,92 [0,80-1,07]
7	rs10954758	Akut	C	0,49	0,40	$1,63 \times 10^{-04}$	1,65 [1,27-2,14]	0,38	0,39	1	0,95 [0,83-1,10]
7	rs10262146	Akut	T	0,33	0,22	$1,23 \times 10^{-05}$	1,84 [1,40-2,41]	0,20	0,21	1	0,93 [0,78-1,10]
16	*rs7203657*	*Akut*	*T*	*0,30*	*0,21*	*$4,78 \times 10^{-06}$*	*1,94 [1,46-2,57]*	*0,24*	*0,21*	*$9,38 \times 10^{-01}$*	*1,20 [1,02-1,41]*
1	rs7546353	Chronisch	C	0,23	0,18	$2,03 \times 10^{-04}$	1,53 [1,22-1,91]	0,19	0,18	1	1,06 [0,92-1,23]
2	rs1387287	Chronisch	T	0,11	0,07	$4,36 \times 10^{-05}$	1,96 [1,42-2,72]	0,07	0,08	1	0,84 [0,68-1,04]
2	rs1517634	Chronisch	G	0,30	0,23	$2,06 \times 10^{-05}$	1,57 [1,27-1,93]	0,26	0,25	1	1,02 [0,90-1,16]
4	rs2250175	Chronisch	A	0,52	0,46	$7,61 \times 10^{-05}$	1,46 [1,21-1,75]	0,45	0,45	1	1,00 [0,90-1,12]
5	rs2030888	Chronisch	T	0,26	0,20	$3,89 \times 10^{-05}$	1,59 [1,27-1,98]	0,20	0,23	1	0,88 [0,77-1,01]
8	rs7465297	Chronisch	G	0,28	0,21	$7,02 \times 10^{-05}$	1,54 [1,24-1,90]	0,22	0,21	1	1,03 [0,90-1,18]
10	*rs17157941*	*Chronisch*	*C*	*0,06*	*0,03*	*$3,50 \times 10^{-05}$*	*2,46 [1,61-3,78]*	*0,03*	*0,04*	*$1,58 \times 10^{-01}$*	*0,63 [0,45-0,87]*
11	rs6578556	Chronisch	C	0,41	0,49	$3,29 \times 10^{-05}$	0,67 [0,56-0,81]	0,48	0,47	1	1,02 [0,91-1,14]
11	rs638234	Chronisch	G	0,19	0,15	$1,91 \times 10^{-04}$	1,58 [1,24-2,02]	0,15	0,16	1	0,98 [0,84-1,14]

15	rs1869133	Chronisch	C	0,24	0,19	$8,34 \times 10^{-05}$	1,55 [1,25-1,93]	0,22	0,21	1	1,08 [0,94-1,24]
19	rs7259348	Chronisch	T	0,21	0,29	$1,70 \times 10^{-04}$	0,66 [0,53-0,82]	0,27	0,29	1	0,92 [0,81-1,04]

3.1.4.1 Power für die Detektion subphänotypischer Assoziationen

Da mit einer Reduzierung des Stichprobenumfangs auch die statistische Teststärke sinkt, mit der Assoziationssignale nachgewiesen werden können, wurde für die Stichprobe B die Teststärke berechnet, die in dieser Stichprobe nach dem Filtern der Fälle nach Subphänotypen für einen hypothetischen SNP besteht. Für einen SNP mit einem OR von 1,3 und einer MAF von 0,2 in den Kontrollen besteht in der gesamten Stichprobe B (1393 Fälle und 2204 Kontrollen) eine statistische Teststärke von 89%, um eine Assoziation von p = 0,05 nachzuweisen. Filtert man die Stichprobe nun, so dass sie nur noch Fälle des *akuten* Subphänotyps enthält (n = 502), reduziert sich die Teststärke auf 61,1%. Filtert man daraufhin die Stichprobe für Fälle des *chronischen* Subphänotyps (n = 891), erreicht man eine Teststärke von 78,9%. Die statistische Teststärke verringert sich also deutlich (*akuter* Subphänotyp: 27,9%; *chronischer* Subphänotyp: 10,1%), wenn nach einem der Subphänotypen gefiltert wird.

3.1.5 Replikation in vier unabhängigen europäischen Populationen

Um die identifizierte Assoziation des SNPs rs479777 zu konkretisieren, wurde der SNP in vier unabhängigen europäischen Stichproben (C-I bis C-IV) (Kapitel 2.1.4) mittels der TaqMan-Technology analysiert. In der deutschen Stichprobe C-I konnte der Marker erfolgreich in 303 Fällen und 281 Kontrollpersonen (Genotypisierungsrate (eng. *call rate*, CR) = 98,6%) genotypisiert werden. In der Stichprobe zeigte der Marker eine signifikante Assoziation (p = 1,33 x 10^{-3}; OR = 0,67; [95% CI] 0,53-0,86) (Tab. 3-3). In der tschechischen Stichprobe (C-II) konnte der SNP erfolgreich in 264 Fällen und 325 Kontrollen (CR = 98,7%) typisiert werden und zeigte ebenfalls eine signifikante Assoziation (p = 2,1 x 10^{-2}; OR = 0,75; [95% CI] 0,58-0,96). Auch in der schwedischen Stichprobe (C-III) (1.027 Fälle, 916 Kontrollen, CR = 96,7%) zeigte der Marker eine signifikante Assoziation (p = 5,83 x 10^{-4}; OR = 0,79; [95% CI] 0,69-0,90). In der Stichprobe C-IV der deutschen Sarkoidose-Familientrio-Stichprobe (342 Trios) wurden 11 Familientrios wegen beobachteter Mendel-Fehler ausgeschlossen. Die durchgeführte Genotypisierung war in 301 Familientrios erfolgreich (CR = 96,9%). Die 301 Familientrios wurden mit dem Transmissions-Ungleichgewichts-Test (eng. *transmission disequilibrium test*, TDT) untersucht: Der SNP zeigte signifikante Assoziation mit Sarkoidose (p = 1,95 x 10^{-3}; OR = 0,68; [95% CI] 0,54-0,87) (Tab. 3-4).

Eine Untersuchung der Stichproben hinsichtlich der Sarkoidose-Subphänotypen zeigte inkonsistente Ergebnisse in den Stichproben C-I, C-III und C-IV. In der Stichprobe C-II konnte keine Subphänotypanalyse durchgeführt werden, da für diese Stichprobe keine Subphänotyp-Informationen vorlagen. Für die Stichproben C-I und C-III wurde die statistische Teststärke zur

Detektion subphänotypischer Assoziationen berechnet, um die Aussagekraft der Ergebnisse zu überprüfen (Tab. 3-5).

Tab. 3-3: Assoziation von rs479777 in den Replikationsstichproben C-I, C-II, C-III. Der SNP rs479777 wurden in drei Replikationsstichproben analysiert. Wenn Subphänotypinformationen für die Stichprobe vorhanden waren, wurde der SNP auch für den jeweiligen Subphänotyp analysiert (Spalte Analysetyp). A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; OR = Quotenverhältnis (engl. *odds ratio*.

Stichprobe	Analysetyp	A1	AF Fälle	AF Kont.	*p*-Wert	OR [95% CI]
C-I	Allg.	C	0,29	0,38	$1{,}33 \times 10^{-03}$	0,67 [0,53-0,86]
C-I	Akut	C	0,25	0,38	$2{,}13 \times 10^{-02}$	0,54 [0,32-0,92]
C-I	Chronisch	C	0,39	0,38	$9{,}56 \times 10^{-01}$	1,01 [0,68-1,51]
C-II	Allg.	C	0,35	0,28	$2{,}10 \times 10^{-02}$	0,75 [0,58-0,96]
C-III	Allg.	C	0,36	0,31	$5{,}83 \times 10^{-04}$	0,79 [0,69-0,90]
C-III	Löfgren	C	0,31	0,36	$2{,}90 \times 10^{-02}$	0,80 [0,66-0,98]

Tab. 3-4: Ergebnisse des Transmissions-Ungleichgewichts-Tests (TDT) in der Stichprobe C-IV für rs479777. Der SNP rs479777 wurde auf Assoziation mit dem allgemeinen Sarkoidosephänotyp getestet. Für Patienten mit Subphänotypinformationen wurde der SNP auch für den jeweiligen Subphänotyp analysiert (Spalte Analysetyp). Es sind die unkorrigierten *p*-Werte des TDT (p_{TDT}) gezeigt. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall (CI) angegeben.

Analysetyp	A1	Übertragen	Nicht-Übertragen	p_{TDT}	OR [95% CI]
Allg.	C	110	161	$1{,}95 \times 10^{-03}$	0,68 [0,54-0,87]
Akut	C	33	33	1	1,00 [0,62-1,62]
Chronisch	C	61	116	$3{,}56 \times 10^{-05}$	0,53 [0,39-0,72]

Tab. 3-5: Ergebnisse der Berechnung der statistischen Teststärke für den SNP rs479777 in den Replikationsstichproben C-I und C-III. Mit den in der Stichprobe B gemessenen MAF der Kontrollen und dem OR von rs479777 wurde die statistische Teststärke berechnet, mit der ein *p*-Wert von 0,05 in der Replikation erreicht werden kann. Dabei wurden in beiden Replikationsstichproben die Fälle nach den Subphänotypen gefiltert.

Stichprobe	Teststärke Analyse akut	Teststärke Analyse chronisch
C-I	10,2%	13,0%
C-III	45,8%	-

In allen untersuchten Replikationsstichproben konnte eine signifikante Assoziation (mit einem OR von 0,67 bis 0,79) von rs479777 mit Sarkoidose nachgewiesen werden. Eine Untersuchung der ORs der Fall-Kontroll-Stichproben (A, B, C-I, C-II, C-III) mittels des Breslow-Day-Tests zeigte keine signifikante Heterogenität der ORs zwischen den untersuchten Stichproben. Eine Metaanalyse aller verfügbaren Fall-Kontroll-Stichproben (A, B, C-I, C-II, C-III) ergibt einen *p*-Wert von $2{,}68 \times 10^{-18}$ und unterstreicht nochmals die Assoziation des Markers mit der Krankheit Sarkoidose.

3.1.6 Feinkartierung der chromosomalen Region *11q13.1* (rs479777)

Basierend auf den GWAS-Ergebnissen und den Rekombinationsraten in der Region wurde eine ca. 180kb große Region um den SNP rs479777 bestimmt, um in dieser Region den zugrunde liegenden genetischen Risikofaktor für Sarkoidose zu finden (Abb. 3-2).

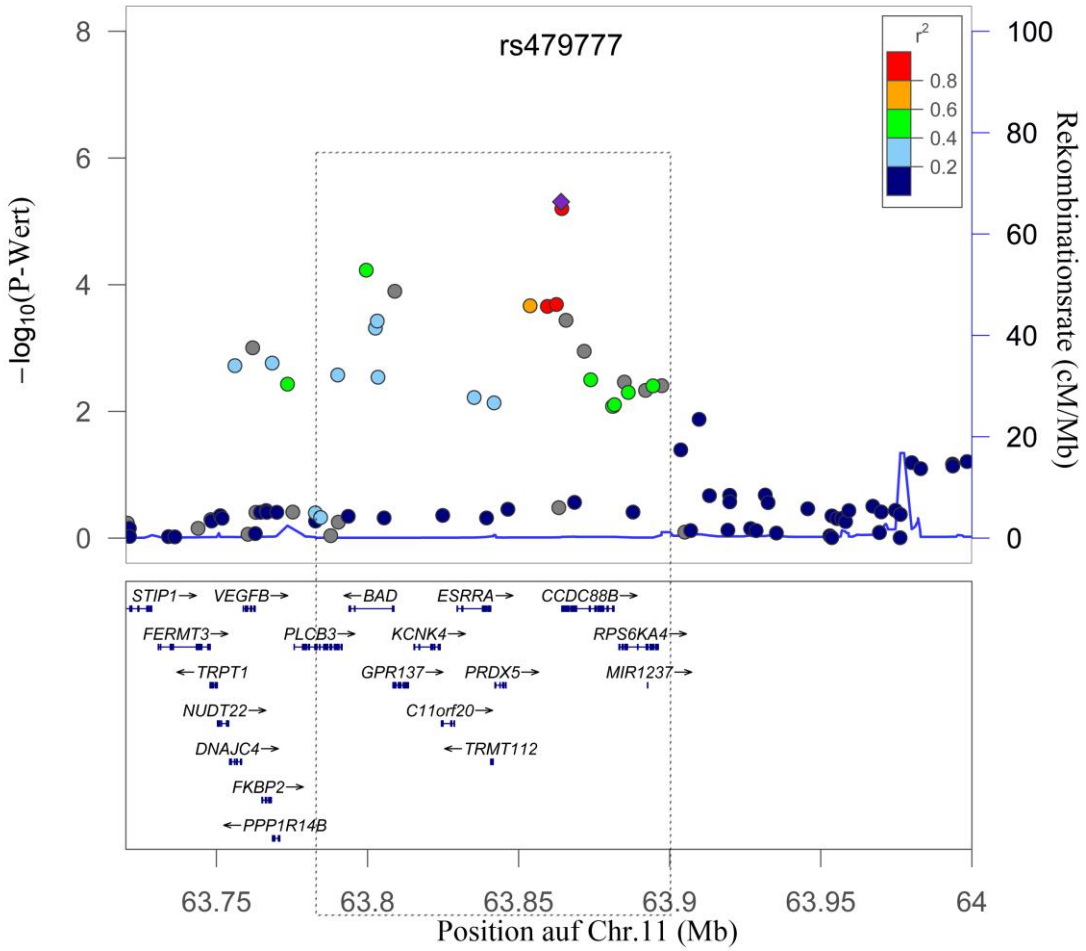

Abb. 3-2: Assoziationssignale des GWAS-Datensatzes in der *11q13.1* Hit-Region. Die chromosomale Position bezieht sich auf das „human reference genome build 18". Die Farbgebung der SNPs gibt die Stärke des Kopplungsungleichgewichtes (LD; in r^2) zwischen dem Referenz-SNP rs479777 und den umgebenden Markern an. Die Rekombinationsrate entstammt der HapMap-Datenbank und die Annotation der Gene stützt sich auf die RefSeq-Datenbank. Die gestrichelte Box zeigt die Region, welche für die Feinkartierung ausgewählt wurde

Um die Region genauer zu charakterisieren, wurden neben dem rs479777 25 *tagging*-SNPs aus der HapMap-Datenbank für die Feinkartierung in der Feinkartierungsstichprobe (Kapitel 2.1.5) ausgewählt. 24 SNPs, 1.810 Fälle (davon 596 *akute* und 1.050 *chronische*) und 2.182 Kontrollpersonen blieben nach Qualitätskontrolle übrig. In der Assoziationsanalyse zeigten 19 Marker eine nominale Assoziation von $p < 0{,}05$, mit den stärksten Assoziationssignalen für die SNPs rs479777 ($p = 3{,}18 \times 10^{-8}$; OR = 0,76; [95% CI] 0,69-0,84), rs11231740 ($p = 4{,}78 \times 10^{-9}$; OR = 1,30; [95% CI] 1,19-1,42) und rs671976 ($p = 1{,}6 \times 10^{-9}$; OR = 1,13; [95% CI] 1,20-1,43) (Tab. 3-6).

Tab. 3-6: Ergebnisse der Feinkartierung der chromosomalen Region *11q13.1*. Ergebnisse der 24 SNPs, die für die Feinkartierung ausgewählt wurden. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Die fett dargestellten SNPs zeigten die stärksten Assoziationssignale in der Feinkartierung. Abkürzungen: 5'-UTR = 5' untranslatierter Bereich (engl. 5' *untranslated region*); AF = Allelfrequenz; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; nonsyn = nicht-synonymer SNP (engl. *nonsynonymous SNP*); OR = Quotenverhältnis (engl. *odds ratio*); syn = synonymer SNP (engl. *synonymous SNP*).

dbSNP ID	Position (bp)	Gen	A1	AF Fälle	AF Kontr.	*p*-Wert	OR [95% CI]
rs3741404	63.999.240	*DNAJC4, i*	G	0,33	0,32	$1{,}76 \times 10^{-01}$	1,07 [0,97-1,17]
rs11231727	64.011.854	*PPP1R14B, us*	T	0,47	0,45	$3{,}82 \times 10^{-02}$	1,10 [1,01-1,20]
rs12798333	64.012.784	*PPP1R14B, i*	A	0,18	0,15	$2{,}55 \times 10^{-05}$	1,29 [1,15-1,45]
rs2244625	64.026.144	*PLCB3, i*	C	0,32	0,29	$6{,}98 \times 10^{-03}$	1,14 [1,04-1,26]
rs671976	**64.046.029**	***BAD, i***	**G**	**0,53**	**0,46**	$\mathbf{1{,}60 \times 10^{-09}}$	**1,31 [1,20-1,43]**
rs11231740	**64.046.641**	***BAD, i***	**T**	**0,52**	**0,45**	$\mathbf{4{,}78 \times 10^{-09}}$	**1,30 [1,19-1,42]**
rs11231741	64.046.885	*BAD, i*	T	0,35	0,32	$4{,}29 \times 10^{-03}$	1,15 [1,04-1,26]
rs2510066	64.052.447	*GPR137, 5'-UTR*	A	0,34	0,39	$3{,}51 \times 10^{-07}$	0,79 [0,72-0,86]
rs953778	64.066.999	*KCNK4, nonsyn*	T	0,15	0,16	$3{,}86 \times 10^{-01}$	0,95 [0,84-1,07]
rs731703	64.078.706	*ESRRA, i*	A	0,35	0,32	$8{,}11 \times 10^{-03}$	1,13 [1,03-1,25]
rs11231749	64.095.178	intergenisch	C	0,36	0,33	$9{,}21 \times 10^{-03}$	1,13 [1,03-1,24]
rs574087	64.102.948	intergenisch	C	0,34	0,40	$3{,}89 \times 10^{-07}$	0,79 [0,72-0,86]
rs499425	64.105.929	*CCDC88B, us*	T	0,34	0,39	$4{,}18 \times 10^{-07}$	0,79 [0,72-0,87]
rs479777	**64.107.477**	***CCDC88B, us***	**C**	**0,30**	**0,36**	$\mathbf{3{,}18 \times 10^{-08}}$	**0,76 [0,69-0,84]**
rs647152	64.109.118	*CCDC88B, nonsyn*	G	0,34	0,39	$1{,}33 \times 10^{-06}$	0,80 [0,73-0,87]
rs685870	64.111.928	*CCDC88B, nonsyn*	T	0,31	0,28	$2{,}67 \times 10^{-02}$	1,12 [1,01-1,23]
rs612448	64.124.515	*CCDC88B, syn*	G	0,44	0,47	$2{,}34 \times 10^{-03}$	0,87 [0,80-0,95]
rs519763	64.124.823	*CCDC88B, 5'-UTR*	T	0,44	0,47	$1{,}98 \times 10^{-03}$	0,87 [0,80-0,95]
rs516124	64.128.423	*RPS6KA4, i*	T	0,36	0,40	$3{,}61 \times 10^{-05}$	0,83 [0,75-0,90]
rs508168	64.135.435	*RPS6KA4, i*	A	0,36	0,40	$6{,}38 \times 10^{-05}$	0,83 [0,76-0,91]
rs11231775	64.144.525	intergenisch	C	0,41	0,42	$1{,}26 \times 10^{-01}$	0,93 [0,85-1,02]
rs6591843	64.147.083	intergenisch	A	0,47	0,49	$1{,}76 \times 10^{-01}$	0,94 [0,86-1,03]
rs750832	64.163.302	intergenisch	C	0,25	0,27	$9{,}22 \times 10^{-02}$	0,92 [0,83-1,01]
rs521319	64.172.031	intergenisch	A	0,33	0,31	$2{,}69 \times 10^{-02}$	1,11 [1,01-1,22]

Eine Übersicht der Region *11q13.1* mit den Assoziationssignalen, den Genen, der Rekombinationsrate und der LD-Struktur wurde auf Basis der in der Feinkartierung gewonnenen Daten und mit Hilfe von HapMap- und RefSeq-Daten erstellt (Abb. 3-3).

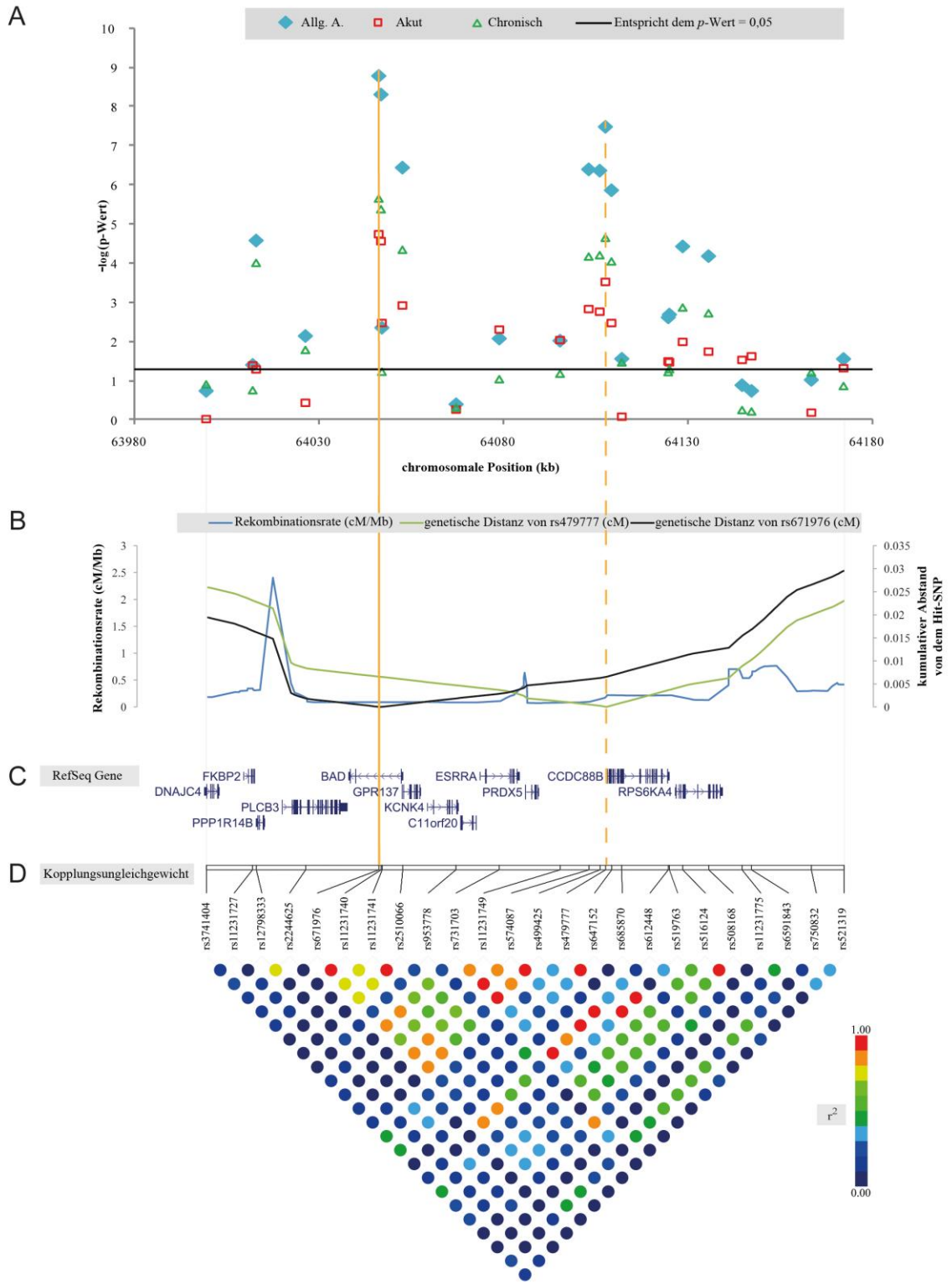

Abb. 3-3: Feinkartierung des neu entdeckten Sarkoidose-Risikolocus auf Chromosom *11q13.1*. (A) Negativ dekadischer Logarithmus der *p*-Werte inklusive der 180kb-Region, in der die Feinkartierung durchgeführt wurde. Die 23 Marker der Feinkartierung sind zusammen mit dem im Rahmen der genomweiten Assoziationsstudie entdeckten SNP rs479777 dargestellt. Der am stärksten assoziierte SNP rs671976 ist durch die durchgezogene orange Linie gekennzeichnet; die gestrichelte orange Linie markiert den SNP rs479777. Die grauen vertikalen Linien markieren die am äußersten Rand der analysierten Region gelegenen SNPs. Die Nukleotidpositionen beziehen sich auf das *„human genome build 19“*. **(B)** Die blaue Linie in dem Diagramm zeigt die Rekombinationsrate (cM/Mb) der untersuchten Region. Die grüne Linie stellt die kumulative genetische Distanz von dem SNP rs479777 in Centimorgan (cM) dar und die schwarze Linie zeigt die kumulative

genetische Distanz von dem SNP rs671976. Alle Daten wurden der HapMap-Datenbank entnommen. **(C)** Die Intron- und Exonstruktur der Gene der Region entsprechend der RefSeq-Datenbank. **(D)** Plot des paarweisen Kopplungsungleichgewichtes zwischen den analysierten Markern in den Kontrollindividuen (Dargestellt als r^2).

Abbildung 3-3 zeigt, dass das Assoziationssignal sich auf eine ungefähr 120 kb große Region beschränkt, begrenzt durch einen Rekombinations-Hotspot bei 64.200 kb und durch ein Fehlen von Assoziationssignalen nach 64.140 kb. Eine Untersuchung der LD-Struktur mit den Kontrollen der Feinkartierungsstichprobe zeigte in dem Locus *11q13.1* eine komplexe LD-Struktur, mit dem am stärksten assoziierten SNP rs671976 befand sich nur ein weiterer stark assoziierter Marker (rs11231740) in hohem LD (r^2 = 0.98), während zwischen dem anderen stark assoziierten SNP (rs479777) ein hohes LD mit den Markern rs2510066, rs574087, rs499425 und rs647152 (r^2 > 0,82) beobachtet wurde. Zwischen den beiden SNPs rs671976 und rs479777 herrscht nur ein mäßiges LD (r^2 = 0,44) in den Kontrollen. Ein unter der Verwendung eines logistischen Regressionsmodells durchgeführter Likelihood-Quotienten-Test zeigte, dass die beiden Marker trotzdem keine unabhängigen Assoziationssignale repräsentieren (Tab. 3-7).

Tab. 3-7: Ergebnisse des Likelihood-Quotienten-Tests für das genotypische Risikomodell jeweils ohne oder mit einem der beiden Marker rs479777 und rs671976. Die Assoziationen der beiden Marker wurden über eine logistische Regression der Ergebnisse des Likelihood-Quotienten-Tests berechnet. Dabei wurde das genotypische Risikomodell für einen Marker alleine und mit dem Effekt des jeweils anderen Markers betrachtet. Der Test wurde für den allgemeinen Phänotyp der Sarkoidose und zusätzlich für die beiden Subphänotypen durchgeführt (Spalte Analysetyp).

	rs479777		rs671976	
Analysetyp	**Marker alleine *p*-Wert**	**Risikomodell mit rs671976 *p*-Wert**	**Marker alleine *p*-Wert**	**Risikomodell mit rs479777 *p*-Wert**
Allg.	$2{,}47 \times 10^{-7}$	$1{,}03 \times 10^{-1}$	$9{,}65 \times 10^{-9}$	$7{,}15 \times 10^{-3}$
Akut	$9{,}33 \times 10^{-4}$	$5{,}06 \times 10^{-1}$	$4{,}99 \times 10^{-5}$	$3{,}04 \times 10^{-2}$
Chronisch	$1{,}09 \times 10^{-4}$	$2{,}56 \times 10^{-1}$	$1{,}08 \times 10^{-5}$	$5{,}32 \times 10^{-3}$

In dem Locus *11q13.1* wurden schon Assoziationen mit den Krankheiten Morbus Crohn, Alopecia areata, ‚primäre biliäre Zirrhose', Lepra und Psoriasis dokumentiert (D. Ellinghaus et al. 2012; Franke et al. 2010a; Jagielska et al. 2012; Mells et al. 2011; Petukhova et al. 2010; F. Zhang et al. 2011a). Bei den Krankheiten Morbus Crohn, Alopecia areata und Psoriasis zeigt der Marker rs694739 das stärkste Assoziationssignal in der Region. Laut HapMap-Daten befindet sich der SNP in hohem LD (r^2 = 0,84) mit rs479777 und wurde daher in der hier durchgeführten Feinkartierung nicht direkt genotypisiert, sondern in dem ursprünglichen Markerset über das LD abgedeckt. Daher wurde der Marker nochmals in der Feinkartierungsstichprobe mittels der TaqMan-Technologie genotypisiert. Der SNP zeigte eine signifikante Assoziation ($p = 3{,}55 \times 10^{-6}$; OR = 0,80; [95% CI] 0,73-0,88), die jedoch schwächer als die Assoziation des SNPs rs671976 ist. Für die Krankheiten ‚primäre biliäre Zirrhose' und Lepra zeigte der Marker rs538147 das stärkste Assoziationssignal in der Region. Der SNP steht

aber weder mit rs479777 oder rs671976 in LD, wurde aber über den *tag*-SNP rs516124 in der Feinkartierung abgedeckt. Da in der Feinkartierung der SNP rs516124 ein sehr viel schwächeres Signal als rs671976 zeigte, wurde der Marker rs538147 nicht nachverfolgt.

3.1.7 *In silico*- und Expressionsanalysen

Die mit Sarkoidose assoziierte Region enthält acht Gene, deren Expressionsregulierung durch kausative Varianten beeinträchtigt werden könnte. Die Region mit den am signifikantesten assoziierten SNPs rs671976, rs11231740, rs479777, rs2510066, rs574087, rs499425 und rs647152 zeigte ein unspezifisches hohes regulatorisches Potential in der UCSC-Datenbank, deshalb wurde für die Region eine *in silico*-Analyse durchgeführt. Dabei wurde die Analyse auf 11 SNPs, die in hohem LD (r^2 > 0,8) zueinander stehen, ausgeweitet und die Marker mit dem NIEHS *SNPinfo*-Webserver überprüft (Xu and Taylor 2009). Für die SNPs rs2510066 und rs663743 wurde ein hohes regulatorisches Potential vorhergesagt, während rs647152 ein mäßiges und rs479777 nur ein ganz schwaches regulatorisches Potential aufwies (Tab. 3-8). Wie weiter in Tabelle 3-8 zu sehen ist, liegen die SNPs rs2510066, rs574087, rs499425, rs479777 und rs663743 in Transkriptionsfaktorbindestellen (TFBS) und rs663743, rs647152 und rs1199046 in Bereichen, in denen Spleißen stattfindet. Die Genvar-Datenbank sagte einen potentiellen *cis*-regulatorischen Effekt der SNPs rs671976, rs11231740, rs2510066, rs574087, rs499425 und rs647152, mit unkorrigierten *p*-Werten von 2,21 x 10^{-4} bis 3,60 x 10^{-7}, auf die Expression des Gens *CCDC88B* (engl. *coiled-coil domain containing 88B*) in lymphoblastoiden Zelllinien voraus (Nica et al. 2011) (Tab. 3-9). Darüber hinaus ist rs647152 ein nicht-synonymer SNP (193 Asp > Glu), der in dem Exon 7 des *CCDC88B*-Gens liegt. Des Weiteren wurde die Region um den SNP rs479777 genauer in der UCSC-Datenbank analysiert (Kapitel 7, Abb. 7-1). Die dort eingebundene ENCODE-Datenbank (Rosenbloom et al. 2012) zeigte in der Region ein stark erhöhtes Level an Acetylierung von Lysin 27 im Histon 3 (H3K27Ac) in Gm12878-Zelllinien (lymphoblastoide Zellen) und NHLF-Zelllinien (normale humane Lungenfibroblasten) und die Bindung einer Reihe von Transkriptionsfaktoren (u.a. ELF1, NF-kappaB, GABP, PAX5, YY1, PU.1 und weitere (siehe Abb. 7-1)) in lymphoblastoiden Zelllinien. Eine genauere Analyse der Region um den SNP rs671976 in der UCSC-Datenbank zeigte weder eine erhöhte Methylierung oder Acetylierung der Histone, es wurde auch keine Bindung (in relevanten Zelllinien) von Transkriptionsfaktoren in der Nähe des Markers beobachtet (Kapitel 7, Abb. 7-2).

Tab. 3-8: NIEHS Prognose von SNP-Funktionen. Abkürzungen: Chr. = Chromosom; nsSNP = nicht-synonymer SNP; TFBS = Transkriptionsfaktorbindestelle; Reg. Potential = Regulationspotential; Konserv. = Konservierung; NA = nicht verfügbar (engl. *not available*).

dbSNP ID	Position (bp)	*p*-Wert	Allele	LD-SNPs	TFBS	Spleißen	nsSNP	Reg. Potential	Konserv.
rs671976	**63.802.605**	**1,60 x 10^{-09}**	**A/G**	**rs671976|rs11231740**	**-**	**-**	**-**	**0**	**0.131**
rs11231740	63.803.217	4,78 x 10^{-09}	C/T	rs11231740|rs671976	-	-	-	0	0.002
rs2510066	63.809.023	3,51 x 10^{-07}	T/C	rs2510066|rs647152| rs479777|rs574087| rs499425	+	-	-	0.298002	0.007
rs694739	63.853.809	NA	G/A	rs647152|rs479777| rs2510066|rs574087| rs499425	-	-	-	0	0.001
rs574087	63.859.524	3,89 x 10^{-07}	G/A	rs574087|rs647152| rs479777|rs2510066| rs499425	+	-	-	NA	0.032
rs499425	63.862.505	4,18 x 10^{-07}	G/A	rs499425|rs647152| rs479777|rs2510066| rs574087	+	-	-	0	0.001
rs479777	**63.864.053**	**3,18 x 10^{-08}**	**C/T**	**rs479777|rs2510066| rs574087|rs499425**	**+**	**-**	**-**	**0.004805**	**0**
rs663743	63.864.311	NA	A/G	rs479777|rs499425	+	+	-	0.401337	0
rs647152	63.865.694	1,33 x 10^{-06}	G/T	rs647152|rs2510066| rs574087|rs499425	-	+	+	0.125346	0.068
rs510372	63.871.713	NA	T/C	rs647152|rs2510066| rs574087|rs499425	-	-	-	NA	0
rs1199046	63.874.702	NA	C/T	rs647152|rs2510066| rs499425	-	+	-	NA	0

Tab. 3-9: Genevar-Datenbank mit prognostizierten *cis*-regulatorischen Effekten auf die *CCDC88B*-Expression. Von den 11 analysierten SNPs zeigten sechs SNPs in der Genevar-Datenbank einen *cis*-regulatorischen Effekt auf die *CCDC88B*-Expression. Der *p*-Wert gibt die mit einem linearen Regressionsmodel errechnete Signifikanz der Assoziation zwischen einem SNP und der veränderten Expression des Gens an.

dbSNP ID	Gen	Chr.	SNP-Position (bp)	Abstand (bp)	*p*-Wert
rs671976	*CCDC88B*	11	63.802.605	61666	$2{,}21 \times 10^{-04}$
rs11231740	*CCDC88B*	11	63.803.217	61054	$2{,}21 \times 10^{-04}$
rs2510066	*CCDC88B*	11	63.809.023	55248	$3{,}60 \times 10^{-07}$
rs574087	*CCDC88B*	11	63.859.524	4747	$3{,}60 \times 10^{-07}$
rs499425	*CCDC88B*	11	63.862.505	1766	$3{,}60 \times 10^{-07}$
rs647152	*CCDC88B*	11	63.865.694	1423	$4{,}77 \times 10^{-07}$

Eine weitere *in silico*-Analyse mit dem Program SNPex (Castelli 2009) in einem unabhängigen Set lymphoblastoider Zelllinien sagte für die SNPs rs671976 und rs11231740 eine *cis*-eQTL-Funktion (eQTL: *expression quantitative trait loci*) für die *KCNK4* (engl. *potassium channel subfamily K member 4*) Genexpression voraus ($p = 4{,}24 \times 10^{-4}$) (Tab. 3-10). Eine weniger signifikante eQTL-Voraussage konnte für zwei Transkripte des *PRDX5*-Gens (engl. *peroxiredoxin 5*) gemacht werden ($p = 1{,}53 \times 10^{-3}$ und $1{,}34 \times 10^{-3}$) (Tab. 3-10).

Tab. 3-10: Ergebnisse der konditionalen Analyse für SNP-Gen-Interaktionen der SNPex-Datenbank. Für die Analyse standen in der SNPex-Datenbank Expressionsdaten von bis zu vier verschiedenen lymphoblastoider Zelllinien zur Verfügung (Zahlen in Klammern hinter dem Gennamen). Der *p*-Wert gibt die Signifikanz der mit einem linearen Regressionsmodel errechneten Assoziation zwischen einem SNP und der veränderten Expression des Gens an. Signifikante *p*-Werte sind in kursiv dargestellt.

	***p*-Wert**							
dbSNP ID	**KCNK4 (1)**	**KCNK4 (2)**	**KCNK4 (3)**	**KCNK4 (4)**	**ESRRA**	**PRDX5 (1)**	**PRDX5 (2)**	**RPS6KA4**
rs671976	*4,24 x 10^{-04}*	2,36 x 10^{-01}	3,20 x 10^{-01}	5,22 x 10^{-01}	8,38 x 10^{-01}	*1,53 x 10^{-03}*	*1,34 x 10^{-03}*	1,76 x 10^{-01}
rs11231740	*4,24 x 10^{-04}*	2,36 x 10^{-01}	3,20 x 10^{-01}	5,22 x 10^{-01}	8,38 x 10^{-01}	*1,53 x 10^{-03}*	*1,34 x 10^{-03}*	1,76 x 10^{-01}
rs2510066	2,48 x 10^{-01}	8,73 x 10^{-01}	7,51 x 10^{-02}	1,62 x 10^{-01}	2,03 x 10^{-01}	2,25 x 10^{-01}	3,48 x 10^{-01}	1,27 x 10^{-01}
rs574087	1,41 x 10^{-01}	9,28 x 10^{-01}	7,89 x 10^{-02}	1,22 x 10^{-01}	1,82 x 10^{-01}	2,24 x 10^{-01}	3,54 x 10^{-01}	7,45 x 10^{-02}
rs499425	2,48 x 10^{-01}	8,73 x 10^{-01}	7,51 x 10^{-02}	1,62 x 10^{-01}	2,03 x 10^{-01}	2,25 x 10^{-01}	3,48 x 10^{-01}	1,27 x 10^{-01}
rs479777	3,35 x 10^{-01}	4,67 x 10^{-01}	2,40 x 10^{-01}	9,72 x 10^{-02}	1,32 x 10^{-01}	3,99 x 10^{-01}	6,03 x 10^{-01}	1,97 x 10^{-01}
rs647152	1,62 x 10^{-01}	8,00 x 10^{-01}	6,16 x 10^{-02}	3,17 x 10^{-01}	3,48 x 10^{-01}	3,02 x 10^{-01}	5,02 x 10^{-01}	1,64 x 10^{-01}

Mittels eines kommerziell verfügbaren cDNA-Gewebe-Panels mit verschiedenen humanen Geweben und Zelllinien wurde die Expression der in der Region befindlichen Gene in den verschiedenen Geweben überprüft (Abb. 3-4). Bei den für die Sarkoidose relevanten Gewebe- und Zelltypen (Lunge und Leukozyten) zeigt die Analyse die Expression von mRNA der Gene *PLCB3* (engl. *phospholipase C,* β *3*), *BAD* (engl. *BCL2-associated agonist of cell death*), *GPR137* (engl. *G protein-coupled receptor 137*), *KCNK4*, *PRDX5* und *CCDC88B*. Während *CCDC88B*-Expression nur in Leukozyten, Milz, Thymus, Pankreas und Lungengewebe nachgewiesen werden konnte, zeigten *BAD*, *GPR137*, *KCNK4* und *PRDX5* Expression in allen getesteten Geweben und Zelllinien.

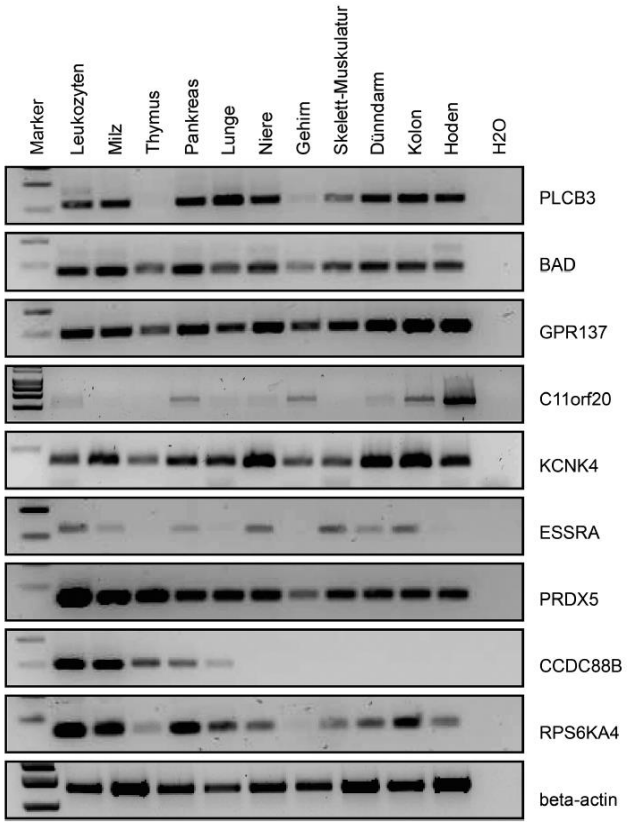

Abb. 3-4: mRNA-Expression der Kandidatengene der assoziierten Region *11q13.1* in verschiedenen humanen Geweben und Zelllinien. Die mRNA-Expression der Gene in der assoziierten Region wurde mittels eines cDNA-Gewebe-Panels untersucht. RT-PCR der gewebe- und zelltypischen Expression der neun Kandidatengene in verschiedenen Gewebe- und Zelltypen, als Kontrolle für die Expression wurde das Haushaltsgen β-*actin* verwendet.

Basierend auf den Expressionsergebnissen und den *in silico*-Analysen wurden die Gene *KCNK4*, *PRDX5* und *CCDC88B* ausgewählt, um in einer weiteren Analyse die differentielle Expression mittels real-time-PCR in bronchoalveolären Lavagen-Proben (BAL) von gesunden Individuen und Sarkoidosepatienten (jeweils n = 12) zu untersuchen. Die quantitative Expressionstudie der Kandidatengene in der BAL-Stichprobe I wurde von Dr. Simone Lipinski durchgeführt. Bei den verwendeten Individuen wurde darauf geachtet, dass sich die Zellzusammensetzungen der Proben

glichen, um Fehler durch zufällige Unterschiede in der Zellzusammensetzung zu vermeiden. Die Analyse der relativen Expression zeigte, dass alle getesteten Gene in den Patienten - verglichen mit jenen der Kontrollpersonen - hoch-reguliert waren (Abb. 3-5). Nur die Expression der zwei Gene *KCNK4* und *PRDX5* zeigte einen signifikanten Unterschied mit dem Wilcoxon-Mann-Whitney-Test zwischen Kontrollen und Sarkoidosepatienten.

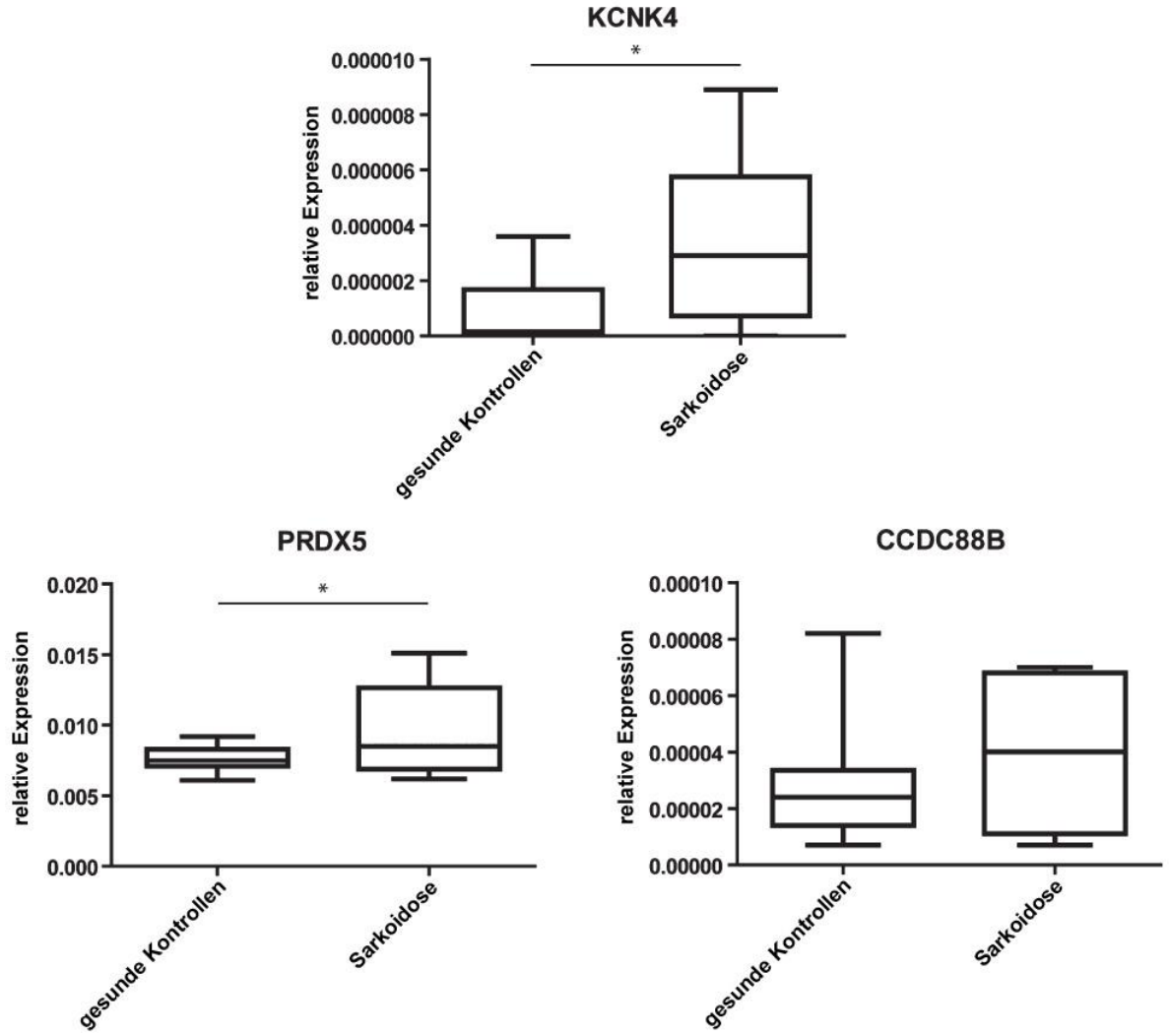

Abb. 3-5: Expression von ausgewählten Genen in Zellen aus bronchoalveolären Lavagen von Sarkoidosepatienten und Kontrollpersonen. Die Proben wurden so ausgewählt, dass sie sich untereinander in ihrem Gehalt an alveolären Makrophagen glichen (bronchoalveoläre Lavagen Stichprobe I). * *p*-Wert < 0,05 in dem Wilcoxon-Mann-Whitney-Test.

Obwohl sich bei *CCDC88B* kein signifikanter Unterschied zwischen den BAL-Proben von Patienten und Kontrollpersonen zeigte, wurde *CCDC88B* wegen der gewebespezifischen Expression und der Voraussage von *cis*-regulatorischen Effekten von mehreren SNPs auf das Gen für eine weiterführende detaillierte Expressionsanalyse ausgewählt. Um den vorhergesagten *cis*-regulatorischen Effekt genauer zu untersuchen, wurde die mRNA-Expression in BAL-Proben (bronchoalveoläre Lavagen Stichprobe II) analysiert. Die genotypspezifische Expressionsstudie in der BAL-Stichprobe II wurde von Prof. Dr. Gernot Zissel und Dr. Kerstin Höhne durchgeführt. Die BAL-Stichprobe II bestand aus

unselektierten BAL-Proben von 27 Sarkoidosepatienten, deren Genotypstatus des SNP-Markers rs671976 mit der Expression des *CCDC88B*-Gens korreliert wurde (Abb. 3-6). Die Analyse der Stichprobe zeigte bei den Proben, welche homozygot für das rs67976 G-Allel waren, ein erhöhtes *CCDC88B* Expressionslevel (p = 0,0187). Bei den verschiedenen Patienten der Stichprobe variierte der prozentuale Anteil der alveolaren Makrophagen von 7 bis 96% und korrelierte nicht mit dem rs671976 Genotyp (durchschnittlicher Anteil der Allele: 51% [AA], 53% [AG] und 53% [GG]).

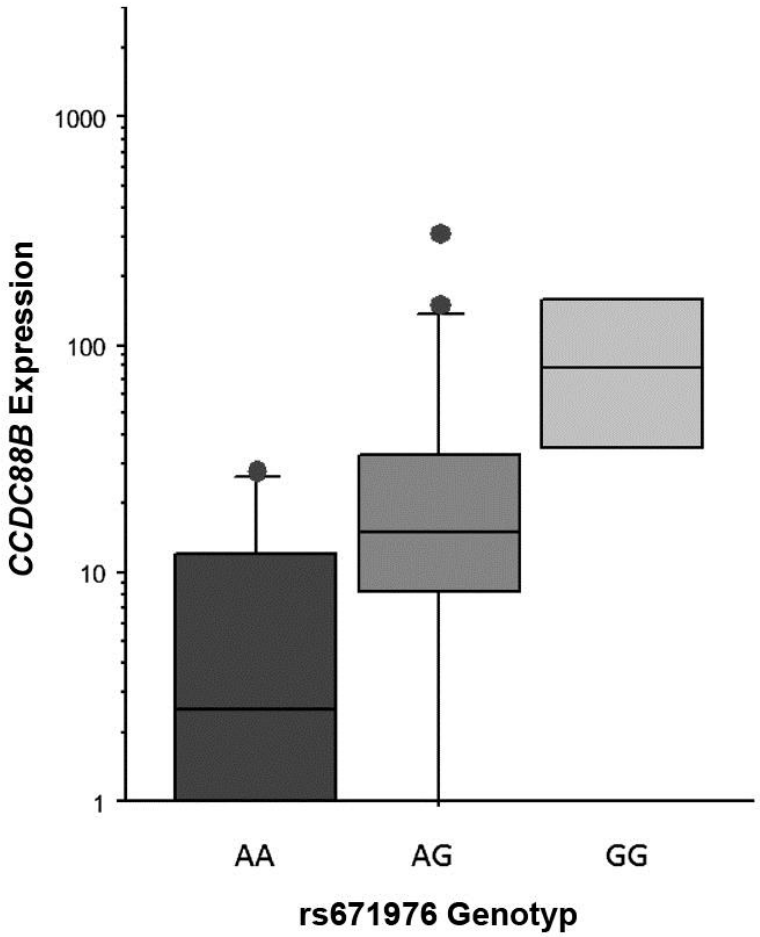

Abb. 3-6: Allel-spezifische Expression der *CCDC88B* mRNA in der bronchoalveoläre Lavagen Stichprobe II. Die Expression der *CCDC88B* mRNA ist positiv mit der Anzahl der rs671976 Risikoallele assoziiert (G Allele; Kruskal-Wallis-Test; p = 0,0187). Die relative Expression ist in einem dimensionsfreien Verhältnis angegeben.

3.2 Identifizierung gemeinsamer genetischer Faktoren der Krankheiten Tuberkulose und Sarkoidose

Um die Krankheiten Tuberkulose und Sarkoidose auf gemeinsame genetische Assoziationen – besonders im Hinblick auf die phänotypische Erscheinungsform der entzündlichen granulomatösen Erkrankung im Allgemeinen – zu überprüfen, standen zwei GWAS-Datensätze zur Verfügung (Kapitel 2.1.2). Dabei wurde ein zweistufiges Studiendesign mit der genomweiten Signifikanz als Signifikanzschwelle verwendet, bei dem in der ersten Stufe (GWAS-Datensätze) alle Marker jedes Individuums auf Assoziation mit der jeweiligen Krankheit getestet wurden (Kapitel 2.2.1.2). In der ersten Stufe wurden vier verschiedene Analysemethoden zur Identifizierung gemeinsamer Assoziationen angewandt (Kapitel 2.2.6.2). In der zweiten Stufe wurden die vielversprechendsten Marker aus der ersten Stufe in weiteren Individuen (Stichproben B, E-I, E-II, C-I, C-II und F) auf Assoziation mit der jeweiligen Krankheit getestet. Dabei fungierten diese Stichproben als Replikationsstichproben, um die in den GWAS-Datensätzen detektierten Assoziationen in zwei Replikationsphasen zu bestätigen. Von allen Stichproben wurden dann die Assoziationsergebnisse der SNPs in einer Metaanalyse zusammengefasst und anhand der genomweiten Signifikanzgrenze beurteilt (siehe Kapitel 2.2.1.2).

3.2.1 Tuberkulose-GWAS

Nach der Qualitätskontrolle der Stichprobe D umfasste der Tuberkulose GWAS-Datensatz Proben von 3.078 Individuen. Dabei galt:

- Individuen mit mehr als 4% fehlender SNP-Daten und einer Heterozygositätsrate > 33% wurden von der Analyse ausgeschlossen (n = 25);
- von unerwarteten Duplikaten oder Verwandten wurde jeweils das Individuum mit der höheren Rate an fehlenden SNP-Daten ausgeschlossen (n = 77);
- nach Überprüfung auf Populationsstratifikation wurden 2 Individuen ausgeschlossen.

Die 3.078 Individuen teilen sich in 1.288 Tuberkulosepatienten und 1.790 gesunde Kontrollpersonen auf. Nach Genotyp-Imputation umfasste der Datensatz 1.358.606 SNPs für weiterführende Analysen. Die genetische Heterogenität des Datensatzes war mit einem *genomic inflation factor* von λ_{GC} = 1,019 gering, und die multidimensionale Skalierungsanalyse bestätigte die erwartete ghanaische Abstammung aller Individuen, wie auch in der dazugehörigen Quantil-Quantil-Abbildung zu erkennen ist (Abb. 3-7).

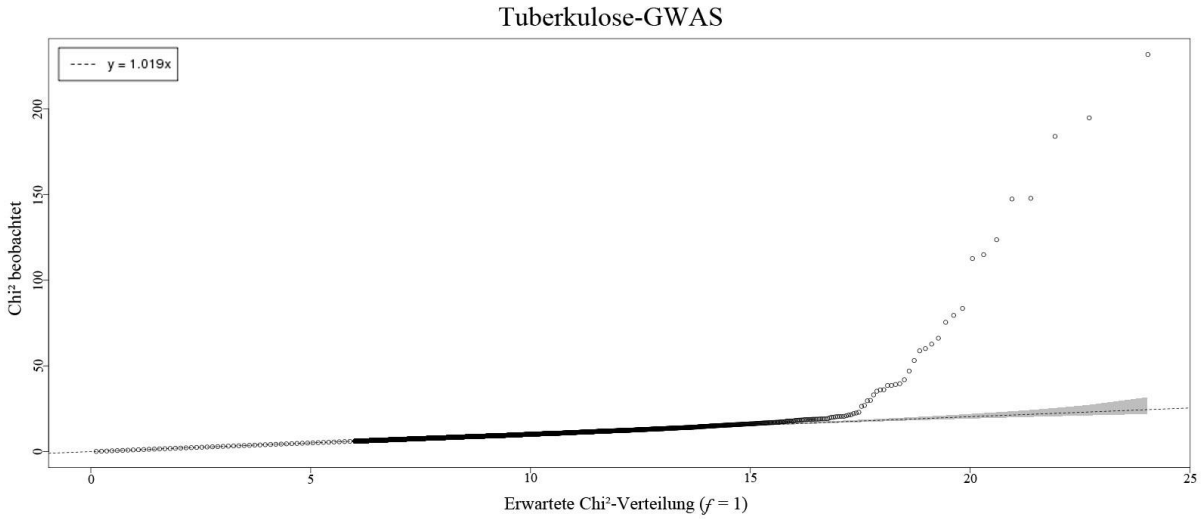

Abb. 3-7: Quantil-Quantil-Diagramm der χ^2-Teststatistik für den Tuberkulose-GWAS. Q-Q-Abbildung der χ^2-Teststatistik mit 95%-Konfidenzbereich (grau unterlegt) für alle SNPs in dem Tuberkulose GWAS, die den Qualitätskriterien entsprachen.

3.2.2 Gemeinsame genetische Faktoren für Tuberkulose und Sarkoidose

Die Assoziationsdaten der GWAS-Datensätze von Tuberkulose und Sarkoidose wurden gemeinsam in vier verschiedenen Ansätzen analysiert (siehe Kapitel 2.2.6.2). Dabei standen nach der Kombination der beiden gefilterten GWAS-Datensätze 949.988 SNPs für die ersten drei nachfolgenden Analysen zur Verfügung und 603.137 SNPs für die ‚*gene ranking*'-Analyse:

1) In der ‚*meta-analysis fixed effects*' wurden 17 SNPs mit einem *p*-Wert < 1 x 10^{-4} für die weiterführenden Analysen ausgewählt (Tab. 3-11).

2) In der ‚*meta-analysis opposite effects*' wurden 13 SNPs mit einem *p*-Wert < 1 x 10^{-4} für die weiterführenden Analysen ausgewählt (Tab. 3-11).

3) In der ‚*LD cluster ranking*'-Analyse wurden 25 SNPs für die weiteren Untersuchungen ausgewählt (Tab. 3-12). (In dem Fall, dass unterschiedliche ‚*lead*-SNPs' hochrangige LD-Gruppen repräsentierten, wurden beide SNPs für die Nachverfolgung ausgewählt.)

4) In der ‚*gene ranking*'-Analyse wurden 8 SNPs für weiterführende Analysen ausgewählt (Tab. 3-12).

Nach dem Ausscheiden von mehrfach detektierten Markern wurden als Ergebnis der vier Analysen insgesamt 52 SNPs für die erste Replikationsphase ausgewählt (Tab. 3-11 und Tab. 3-12). Die Allelfrequenzen und die 95% Konfidenzintervalle der 52 SNPs sind in Kapitel 7 in den Tabellen 7-1 und 7-2 aufgeführt.

Tab. 3-11: Ergebnisse für die 30 SNPs mit den höchsten Rängen aus den beiden Metaanalysen der GWAS-Daten. Der erste Abschnitt der Tabelle zeigt die 17 Marker, die mit *meta-analysis fixed effects* identifiziert wurden, der zweite Abschnitt die 13 Marker mit *meta-analysis opposite effects*. Die Spalten ‚Metaanalyse fixed effects' und ‚Metaanalyse opposite effects' zeigen den *p*-Wert und das OR der durchgeführten Metaanalysen. In den Spalten Sarkoidose und Tuberkulose sind die *p*-Werte und allelischen ORs der SNPs aus den GWAS-Datensätzen dargestellt. Das A1 in den eckigen Klammern hinter dem OR bezeichnet das seltenere Allel des SNPs in den Kontrollen der GWAS-Daten. Die Nukleotidpositionen beziehen sich auf das „*human genome build 18*" Abkürzungen: Chr. = Chromosom; e = exonisch; i = intronisch; OR = Quotenverhältnis (engl. *odds ratio*).

				Metaanalyse fixed effects		**Sarkoidose**		**Tuberkulose**	
Chr.	**Position (bp)**	**dbSNP ID**	**Gen**	***p*-Wert**	**OR**	***p*-Wert**	**OR [A1]**	***p*-Wert**	**OR [A1]**
12	114202449	rs1732581	intergenisch	$7{,}93 \times 10^{-06}$	0,70	$6{,}66 \times 10^{-04}$	0,75 [T]	$3{,}36 \times 10^{-04}$	0,45 [T]
11	115566825	rs2621483	intergenisch	$1{,}21 \times 10^{-05}$	1,35	$7{,}76 \times 10^{-03}$	1,37 [A]	$5{,}12 \times 10^{-04}$	1,34 [A]
1	22749827	rs11584687	intergenisch	$1{,}52 \times 10^{-05}$	0,83	$3{,}59 \times 10^{-03}$	0,79 [G]	$1{,}10 \times 10^{-03}$	0,84 [G]
2	26816882	rs11126681	intergenisch	$2{,}07 \times 10^{-05}$	0,78	$1{,}77 \times 10^{-03}$	0,77 [G]	$3{,}72 \times 10^{-03}$	0,79 [G]
4	188322475	rs1471921	intergenisch	$3{,}09 \times 10^{-05}$	0,81	$6{,}28 \times 10^{-03}$	0,74 [C]	$1{,}04 \times 10^{-03}$	0,83 [T]
17	75845351	rs9915508	intergenisch	$3{,}21 \times 10^{-05}$	0,77	$4{,}67 \times 10^{-03}$	0,76 [T]	$2{,}26 \times 10^{-03}$	1,28 [G]
10	38511145	rs2505202	*LOC100129055, i*	$3{,}26 \times 10^{-05}$	0,82	$3{,}22 \times 10^{-03}$	0,79 [C]	$2{,}74 \times 10^{-03}$	0,84 [C]
9	116703705	rs3181200	*TNFSF8, e*	$3{,}32 \times 10^{-05}$	0,80	$5{,}61 \times 10^{-03}$	0,80 [T]	$1{,}99 \times 10^{-03}$	1,27 [G]
12	116741236	rs10850905	*KSR2, i*	$4{,}21 \times 10^{-05}$	0,73	$5{,}11 \times 10^{-03}$	0,73 [G]	$2{,}82 \times 10^{-03}$	0,73 [G]
1	15733390	rs11583306	*DNAJC16, i*	$4{,}31 \times 10^{-05}$	0,83	$1{,}54 \times 10^{-03}$	0,74 [T]	$3{,}44 \times 10^{-03}$	0,86 [T]
10	118447895	rs758362	*HSPA12A, i*	$4{,}43 \times 10^{-05}$	0,82	$4{,}79 \times 10^{-03}$	1,28 [A]	$2{,}48 \times 10^{-03}$	1,19 [A]
2	187249212	rs10931252	*ITGAV, i*	$4{,}85 \times 10^{-05}$	0,78	$5{,}76 \times 10^{-03}$	0,76 [T]	$2{,}65 \times 10^{-03}$	0,79 [T]
1	57116750	rs2300955	*C8A, i*	$6{,}88 \times 10^{-05}$	0,78	$3{,}03 \times 10^{-03}$	0,72 [A]	$5{,}64 \times 10^{-03}$	0,80 [A]
13	32681621	rs9527209	*STARD13, i*	$9{,}08 \times 10^{-05}$	0,69	$3{,}54 \times 10^{-03}$	0,74 [A]	$2{,}61 \times 10^{-03}$	0,51 [A]
9	116705008	rs3181374	*TNFSF8, e*	$9{,}61 \times 10^{-05}$	0,81	$5{,}83 \times 10^{-03}$	0,80 [G]	$5{,}77 \times 10^{-03}$	0,81 [A]
10	38457199	rs1548255	intergenisch	$9{,}61 \times 10^{-05}$	0,83	$4{,}96 \times 10^{-03}$	0,79 [C]	$5{,}42 \times 10^{-03}$	0,85 [C]
17	73350621	rs8080480	intergenisch	$9{,}71 \times 10^{-05}$	1,19	$1{,}80 \times 10^{-03}$	1,29 [T]	$8{,}61 \times 10^{-03}$	1,15 [C]
				Metaanalyse opposite effects		**Sarkoidose**		**Tuberkulose**	
Chr.	**Position (bp)**	**dbSNP ID**	**Gen**	***p*-Wert**	**OR**	***p*-Wert**	**OR [A1]**	***p*-Wert**	**OR [A1]**
6	32520458	rs7194	*HLA-DRA, e*	$9{,}04 \times 10^{-07}$	1,25	$1{,}00 \times 10^{-05}$	1,44 [G]	$2{,}83 \times 10^{-03}$	0,85 [G]
6	32519013	rs8084	*HLA-DRA, e*	$7{,}19 \times 10^{-06}$	1,23	$3{,}44 \times 10^{-05}$	1,41 [A]	$8{,}48 \times 10^{-03}$	0,86 [A]
5	78853379	rs6884494	intergenisch	$1{,}03 \times 10^{-05}$	0,76	$1{,}69 \times 10^{-04}$	0,57 [A]	$1{,}72 \times 10^{-03}$	1,23 [A]
10	81891702	rs745182	*PLAC9, i*	$1{,}78 \times 10^{-05}$	1,24	$2{,}01 \times 10^{-04}$	1,36 [C]	$9{,}61 \times 10^{-03}$	0,85 [T]
18	41186112	rs11659929	*SLC14A2, i*	$1{,}96 \times 10^{-05}$	1,21	$9{,}68 \times 10^{-04}$	1,32 [C]	$2{,}96 \times 10^{-03}$	0,85 [C]
7	22707984	rs4321884	intergenisch	$3{,}29 \times 10^{-05}$	1,21	$5{,}01 \times 10^{-03}$	1,25 [A]	$1{,}92 \times 10^{-03}$	0,84 [A]
2	137290355	rs778199	intergenisch	$3{,}75 \times 10^{-05}$	1,22	$1{,}97 \times 10^{-03}$	1,29 [T]	$4{,}54 \times 10^{-03}$	0,84 [T]
2	137316119	rs778193	intergenisch	$4{,}05 \times 10^{-05}$	1,24	$4{,}11 \times 10^{-03}$	1,26 [T]	$3{,}21 \times 10^{-03}$	0,82 [T]
1	150466124	rs1390488	intergenisch	$4{,}15 \times 10^{-05}$	1,23	$1{,}20 \times 10^{-03}$	1,42 [T]	$3{,}33 \times 10^{-03}$	0,85 [C]
3	194020854	rs6444661	*MB21D2, i*	$5{,}77 \times 10^{-05}$	0,76	$3{,}23 \times 10^{-03}$	0,78 [T]	$5{,}75 \times 10^{-03}$	1,35 [T]
5	78839534	rs11948804	*HOMER1, i*	$6{,}04 \times 10^{-05}$	0,75	$2{,}14 \times 10^{-04}$	0,54 [T]	$7{,}08 \times 10^{-03}$	1,23 [T]
11	60472288	rs566353	*SLC15A3, i*	$7{,}41 \times 10^{-05}$	1,30	$3{,}16 \times 10^{-03}$	1,27 [T]	$7{,}16 \times 10^{-03}$	0,74 [T]
6	109858294	rs9480957	*PPIL6, i*	$8{,}89 \times 10^{-05}$	1,21	$1{,}26 \times 10^{-03}$	1,31 [A]	$1{,}18 \times 10^{-02}$	0,86 [A]

Tab. 3-12: Ergebnisse für die 33 SNPs mit den höchsten Rängen aus der *LD cluster ranking*-Analyse und der *gene ranking*-Analyse. Der erste Abschnitt der Tabelle zeigt die 25 Marker, die mit der *LD cluster ranking*-Analyse identifiziert wurden, der zweite Abschnitt die 8 Marker der *gene ranking*-Analyse. In den Spalten Sarkoidose und Tuberkulose sind die *p*-Werte und allelischen ORs der SNPs aus den GWAS-Datensätzen dargestellt. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen der GWAS-Daten. Abkürzungen: Chr. = Chromosom; e = exonisch; i = intronisch; OR = Quotenverhältnis (engl. *odds ratio*).

LD cluster ranking		**Sarkoidose**				**Tuberkulose**			
Chr.	**Gen**	**dbSNP ID**	**A1**	***p*-Wert**	**OR**	**dbSNP ID**	**A1**	***p*-Wert**	**OR**
6	intergenisch	rs6919855	C	$1{,}02 \times 10^{-03}$	0,75	rs9268880	T	$6{,}32 \times 10^{-05}$	1,27
12	intergenisch	rs1732581	T	$6{,}66 \times 10^{-04}$	0,75	rs1732581	T	$3{,}36 \times 10^{-04}$	2,21
14	intergenisch	rs11844637	C	$7{,}53 \times 10^{-04}$	0,77	rs12431541	A	$1{,}24 \times 10^{-03}$	1,22
5	intergenisch	rs6884494	A	$1{,}69 \times 10^{-04}$	1,76	rs6884494	A	$1{,}72 \times 10^{-03}$	0,81
10	*LOC100129055, i*	rs1740736	G	$2{,}72 \times 10^{-03}$	1,28	rs2505202	C	$2{,}74 \times 10^{-03}$	0,84
6	*HLA-DRA, e*	rs7194	G	$1{,}00 \times 10^{-05}$	0,70	rs7194	G	$2{,}83 \times 10^{-03}$	1,17
14	intergenisch	rs7148429	A	$1{,}82 \times 10^{-03}$	0,61	rs8007404	A	$2{,}89 \times 10^{-03}$	0,83
18	*SLC14A2, i*	rs11659929	C	$9{,}68 \times 10^{-04}$	0,76	rs11659929	C	$2{,}96 \times 10^{-03}$	0,85
2	intergenisch	rs778198	T	$2{,}68 \times 10^{-03}$	1,27	rs778193	T	$3{,}21 \times 10^{-03}$	1,22
1	intergenisch	rs4845741	A	$7{,}91 \times 10^{-04}$	0,70	rs1390488	C	$3{,}33 \times 10^{-03}$	0,85
12	intergenisch	rs12304940	C	$1{,}38 \times 10^{-03}$	1,42	rs2190733	C	$3{,}39 \times 10^{-03}$	1,19
1	*AGMAT, e* *DNAJC16, i*	rs11580170	T	$1{,}12 \times 10^{-03}$	1,35	rs11583306	T	$3{,}44 \times 10^{-03}$	1,17
2	intergenisch	rs828270	C	$1{,}61 \times 10^{-03}$	0,77	rs11126681	G	$3{,}72 \times 10^{-03}$	1,26
8	intergenisch	rs10505536	A	$9{,}48 \times 10^{-04}$	1,32	rs2099292	T	$3{,}83 \times 10^{-03}$	0,79
2	intergenisch	rs778199	T	$1{,}97 \times 10^{-03}$	1,29	rs778199	T	$4{,}54 \times 10^{-03}$	1,19
Gene ranking		**Sarkoidose**				**Tuberkulose**			
Chr.	**Gen**	**dbSNP ID**	**A1**	***p*-Wert**	**OR**	**dbSNP ID**	**A1**	***p*-Wert**	**OR**
10	*LIPA, i*	rs1960574	T	$8{,}05 \times 10^{-04}$	1,32	rs1029074	A	$5{,}33 \times 10^{-04}$	0,79
3	*IL17RD, i*	rs747088	T	$8{,}65 \times 10^{-04}$	0,64	rs17057784	A	$5{,}64 \times 10^{-05}$	0,80
5	*ADAMTS16, i*	rs1560064	A	$9{,}06 \times 10^{-04}$	0,69	rs2913633	T	$2{,}29 \times 10^{-04}$	0,82
17	*MYO1D, i*	rs2519866	C	$3{,}91 \times 10^{-04}$	1,44	rs225214	T	$3{,}65 \times 10^{-05}$	0,80

3.2.3 Erste Replikationsphase

Für die Bestätigung der Assoziation der durch die vier verschiedenen Analysen der GWAS-Datensätze identifizierten Kandidaten-SNPs standen die unabhängigen Stichproben B, E-I und E-II zur Verfügung. Die Sarkoidose-Stichprobe B umfasste nach Qualitätsfilterung (siehe Kapitel 2.2.5) 3.623 Individuen, hiervon 1.486 Sarkoidosepatienten. Nach der Qualitätskontrolle (Kapitel 2.2.5) enthielt die Tuberkulose-Stichprobe E-I 2.454 Individuen, davon 780 Tuberkulosepatienten.

Die 52 Kandidaten-SNPs wurden mit der Sequenom®-Technologie in den Stichproben B und E-I genotypisiert (Tab. 3-13 und Tab. 3-14). Wegen eines zu hohen Dimer-Potentials für den Vorwärts-Extensions-Primer des SNPs rs7148429 war es nicht möglich diesen SNPs zu genotypisieren und der SNP wurde von der Analyse ausgeschlossen. In der Stichprobe B betrug die Genotypisierungsrate (CR) 93,63%, und 49 SNPs wurden erfolgreich genotypisiert; in der Stichprobe E-I betrug die Genotypisierungsrate 99,83% und 48 SNPs wurden erfolgreich genotypisiert.

Tab. 3-13: Ergebnisse der ersten Replikationsphase für die Kandidaten-SNPs in der Sarkoidosestichprobe B. Nominell signifikante Ergebnisse (*p*-Wert < 0,05) sind fett hervorgehoben. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; Chr. = Chromosom; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; NA = nicht verfügbar (engl. *not available*); OR = Quotenverhältnis (engl. *odds ratio*).

Chr.	dbSNP ID	Position (bp)	A1	AF Fälle	AF Kont.	*p*-Wert	OR [95% CI]
1	rs11583306	15733390	T	0,2255	0,2447	$6{,}00 \times 10^{-02}$	0,90 [0,80-1,01]
1	rs11580170	15782331	T	0,2515	0,2692	$9{,}46 \times 10^{-02}$	0,91 [0,82-1,02]
1	rs11584687	22749827	G	0,4594	0,4593	$9{,}97 \times 10^{-01}$	1,00 [0,91-1,10]
1	rs2300955	57116750	T	0,1598	0,1769	$5{,}77 \times 10^{-02}$	0,88 [0,78-1,00]
1	rs4845741	150450098	A	0,1594	0,1515	$3{,}57 \times 10^{-01}$	1,06 [0,93-1,21]
1	rs1390488	150466124	T	0,1618	0,1524	$2{,}80 \times 10^{-01}$	1,07 [0,94-1,22]
2	rs11126681	26816882	G	0,421	0,4215	$9{,}70 \times 10^{-01}$	1,00 [0,91-1,10]
2	rs828270	26823613	C	0,4201	0,4187	$9{,}05 \times 10^{-01}$	1,01 [0,91-1,11]
2	rs778199	137290355	T	0,3514	0,349	$8{,}37 \times 10^{-01}$	1,01 [0,92-1,12]
2	rs778198	137312836	T	0,4308	0,4181	$2{,}83 \times 10^{-01}$	1,05 [0,96-1,16]
2	rs778193	137316119	A	0,4343	0,4234	$3{,}60 \times 10^{-01}$	1,05 [0,95-1,15]
2	rs10931252	187249212	T	NA	0,2049	NA	NA
3	rs17057784	57155442	A	0,1405	0,1503	$2{,}46 \times 10^{-01}$	0,92 [0,81-1,06]
3	rs6444661	194020854	T	0,4799	0,494	$2{,}41 \times 10^{-01}$	0,95 [0,86-1,04]
4	rs1471921	188322475	C	0,1579	0,1681	$2{,}50 \times 10^{-01}$	0,93 [0,82-1,05]
5	rs1560064	5242203	T	0,1475	0,1334	$9{,}08 \times 10^{-02}$	1,12 [0,98-1,29]
5	rs2913633	5289356	C	0,4406	0,4488	$4{,}97 \times 10^{-01}$	0,97 [0,88-1,06]
5	rs11948804	78839534	T	0,0914	0,08805	$6{,}25 \times 10^{-01}$	1,04 [0,88-1,23]
5	rs6884494	78853379	A	0,09614	0,09272	$6{,}26 \times 10^{-01}$	1,04 [0,89-1,22]
6	**rs8084**	**32519013**	**A**	**0,5332**	**0,4398**	$\mathbf{6{,}90 \times 10^{-15}}$	**1,46 [1,32-1,60]**
6	**rs6919855**	**32536989**	**C**	**0,391**	**0,3349**	$\mathbf{1{,}22 \times 10^{-06}}$	**1,28 [1,16-1,41]**
6	**rs9268880**	**32539336**	**T**	**0,3915**	**0,3371**	$\mathbf{2{,}45 \times 10^{-06}}$	**1,27 [1,15-1,40]**
6	rs9480957	109858294	A	0,3347	0,3299	$6{,}74 \times 10^{-01}$	1,02 [0,92-1,13]
7	rs4321884	22707984	A	0,5108	0,4888	$6{,}67 \times 10^{-02}$	1,09 [0,99-1,20]
8	rs2099292	130427095	A	0,3493	0,3365	$2{,}64 \times 10^{-01}$	1,06 [0,96-1,17]
8	rs10505536	130436462	A	0,3507	0,3359	$1{,}92 \times 10^{-01}$	1,07 [0,97-1,18]
9	rs3181200	116703705	A	0,4485	0,4648	$1{,}73 \times 10^{-01}$	0,94 [0,85-1,03]
9	rs3181374	116705008	C	0,4546	0,47	$1{,}98 \times 10^{-01}$	0,94 [0,86-1,03]
10	rs1548255	38457199	C	0,4236	0,4265	$8{,}03 \times 10^{-01}$	0,99 [0,90-1,09]
10	rs2505202	38511145	C	0,4709	0,4643	$5{,}83 \times 10^{-01}$	1,03 [0,93-1,13]
10	rs1740736	38538915	G	0,4733	0,4658	$5{,}32 \times 10^{-01}$	1,03 [0,94-1,13]
10	**rs745182**	**81891702**	**G**	**0,4705**	**0,4198**	$\mathbf{2{,}07 \times 10^{-05}}$	**1,23 [1,12-1,35]**
10	rs1029074	90989016	G	0,2242	0,2287	$6{,}55 \times 10^{-01}$	0,97 [0,87-1,09]
10	rs1960574	91014005	A	0,4109	0,4082	$8{,}19 \times 10^{-01}$	1,01 [0,92-1,11]
10	**rs758362**	**118447895**	**T**	**0,304**	**0,2808**	$\mathbf{3{,}32 \times 10^{-02}}$	**1,12 [1,01-1,24]**
11	rs566353	60472288	A	0,4597	0,4482	$3{,}34 \times 10^{-01}$	1,05 [0,95-1,15]
11	rs2621483	115566825	A	0,1229	0,121	$8{,}12 \times 10^{-01}$	1,02 [0,88-1,18]
12	**rs2190733**	**66729518**	**C**	**0,1562**	**0,1832**	$\mathbf{2{,}90 \times 10^{-03}}$	**0,83 [0,73-0,94]**
12	**rs12304940**	**66744148**	**C**	**0,1562**	**0,1816**	$\mathbf{4{,}84 \times 10^{-03}}$	**0,83 [0,73-0,95]**
12	rs1732581	114202449	T	0,382	0,3599	$5{,}61 \times 10^{-02}$	1,10 [1,00-1,21]
12	rs10850905	116741236	G	0,311	0,3005	$3{,}44 \times 10^{-01}$	1,05 [0,95-1,16]
13	rs9527209	32681621	A	0,1719	0,1662	$5{,}28 \times 10^{-01}$	1,04 [0,92-1,18]
14	rs12431541	29090267	A	0,4638	0,4514	$3{,}02 \times 10^{-01}$	1,05 [0,96-1,16]
14	rs11844637	29093372	C	0,4664	0,4532	$2{,}69 \times 10^{-01}$	1,06 [0,96-1,16]
14	rs8007404	58524255	A	0,05593	0,05302	$5{,}92 \times 10^{-01}$	1,06 [0,86-1,30]
17	rs2519866	27859883	G	0,3206	0,3138	$5{,}44 \times 10^{-01}$	1,03 [0,93-1,14]
17	rs225214	27920869	G	0,3977	0,3758	$6{,}13 \times 10^{-02}$	1,10 [1,00-1,21]
17	rs8080480	73350621	T	0,3555	0,3381	$1{,}29 \times 10^{-01}$	1,08 [0,98-1,19]
17	rs9915508	75845351	T	0,4956	0,4971	$8{,}98 \times 10^{-01}$	0,99 [0,90-1,09]
18	rs11659929	41186112	C	0,3378	0,3337	$7{,}23 \times 10^{-01}$	1,02 [0,92-1,13]

Tab. 3-14: Ergebnisse der ersten Replikationsphase für die Kandidaten-SNPs in der Tuberkulosestichprobe E-I. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; Chr. = Chromosom; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; OR = Quotenverhältnis (engl. *odds ratio*).

Chr.	dbSNP ID	Position (bp)	A1	AF Fälle	AF Kont.	*p*-Wert	OR [95% CI]
1	rs11583306	15733390	T	0,4299	0,4353	$7{,}19 \times 10^{-01}$	0,98 [0,87-1,10]
1	rs11580170	15782331	C	0,4472	0,4292	$2{,}36 \times 10^{-01}$	1,08 [0,95-1,22]
1	rs11584687	22749827	G	0,3901	0,3997	$5{,}23 \times 10^{-01}$	0,96 [0,85-1,09]
1	rs2300955	57116750	T	0,1391	0,1329	$5{,}55 \times 10^{-01}$	1,05 [0,89-1,26]
1	rs4845741	150450098	G	0,5026	0,4916	$4{,}76 \times 10^{-01}$	1,05 [0,93-1,18]
1	rs1390488	150466124	C	0,5006	0,4809	$1{,}98 \times 10^{-01}$	1,08 [0,96-1,22]
2	rs11126681	26816882	G	0,1207	0,1373	$1{,}10 \times 10^{-01}$	0,86 [0,72-1,03]
2	rs828270	26823613	G	0,2099	0,2158	$6{,}39 \times 10^{-01}$	0,97 [0,83-1,12]
2	rs778199	137290355	T	0,2327	0,2507	$1{,}71 \times 10^{-01}$	0,91 [0,79-1,04]
2	rs778198	137312836	T	0,2683	0,2867	$1{,}84 \times 10^{-01}$	0,91 [0,80-1,04]
2	rs778193	137316119	A	0,2673	0,2858	$1{,}78 \times 10^{-01}$	0,91 [0,80-1,04]
3	rs17057784	57155442	A	0,3615	0,3751	$3{,}61 \times 10^{-01}$	0,94 [0,83-1,07]
3	rs6444661	194020854	T	0,1842	0,1801	$7{,}29 \times 10^{-01}$	1,03 [0,88-1,20]
4	rs1471921	188322475	T	0,3113	0,32	$5{,}43 \times 10^{-01}$	0,96 [0,84-1,09]
5	rs1560064	5242203	T	0,2276	0,2427	$2{,}47 \times 10^{-01}$	0,92 [0,80-1,06]
5	rs2913633	5289356	T	0,3723	0,3758	$8{,}11 \times 10^{-01}$	0,98 [0,87-1,12]
5	rs11948804	78839534	T	0,2026	0,189	$2{,}62 \times 10^{-01}$	1,09 [0,94-1,27]
5	rs6884494	78853379	A	0,3766	0,3731	$8{,}15 \times 10^{-01}$	1,02 [0,90-1,15]
6	rs8084	32519013	A	0,4474	0,4464	$9{,}51 \times 10^{-01}$	1,00 [0,89-1,13]
6	rs6919855	32536989	C	0,316	0,3113	$7{,}45 \times 10^{-01}$	1,02 [0,90-1,16]
6	rs9268880	32539336	T	0,2968	0,2921	$7{,}37 \times 10^{-01}$	1,02 [0,90-1,17]
6	rs9480957	109858294	A	0,2693	0,2772	$5{,}64 \times 10^{-01}$	0,96 [0,84-1,10]
7	rs4321884	22707984	A	0,3288	0,3412	$3{,}94 \times 10^{-01}$	0,95 [0,83-1,08]
8	rs2099292	130427095	A	0,09949	0,1077	$3{,}82 \times 10^{-01}$	0,92 [0,75-1,12]
8	rs10505536	130436462	A	0,113	0,1188	$5{,}55 \times 10^{-01}$	0,94 [0,78-1,14]
9	rs3181200	116703705	C	0,1354	0,1398	$6{,}81 \times 10^{-01}$	0,96 [0,81-1,15]
9	rs3181374	116705008	T	0,133	0,139	$5{,}72 \times 10^{-01}$	0,95 [0,80-1,13]
10	rs1548255	38457199	C	0,4286	0,4168	$4{,}38 \times 10^{-01}$	1,05 [0,93-1,19]
10	rs2505202	38511145	T	0,4852	0,4979	$4{,}09 \times 10^{-01}$	0,95 [0,84-1,07]
10	rs1740736	38538915	A	0,4814	0,4919	$4{,}92 \times 10^{-01}$	0,96 [0,85-1,08]
10	rs745182	81891702	A	0,5148	0,4883	$8{,}48 \times 10^{-02}$	1,11 [0,99-1,25]
10	rs1960574	91014005	A	0,3944	0,3703	$1{,}07 \times 10^{-01}$	1,11 [0,98-1,25]
10	rs758362	118447895	T	0,4858	0,4823	$8{,}21 \times 10^{-01}$	1,01 [0,90-1,14]
11	rs566353	60472288	A	0,07885	0,07527	$6{,}60 \times 10^{-01}$	1,05 [0,84-1,32]
11	rs2621483	115566825	A	0,2015	0,1812	$8{,}98 \times 10^{-02}$	1,14 [0,98-1,33]
12	rs2190733	66729518	C	0,2724	0,2838	$4{,}11 \times 10^{-01}$	0,95 [0,83-1,08]
12	rs12304940	66744148	C	0,2619	0,2726	$4{,}34 \times 10^{-01}$	0,95 [0,83-1,09]
12	rs1732581	114202449	T	0,01733	0,02002	$5{,}22 \times 10^{-01}$	0,86 [0,55-1,36]
12	rs10850905	116741236	G	0,1635	0,1626	$9{,}38 \times 10^{-01}$	1,01 [0,86-1,18]
13	rs9527209	32681621	A	0,01348	0,01884	$1{,}78 \times 10^{-01}$	0,71 [0,43-1,17]
14	rs12431541	29090267	A	0,2263	0,2191	$5{,}71 \times 10^{-01}$	1,04 [0,90-1,20]
14	rs11844637	29093372	C	0,3729	0,3526	$1{,}67 \times 10^{-01}$	1,09 [0,96-1,24]
14	rs8007404	58524255	A	0,2788	0,292	$3{,}43 \times 10^{-01}$	0,94 [0,82-1,07]
17	rs2519866	27859883	G	0,4711	0,467	$7{,}90 \times 10^{-01}$	1,02 [0,90-1,15]
17	rs225214	27920869	A	0,4273	0,4088	$2{,}22 \times 10^{-01}$	1,08 [0,96-1,22]
17	rs8080480	73350621	C	0,4589	0,4596	$9{,}60 \times 10^{-01}$	1,00 [0,88-1,13]
17	rs9915508	75845351	G	0,2513	0,2533	$8{,}80 \times 10^{-01}$	0,99 [0,86-1,14]
18	rs11659929	41186112	C	0,4537	0,456	$8{,}84 \times 10^{-01}$	0,99 [0,88-1,12]

In der Stichprobe E-I zeigte keiner der getesteten Marker eine nominelle Assoziation, daher wurde die Teststärke der Stichprobe mit einem hypothetischen SNP berechnet. Für einen SNP mit einem OR von 1,3 und einer MAF von 0,2 in den Kontrollen besteht in der gesamten Stichprobe E-I eine

statistische Teststärke von 71,6%, um eine Assoziation von $p = 0,05$ nachzuweisen. Die Stichprobe E-I verfügt damit über keine ausreichende Teststärke, um die in der GWAS beobachteten Assoziationen zu detektieren.

Für jeden Marker wurde nun eine Metaanalyse aller GWAS-Ergebnisse und der Ergebnisse der ersten Replikationsphase durchgeführt. Um falsch-positive Ergebnisse weitestgehend ausschließen zu können, wurde ein konservativer Schwellwert von einem Gesamtassoziationssignal (p_{META}) < 5 x 10^{-5} für die Markerauswahl der zweiten Replikationsphase festgelegt. Von den 48 erfolgreich genotypisieren SNPs zeigten sieben Marker ein Assoziationssignal von $p_{META} < 5 \times 10^{-5}$ (Tab. 3-16).

Von diesen sieben Markern erfüllten stromabwärts des Interferon-Gamma-Gens (*IFNG*) zwei SNPs (rs2190733 und rs12304940) die gesetzten Kriterien, um in der zweiten Replikationsphase genotypisiert zu werden. Da die SNPs jedoch in beiden Krankheiten (und damit auch Populationen) im starken LD zueinander standen (r^2 = 0,95 in Tuberkulose und r^2 = 0,989 in Sarkoidose) und der Marker rs2190733 ein besseres Ergebnis in der ersten Replikationsphase zeigte (Tab. 3-13 und Tab. 3-14), wurde nur der SNP rs2190733 für die nachfolgenden Replikationen ausgewählt wegen des besseren Assoziationsergebnisses (rs2190733: Sarkoidose $p = 2,90 \times 10^{-03}$; Tuberkulose $p = 4,11 \times 10^{-01}$; rs12304940: Sarkoidose $p = 4,84 \times 10^{-03}$; Tuberkulose $p = 4,34 \times 10^{-01}$).

Drei Marker in der HLA-Region zeigten im Rahmen der Analysen eine Assoziation mit Tuberkulose und Sarkoidose. Für beide Krankheiten sind Assoziationen mit Varianten in der HLA-Region bekannt (Adrianto et al. 2012; Foley et al. 2001; Levin et al. 2013; Magira et al. 2012; Rossman et al. 2003; Suzuki et al. 2012; Y. Wang et al. 2012), daher wurde von diesen drei SNPs nur der am stärksten assoziierte SNP (rs8084) für die zweite Replikationsphase ausgesucht. Insgesamt wurden also vier der sieben Marker für die Weiterverfolgung mittels Genotypisierung in den weiteren Replikationsstichproben ausgewählt.

Eine genauere Betrachtung der Ergebnisse zeigte eine zusätzliche, nur für Tuberkulose bestehende starke Assoziation eines SNPs (rs225214) mit der Krankheit ($p_{GWAS} = 3,65 \times 10^{-5}$ und $p_{Repl.I} = 2,22 \times 10^{-1}$). Die Auswahl dieses SNPs für weitere Untersuchungen folgt nicht den oben erwähnten Selektionskriterien (kombinierter *p*-Wert von < 5 x 10^{-5} in der Metaanalyse der Stichproben A, B, D und E-I), eine Metaanalyse der Tuberkulose-GWAS-Daten und der Daten der ersten Replikationsphase zeigte jedoch, dass dieser SNP die stärkste Assoziation aller 50 Marker mit Tuberkulose zeigte (rs225214, $p = 8,33 \times 10^{-5}$) (Tab. 3-15). Der SNP unterschritt als einziger Marker den Wert von $p < 10^{-4}$ in der Metaanalyse der Tuberkulosedaten. Für die zweite Replikationsphase wurde der SNP rs225214 als mögliche nur die Tuberkulose betreffende Assoziation nachverfolgt.

Tab. 3-15: Ergebnisse des GWAS und der ersten Replikationsphase für die Kandidaten-SNPs bei der Tuberkulose. Die Spalte Analyse bezeichnet die Analyseform, mit der der jeweilige Marker identifiziert wurde. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Die Spalte p_{META} zeigt den kombinierten *p*-Wert der Metaanalyse von der GWAS- und der ersten Replikationsphase. SNPs mit einem kombinierten *p*-Wert < 10^{-4} sind fett hervorgehoben. Abkürzungen: Chr. = Chromosom; e = exonisch; G.-rank = *gene ranking*-Analyse; i = intronisch; LD-rank = *LD cluster ranking*-Analyse; Meta. f. = *meta-analysis fixed effects*; Meta. o. = *meta-analysis opposite effects*; Repl.I = erste Replikationsphase.

Chr.	dbSNP ID	Position (bp)	Gen	Analyse	A1	p_{GWAS}	$p_{Repl.I}$	p_{META}
1	rs11583306	15733390	*DNAJC16*, *i*	LD-rank.	T	$3{,}44 \times 10^{-03}$	$7{,}19 \times 10^{-01}$	$1{,}43 \times 10^{-02}$
1	rs11580170	15782331	*AGMAT*, *e*	LD-rank.	C	$4{,}31 \times 10^{-02}$	$2{,}36 \times 10^{-01}$	$2{,}15 \times 10^{-02}$
1	rs11584687	22749827	intergenisch	Meta. f.	G	$1{,}10 \times 10^{-03}$	$5{,}23 \times 10^{-01}$	$3{,}73 \times 10^{-03}$
1	rs2300955	57116750	*C8A*, *i*	Meta. f.	A	$5{,}64 \times 10^{-03}$	$5{,}55 \times 10^{-01}$	$9{,}28 \times 10^{-02}$
1	rs4845741	150450098	intergenisch	LD-rank.	G	$1{,}61 \times 10^{-01}$	$4{,}76 \times 10^{-01}$	$1{,}30 \times 10^{-01}$
1	rs1390488	150466124	intergenisch	LD-rank.	C	$3{,}33 \times 10^{-03}$	$1{,}98 \times 10^{-01}$	$2{,}44 \times 10^{-03}$
2	rs11126681	26816882	intergenisch	LD-rank.	G	$3{,}72 \times 10^{-03}$	$1{,}10 \times 10^{-01}$	$1{,}20 \times 10^{-03}$
2	rs828270	26823613	intergenisch	LD-rank.	G	$5{,}99 \times 10^{-01}$	$6{,}39 \times 10^{-01}$	$9{,}20 \times 10^{-01}$
2	rs778199	137290355	intergenisch	LD-rank.	T	$4{,}54 \times 10^{-03}$	$1{,}71 \times 10^{-01}$	$2{,}27 \times 10^{-03}$
2	rs778198	137312836	intergenisch	LD-rank.	T	$7{,}45 \times 10^{-03}$	$1{,}84 \times 10^{-01}$	$3{,}83 \times 10^{-03}$
2	rs778193	137316119	intergenisch	LD-rank.	T	$3{,}21 \times 10^{-03}$	$1{,}78 \times 10^{-01}$	$2{,}39 \times 10^{-03}$
3	rs17057784	57155442	*IL17RD, i*	G.-rank.	A	$5{,}64 \times 10^{-05}$	$3{,}61 \times 10^{-01}$	$1{,}48 \times 10^{-02}$
3	rs6444661	194020854	*MB21D2*, *i*	Meta. o.	T	$5{,}75 \times 10^{-03}$	$7{,}29 \times 10^{-01}$	$5{,}50 \times 10^{-02}$
4	rs1471921	188322475	intergenisch	Meta. f.	T	$1{,}04 \times 10^{-03}$	$5{,}43 \times 10^{-01}$	$3{,}48 \times 10^{-02}$
5	rs1560064	5242203	*ADAMTS16, i*	G.-rank.	A	$1{,}94 \times 10^{-02}$	$2{,}47 \times 10^{-01}$	$1{,}12 \times 10^{-02}$
5	rs2913633	5289356	*ADAMTS16, i*	G.-rank.	T	$2{,}29 \times 10^{-04}$	$8{,}11 \times 10^{-01}$	$8{,}73 \times 10^{-03}$
5	rs11948804	78839534	*HOM x 10R1*, *i*	Meta. o.	T	$7{,}08 \times 10^{-03}$	$2{,}62 \times 10^{-01}$	$2{,}76 \times 10^{-01}$
5	rs6884494	78853379	intergenisch	LD-rank.	A	$1{,}72 \times 10^{-03}$	$8{,}15 \times 10^{-01}$	$5{,}03 \times 10^{-02}$
6	rs8084	32519013	*HLA-DRA*, *e*	Meta. o.	A	$8{,}48 \times 10^{-03}$	$9{,}51 \times 10^{-01}$	$4{,}48 \times 10^{-02}$
6	rs6919855	32536989	intergenisch	LD-rank.	C	$2{,}03 \times 10^{-04}$	$7{,}45 \times 10^{-01}$	$2{,}82 \times 10^{-03}$
6	rs9268880	32539336	intergenisch	LD-rank.	T	$6{,}32 \times 10^{-05}$	$7{,}37 \times 10^{-01}$	$1{,}31 \times 10^{-03}$
6	rs9480957	109858294	*PPIL6*, *i*	Meta. o.	A	$1{,}18 \times 10^{-02}$	$5{,}64 \times 10^{-01}$	$2{,}19 \times 10^{-02}$
7	rs4321884	22707984	intergenisch	Meta. o.	A	$1{,}92 \times 10^{-03}$	$3{,}94 \times 10^{-01}$	$3{,}66 \times 10^{-03}$
8	rs2099292	130427095	intergenisch	LD-rank.	T	$3{,}83 \times 10^{-03}$	$3{,}82 \times 10^{-01}$	$8{,}94 \times 10^{-02}$
8	rs10505536	130436462	intergenisch	LD-rank.	A	$5{,}43 \times 10^{-03}$	$5{,}55 \times 10^{-01}$	$1{,}33 \times 10^{-01}$
9	rs3181200	116703705	*TNFSF8*, *e*	Meta. f.	G	$1{,}99 \times 10^{-03}$	$6{,}81 \times 10^{-01}$	$3{,}48 \times 10^{-02}$
9	rs3181374	116705008	*TNFSF8*, *e*	Meta. f.	A	$5{,}77 \times 10^{-03}$	$5{,}72 \times 10^{-01}$	$7{,}78 \times 10^{-02}$
10	rs1548255	38457199	intergenisch	Meta. f.	C	$5{,}42 \times 10^{-03}$	$4{,}38 \times 10^{-01}$	$1{,}45 \times 10^{-01}$
10	rs2505202	38511145	*LOC100129055*, *i*	LD-rank.	C	$2{,}74 \times 10^{-03}$	$4{,}09 \times 10^{-01}$	$1{,}11 \times 10^{-01}$
10	rs1740736	38538915	*LOC100129055*, *i*	LD-rank.	G	$3{,}80 \times 10^{-03}$	$4{,}92 \times 10^{-01}$	$1{,}06 \times 10^{-01}$
10	rs745182	81891702	*PLAC9*, *i*	Meta. o.	T	$9{,}61 \times 10^{-03}$	$8{,}48 \times 10^{-02}$	$2{,}25 \times 10^{-03}$
10	rs1960574	91014005	*LIPA, i*	G.-rank.	T	$7{,}14 \times 10^{-01}$	$1{,}07 \times 10^{-01}$	$1{,}86 \times 10^{-01}$
10	rs758362	118447895	*HSPA12A*, *i*	Meta. f.	A	$2{,}48 \times 10^{-03}$	$8{,}21 \times 10^{-01}$	$1{,}87 \times 10^{-02}$
11	rs566353	60472288	*SLC15A3*, *i*	Meta. o.	T	$7{,}16 \times 10^{-03}$	$6{,}60 \times 10^{-01}$	$2{,}60 \times 10^{-02}$
11	rs2621483	115566825	intergenisch	Meta. f.	A	$5{,}12 \times 10^{-04}$	$8{,}98 \times 10^{-02}$	$3{,}14 \times 10^{-04}$
12	rs2190733	66729518	intergenisch	LD-rank.	C	$3{,}39 \times 10^{-03}$	$4{,}11 \times 10^{-01}$	$5{,}75 \times 10^{-03}$
12	rs12304940	66744148	intergenisch	LD-rank.	C	$8{,}40 \times 10^{-03}$	$4{,}34 \times 10^{-01}$	$2{,}09 \times 10^{-02}$
12	rs1732581	114202449	intergenisch	Meta. f.	T	$3{,}36 \times 10^{-04}$	$5{,}22 \times 10^{-01}$	$2{,}46 \times 10^{-03}$
12	rs10850905	116741236	*KSR2*, *i*	Meta. f.	G	$2{,}82 \times 10^{-03}$	$9{,}38 \times 10^{-01}$	$7{,}31 \times 10^{-02}$
13	rs9527209	32681621	*STARD13*, *i*	Meta. f.	A	$2{,}61 \times 10^{-03}$	$1{,}78 \times 10^{-01}$	$1{,}65 \times 10^{-03}$
14	rs12431541	29090267	intergenisch	LD-rank.	A	$1{,}24 \times 10^{-03}$	$5{,}71 \times 10^{-01}$	$4{,}53 \times 10^{-03}$
14	rs11844637	29093372	intergenisch	LD-rank.	C	$1{,}16 \times 10^{-02}$	$1{,}67 \times 10^{-01}$	$4{,}82 \times 10^{-03}$
14	rs8007404	58524255	intergenisch	LD-rank.	A	$2{,}89 \times 10^{-03}$	$3{,}43 \times 10^{-01}$	$1{,}25 \times 10^{-01}$
17	**rs225214**	**27920869**	***MYO1D, i***	**G.-rank.**	**T**	$\mathbf{3{,}65 \times 10^{-05}}$	$\mathbf{2{,}22 \times 10^{-01}}$	$\mathbf{8{,}33 \times 10^{-05}}$
17	rs8080480	73350621	intergenisch	Meta. f.	C	$8{,}61 \times 10^{-03}$	$9{,}60 \times 10^{-01}$	$4{,}26 \times 10^{-02}$
17	rs9915508	75845351	intergenisch	Meta. f.	G	$2{,}26 \times 10^{-03}$	$8{,}80 \times 10^{-01}$	$6{,}00 \times 10^{-02}$
18	rs11659929	41186112	*SLC14A2*, *i*	Meta. o.	C	$2{,}96 \times 10^{-03}$	$8{,}84 \times 10^{-01}$	$3{,}11 \times 10^{-02}$

3.2.3.1 Genotypisierung von Kandidaten-SNPs in zusätzlichen ghanaischen Kontrollpersonen

Die nach der Analyse der ersten Replikationsphase ausgewählten vier gemeinsamen Kandidaten-SNPs und der eine tuberkulosespezifische SNP wurden zusätzlich in der Stichprobe E-II mit der TaqMan®-Technologie genotypisiert. Die zusätzlichen Kontrollen aus Ghana wurden erst sehr spät zur Verfügung gestellt, so dass nur noch die fünf Kandidaten-SNPs in dieser Stichprobe genotypisiert werden konnten. Nach Qualitätskontrolle (Kapitel 2.2.5) konnten 925 Kontrollpersonen der Stichprobe erfolgreich genotypisiert werden. Die Ergebnisse der zusätzlichen Genotypisierung sind in der Tabelle 3-16 mit einem Sternchen versehen. Die Ergebnisse der ersten Replikationsphase ohne die zusätzliche Stichprobe sind in den Tabellen 3-14 und 3-15 zu finden. Die Ergänzung der Replikationsstichprobe E-I mit weiteren Kontrollpersonen (E-II) verbesserte die Teststärke der Stichprobe bei dem hypothetischen SNP (OR = 1,3; MAF = 0,2) von 71,6% auf 76,8%, um eine Assoziation von p = 0,05 nachzuweisen. Durch die Ergänzung der Stichprobe zeigte der tuberkulosespezifische SNP rs225214 eine nominell signifikante Assoziation ($p = 2 \times 10^{-3}$). Bei den anderen Kandidaten-SNPs konnten trotz der zusätzlichen Kontrollen keine nominell signifikanten Assoziationen beobachtet werden (Tab. 3-16).

Tab. 3-16: Ergebnisse der GWAS-Phase und der ersten Replikationsphase, sowie kombinierte *p*-Werte für die entsprechenden Phasen der beiden Krankheiten. Die Spalte Analyse bezeichnet die Analyseform, mit der der jeweilige Marker identifiziert wurde. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Die Spalte SA + TB GWAS zeigt den kombinierten *p*-Wert aus den beiden GWAS-Datensätzen und die Spalte SA +TB Repl.I den kombinierten *p*-Wert der beiden Replikationsstichproben. Die *p*-Werte der Tuberkulosereplikation, die mit einem Sternchen versehen sind, beinhalten die zusätzlichen ghanaischen Kontrollpersonen (Stichprobe E-II). Die Spalte Meta zeigt den kombinierten *p*-Wert der Metaanalyse von den GWAS- und den ersten Replikationsphasen. Die Kandidaten-SNPs sind fett und der TB-spezifische SNP fett und kursiv hervorgehoben. Abkürzungen: Chr. = Chromosom; e = exonisch; G.-rank = *gene ranking*-Analyse; i = intronisch; LD-rank = *LD cluster ranking*-Analyse; Meta. f. = *meta-analysis fixed effects*; Meta. o. = *meta-analysis opposite effects*; NA = nicht verfügbar (engl. *not available*); Repl.I = erste Replikationsphase; SA = Sarkoidose; TB = Tuberkulose.

						SA GWAS	TB GWAS	SA + TB GWAS	SA Repl.I	TB Repl.I	SA + TB Repl.I	Meta
Chr.	**dbSNP ID**	**Position (bp)**	**Gen**	**Analyse**	**A1**	p_{GWAS}	p_{GWAS}	$p_{komb.}$	p_{Repl}	p_{Repl}	$p_{komb.}$	p_{META}
1	rs11583306	15733390	*DNAJC16, i*	LD-rank.	T	$1{,}54 \times 10^{-03}$	$3{,}44 \times 10^{-03}$	$4{,}31 \times 10^{-05}$	$6{,}00 \times 10^{-02}$	$7{,}19 \times 10^{-01}$	$1{,}03 \times 10^{-01}$	$7{,}96 \times 10^{-05}$
1	rs11580170	15782331	*AGMAT, e*	LD-rank.	T	$1{,}12 \times 10^{-03}$	$4{,}31 \times 10^{-02}$	$6{,}62 \times 10^{-04}$	$9{,}46 \times 10^{-02}$	$2{,}36 \times 10^{-01}$	$4{,}17 \times 10^{-02}$	$1{,}64 \times 10^{-04}$
1	rs11584687	22749827	intergenisch	Meta. f.	G	$3{,}59 \times 10^{-03}$	$1{,}10 \times 10^{-03}$	$1{,}52 \times 10^{-05}$	$9{,}97 \times 10^{-01}$	$5{,}23 \times 10^{-01}$	$6{,}97 \times 10^{-01}$	$1{,}86 \times 10^{-03}$
1	rs2300955	57116750	*C8A, i*	Meta. f.	T	$3{,}03 \times 10^{-03}$	$5{,}64 \times 10^{-03}$	$6{,}88 \times 10^{-05}$	$5{,}77 \times 10^{-02}$	$5{,}55 \times 10^{-01}$	$2{,}34 \times 10^{-01}$	$5{,}87 \times 10^{-04}$
1	rs4845741	150450098	intergenisch	LD-rank.	A	$7{,}91 \times 10^{-04}$	$1{,}61 \times 10^{-01}$	$7{,}38 \times 10^{-01}$	$3{,}57 \times 10^{-01}$	$4{,}76 \times 10^{-01}$	$9{,}18 \times 10^{-01}$	$7{,}65 \times 10^{-01}$
1	rs1390488	150466124	intergenisch	LD-rank.	T	$1{,}20 \times 10^{-03}$	$3{,}33 \times 10^{-03}$	$4{,}15 \times 10^{-05}$	$2{,}80 \times 10^{-01}$	$1{,}98 \times 10^{-01}$	$9{,}33 \times 10^{-02}$	$6{,}64 \times 10^{-05}$
2	rs11126681	26816882	intergenisch	LD-rank.	G	$1{,}77 \times 10^{-03}$	$3{,}72 \times 10^{-03}$	$2{,}07 \times 10^{-05}$	$9{,}70 \times 10^{-01}$	$1{,}10 \times 10^{-01}$	$4{,}37 \times 10^{-01}$	$1{,}46 \times 10^{-03}$
2	rs828270	26823613	intergenisch	LD-rank.	C	$1{,}61 \times 10^{-03}$	$5{,}99 \times 10^{-01}$	$1{,}94 \times 10^{-02}$	$9{,}05 \times 10^{-01}$	$6{,}39 \times 10^{-01}$	$7{,}20 \times 10^{-01}$	$2{,}30 \times 10^{-01}$
2	rs778199	137290355	intergenisch	LD-rank.	T	$1{,}97 \times 10^{-03}$	$4{,}54 \times 10^{-03}$	$3{,}75 \times 10^{-05}$	$8{,}37 \times 10^{-01}$	$1{,}71 \times 10^{-01}$	$3{,}44 \times 10^{-01}$	$7{,}25 \times 10^{-04}$
2	rs778198	137312836	intergenisch	LD-rank.	T	$2{,}68 \times 10^{-03}$	$7{,}45 \times 10^{-03}$	$7{,}34 \times 10^{-01}$	$2{,}83 \times 10^{-01}$	$1{,}84 \times 10^{-01}$	$9{,}06 \times 10^{-01}$	$8{,}99 \times 10^{-01}$
2	rs778193	137316119	intergenisch	LD-rank.	A	$4{,}11 \times 10^{-03}$	$3{,}21 \times 10^{-03}$	$4{,}05 \times 10^{-05}$	$3{,}60 \times 10^{-01}$	$1{,}78 \times 10^{-01}$	$1{,}29 \times 10^{-01}$	$2{,}20 \times 10^{-04}$
2	rs10931252	187249212	*ITGAV, i*	Meta. f.	T	$5{,}76 \times 10^{-03}$	$2{,}65 \times 10^{-03}$	$4{,}85 \times 10^{-05}$	NA	NA	NA	NA
3	rs17057784	57155442	*IL17RD, i*	G.-rank.	A	$6{,}33 \times 10^{-01}$	$5{,}64 \times 10^{-05}$	$7{,}26 \times 10^{-04}$	$2{,}46 \times 10^{-01}$	$3{,}61 \times 10^{-01}$	$1{,}45 \times 10^{-01}$	$2{,}09 \times 10^{-01}$
3	rs6444661	194020854	*MB21D2, i*	Meta. o.	T	$3{,}23 \times 10^{-03}$	$5{,}75 \times 10^{-03}$	$5{,}77 \times 10^{-05}$	$2{,}41 \times 10^{-01}$	$7{,}29 \times 10^{-01}$	$2{,}35 \times 10^{-01}$	$1{,}84 \times 10^{-03}$
4	rs1471921	188322475	intergenisch	Meta. f.	C	$6{,}28 \times 10^{-03}$	$1{,}04 \times 10^{-03}$	$3{,}09 \times 10^{-05}$	$2{,}50 \times 10^{-01}$	$5{,}43 \times 10^{-01}$	$6{,}95 \times 10^{-01}$	$1{,}75 \times 10^{-03}$
5	rs1560064	5242203	*ADAMTS16, i*	G.-rank.	T	$9{,}06 \times 10^{-04}$	$1{,}94 \times 10^{-02}$	$6{,}43 \times 10^{-01}$	$9{,}08 \times 10^{-02}$	$2{,}47 \times 10^{-01}$	$6{,}66 \times 10^{-01}$	$9{,}99 \times 10^{-01}$
5	rs2913633	5289356	*ADAMTS16, i*	G.-rank.	C	$6{,}14 \times 10^{-02}$	$2{,}29 \times 10^{-04}$	$4{,}89 \times 10^{-02}$	$4{,}97 \times 10^{-01}$	$8{,}11 \times 10^{-01}$	$6{,}92 \times 10^{-01}$	$1{,}15 \times 10^{-01}$
5	rs11948804	78839534	*HOMER1, i*	Meta. o.	T	$2{,}14 \times 10^{-04}$	$7{,}08 \times 10^{-03}$	$6{,}04 \times 10^{-05}$	$6{,}25 \times 10^{-01}$	$2{,}62 \times 10^{-01}$	$2{,}47 \times 10^{-01}$	$1{,}08 \times 10^{-01}$
5	rs6884494	78853379	intergenisch	LD-rank.	A	$1{,}69 \times 10^{-04}$	$1{,}72 \times 10^{-03}$	$1{,}03 \times 10^{-05}$	$6{,}26 \times 10^{-01}$	$8{,}15 \times 10^{-01}$	$6{,}27 \times 10^{-01}$	$1{,}71 \times 10^{-02}$
6	**rs8084**	32519013	***HLA-DRA, e***	**Meta. o.**	**A**	$\mathbf{3{,}44 \times 10^{-05}}$	$\mathbf{8{,}48 \times 10^{-03}}$	$\mathbf{7{,}19 \times 10^{-06}}$	$\mathbf{6{,}90 \times 10^{-15}}$	$\mathbf{6{,}47 \times 10^{-01}}$*	$\mathbf{3{,}47 \times 10^{-10}}$	$\mathbf{1{,}35 \times 10^{-14}}$
6	rs6919855	32536989	intergenisch	LD-rank.	C	$1{,}02 \times 10^{-03}$	$2{,}03 \times 10^{-04}$	$8{,}52 \times 10^{-07}$	$1{,}22 \times 10^{-06}$	$7{,}45 \times 10^{-01}$	$4{,}68 \times 10^{-05}$	$3{,}71 \times 10^{-10}$
6	rs9268880	32539336	intergenisch	LD-rank.	T	$1{,}37 \times 10^{-03}$	$6{,}32 \times 10^{-05}$	$3{,}15 \times 10^{-07}$	$2{,}45 \times 10^{-06}$	$7{,}37 \times 10^{-01}$	$6{,}84 \times 10^{-05}$	$2{,}64 \times 10^{-10}$
6	rs9480957	109858294	*PPIL6, i*	Meta. o.	A	$1{,}26 \times 10^{-03}$	$1{,}18 \times 10^{-02}$	$8{,}89 \times 10^{-05}$	$6{,}74 \times 10^{-01}$	$5{,}64 \times 10^{-01}$	$4{,}90 \times 10^{-01}$	$2{,}16 \times 10^{-03}$
7	**rs4321884**	22707984	**intergenisch**	**Meta. o.**	**A**	$\mathbf{5{,}01 \times 10^{-03}}$	$\mathbf{1{,}92 \times 10^{-03}}$	$\mathbf{3{,}29 \times 10^{-05}}$	$\mathbf{6{,}67 \times 10^{-02}}$	$\mathbf{4{,}42 \times 10^{-01}}$*	$\mathbf{5{,}53 \times 10^{-02}}$	$\mathbf{3{,}79 \times 10^{-05}}$
8	rs2099292	130427095	intergenisch	LD-rank.	A	$1{,}21 \times 10^{-03}$	$3{,}83 \times 10^{-03}$	$1{,}48 \times 10^{-05}$	$2{,}64 \times 10^{-01}$	$3{,}82 \times 10^{-01}$	$5{,}44 \times 10^{-01}$	$1{,}71 \times 10^{-03}$
8	rs10505536	130436462	intergenisch	LD-rank.	A	$9{,}48 \times 10^{-04}$	$5{,}43 \times 10^{-03}$	$1{,}56 \times 10^{-05}$	$1{,}92 \times 10^{-01}$	$5{,}55 \times 10^{-01}$	$3{,}81 \times 10^{-01}$	$1{,}43 \times 10^{-03}$
9	rs3181200	116703705	*TNFSF8, e*	Meta. f.	A	$5{,}61 \times 10^{-03}$	$1{,}99 \times 10^{-03}$	$3{,}32 \times 10^{-05}$	$1{,}73 \times 10^{-01}$	$6{,}81 \times 10^{-01}$	$3{,}16 \times 10^{-01}$	$8{,}27 \times 10^{-04}$
9	rs3181374	116705008	*TNFSF8, e*	Meta. f.	C	$5{,}83 \times 10^{-03}$	$5{,}77 \times 10^{-03}$	$9{,}61 \times 10^{-05}$	$1{,}98 \times 10^{-01}$	$5{,}72 \times 10^{-01}$	$3{,}87 \times 10^{-01}$	$2{,}04 \times 10^{-03}$
10	rs1548255	38457199	intergenisch	Meta. f.	C	$4{,}96 \times 10^{-03}$	$5{,}42 \times 10^{-03}$	$9{,}61 \times 10^{-05}$	$8{,}03 \times 10^{-01}$	$4{,}38 \times 10^{-01}$	$7{,}80 \times 10^{-01}$	$2{,}87 \times 10^{-02}$

10	rs2505202	38511145	*LOC100129055, i*	LD-rank.	C	3,22 x 10^{-03}	2,74 x 10^{-03}	3,26 x 10^{-05}	5,83 x 10^{-01}	4,09 x 10^{-01}	3,45 x 10^{-01}	6,42 x 10^{-02}
10	rs1740736	38538915	*LOC100129055, i*	LD-rank.	G	2,72 x 10^{-03}	3,80 x 10^{-03}	4,17 x 10^{-05}	5,32 x 10^{-01}	4,92 x 10^{-01}	3,55 x 10^{-01}	6,78 x 10^{-02}
10	**rs745182**	81891702	***PLAC9, i***	**Meta. o.**	**G**	**2,01 x 10^{-04}**	**9,61 x 10^{-03}**	**1,78 x 10^{-05}**	**2,07 x 10^{-05}**	**9,75 x 10^{-01}***	**1,04 x 10^{-03}**	**1,97 x 10^{-07}**
10	rs1029074	90989016	*LIPA, i*	G.-rank.	G	3,10 x 10^{-01}	5,33 x 10^{-04}	6,00 x 10^{-04}	6,55 x 10^{-01}	NA	NA	NA
10	rs1960574	91014005	*LIPA, i*	G.-rank.	A	8,05 x 10^{-04}	7,14 x 10^{-01}	1,32 x 10^{-01}	8,19 x 10^{-01}	1,07 x 10^{-01}	2,46 x 10^{-01}	9,22 x 10^{-01}
10	rs758362	118447895	*HSPA12A, i*	Meta. f.	T	4,79 x 10^{-03}	2,48 x 10^{-03}	4,43 x 10^{-05}	3,32 x 10^{-02}	8,21 x 10^{-01}	7,73 x 10^{-02}	7,30 x 10^{-05}
11	rs566353	60472288	*SLC15A3, i*	Meta. o.	A	3,16 x 10^{-03}	7,16 x 10^{-03}	7,41 x 10^{-05}	3,34 x 10^{-01}	6,60 x 10^{-01}	2,86 x 10^{-01}	1,88 x 10^{-03}
11	rs2621483	115566825	intergenisch	Meta. f.	A	7,76 x 10^{-03}	5,12 x 10^{-04}	1,21 x 10^{-05}	8,12 x 10^{-01}	8,98 x 10^{-02}	1,81 x 10^{-01}	1,78 x 10^{-04}
12	**rs2190733**	66729518	**intergenisch**	**LD-rank.**	**C**	**3,51 x 10^{-03}**	**3,39 x 10^{-03}**	**7,26 x 10^{-05}**	**2,90 x 10^{-03}**	**3,57 x 10^{-01}**	**5,80 x 10^{-03}**	**2,63 x 10^{-06}**
12	rs12304940	66744148	intergenisch	LD-rank.	C	1,38 x 10^{-03}	8,40 x 10^{-03}	5,84 x 10^{-05}	4,84 x 10^{-03}	4,34 x 10^{-01}	9,45 x 10^{-03}	8,12 x 10^{-06}
12	rs1732581	114202449	intergenisch	Meta. f.	T	6,66 x 10^{-04}	3,36 x 10^{-04}	7,93 x 10^{-06}	5,61 x 10^{-02}	5,22 x 10^{-01}	8,18 x 10^{-02}	4,05 x 10^{-01}
12	rs10850905	116741236	*KSR2, i*	Meta. f.	G	5,11 x 10^{-03}	2,82 x 10^{-03}	4,21 x 10^{-05}	3,44 x 10^{-01}	9,38 x 10^{-01}	4,06 x 10^{-01}	1,87 x 10^{-01}
13	rs9527209	32681621	*STARD13, i*	Meta. f.	A	3,54 x 10^{-03}	2,61 x 10^{-03}	9,08 x 10^{-05}	5,28 x 10^{-01}	1,78 x 10^{-01}	7,79 x 10^{-01}	5,45 x 10^{-02}
14	rs12431541	29090267	intergenisch	LD-rank.	A	1,45 x 10^{-03}	1,24 x 10^{-03}	5,33 x 10^{-01}	3,02 x 10^{-01}	5,71 x 10^{-01}	2,39 x 10^{-01}	1,92 x 10^{-01}
14	rs11844637	29093372	intergenisch	LD-rank.	C	7,53 x 10^{-04}	1,16 x 10^{-02}	8,45 x 10^{-01}	2,69 x 10^{-01}	1,67 x 10^{-01}	8,55 x 10^{-02}	1,53 x 10^{-01}
14	rs8007404	58524255	intergenisch	LD-rank.	A	5,50 x 10^{-03}	2,89 x 10^{-03}	1,42 x 10^{-04}	5,92 x 10^{-01}	3,43 x 10^{-01}	6,12 x 10^{-01}	2,20 x 10^{-02}
17	rs2519866	27859883	*MYO1D, i*	G.-rank.	G	3,91 x 10^{-04}	NA	NA	5,44 x 10^{-01}	NA	NA	NA
17	***rs225214***	27920869	***MYO1D, i***	***G.-rank.***	***G***	***2,82 x 10^{-01}***	***3,65 x 10^{-05}***	***3,85 x 10^{-03}***	***6,13 x 10^{-02}***	***2,00 x 10^{-03}****	***5,09 x 10^{-01}***	***2,20 x 10^{-02}***
17	rs8080480	73350621	intergenisch	Meta. f.	T	1,80 x 10^{-03}	8,61 x 10^{-03}	9,71 x 10^{-05}	1,29 x 10^{-01}	9,60 x 10^{-01}	2,25 x 10^{-01}	4,90 x 10^{-04}
17	rs9915508	75845351	intergenisch	Meta. f.	T	4,67 x 10^{-03}	2,26 x 10^{-03}	3,21 x 10^{-05}	8,98 x 10^{-01}	8,80 x 10^{-01}	9,82 x 10^{-01}	2,46 x 10^{-02}
18	rs11659929	41186112	*SLC14A2, i*	Meta. o.	C	9,68 x 10^{-04}	2,96 x 10^{-03}	1,96 x 10^{-05}	7,23 x 10^{-01}	8,84 x 10^{-01}	7,17 x 10^{-01}	8,20 x 10^{-01}

3.2.4 Zweite Replikationsphase

Um die beobachteten Assoziationen weiter zu verifizieren, wurden die vier Kandidaten-SNPs für beide Krankheiten und der eine tuberkulosespezifische SNP in der zweiten Replikationsphase in drei unabhängigen Replikationsstichproben (Stichproben C-I, C-II und F) mittels der TaqMan®-Technologie genotypisiert. In der deutschen Stichprobe C-I konnten die Marker erfolgreich in 284 Sarkoidosepatienten und 279 Kontrollpersonen genotypisiert werden (CR = 100%). Der Marker rs745182 zeigte eine nominell signifikante Assoziation mit Sarkoidose in dieser Stichprobe (p = 4,27 x 10^{-2}) (Tab. 3-17). Die tschechische Sarkoidose-Stichprobe C-II umfasste nach Qualitätskontrolle 254 Patienten und 306 Kontrollindividuen (CR = 100%). Zwei Marker (rs8084, p = 2,29 x 10^{-3} und rs745182, p = 3,9 x 10^{-4}) zeigten in dieser Stichprobe eine nominell signifikante Assoziation mit Sarkoidose (Tab. 3-17). Die Allelfrequenzen der Marker sind in Tabelle 7-3 aufgeführt.

Die südafrikanische Tuberkulosestichprobe F enthielt nach der Qualitätskontrolle 386 Tuberkulosepatienten und 372 Kontrollpersonen (CR = 100%). Keiner der getesteten Marker zeigte in dieser Stichprobe eine signifikante Assoziation mit Tuberkulose (Tab. 3-18). Die dazugehörigen Allelfrequenzen sind in Tabelle 7-4 dargestellt.

Tab. 3-17: Assoziation der Kandidaten-SNPs in den Replikationsstichproben C-I und C-II. Nominell signifikante Ergebnisse (*p*-Wert < 0,05) sind fett hervorgehoben. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: Chr. = Chromosom; CI = Konfidenzintervall (engl. *confidence interval*); LD-rank = *LD cluster ranking*-Analyse; Meta. f. = *meta-analysis fixed effects*; Meta. o. = *meta-analysis opposite effects*; OR = Quotenverhältnis (engl. *odds ratio*); Repl.II = zweite Replikationsphase; us = stromaufwärts (engl. *upstream*).

Zweite Replikationsphase Sarkoidose (Repl.II)					Stichprobe C-I		Stichprobe C-II	
Chr.	**dbSNP ID**	**Gen**	**Analyse**	**A1**	***p*-Wert**	**OR [95% CI]**	***p*-Wert**	**OR [95% CI]**
6	rs8084	*HLA-DRA, e*	Meta. o.	A	9,87 x 10^{-02}	1,22 [0,96-1,54]	**2,29 x 10^{-03}**	**1,45 [1,14-1,83]**
7	rs4321884	*IL6, us*	Meta. o.	A	7,65 x 10^{-01}	1,04 [0,82-1,31]	2,74 x 10^{-01}	0,88 [0,69-1,11]
10	rs745182	*PLAC9, i*	Meta. o.	C	**4,27 x 10^{-02}**	**1,28 [1,01-1,62]**	**3,90 x 10^{-04}**	**1,53 [1,21-1,94]**
12	rs2190733	*IFNG, us*	LD-rank.	C	2,13 x 10^{-01}	1,21 [0,90-1,64]	3,11 x 10^{-01}	0,85 [0,61-1,17]

Tab. 3-18: Assoziation der Kandidaten-SNPs in der Replikationsstichprobe F. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: Chr. = Chromosom; CI = Konfidenzintervall (engl. *confidence interval*); G.-rank = *gene ranking*-Analyse; LD-rank = *LD cluster ranking*-Analyse; Meta. f. = *meta-analysis fixed effects*; Meta. o. = *meta-analysis opposite effects*; OR = Quotenverhältnis (engl. *odds ratio*); Repl.II = zweite Replikationsphase; us = stromaufwärts (engl. *upstream*).

Zweite Replikationsphase Tuberkulose (Repl.II)					Stichprobe F	
Chr.	**dbSNP ID**	**Gen**	**Analyse**	**A1**	***p*-Wert**	**OR [95% CI]**
6	rs8084	*HLA-DRA, e*	Meta. o.	A	6,64 x 10^{-01}	0,96 [0,85-1,28]
7	rs4321884	*IL6, us*	Meta. o.	A	5,12 x 10^{-02}	0,79 [1,00-1,61]
10	rs745182	*PLAC9, i*	Meta. o.	C	7,66 x 10^{-01}	1,03 [0,84-1,27]
12	rs2190733	*IFNG, us*	LD-rank.	C	5,52 x 10^{-01}	0,92 [0,70-1,21]
17	rs225214	*MYO1D, i*	G.-rank.	T	4,57 x 10^{-01}	0,93 [0,75-1,14]

Um die Aussagekraft der zweiten Replikationsphase zu überprüfen, wurde für alle fünf getesteten Marker eine Berechnung der statistischen Teststärke in den Replikationsstichproben C-I, C-II und F durchgeführt. Mit den in der ersten Replikationsphase erhaltenen ORs und Frequenzen des MAFs der Marker wurde die erreichbare Teststärke in der zweiten Replikationsphase berechnet (Tab. 3-19 und Tab. 3-20). Nur für einen SNP (rs8084) konnte in den Sarkoidose-Replikationsstichproben C-I und C-II eine Teststärke von über 50% für einen *p*-Wert von 0,05 erreicht werden. Aufgrund der kleinen Stichprobengrößen der Replikationsstichproben der zweiten Phase war eine statistische Teststärke zur Überprüfung der Marker auf Assoziation nicht gegeben. Daher können also die negativen Ergebnisse in der zweiten Replikationsphase eine Assoziation der einzelnen Marker nicht widerlegen. Wurden dennoch die Ergebnisse der zweiten Replikationsphase berücksichtigt, zeigten zwei Marker eine genomweite Signifikanz in einer Metaanalyse aller verwendeten Stichproben (A, B, C-I, C-II, D, E-I, E-II und F): rs8084 (HLA-DRA, $p = 2{,}52 \times 10^{-16}$) und rs745182 (*PLAC9*, $p = 1{,}26 \times 10^{-9}$) (Tab. 3-21). Eine Zusammenfassung der Ergebnisse der Marker mit dazugehörigen *odds ratios* ist in Tabelle 7-5 dargestellt.

Tab. 3-19: Errechnete Teststärke für die Replikationsstichproben C-I und C-II. Abkürzungen: MAF = Allelfrequenz des seltenen Allels (engl. *minor allele frequency*); OR = Quotenverhältnis (engl. *odds ratio*); Repl.I = erste Replikationsphase.

Sarkoidose	Stichprobe B (Repl.I)		Stichprobe C-I	Stichprobe C-II
dbSNP ID	$MAF_{Kontrollen}$	OR	Power für $p = 0{,}05$	Power für $p = 0{,}05$
rs8084	0,4398	1,455	62,30%	62,20%
rs4321884	0,4888	1,092	8,30%	8,30%
rs745182	0,4198	1,228	23,60%	23,60%
rs2190733	0,1832	0,825	14,20%	13,80%
rs225214	0,3758	1,097	8,50%	8,50%

Tab. 3-20: Errechnete Teststärke für die Replikationsstichprobe F. Abkürzungen: MAF = Allelfrequenz des seltenen Allels (engl. *minor allele frequency*); OR = Quotenverhältnis (engl. *odds ratio*); Repl.I = erste Replikationsphase.

Tuberkulose	Stichprobe E-I (Repl.I)		Stichprobe F
dbSNP ID	$MAF_{Kontrollen}$	OR	Power für $p = 0{,}05$
rs8084	0,4464	1,004	5,00%
rs4321884	0,3412	0,946	6,50%
rs745182	0,4883	1,112	11,30%
rs2190733	0,2838	0,945	6,40%
rs225214	0,4088	1,079	8,10%

Tab. 3-21: Assoziationsergebnisse der Kandidaten-SNPs der kombinierten Metaanalyse aller Stichproben beider Krankheiten. Die *p*-Werte der Kandidaten-SNPs sind nach Krankheiten sortiert für die jeweilige Stichprobe abgebildet. Die Spalte p_{META} zeigt die *p*-Werte der Metaanalyse aller verwendeten Stichproben. Abkürzungen: e = exonisch; i = intronisch; Repl.I = erste Replikationsphase; Repl.II = zweite Replikationsphase; SA = Sarkoidose; TB = Tuberkulose; us = stromaufwärts (engl. *upstream*).

		SA				TB			
dbSNP ID	**Gen**	**Stichprobe A** p_{GWAS}	**Stichprobe B** $p_{Repl.I}$	**Stichprobe C-I** $p_{Repl.II}$	**Stichprobe C-II** $p_{Repl.II}$	**Stichprobe D** p_{GWAS}	**Stichproben E-I & E-II** $p_{Repl.I}$	**Stichprobe F** $p_{Repl.II}$	p_{META}
rs8084	*HLA-DRA, e*	$3{,}44 \times 10^{-05}$	$6{,}90 \times 10^{-15}$	$9{,}87 \times 10^{-02}$	$2{,}29 \times 10^{-03}$	$8{,}48 \times 10^{-03}$	$6{,}47 \times 10^{-01}$	$6{,}64 \times 10^{-01}$	$2{,}52 \times 10^{-16}$
rs4321884	*IL6, us*	$5{,}01 \times 10^{-03}$	$6{,}67 \times 10^{-02}$	$7{,}65 \times 10^{-01}$	$2{,}74 \times 10^{-01}$	$1{,}92 \times 10^{-03}$	$4{,}42 \times 10^{-01}$	$5{,}12 \times 10^{-02}$	$1{,}04 \times 10^{-05}$
rs745182	*PLAC9, i*	$2{,}01 \times 10^{-04}$	$2{,}07 \times 10^{-05}$	$4{,}27 \times 10^{-02}$	$3{,}90 \times 10^{-04}$	$9{,}61 \times 10^{-03}$	$9{,}75 \times 10^{-01}$	$7{,}66 \times 10^{-01}$	$1{,}26 \times 10^{-09}$
rs2190733	*IFNG, us*	$3{,}51 \times 10^{-03}$	$2{,}90 \times 10^{-03}$	$2{,}13 \times 10^{-01}$	$3{,}11 \times 10^{-01}$	$9{,}61 \times 10^{-03}$	$3{,}57 \times 10^{-01}$	$5{,}52 \times 10^{-01}$	$8{,}49 \times 10^{-06}$
rs225214	*MYO1D, i*	-	-	-	-	$3{,}65 \times 10^{-05}$	$2{,}00 \times 10^{-03}$	$4{,}57 \times 10^{-01}$	$1{,}50 \times 10^{-06}$

Das in den Analysen detektierte Signal von rs745182 innerhalb des intronischen Bereichs des *PLAC9*-Gens liegt in unmittelbarer Nähe des Gens *ANXA11*. Für Sarkoidose wurde bereits eine Assoziation des *ANXA11*-Gens mit der Krankheit in mehreren Publikationen beschrieben (Cozier et al. 2012; Hofmann et al. 2008; Levin et al. 2013; Li et al. 2010; Mrazek et al. 2011). Um zu überprüfen, ob der identifizierte SNP rs745182 ein neues, von der *ANXA11*-Assoziation unabhängiges Signal darstellt, wurde eine logistische Regressionsanalyse mit dem Programm PLINK (Purcell et al. 2007) durchgeführt. In der Stichprobe A wurde zunächst eine multivariate logistische Regressionsanalyse ohne Kovariate durchgeführt. Dann wurde jeweils einmal auf den genetischen Effekt von rs745182 und einmal auf den Effekt des von Hofmann *et al.* (2008) identifizierten nicht-synonymen *ANXA11*-SNPs rs1049550 konditioniert (Tab. 3-22). Die genetische Assoziation von rs745182 behielt ihre Signifikanz bei (p = 0,0499), nachdem mit dem additiven Modell auf den Effekt von rs1049550 konditioniert wurde. Wird bei dem SNP rs1049550 auf die Effekte von rs745182 konditioniert, ist ebenfalls noch eine signifikante Assoziation (p = 8,18 x 10^{-03}) zu beobachten. Damit stellt der SNP rs745182 (*PLAC9*) ein neues (von rs1049550 unabhängiges) genetisches Assoziationssignal für Sarkoidose dar.

Tab. 3-22: Logistische Regressionsanalyse der SNPs rs745182 und rs1049550 in 2.139 Individuen der Stichprobe A. Um zu testen, ob die genetischen Effekte der beiden SNPs unabhängig sind, wurden logistische Regressionsanalysen zwischen den Effekten der beiden SNPs durchgeführt. Dabei wurde die Analyse ohne und mit Kovariaten durchgeführt, um zu testen, ob der entsprechende SNP einen unabhängigen genetischen Effekt aufweist.

dbSNP ID	A1	Kovariate	OR [95% CI]	asymptotischer *p*-Wert
rs745182	C	*keine*	1,43 [1,25-1,64]	2,79 x 10^{-07}
	C	rs1049550	1,21 [1,00-1,45]	4,99 x 10^{-02}
rs1049550	A	*keine*	0,67 [0,58-0,78]	8,19 x 10^{-08}
	A	rs745182	0,77 [0,63-0,93]	8,18 x 10^{-03}

Eine Studie von Ansari *et al.* (2011) konnte additive und subtraktive Effekte von verschiedenen *IFNG* Polymorphismen im Zusammenhang mit weiteren Zytokinen (u.a. *IL6*) und den in den Zytokinen auftretenden Polymorphismen für die Entwicklung von pulmonaler Tuberkulose nachweisen. In den hier durchgeführten Analysen wurden für Sarkoidose und Tuberkulose zwei mögliche Assoziationen in der Nähe des *IFNG*-Gens (rs2190733) und des Interleukin6-Gens (*IL6*; rs4321884) identifiziert. Daher wurden die beiden SNPs auf mögliche epistatische Effekte zwischen ihnen überprüft (Tab. 3-23). Die Analyse auf Epistase zwischen den SNPs rs2190733 und rs4321884 zeigte, dass keine epistatischen Effekte zwischen den SNPs herrschen.

Tab. 3-23: Analyse der Epistase zwischen rs4321884 und rs2190733 in Sarkoidose- und Tuberkulose-GWAS-Daten.

dbSNP ID 1	dbSNP ID 2	OR der Interaktion	asymptotischer *p*-Wert
Sarkoidose			
rs4321884	rs2190733	0,97	0,8161
Tuberkulose			
rs4321884	rs2190733	0,93	0,4306

3.2.5 *In silico*-Analysen

Die Replikation der Ergebnisse in den zwei Replikationsphasen war in der zweiten Replikationsphase zum Teil nicht erfolgreich. Die Berechnung der statistischen Teststärken der für die zweite Replikationsphase verwendeten Stichproben zeigte nämlich, dass diese keine ausreichende Teststärke für eine Replikation der in der ersten Replikationsphase bestätigten Ergebnisse hatten. Es standen auch keine weiteren Stichproben zur Replikation der Kandidaten-SNPs zur Verfügung. Auf Grund dieser Tatsache wurden die Marker *in silico* auf mögliche Funktionen analysiert. Mit der UCSC-Datenbank wurden die Regionen der Marker hinsichtlich potentieller Transkriptionsfaktorbindestellen (TFBS) und ihres regulatorischen Potentials, unter anderem mit Hilfe der ENCODE-Datenbank (Rosenbloom et al. 2012), überprüft. Außerdem wurden die Kandidaten-SNPs mit dem *SNPinfo*-Webserver (Xu and Taylor 2009) des „National Institute of Environmental Health Sciences“ (NIEHS) auf mögliche regulatorische Funktionen überprüft. Des Weiteren wurden die Kandidaten-SNP-Regionen mit der *„eQTL resources at the Pritchard lab*“-Datenbank (J. Pritchard) auf Marker mit regulatorischen Effekten überprüft.

3.2.5.1 SNP rs8084

Der SNP rs8084 liegt im dritten Exon des *HLA-DRA* und ist, je nach Transktiptvariante des Gens, ein spaSNP (engl. *splice acceptor variant*) mit einem möglichen Effekt auf das Spleißen oder aber ein synonymer SNP hinsichtlich der Aminosäuresequenz des Proteins (Kapitel 7, Abb.7-3).

Eine *in silico*-Analyse des Markers mit dem NIEHS *SNPinfo*-Webserver zeigte ein unterschiedliches LD (r^2 > 0,8) in dieser Region für Europäer (CEU) und Afrikaner (YRI) (Tab. 3-24). Der SNP rs8084 ist jedoch hoch konserviert und besitzt laut der Datenbank ein hohes regulatorisches Potential (Tab. 3-24). In der afrikanischen Population wurde des Weiteren für den LD-SNP rs7192 ein hohes regulatorisches Potential vorhergesagt. Die SNPs rs7192 und rs7194 liegen in Bereichen, in denen Spleißen stattfindet, wobei rs7192 zusätzlich ein nicht-synonymer SNP ist. Eine Analyse der Region (Chr6: 32.518.262 - 32.519.764 bp) des Kandidaten-SNPs rs8084 mit der *eQTL resources*-Datenbank sagte einen potentiellen *cis*-regulatorischen Effekt für die Marker rs8084, rs7192, rs2239804 und rs9268659 auf diverse HLA-Gene voraus (Brown et al. 2012; Montgomery et al. 2010; Stranger et al. 2007; Veyrieras et al. 2008) (Kapitel 7, Abb. 7-4).

Tab. 3-24: NIEHS Prognose von SNP-Funktionen. Die SNP-Funktionen-Vorhersage wurde jeweils für europäische und afrikanische Populationen durchgeführt. Dabei wurde der Kandidaten-SNP und alle Marker, die sich zusammen mit dem Kandidaten-SNP in der jeweiligen Population in einem LD von $r^2 > 0{,}8$ befinden, analysiert. Abkürzungen: Chr. = Chromosom; nsSNP = nicht-synonymer SNP; TFBS = Transkriptionsfaktorbindestelle; Reg. Potential = Regulationspotential; Konserv. = Konservierung; NA = nicht verfügbar (engl. *not available*).

dbSNP ID	Chr.	Position (bp)	Allele	LD-SNP	TFBS	Spleißen	nsSNP	Reg. Potential	Konserv.
CEU Population									
rs3130309	6	32323450	A/G	rs8084	-	-	-	NA	0
rs3763327	6	32521808	C/G	rs8084	-	-	-	0	0
rs8084	6	32519013	A/C	rs8084	-	-	-	0,386051	0,588
rs9268659	6	32518919	C/T	rs8084	-	-	-	0,182211	0,077
YRI Population									
rs2213585	6	32521128	G/A	rs8084	-	-	-	0	0
rs2213586	6	32521072	G/A	rs8084	-	-	-	0	0
rs2227139	6	32521437	G/A	rs8084	-	-	-	0	0
rs2239803	6	32519811	T/C	rs8084	-	-	-	0	0
rs2239804	6	32519501	T/C	rs8084	-	-	-	0,000874	0,341
rs3763327	6	32521808	C/G	rs8084	-	-	-	0	0
rs4935356	6	32520366	T/G/A	rs8084	-	-	-	0	0
rs6926374	6	32517283	A/G	rs8084	-	-	-	0	0
rs7192	6	32519624	G/T	rs8084	-	+	+	0,563105	0
rs7194	6	32520458	A/G	rs8084	-	+	-	0	0
rs7195	6	32520517	A/G	rs8084	-	-	-	0,058567	0
rs8084	6	32519013	A/C	rs8084	-	-	-	0,386051	0,588
rs9268658	6	32518694	A/G	rs8084	-	-	-	0	0
rs9268659	6	32518919	C/T	rs8084	-	-	-	0,182211	0,077

3.2.5.2 SNP rs4321884

Der Marker rs4321884 ist stromaufwärts des *IL6*-Gens in einer intergenischen Region gelegen. In der ENCODE-Spur der UCSC-Datenbank zeigte sich ein leicht erhöhtes H3K4Me1-Level (monomethyliertes Lysin 4 im Histon H3) in Gm12878- (lymphoblastoide Zelllinie) und NHLF-Zelllinien (normale humane Lungenfibroblasten) in der Region des SNPs (Kapitel 7, Abb. 7-5). Die „*Transcription Factor ChIP-Seq*" von ENCODE zeigte die Bindung der Transkriptionsfaktoren in unmittelbarer Nähe des Markers in Gm12878-Zellen und in K562-Zellen (myeloische Leukämie-Zelllinie; diese Zellen weisen Ähnlichkeiten zu Neutrophilen auf). Eine *in silico*-Analyse des Markers mit dem NIEHS *SNPinfo*-Webserver sagte keine regulatorische Funktionen oder eine Konservierung für den Marker voraus. Eine Analyse der Region (Chr7: 22.707.233 - 22.708.735 bp) um den Marker rs4321884 mit der *eQTL resources*-Datenbank sagte einen potentiellen *cis*-regulatorischen Effekt für den SNP auf ein bisher unbekanntes Gen voraus (Montgomery et al. 2010) (Kapitel 7, Abb. 7-6).

3.2.5.3 SNP rs745182

Der SNP rs745182 liegt im intronischen Bereich des *PLAC9*-Gens, in einem Gebiet mit leicht erhöhtem H3K4Me1-Level in Gm12878- und NHLF-Zelllinien. Stromabwärts des Markers zeigte die Analyse mit der ENCODE-Datenbank die Bindung einer Reihe von Transkriptionsfaktoren (Kapitel 7, Abb. 7-7).

Eine Analyse mit NIEHS *SNPinfo* zeigte ein leichtes regulatorisches Potential für zwei LD-SNPs in der CEU-Population (rs10887581 und rs2236558) und sagte ein hohes regulatorisches Potential für LD-SNP rs2819874 voraus (Tab. 3-25). Die Abfrage der *eQTL resources*-Datenbank lieferte für keinen Marker in der Region (Chr10: 81.890.951 - 81.892.453 bp) ein Ergebnis.

Tab. 3-25: NIEHS-Prognose von SNP-Funktionen. Die SNP-Funktionen-Vorhersage wurde jeweils für europäische und afrikanische Populationen durchgeführt. Dabei wurden der Kandidaten-SNP und alle Marker, die sich zusammen mit dem Kandidaten-SNP in der jeweiligen Population in einem LD von r^2 > 0,8 befinden, analysiert. Abkürzungen: Chr. = Chromosom; nsSNP = nicht-synonymer SNP; TFBS = Transkriptionsfaktorbindestelle; Reg. Potential = Regulationspotential; Konserv. = Konservierung; NA = nicht verfügbar (engl. *not available*).

dbSNP ID	Chr.	Position (bp)	Allele	LD-SNP	TFBS	Spleißen	nsSNP	Reg. Potential	Konserv.
CEU Population									
rs10887553	10	81898653	G/T	rs745182	-	-	-	NA	0
rs10887581	10	81927203	C/T	rs745182	-	-	-	0,166747	0
rs2236558	10	81913185	T/G	rs745182	-	-	-	0,187618	0
rs2819874	10	81894618	G/A	rs745182	-	-	-	0,435595	0
rs2819950	10	81925259	G/A	rs745182	-	-	-	0	0
rs745182	10	81891702	T/C	rs745182	-	-	-	0	0
YRI Population									
rs745182	10	81891702	T/C	rs745182	-	-	-	0	0

3.2.5.4 SNP rs2190733

Der SNP rs2190733 liegt in einer intergenischen Region stromabwärts des Interferon-Gamma-Gens (*IFNG*). Die Analyse mit der ENCODE-Datenbank zeigte, dass der Marker in der Nähe einer TFBS liegt, an der unter anderem die Transkriptionsfaktoren NF-kappaB und PU.1 binden (Kapitel 7, Abb. 7-8). Des Weiteren befindet sich stromabwärts des Markers eine *IFNG* antisense-RNA (AK124066). Die *in silico*-Analyse mit NIEHS *SNPinfo* sagte keine regulatorische Funktionen oder eine Konservierung für den Marker oder seine LD-SNPs voraus, auch die Analyse mit der *eQTL resources*-Datenbank zeigte keine eQTL-Daten.

3.2.5.5 SNP rs225214

Der tuberkulosespezifische SNP rs225214 liegt im intronischen Bereich des *MYO1D*-Gens. In der ENCODE-Spur zeigte sich ein leicht erhöhtes Methylierungslevel (H3K4Me1) und ein stark erhöhtes Acetylierungslevel (H3K27Ac) in Gm12878- und NHLF-Zelllinien in der Region des Markers. Die ‚*Transcription Factor ChIP-seq*' von ENCODE zeigte, dass in direkter Nachbarschaft des SNPs eine Reihe von Transkriptionsfaktoren binden (u.a. PU.1) (Kapitel 7, Abb. 7-9). Eine Analyse mit dem NIEHS *SNPinfo*-Webserver zeigte auch ein hohes regulatorisches Potential für rs225214 und ein etwas schwächeres Signal für den LD-SNP rs225212 (Tab. 3-26). Die Analyse mit der *eQTL resources*-Datenbank sagte für den SNP rs225214 eine eQTL-Funktion für das *PSMD11*-Gen und für rs225212

eine eQTL-Funktion für *PSMD11* und *C17orf79* voraus (Montgomery et al. 2010) (Kapitel 7, Abb. 7-10).

Tab. 3-26: NIEHS-Prognose von SNP-Funktionen. Die SNP-Funktionen-Vorhersage wurde für die afrikanische Population durchgeführt. Dabei wurde der Kandidaten-SNP und alle Marker, die sich mit dem Kandidaten-SNP in einem LD von $r^2 > 0,8$ befinden, analysiert. Abkürzungen: Chr. = Chromosom; nsSNP = nicht-synonymer SNP; TFBS = Transkriptionsfaktorbindestelle; Reg. Potential = Regulationspotential; Konserv. = Konservierung.

dbSNP ID	Chr.	Position	Allele	LD-SNP	TFBS	Spleißen	nsSNP	Reg. Potential	Konserv.
YRI Population									
rs225212	17	27920568	T/G/C	rs225214	-	-	-	0,077929	0
rs225214	17	27920869	T/C	rs225214	-	-	-	0,109707	0

4 Diskussion

Heutzutage ist es gesicherte Erkenntnis, dass viele Krankheiten eine genetische Grundlage haben. Neben den Erkrankungen mit monogenetischen Ursachen stehen vor allem Krankheiten mit polygenetischen Ursachen im Fokus der Forschung. Letztere sind nicht leicht zu untersuchen, da jedes einzelne Gen häufig nur einen kleinen Effekt auf den Phänotyp hat und sich die genetische Ursache der Krankheit erst durch die Aufsummierung der Effekte oder durch eine spezifische Kombination von Effekten der einzelnen Gene bedingt. Bei komplexen Krankheiten spielen neben den genetischen Faktoren aber auch verschiedene andere Faktoren wie Umwelteinflüsse, Lebensgewohnheiten und ihre Wechselwirkung untereinander eine Rolle. Auch bei vielen Infektionskrankheiten tragen neben den Umwelteinflüssen genetische Faktoren zur Krankheitsentstehung bei.

Die entzündlichen Lungenkrankheiten Sarkoidose und Tuberkulose sind solche Krankheiten, die unter anderem auf genetischen Ursachen beruhen. Da Entzündungsreaktionen zum großen Teil vom Immunsystem vermittelt sind (zusammengefasst in Raj et al. 2013), wurden als genetische Ursache entzündlicher Krankheiten bisher hauptsächlich Gene des HLA-Systems (humanes Leukozytenantigen-System) identifiziert. Bei Entzündungen wird ein hoch koordiniertes Genexpressionsprogramm aktiviert, welches spezifisch für den initialen Stimulus und den Zelltyp ist (zusammengefasst in Natoli et al. 2011). Mutationen in den daran beteiligten Genen können also spezifisch zur Suszeptibilität gegenüber bestimmten Entzündungskrankheiten beitragen.

Die Erblichkeit der Krankheiten Sarkoidose und Tuberkulose wird als sehr hoch eingeschätzt (Sarkoidose: ~66% (Sverrild et al. 2008); Tuberkulose: bis zu ~71% (zusammengefasst in Moller and Hoal 2010)). Diese Zahlen müssen jedoch auch kritisch betrachtet werden, da in den Erblichkeitsberechnungen die nicht-genetischen Faktoren häufig nicht entsprechend berücksichtigt werden (zusammengefasst in Vineis and Pearce 2011). Obwohl für beide Krankheiten schon eine Vielzahl an Risikoloci identifiziert werden konnte, lässt sich anhand der bisher identifizierten Loci die genetische Komponente der Krankheiten noch nicht vollständig erklären.

SNPs können durch Veränderung der DNA-Sequenz eine Reihe von verschiedenen Auswirkungen haben. SNPs, die zu einem Aminosäureaustausch, zu Stoppcodons, Leserasterverschiebungen oder auch zu Veränderungen im Spleißmuster führen, können heutzutage einfach erkannt werden und ihre Auswirkungen auf die Proteine sind normalerweise offensichtlich. Doch auch SNPs in nicht-codierenden DNA-Bereichen können in Zielsequenzen von Transkriptionsfaktoren (oder anderen regulatorischen Proteinen) liegen, zu einer veränderten Bindeaffinität von Transkriptionsfaktoren führen und somit die Transkription von Genen beeinflussen (zusammengefasst in

Stamatoyannopoulos 2012). Kann das durch die Mutation betroffene Gen identifiziert werden, wird es möglich für die Krankheit relevante Signaltransduktionswege aufzudecken. Mit dem Wissen über die von der Signaltransduktionskette betroffenen Mechanismen könnten dann Medikamente entwickelt werden, die die gestörten Mechanismen reparieren oder Ersatzfunktionen übernehmen. Von der Identifizierung einer Mutation bis zur Entwicklung einer Therapie ist es ein weiter Weg, doch ohne diese genetische Grundlagenforschung ist es kaum möglich effektive Therapien zu entwickeln.

4.1 Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen

Die Ergebnisse dieses Teils der Arbeit wurden bereits 2012 in dem Journal „*American Journal of Respiratory and Critical Care Medicine*“ veröffentlicht (Fischer et al. 2012).

Für die Sarkoidose konnten folgende genetischen Suszeptibilitätsloci identifiziert werden: *ANXA11* (Hofmann et al. 2008), *BTNL2* (Rybicki et al. 2005; Valentonyte et al. 2005), *CCR2* (Spagnolo et al. 2003), *IL23R* (Fischer et al. 2011), *MHC2TA* (Grunewald et al. 2010), *NOTCH4* (Adrianto et al. 2012), *OS9* (Hofmann et al. 2013), *RAB23* (Hofmann et al. 2011) und außerdem verschiedene *HLA*-Haplotypen (Grutters et al. 2003; Schurmann et al. 2001), jedoch kann mit den bisher bekannten Loci das Erkrankungsrisiko nicht vollständig erklärt werden. Es ist also sehr wahrscheinlich, dass weitere genetische Risikofaktoren für Sarkoidose existieren.

Für den Krankheitsverlauf der Sarkoidose spielt es eine entscheidende Rolle, welcher Subphänotyp beim Patienten auftritt. Bei einem *akuten* Verlauf der Sarkoidose treten heftige Symptome auf, heilen jedoch nach spätestens zwei Jahren von selbst wieder ab. Bei einem *chronischen* Verlauf treten zunächst nur schwache Symptome auf, die sich jedoch im Laufe der Zeit verstärken, ohne Behandlung nicht abheilen und in schweren Fällen sogar zum Tod führen können (American Thoracic Society 1999; Baughman et al. 1997; Baughman et al. 2003; Chappell et al. 2000; Neville et al. 1983; J. M. Reich 2002) (siehe Kapitel 1.2.1). Es ist somit von klinischer Relevanz, die genetischen Faktoren zu kennen, die zu der Ausprägung des jeweiligen Subphänotyps beitragen.

Für die meisten der in dieser Studie analysierten Patienten standen Informationen über ihren Subphänotypstatus zur Verfügung, diese wurden genutzt, um die Sarkoidose-Subphänotypen (*akuter* Phänotyp und *chronischer* Phänotyp) einzeln zu analysieren und auf diese Weise subphänotypspezifische genetische Assoziationen zu detektieren.

Einige der bisher für Sarkoidose identifizierten Suszeptibilitätsloci konnten schon einem der beiden Subphänotypen zugeordnet werden: Mit dem *akuten* Subphänotyp scheinen die Risikoloci *HLA-DRB1*03* (Grunewald and Eklund 2009), *DRB1*0301* (Sato et al. 2010), *DQB1*0201* (Grutters et al. 2003), *CCR2 (Spagnolo et al. 2003)*, *MHC2TA* (Grunewald et al. 2010) und *OS9* (Hofmann et al. 2013)

assoziiert zu sein, während die Risikoloci *BTNL2* (Li et al. 2006; Rybicki et al. 2005) und *IL23R* (Fischer et al. 2011) mit dem *chronischen* Subphänotypen assoziiert werden.

Mit dem in dieser Arbeit verwendeten Affymetrix 6.0 Sarkoidose-GWAS-Datensatz, der größten europäischen Sarkoidose-GWAS, konnte schon der genetische Suszeptibilitätslocus *OS9* (Hofmann et al. 2013) identifiziert werden. Die Teststärke und Markerdichte dieses Datensatzes wurde im Rahmen dieser Arbeit nochmals mittels der bioinformatischen Methode der Imputation erhöht und somit die Genotypdaten von 1.294.967 SNPs auf weitere genetische Assoziationen mit der Krankheit analysiert.

Die dabei identifizierte Assoziation des SNPs rs479777 wurde erfolgreich in vier Stichproben aus Deutschland, der Tschechischen Republik und Schweden repliziert und konnte somit die Assoziation in unabhängigen europäischen Populationen bestätigen. Eine genauere Betrachtung des SNPs bei den *akuten* (OR = 0,79) und *chronischen* Subphänotypen (OR = 0,78) in der Stichprobe B zeigte keinen signifikanten Unterschied zwischen den Subphänotypen und macht so deutlich, dass es sich dabei nicht um eine subphänotypspezifische Assoziation handelt.

Der SNP rs479777 liegt einige hundert Basenpaare stromaufwärts des *CCDC88B*-Gens in einem Bereich mit erhöhter Acetylierung (H3K27Ac). Die ENCODE-Datenbank zeigte an dieser Stelle die Bindung einer Reihe verschiedener Transkriptionsfaktoren (Kapitel 7, Abb. 7-1). Durch einen Basenaustausch könnte die Bindeaffinität für die dort bindenden Transkriptionsfaktoren verändert sein. In einigen Studien konnte nachgewiesen werden, dass SNPs zu einer Veränderung der Bindeaffinität von Transkriptionsfaktoren führen und somit auch die Expression von Genen beeinflussen können (Kasowski et al. 2010; Prokunina et al. 2002; Szalai et al. 2005; Tokuhiro et al. 2003). In der sich bei dem Marker befindenden TFBS binden unter anderem die Transkriptionsfaktoren NF-kappaB und PU.1 in lymphoblastoiden Zelllinien. Die Rolle von NF-kappaB in vielen Bereichen der Immunantwort ist hinlänglich bekannt (zusammengefasst in Hayden et al. 2006) und es konnte in einer Studie gezeigt werden, dass durch die Bindung von PU.1 in dem *CCDC88B* Promotor ein PU.1 abhängiger, auf TLR4 reagierender Promotor (engl. *PU.1-dependent TLR4-responsive promoter*) entsteht (Escoubet-Lozach et al. 2011). Über TLR4-Signalwege wird in Makrophagen eine Vielzahl von Genen aktiviert, die für die anti-mikrobielle Aktivität und für die Initiation von sekundären Entzündungssignalwegen eine Rolle spielen (Escoubet-Lozach et al. 2011). Die *in silico*-Analyse mit der NIEHS-Datenbank identifizierte einen im starken LD stehenden SNP rs663743 (r^2 = 0,92), der ebenfalls in einer TFBS liegt. Des Weiteren sagte die Datenbank einen Einfluss dieses Markers auf Spleiß-Vorgänge voraus (Tab. 3-8). Der SNP rs479777 oder der LD-SNP rs663743 könnten somit in einer Immunantwort die Expression der Gene beeinflussen, die von den dort bindenden Transkriptionsfaktoren reguliert werden.

Interessanterweise führt das seltenere Allel des Markers rs479777 dazu, dass sich das Risiko an Sarkoidose zu erkranken verringert (OR = 0,76), d.h. dass eine Mutation an dieser Stelle vielleicht anderen krankheitsauslösenden Effekten entgegenwirken kann.

Für den im Rahmen der Feinkartierung identifizierten SNP rs671976 konnte gezeigt werden, dass das seltenere Allel mit einem erhöhten Erkrankungsrisiko für Sarkoidose assoziiert ist (OR = 1,31). Der SNP rs671976 liegt im intronischen Bereich des *BAD* Gens, diese Region zeigte keine erhöhten Methylierungs- oder Acetylierungsmuster (H3K4Me1, H3K4Me3 und H3K27Ac) in den Zelllinien Gm12878 und NHLF (Kapitel 7, Abb. 7-2). Die ChIP-Sequenzierung von ENCODE zeigte keine Transkriptionsfaktor-Bindung in relevanten Zelllinien in der Region. Auch die Analyse mit der NIEHS *in silico*-Datenbank konnte für den SNP keine funktionelle Relevanz vorhersagen (Tab. 3-8). Diese Ergebnisse deuten darauf hin, dass der Marker selbst nicht die kausative Variante für die Krankheit Sarkoidose darstellt, sondern mit der kausativen Variante im starken LD steht und diese durch weitere Untersuchungen identifiziert werden muss.

Analysen mit der Genevar-Datenbank identifizierten den SNP rs671976 jedoch als eQTL für die *CCDC88B*-Genexpression (Tab. 3-9). Laut einer Studie wird *CCDC88B* vorrangig in Endothelzellen und Makrophagen transkribiert (Matsushita et al. 2011), doch in dem Gewebepanel, welches im Rahmen dieser Arbeit durchgeführt wurde, konnte auch eine mRNA-Expression in der Milz, im Thymus, im Pankreas und in der Lunge beobachtet werden (Abb. 3-4). In der bronchoalveoläre Lavagen-Stichprobe II war die *CCDC88B* mRNA-Expression mit dem rs671976 Genotyp, nicht aber mit der prozentualen Anzahl der alveolären Makrophagen in den Proben korreliert, wobei sich die Expression der *CCDC88B*-mRNA mit dem „G"-Allel des SNPs erhöht (Abb. 3-6). Diese Ergebnisse stützen die Voraussage der *in silico*-Datenbanken, dass der SNP rs671976 ein eQTL für die *CCDC88B*-Expression ist. Wegen des geringen Stichprobenumfangs von nur 27 Patienten ist dieses Ergebnis jedoch kritisch zu betrachten und weitere Untersuchungen sind nötig, um die eQTL-Funktion des SNPs zu verifizieren.

Eine Expressionsanalyse mit der bronchoalveolären Lavagen-Stichprobe I zeigte eine signifikant erhöhte mRNA-Expression von *KCNK4* und *PRDX5* in einem Vergleich von Sarkoidosepatienten mit gesunden Kontrollen (Abb. 3-5). In dieser Lavagen-Stichprobe wurden die Proben hinsichtlich der Proportion der alveolären Makrophagen abgestimmt, um eine möglichst vergleichbare Ausgangsposition zu schaffen. Für *CCDC88B* konnte zwar ebenfalls eine erhöhte mRNA-Expression festgestellt werden, aber sie zeigte keinen signifikanten Unterschied zu den gesunden Kontrollen. Wegen der sehr kleinen Stichprobengröße (jeweils n = 12) und dem unbekannten genetischen Hintergrund der Proben kann eine in Sarkoidosepatienten erhöhte *CCDC88B*-Expression allerdings auch nicht ausgeschlossen werden.

Insgesamt deuten die in dem Zusammenhang mit den beiden SNPs rs479777 und rs671976 durchgeführten Analysen darauf hin, dass *CCDC88B* ein vielversprechender Kandidat in der Krankheitsätiologie der Sarkoidose ist und als interessanter Ausgangspunkt für weiterführende funktionelle Studien dienen kann. Ein weiterer interessanter Aspekt ist, dass es möglich ist, beide SNPs zusammen auch auf funktioneller Ebene in einen Kontext mit der *CCDC88B*-Expression zu setzen. In der bronchoalveoläre Lavagen-Stichprobe II zeigte sich, dass die Expression der *CCDC88B*-mRNA bei homozygoten Trägern des „G"-Allels des SNPs rs671976 erhöht ist. Der SNP rs479777 liegt in einer TFBS direkt stromaufwärts des *CCDC88B*-Gens, eine Mutation hier könnte zu einer verminderten Bindung von Transkriptionsfaktoren führen und damit vielleicht die *CCDC88B*-Expression herabsetzen. Damit würde sich die protektive Wirkung des SNPs (OR = 0,76) mit der verminderten Expression von *CCDC88B*-mRNA erklären und einen weiteren Hinweis auf eine mögliche Relevanz von *CCDC88B* in der Ätiologie der Sarkoidose liefern.

Das von dem Gen codierte Protein „GRP78-interacting protein induced by ER-stress" (Gipie) (oder auch „Coiled-coil domain-containing protein 88B") interagiert, laut einer Studie von Matsushita *et al.* (2011), mit dem GRP78- und IRE1-Protein und stabilisiert die Bindung zwischen den beiden Proteinen. Gipie Expression wird durch ER-Stress hervorgerufen, unterdrückt den IRE1-JNK-Signalweg (engl. *c-Jun N-terminal kinase*; JNK) und verhindert durch ER-Stress hervorgerufene Apoptose in Endothelzellen. Bei Makrophagen in Tuberkulose-Granulomen wurde eine erhöhte Aktivierung von Genen des ER-Stress-Signalwegs festgestellt. Der ER-Stress wird hauptsächlich in Makrophagen, die sich in den nekrotischen Bereichen der Granulome befinden, beobachtet und führt dort wohl zu einer vermehrten Apoptose der Zellen (Seimon et al. 2010). Eine erhöhte *CCDC88B* Expression könnte in der Sarkoidose dazu führen, dass in den Granulomen der Sarkoidose nur sehr selten die Apoptose von Zellen beobachtet wird. Des Weiteren wurden in Immunpräzipitationsversuchen die Peptide Sortierungs-Nexin (engl. *sorting nexin*), Lymphozytenspezifisches Protein 1 (engl. *lymphocyte-specific protein 1*), Nukleolin und Cathepsin Z als Interaktionspartner von Gipie identifiziert (Matsushita et al. 2011). Eine genauere funktionelle Charakterisierung des Gipie-Proteins kann vielleicht enthüllen, welche biologische Relevanz Gipie in der Sarkoidose hat. Bisher ist ein funktioneller Zusammenhang jedenfalls noch rein spekulativ.

Die *in silico*-Analysen mit der SNPexp-Datenbank prognostizierten einen möglichen eQTL-Effekt des SNPs rs671976 und seines LD-SNPs rs11231740 (r^2 = 0,98) auf die Gene *KCNK4* und *PRDX5* (Tab. 3-10). Bei beiden Genen wurde die mRNA-Expression in dem Gewebepanel in allen untersuchten Geweben nachgewiesen (Abb. 3-5) und beide zeigten eine erhöhte Expression in Sarkoidosepatienten im Vergleich mit gesunden Kontrollpersonen (Abb. 3-5).

KCNK4 (Kaliumkanal Subfamilie K Mitglied 4, auch TWIK-verwandter Arachidonsäure stimulierter K^+-Kanal) kodiert für einen mechanosensitiven Zweiporen K^+-Kanal (Lesage et al. 2000) und scheint, zumindest in der Maus, eine Rolle in der Mechanosensorik der Luftröhre zu spielen (Lembrechts et al. 2011). Obwohl eine erhöhte Expression von *KCNK4* in Sarkoidosepatienten festgestellt werden konnte (Abb. 3-5), scheint anhand der Proteinfunktion nach aktuellem Wissensstand eine Verbindung mit der Krankheit Sarkoidose sehr unwahrscheinlich.

PRDX5 (Peroxiredoxin 5) ist ein Antioxidationsenzym, welches Zellen bei endogenem und exogenem Peroxid-Stress schützt und somit den Zelltod durch Peroxid-Exposition vermindern kann (Dubuisson et al. 2004; Knoops et al. 2011). Während einer akuten Entzündung in der Lunge von Ratten wurde zum ersten Mal eine Hochregulierung der *PRDX5*-Expression beobachtet (Knoops et al. 1999). Studien haben gezeigt, dass diese erhöhte *PRDX5*-Expression während einer Entzündung nur zu einem geringen Teil durch eine Hochregulierung der Expression in den Epithelzellen bedingt ist. Hauptsächlich ist diese durch eingewanderte Neutrophile und Monozyten verursacht, die hohe Level des Peroxiredoxins exprimieren (Cheah et al. 2009; Krutilina et al. 2006). Entsprechende *in vitro*-Versuche haben gezeigt, dass die *PRDX5*-Expression in LPS (Lipopolysaccharide) und IFN-γ exponierten primären Makrophagen stark erhöht ist (Abbas et al. 2009; Diet et al. 2007).

Weitere Studien stellen fest, dass eine geminderte „oxidative Burst"-Funktion (Freisetzung von reaktiven Sauerstoffspezies) zur Granulomaformation in Patienten mit septischer Granulomatose beiträgt (Baehner and Nathan 1967). Eine Deregulation von *PRDX5* könnte also Auswirkungen auf den Mechanismus des „oxidativen Bursts" haben, der von Makrophagen und anderen phagozytischen Zellen für die intrazelluläre Verdauung von Fremdkörpern genutzt wird. Die Rolle von alveolären Makrophagen für die Granulomaformation in der Sarkoidose ist bekannt (Agostini et al. 2000; Bachwich et al. 1986; Hance et al. 1985; Robinson et al. 1985; Silva et al. 2013) und eine Fehlregulation von Makrophagenfunktionen könnte daher zu dem Phänotyp Sarkoidose beitragen. Wenn die Deregulation von *PRDX5* eine Auswirkung auf die produzierten Mengen von reaktiven Sauerstoffspezies der Makrophagen und anderen phagozytischen Zellen hat, besteht die Möglichkeit, dass dies, wegen einer verminderten intrazellulären Verdauung von Krankheitserregern, zur Krankheitsentstehung durch die Granulomaformation führt. Es ist aber auch möglich, dass der oxidative Stress, der als Merkmal granulomatöser Entzündungen beobachtet wird, seine Ursache in der Fehlregulation von *PRDX5* hat. Weitere Untersuchungen müssen zeigen, ob das Gen *PRDX5* sich als Sarkoidose-Risikofaktor bestätigt und ob die vermuteten funktionellen Zusammenhänge mit der Krankheit der Wirklichkeit entsprechen.

Die Ergebnisse der Feinkartierung, der Expressionsstudien und die Auswertung der verfügbaren Literatur weisen darauf hin, dass *CCDC88B*, *KCNK4* und *PRDX5* vielversprechende Kandidaten für das

der assoziierten Region zugrunde liegende Risikogen für Sarkoidose sein könnten. Allerdings werden weitere funktionelle Studien nötig sein, um das spezifische Risikogen exakt zu identifizieren und um den krankheitsauslösenden Mechanismus aufzudecken.

Die Ergebnisse dieser Arbeit liefern einen deutlichen Beweis für eine Assoziation von SNP-Markern auf Chromosom *11q13.1* mit der Krankheit Sarkoidose in Europäern. In der HapMap-Datenbank zeigt sich ein sehr inkonsistentes Bild der Allelfrequenzen der SNPs rs479777 und rs671976 in verschiedenen Populationen, und da im Rahmen dieser Arbeit die Ergebnisse nur in europäisch stämmigen Populationen repliziert wurden, sind nun weitere Analysen dieses Locus in nicht-europäischen Stichproben nötig, um zu ermitteln, ob dieser neue Risikolocus bei Patienten auch unabhängig von ihrer ethnischen Herkunft zu Prädisposition für Sarkoidose beiträgt. Weitere Untersuchungen sind nötig, um die dem Signal zugrunde liegenden Risikovarianten und ihre funktionelle Rolle in der Krankheit zu finden und den Kontext möglicher regulatorischer Auswirkungen herzustellen.

Ziel dieser Arbeit war es auch, den heterogenen klinischen Phänotyp der Sarkoidose durch die Aufdeckung der genetischen Unterschiede zwischen den Subphänotypen weiter zu differenzieren. Im Rahmen der Untersuchungen für diese Arbeit konnte kein subphänotypspezifischer Risikolocus identifiziert werden. Obwohl für die Analysen die größte europäische Sarkoidose-GWAS und Validierungsstichprobe zur Verfügung stand, wurde die Individuenzahl durch das Filtern nach den Subphänotypen substanziell in der GWAS- und Validierungsstichprobe reduziert. Durch diese Reduzierung der getesteten Individuen verminderte sich die statistische Teststärke der Datensätze für die Subphänotyp-Analysen deutlich gegenüber einer Analyse des kompletten Sarkoidose-Datensatzes (*akuter* Subphänotyp: um 27,9%; *chronischer* Subphänotyp: um 10,1%; siehe Kapitel 3.1.4.1). Die Ergebnisse dieser Studie sprechen somit nicht gegen ein Vorhandensein von weiteren subphänotypischen genetischen Risikofaktoren, sondern zeigen lediglich, dass weiterer Bedarf an genetischen Analysen von sorgsam diagnostizierten Sarkoidose-Stichproben herrscht.

Bei der Analyse des imputierten Sarkoidose-GWAS-Datensatzes konnte auf dem Chromosom *11q13.1* ein neuer genomweiter signifikanter Risikolocus für die Krankheit Sarkoidose identifiziert werden. Interessanterweise konnte in weiteren Studien gezeigt werden, dass der Locus 11q13.1 mit mehreren Entzündungskrankheiten (Morbus Crohn, Alopecia areata, Primär biliäre Zirrhose, Lepra, Psoriasis) assoziiert ist (D. Ellinghaus et al. 2012; Jagielska et al. 2012; Mells et al. 2011; Petukhova et al. 2010; F. Zhang et al. 2011a). Diese gemeinsame Assoziation an diesem Locus bestärkt die Vermutung, dass grundsätzlich eine gemeinsame genetische Basis für die entzündlichen Krankheiten Morbus Crohn, Lepra, Psoriasis und Sarkoidose besteht (Fischer et al. 2011; Franke et al. 2008; H. S. Kim et al. 2011; F. Zhang et al. 2009; F. Zhang et al. 2011a; Zhu et al. 2012). Nach dem bisherigen

Kenntnisstand ist dies der erste gemeinsame Risikolocus von Sarkoidose, Alopecia areata und ‚Primär biliärer Zirrhose' außerhalb der HLA-Region. Die GWAS-Resultate für die anderen Krankheiten zeigten ihre stärksten Assoziationen in dieser Region bei den SNPs rs694739 und rs538147, welche stromaufwärts und stromabwärts des *CCDC88B* Gens liegen, während die Feinkartierung in der Sarkoidose die stärksten Assoziationssignale bei den SNPs rs479777 und rs671976 aufzeigte. In dieser Region herrscht jedoch eine komplexe LD-Struktur und weitere Studien müssen klären, ob die bisher in dieser Region identifizierten Signale alle ihren Ursprung in der gleichen kausativen Variante haben oder ob in dem Locus krankheitspezifische Assoziationen bestehen. Obwohl für die Krankheiten Morbus Crohn, Psoriasis, Lepra und Alopecia areata, hauptsächlich wegen ihrer genomischen Nähe, die Gene *PRDX5*, *ESRRA* und *RPS6KA4* als die zugrunde liegenden Risikofaktoren postuliert wurden, deuten die im Rahmen dieser Arbeit durchgeführten Expressions- und *in silico*-Analysen daraufhin, dass für Sarkoidose das Gen *CCDC88B* der zugrunde liegende Risikofaktor ist.

4.2 Identifizierung gemeinsamer genetischer Faktoren der Krankheiten Tuberkulose und Sarkoidose

Wie einleitend bereits erwähnt, zeigen die Krankheiten Tuberkulose und Sarkoidose in ihrer phänotypischen Ausprägung viele Überschneidungen (Kapitel 1.3). Auch konnten Studien bereits zeigen, dass bei beiden Entzündungskrankheiten eine Mutation, wenn auch an verschiedenen Stellen, in dem Gen *BTNL2* zur Prädisposition für die Krankheiten beiträgt (Lian et al. 2010; Valentonyte et al. 2005). Auch in weiteren Entzündungskrankheiten, die ähnliche Entzündungsreaktionen zeigen (Morbus Crohn, Colitis ulcerosa, Lepra), konnten Studien gemeinsame genetische Risikoloci (z.B. *IL23R*) aufdecken (Ali et al. 2013; Duerr et al. 2006; Fischer et al. 2011). Histologische Untersuchungen zeigen, dass die Granulome der Sarkoidose und Tuberkulose einen weitgehend identischen Aufbau besitzen (zusammengefasst in Baughman et al. 2011; Ramakrishnan 2012). Dies und auch der fast identische Befall von Organen (Lunge, Lymphknoten, Leber, Milz, Nervensystem, u.a. (Costabel 2001; Golden and Vikram 2005)) legen die Vermutung nahe, dass die beiden Krankheiten sich weitere gemeinsame Risikoloci auch außerhalb der HLA-Region teilen.

Für die Überprüfung dieser Hypothese stand, neben dem schon für die Sarkoidose-Analyse genutzten GWAS-Datensatz, ein Tuberkulose-GWAS-Datensatz einer ghanaischen Stichprobe zu Verfügung. Daraus ergab sich jedoch ein Problem in der Datenanalyse, da Afrikaner und Europäer eine unterschiedliche genetische Struktur in ihren Genomen besitzen. Studien konnten zeigen, dass es problematisch ist, genetische Signale, die in anderen (z.B. europäischen) Populationen identifiziert wurden, in afrikanischen Populationen zu replizieren (Tarazona-Santos and Tishkoff 2005; Tishkoff et al. 1996). Da in Afrika über einen längeren Zeitraum als in anderen Populationen Gelegenheit zu

Rekombinationsereignissen bestand, haben in afrikanischen Populationen auch deutlich mehr Rekombinationsereignisse stattgefunden (zusammengefasst in Campbell and Tishkoff 2008). Die Ergebnisse einiger Studien deuten darauf hin, dass Rekombinations-Hotspots (genomische Bereiche mit einer erhöhten Rekombinationsrate) populationsspezifisch sind (Clark et al. 2007; Conrad et al. 2006; Crawford et al. 2004), und es wurde nachgewiesen, dass Afrikaner und Afroamerikaner eine höhere Anzahl an Rekombinations-Hotspots als Nicht-Afrikaner besitzen (Clark et al. 2007; Crawford et al. 2004; Hinch et al. 2011). Außerdem konnte gezeigt werden, dass Afrikaner sehr viel kürzere Haplotypblöcke haben und sich auch die haplotypidentifizierenden SNPs (engl. *haplotype tag SNP*, htSNP) stark von anderen Populationen abweichen (Tarazona-Santos and Tishkoff 2005; Tishkoff et al. 1996). Selbst die afrikanischen Subpopulationen unterscheiden sich in dieser Hinsicht, durch die spezifischen demographischen Historien der Subpopulationen in den verschiedenen geographischen Regionen von Afrika, zum Teil stark voneinander (Tishkoff et al. 1996). Daraus ergab sich für diese Arbeit das mögliche Problem, dass ein SNP-Marker, der in dem deutschen GWAS-Datensatz (Sarkoidose-GWAS) ein Assoziationssignal zeigt, nicht zwingend ebenfalls ein Assoziationssignal in dem afrikanischen GWAS-Datensatz (Tuberkulose-GWAS) zeigen muss, obwohl in beiden Populationen der gleiche Locus assoziiert ist. Aufgrund der unterschiedlichen genomischen Varianz in den beiden Populationen besteht eine hohe Wahrscheinlichkeit, dass die kausative Variante in den jeweiligen Populationen von unterschiedlichen htSNPs repräsentiert wird. Um diese Möglichkeit zu berücksichtigen, wurden die beiden GWAS-Datensätze mit vier verschiedenen Methoden analysiert: 1) *meta-analysis fixed effects* 2) *meta-analysis opposite effects* 3) *LD cluster ranking* 4) *gene ranking*.

Neben der „normalen“ Metaanalyse (‚*meta-analysis fixed effects*‘), bei der die beiden GWAS-Datensätze auf in beiden Krankheiten assoziierten SNPs mit gleichem Effekt überprüft wurden (Kapitel 2.2.6.2), wurden die Datensätze noch mit drei weiteren Analysemethoden untersucht.

Die Assoziationssignale eines Markers heben sich in einer Metaanalyse gegenseitig auf, wenn die *odds ratios* in den einzelnen Datensätzen nicht die gleiche Effektrichtung aufweisen. Da der Hauptanteil der untersuchten SNPs wahrscheinlich eher einen genetischen Marker anstatt einer kausativen Variante darstellt und die Möglichkeit besteht, dass in verschiedenen Populationen verschiedene Allele dieser Marker einen positiven Zusammenhang mit den kausativen Varianten zeigen, wurde der Effekt (protektiv oder Risiko erhöhend) eines SNP-Allels auf die Krankheit vernachlässigt. Dazu wurde eine Metaanalyse durchgeführt (‚*meta-analysis opposite effects*‘), bei der das *odds ratio* der Marker des einen Datensatzes (Tuberkulose-GWAS) umgekehrt wurde (siehe Kapitel 2.2.6.2). Mit dieser Analysemethode konnten 13 SNPs identifiziert werden (Tab. 3-11).

Mit dem ‚*LD cluster ranking*‘-Ansatz wurde versucht dem Problem der unterschiedlichen LD-Struktur in beiden Populationen Rechnung zu tragen. Durch die unterschiedliche genomische Struktur ist es

möglich, dass sich assoziierte Marker in den beiden Populationen in ihrer genauen Position unterscheiden, aber trotzdem ein Signal für den gleichen Locus darstellen. Anhand der GWAS-Daten wurden jeweils die SNPs in LD-Gruppen angeordnet und einem „lead"-SNP zugeordnet (siehe Kapitel 2.2.6.2). Die beiden Datensätze wurden nun auf gemeinsame SNPs in diesen LD-Gruppen untersucht und es wurden (kleine) definierte Regionen mit einem Assoziationssignal in beiden Datensätzen detektiert. 25 SNPs in 15 unterschiedlichen Loci konnten mit dieser Analysemethode identifiziert werden (Tab. 3-12).

Mit dem ‚*gene ranking*'-Ansatz soll die biologische Relevanz von Genen im Zusammenhang mit der Krankheitsentstehung berücksichtigt werden. Dieser Ansatz dient also dazu populationsspezifische Mutationen mit unterschiedlichen genomischen Positionen, die in einer großen Distanz zueinander liegen, zu identifizieren (Kapitel 2.2.6.2). Dabei können diese Mutationen trotzdem das gleiche Gen betreffen und somit eine funktionelle Relevanz bei der Krankheitsentstehung (oder aber auch bei der Abgrenzung) von Sarkoidose und Tuberkulose haben. Durch das ‚*gene ranking*' werden Gene identifiziert, die in den jeweiligen Datensätzen mit der Krankheit assoziiert sind, bei denen die assoziierte Mutation sich jedoch aufgrund der unterschiedlichen Populationen an völlig unterschiedlichen Regionen des Gens befindet. Mit dem ‚*gene ranking*' wird ein eher funktioneller Ansatz verfolgt, um gemeinsame Risikogene für die beiden Krankheiten zu detektieren. Mit dieser Analysemethode wurden 8 SNPs in vier verschiedenen Genen identifiziert (Tab. 3-12).

Mit den vier verschiedenen Analysen wurden insgesamt 52 SNPs identifiziert, die potentielle Kandidaten für gemeinsame Risikoloci darstellen. 36 von 52 SNPs konnten dabei mit den drei zusätzlichen Methoden detektiert werden, dies zeigt, dass diese Methoden durchaus ein probates Mittel für die Analyse von genetisch diversen Populationen sein können.

In einer Metaanalyse aller verwendeten Stichproben konnten zwei SNPs (rs8084 und rs745182) mit einem genomweit signifikanten *p*-Wert von $< 5 \times 10^{-8}$ identifiziert werden (Tab. 3-21). Die *p*-Werte dieser beiden SNPs erreichten jedoch in den Tuberkulose-Stichproben E-I, E-II und F keine Signifikanz ($p > 0{,}05$) (Tab. 3-21); diese Stichproben besitzen jedoch jeweils eine Teststärke von < 80%, daher kann nicht vollständig ausgeschlossen werden, dass die beiden Marker echte gemeinsame Risikoloci für Sarkoidose und Tuberkulose darstellen.

Der erste Marker (rs8084) liegt im exonischen Bereich des *HLA-DRA*-Gens in der HLA-Region auf Chromosom 6. Der *p*-Wert in den Tuberkulose-Stichproben E-I und E-II ist nicht signifikant, eine Berechnung der statistischen Teststärke für den Marker in der Stichprobe zeigte jedoch eine Teststärke von nur 42,7%, es existiert also die Möglichkeit, dass eine durchaus vorhandene Assoziation nicht detektiert werden konnte. Die Berechnung der Teststärke in der für die zweite

Replikationsphase genutzten Stichprobe F zeigte eine nur sehr geringe Teststärke für den Marker (Tab. 3-20) und machte deutlich, dass die negativen Ergebnisse somit keine statistische Aussagekraft haben. Die genaue Betrachtung der einzelnen *p*-Werte des Markers zeigt aber auch eine deutlich stärkere Assoziation mit der Krankheit Sarkoidose und der SNP rs8084 (*HLA-DRA*) erreicht alleine in den Sarkoidose-Stichproben eine genomweite Signifikanz ($p = 7{,}63 \times 10^{-21}$). Während die Replikation in der deutschen Stichprobe (Stichprobe C-I) kein signifikantes Ergebnis lieferte, konnte eine signifikante Assoziation in der tschechischen Stichprobe (Stichprobe C-II) bestätigt werden (Tab. 3-17). Die Berechnung der Teststärken der beiden in der zweiten Replikationsphase genutzten Stichproben zeigte, dass beide nur eine Teststärke von ~60% für den Marker besitzen (Tab. 3-19). Das negative Ergebnis in der deutschen Stichprobe kann somit die Assoziation des SNPs nicht widerlegen, da die Wahrscheinlichkeit, die gefundene Assoziation zu replizieren, nur 62,2% beträgt.

Die Ergebnisse stützen daher die bisher in verschiedenen Studien gemachten Beobachtungen einer Assoziation von Sarkoidose mit dem *HLA-DRA*-Locus (Adrianto et al. 2012; Hofmann et al. 2008; Levin et al. 2013). Der SNP rs8084 selbst führt zu einem synonymen Basenaustausch im Exon des HLA-Klasse-II-Histokompatibilitäts-Antigens (engl. *HLA class II histocompatibility antigen*) und hat daher keine direkte Auswirkung auf das gebildete Protein. Die *in silico*-Analysen des SNPs und seiner LD-Marker weisen auf ein hohes regulatorisches Potential für den SNP rs8084 und rs9268659 in der CEU-Population auf (Tab. 3-24). Auch die „*eQTL resources at the Prichard lab*"-Datenbank weist auf ein regulatorisches Potential der beiden SNPs auf diverse in der Region liegende HLA-Gene hin (*HLA-DRA*, *DRB1*, *DRB5*, *DQA1*, *DQA2*, *DQB1*) (Kapitel 7, Abb. 7-4). Die gefundene Assoziation mit dem SNP rs8084 unterstützt die bereits beobachtete genetische Relevanz des *HLA-DRA*-Locus für die Entstehung der Sarkoidose. Die Frage, ob dieser Locus auch im Zusammenhang mit der Entstehung der Tuberkulose steht, kann mit den in dieser Arbeit generierten Ergebnissen nicht weiter geklärt werden.

In anderen Studien konnten allerdings Assoziationen von Tuberkulose mit dem *HLA-DRB1* Allel und den Allelen *DQA1* und *DQB1* nachgewiesen werden (Amirzargar et al. 2004; Magira et al. 2012; Shi et al. 2011) und die *in silico*-Analyse sagt eine eQTL-Funktion der SNPs rs8084 und rs9268659 auf diese HLA-Allele voraus. LD-Strukturen in der HLA-Region sind sehr komplex und zeigen besonders starke Unterschiede in verschiedenen Populationen (Evseeva et al. 2010). Dies macht es sehr schwierig ein für beide Krankheiten zugrunde liegendes Signal in dieser Region genau zu lokalisieren, und es besteht die Möglichkeit, dass der SNP rs8084 in den beiden Populationen jeweils auf unterschiedliche mit den Krankheiten assoziierte Loci hinweist. Der SNP rs8084 konnte auch schon mit weiteren Entzündungskrankheiten in Verbindung gebracht werden (Lee et al. 2012a; Lee et al. 2012b; G. G. Song et al. 2013), doch sind weitere eingehende Untersuchungen sind nötig, um zu

klären, ob dieser Marker auch einen gemeinsamen Risikolocus für Sarkoidose und Tuberkulose darstellt.

Der zweite Marker (rs745182) mit einer genomweiten Signifikanz in der kombinierten Metaanalyse liegt im intronischen Bereich des *PLAC9* Gens auf Chromosom 10. Die Assoziation mit Tuberkulose konnte in den Stichproben E-I, E-II und F nicht bestätigt werden, doch zeigte der SNP zeigte auch in den Sarkoidosestichproben eine genomweite signifikante Assoziation allein mit Sarkoidose (rs745182, p = 1,99 x 10^{-11}, OR 1,29). Da das Gen *PLAC9* in direkter Nachbarschaft zu dem Gen *ANXA11* liegt, welches ein bekanntes Sarkoidose Risikogen ist (Hofmann et al. 2008; Levin et al. 2013), wurde eine logistische Regressionsanalyse mit dem publizierten nsSNP rs1049550 (Hofmann et al. 2008) und mit rs745182 durchgeführt (Tab. 3-22). Nachdem auf den genetischen Effekt von rs1049550 konditioniert wurde, zeigten die Ergebnisse der Analyse für rs745182 einen signifikanten p-Wert (p = 0,0499) bei Nutzung des additiven Modells. Auch das genetische Assoziationssignal von rs1049550 zeigte nach der Konditionierung auf den Effekt von rs745182 weiterhin eine signifikante Assoziation (p = 8.18 x 10^{-3}). Damit konnte mit der Analyse gezeigt werden, dass das in der Sarkoidose detektierte Assoziationssignal von rs745182 unabhängig von dem bereits publizierten nsSNP rs1049550 ist. Die nach der Konditionierung beobachtete Assoziation von rs745182 liegt jedoch sehr nahe an dem Signifikanzschwellenwert von α = 0,05. Weitere Analysen in zusätzlichen Sarkoidosestichproben sollten daher durchgeführt werden, um zu bestätigen, dass der SNP ein neues unabhängiges Assoziationssignal darstellt.

In dieser Arbeit konnte die Assoziation des SNPs rs745182 in den Stichproben C-I und C-II bestätigt werden und wegen der nach der Konditionierung erhaltenen Ergebnisse stellt der SNP somit ein neues genetisches Assoziationssignal für die Sarkoidose dar. Die *in silico*-Analysen zeigten keine erhöhte Methylierung oder Acetylierung im Bereich des Markers rs745182, die NIEHS-Datenbank sagte jedoch ein hohes regulatorisches Potential für den LD-SNP rs2819874 voraus (Tab. 3-25). Eine Betrachtung der SNP-Region im UCSC-Browser zeigte, dass sich in direkter Nachbarschaft des SNPs Bindestellen für diverse Transkriptionsfaktoren befinden (Kapitel 7, Abb. 7-7). Zusätzliche logistische Regressionsanalysen und Feinkartierungen in weiteren Stichproben sind aber nötig, um dieses neue genetische Assoziationssignal für Sarkoidose in der Region einzuengen und potentielle funktionelle Varianten aufzudecken.

Da beide Marker (rs8084 und rs745182) in den Tuberkulose-Stichproben (E-I, E-II und F) keine Assoziation unter dem Schwellenwert (p < 0,05) zeigten, konnte für die SNPs rs8084 und rs745182 keine Assoziation mit Tuberkulose nachgewiesen werden.

Die Marker rs4321884, rs2190733 und rs758362 zeigten in der kombinierten Metaanalyse keine genomweite Signifikanz (Tab. 3-21), jedoch offenbaren die SNPs starke Assoziationen in der

kombinierten Metaanalyse und es ist möglich, dass die beobachteten Signale gemeinsame Suzeptibilitätsloci für Sarkoidose und Tuberkulose repräsentieren. Beide Replikationsstichproben der Tuberkulose weisen eine Teststärke von < 80% auf, daher können die in der Tuberkulose-GWAS vorhandene Assoziationen in den Replikationen nicht zweifelsfrei Identifiziert werden und die erhaltenen Werte haben keine statistisch abgesicherte Aussagekraft. Die Frage eines angemessenen Signifikanzschwellenwertes in genomweiten Assoziationsstudien wurde in der Literatur diskutiert (Clarke et al. 2011; Dudbridge and Gusnanto 2008; Hoggart et al. 2008; Pe'er et al. 2008) und der von Risch und Merikangas postulierte Schwellenwert ($p \leq 5 \times 10^{-8}$) ist der heutzutage am weitesten verbreitete und akzeptierte (N. Risch and Merikangas 1996). Dieser Schwellenwert basiert auf der Bonferroni-Korrektur unter der Annahme von 10^6 unabhängigen Tests, da bei heutigen GWAS-Analysen von einer SNP-Anzahl zwischen 10^5 und 10^6 ausgegangen wird. Wenn eine so hohe Anzahl von unabhängigen Tests durchgeführt wird, ist es natürlich nötig, stringente Schwellenwerte zu nutzen, um falsch-positive Assoziationen zu vermeiden. Die Bonferroni-Korrektur ist eine sehr konservative Methode, um für multiples Testen zu korrigieren. Eine weniger konservative Methode (Benjamini-Hochberg Methode) basiert auf der *False Discovery Rate* (FDR) (Benjamini and Hochberg 1995). Für 2,5 Millionen getestete SNPs bedeutet dies, dass bei einem *p*-Wert von 1×10^{-7} die erwartete Anzahl an falsch-positiven Ergebnissen = oder < 1 ist. Daraus ergibt sich, dass der allgemein akzeptierte Schwellenwert von $p = 5 \times 10^{-8}$ möglicherweise ein wenig zu konservativ ist und Ergebnisse mit einem genomweiten *p*-Wert von $< 10^{-7}$ durchaus echte Assoziationssignale darstellen können (Broer et al. 2013; Gordon et al. 2007).

Es besteht also die Möglichkeit, dass die Marker rs4321884 (nahe *IL6*) und rs2190733 (nahe *IFNG*) gemeinsame Suzeptibilitätsloci für Sarkoidose und Tuberkulose darstellen, da diese stark mit den beiden Krankheiten assoziiert sind und die Assoziationssignale der Marker gleichmäßig in den Stichproben A und D der Krankheiten verteilt sind.

Zwei der Assoziationssignale liegen in der Nachbarschaft von Genen, die für Zytokine codieren. Zytokine spielen eine wichtige Rolle bei immunologischen Reaktionen, damit stellen die hier detektierten Assoziationen sehr interessante Loci für genetische Auslöser der beiden Krankheiten dar.

Der Marker rs4321884 liegt in einer intergenischen Region stromaufwärts des Interleukin-6-Gens (*IL6*). Eine Betrachtung der Region im UCSC-Browser zeigte, dass der SNP in einem aktiven Bereich des Genoms liegt und in direkter Nachbarschaft des SNPs sich Bindestellen für diverse Transkriptionsfaktoren befinden (Kapitel 7, Abb. 7-5). Des Weiteren wird für den Marker ein potentieller regulatorischer Effekt auf ein unbekanntes Gen vorausgesagt (Abb. 7-6) (Montgomery et al. 2010), es kann also spekuliert werden, ob diese Region in der Regulation der Transkription des in

der Nähe liegenden Gens *IL6* eine Rolle spielt. Das IL-6 Protein ist ein proinflammatorisches Zytokin, welches von T-Zellen und Makrophagen abgesondert wird und Immunantworten stimuliert, die zu lokalen Entzündungen und zur Granulomformation beitragen (Heinrich et al. 2003; Nemeth et al. 2011; Tanaka et al. 2012). Auch konnten Studien zeigen, dass IL-6 bei Mäusen an der Kontrolle der Balance von T-Zellen und regulatorischen T-Zellen (T_{Reg}) beteiligt ist (Doganci et al. 2005). Interessanterweise herrscht auch bei Sarkoidosepatienten ein Ungleichgewicht an T_{Reg}-Zellen im Vergleich zu gesunden Kontrollen (Miyara et al. 2006; Taflin et al. 2009). Ebenso wurde auch eine Veränderung der T_{Reg}-Zellpopulation in Tuberkulosepatienten im Vergleich zu gesunden Kontrollen festgestellt (Miyara et al. 2006). Es scheint also, dass eine Veränderung der T_{Reg}-Zellpopulation in enger Verbindung mit Immunantworten steht, die eine Rolle in der Granulomformation spielen. Die Rolle von IL-6 in der Pathogenese der beiden Krankheiten wird durch Beobachtungen einer erhöhten *IL6* Genexpression bei pulmonaler Tuberkulose (Unsal et al. 2005; Y. Zhang et al. 1994) und durch die Feststellung von erhöhten IL-6 Konzentrationen in Lungenproben von Sarkoidosepatienten (Minshall et al. 1997) unterstrichen.

Der SNP rs2190733 liegt in einem intergenischen Bereich auf dem Chromosom 12 stromabwärts des Interferon Gamma Gens (*IFNG*). Laut dem UCSC-Browser liegt der SNP in der Nähe einer TFBS mit erhöhter Methylierung (H3K4Me1) und Acetylierung (H3K27Ac) (Kapitel 7, Abb. 7-8). In der TFBS binden in lymphoblastoiden Zelllinien die für viele Entzündungsmechanismen wichtigen Transkriptionsfaktoren NF-kappaB und PU.1 (TLR4-Antwort). Stromabwärts des Markers befindet sich eine IFNG antisense-RNA (AK124066). Die Funktion dieser antisense-lncRNA (lange nicht-codierende RNA; engl. *long non-coding RNA*) ist bisher nicht bekannt. Mittlerweile wurden viele ncRNAs im menschlichen Genom entdeckt (Derrien et al. 2012; Guttman et al. 2009) und einigen von ihnen konnten funktionelle Rollen in der Regulation der Genexpression auf dem Level der Chromatinmodifikation, Transkription und der posttranskriptionalen Modifizierung nachgewiesen werden (Dinger et al. 2009; Mattick and Gagen 2001; Mercer et al. 2009). Es kann also spekuliert werden, ob diese antisense-lncRNA eine Rolle bei der Regulierung der *IFNG*-Expression spielt. Die sich in der Nähe des SNPs befindende TFBS könnte, aufgrund ihrer räumlichen Nähe, die Transkription der antisense-lncRNA beeinflussen. Diese Vermutungen müssten jedoch durch weiterführende Experimente überprüft werden.

Das Interferon-Gamma (IFN-γ) Zytokin spielt eine wichtige Rolle in Mechanismen der angeborenen und adaptiven Immunantwort und wird von T-Helferzellen, zytotoxischen T-Zellen und natürlichen Killerzellen abgesondert (Schoenborn and Wilson 2007). IFN-γ spielt eine Schlüsselrolle in der Aktivierung zellulärer Immunantworten im Zusammenhang mit der anti-mykobakteriellen Immunität (Flynn et al. 1993; Ottenhoff et al. 2002; van de Vosse et al. 2004). Dies macht IFN-γ zu einem der

essentiellen Zytokine für die Wirtskontrolle der *M. tuberculosis*-Proliferation. Des Weiteren konnte eine Assoziation mit Polymorphismen in dem *IFNG*-Gen (hauptsächlich rs2430561) mit der Suzeptibilität für Tuberkulose in verschiedenen genetischen Studien nachgewiesen werden (Ding et al. 2008; Pacheco et al. 2008; Tso et al. 2005; Vallinoto et al. 2010; J. Wang et al. 2010). Während für Sarkoidose bisher keine genetischen Studien von einer Assoziation mit *IFNG* berichten, konnten allerdings andere Studien zeigen, dass Patienten mit *chronischer* Sarkoidose höhere Serum-Interferon-Gamma-Level haben als gesunde Individuen (Prior and Haslam 1991), obwohl in der Krankheitsätiologie der Sarkoidose bisher der Nachweis für eine Beteiligung von Mykobakterien fehlt. Verschiedene Kombinationen von *IFNG*- und *IL6*-Polymorphismen hängen offenbar mit dem Schutz vor oder der Anfälligkeit für Tuberkulose zusammen (Ansari et al. 2011), was zusätzlich die Bedeutung der beiden Gene in der humanen Immunantwort gegen Mykobakterien unterstreicht. Ein Test auf Epistase bei den Markern rs4321884 und rs2190733 zeigte jedoch keine Interaktion zwischen den beiden SNPs (Tab. 3-23).

Die hier beobachteten Assoziationen der SNPs rs4321884 und rs2190733 (auch wenn diese einige Kilo-Basenpaare stromaufwärts oder stromabwärts der Gene *IFNG* und *IL6* liegen) mit den beiden Krankheiten liefern weitere Hinweise auf eine Beteiligung der Gene oder ihrer regulatorischen Regionen für die Krankheitsentwicklung. Diese Ergebnisse benötigen jedoch zusätzliche Replikationen in weiteren Stichproben. Eine Feinkartierung der beiden Regionen sollte dann zeigen, ob diese Regionen mit der Suszeptibilität für Sarkoidose und Tuberkulose in Verbindung stehen.

Im Rahmen der Analyse von gemeinsamen Risikoloci für Sarkoidose und Tuberkulose wurde auch ein tuberkulosespezifischer Marker identifiziert. Dieser Marker rs225214 liegt in der intronischen Region des *MYO1D*-Gens; zwar erreichte der SNP keine genomweite Signifikanz in den Tuberkulosestichproben, doch deutet die detektierte Assoziation (p = 1,5 x 10^{-6}, OR = 1,19) auf einen möglichen neuen Risikolocus für Tuberkulose hin. In dem Bereich des Markers besteht ein leicht erhöhtes Methylierungsmuster (H3K4Me1) und eine starke Acetylierung (H3K27Ac) in lymphoblastoiden Zellen und in Lungenfibroblasten (Kapitel 7, Abb. 7-9). Die Region ist eine TFBS für eine Reihe von Transkriptionsfaktoren (unter anderem auch PU.1), was auf ein hohes regulatorisches Potential dieser Region hindeutet (Abb. 7-9). Des Weiteren zeigte die Analyse mit der NIEHS *SNPinfo*-Datenbank ein hohes regulatorisches Potential für den Marker und seinen LD-SNP rs225212 (Tab. 3-26). Das unkonventionelle Myosin-1D-Protein (engl. *unconventional myosin-Id*) ist wie andere unkonventionelle Myosine ein Aktin-basiertes Motormolekül mit ATPase-Aktivität und dient intrazellulären Bewegungen. Die stark divergenten Schwanzregionen enthalten Proteininteraktionsdomänen, welche wahrscheinlich Membrankompartimente binden und diese dann an Aktinfilamenten entlang bewegen. Somit kann vermutet werden, dass das unkonventionelle

Myosin 1D Protein eine Rolle bei der Phagozytose des *M. tuberculosis* durch Makrophagen spielt, in dem Review von Ng (2010) wird Myosin 1D zumindest im Zusammenhang mit der Autophagozytose erwähnt. Dieser Zusammenhang ist aber noch rein spekulativ und weitere Forschung zu der Funktion des Myosin 1D ist nötig, um die Funktion dieses Myosins aufzuklären. Die „*eQTL resources browser*"-Datenbank prognostiziert eine regulatorische Funktion des Markers rs225214 auf das Gen *PSMD11* und für rs225212 auf die Gene *PSMD11* und *C17orf79* (Kapitel 7, Abb. 7-10). Das Gen *PSMD11* codiert für die nicht-ATPase-aktive Proteasome-26S-Untereinheit 11 (engl. *proteasome 26S subunit, non-ATPase, 11*), doch welchen funktionellen Zusammenhang dieses Protein mit Tuberkulose haben könnte, ist nicht bekannt. Genetische Studien der Tuberkulose hatten bisher Schwierigkeiten, krankheitsassoziierte Marker mit ausreichender Signifikanz zu identifizieren, was möglicherweise seine Ursache darin hat, dass bei multifaktoriellen Krankheiten die genetischen Komponenten, welche zum Ausbruch der Krankheit führen, durch seltene Suszeptibilitäts-Allele mit geringem Effekt verursacht werden, die an vielen verschiedenen Loci liegen (Di Rienzo 2006; J. K. Pritchard 2001). Der Marker stellt daher einen interessanten Kandidaten-SNP dar, dessen Assoziation in Replikationen mit weiteren Tuberkulosestichproben überprüft werden muss.

4.3 Fazit

Bisher gibt es ein Problem bei der Erklärung der genetischen Ursachen komplexer Krankheiten: Sie weisen häufig eine hohe Erblichkeit (45 - 80%) auf (Barrett et al. 2008; E. Ellinghaus et al. 2010; Harley et al. 2008; Sverrild et al. 2008), aber mit den bisher identifizierten krankheitsverursachenden genetischen Varianten kann nur ein Bruchteil dieser Erblichkeit erklärt werden. Es scheint daher, dass sich bei diesen Krankheiten eine sehr komplexe genetische Architektur zeigt und die beobachteten Phänotypen durch die gemeinsame Aktion sehr vieler Loci mit kleinen Effekten, d.h. meist seltener Varianten herbeigeführt werden (Valdar et al. 2006). Protein-Protein-Interaktionen und Protein-DNA-Interaktionen wirken außerdem in den Transkriptionsnetzwerken und erschweren es, das Genotyp-Phänotyp-Verhältnis einfach über eine Addition von unabhängigen genetischen Effekten zu erklären (zusammengefasst in Beyer et al. 2007; Gerstein et al. 2012; Stamatoyannopoulos 2012). Zusätzlich spielen bei den meisten Krankheiten auch die Gen-Umwelt-Interaktionen eine große Rolle, d.h. bestimmte genetische Variationen können erst unter bestimmten Umwelteinflüssen eine Rolle spielen (Rava et al. 2013). Für komplexe Krankheiten ist es zudem schwer die relevanten Umwelteinflüsse zu definieren und aufzuspüren. Auch ist sehr wenig darüber bekannt, inwiefern und inwieweit Umwelteinflüsse die verschiedenen allelischen Effekte modifizieren. Neu entdeckte Regulationsebenen basierend auf nicht-codierenden microRNAs (miRNAs) oder lncRNAs haben das Potential, der Regulation der Expression von Genen ein neues Level an Komplexität hinzuzufügen

(Derrien et al. 2012; Dinger et al. 2009; Guttman et al. 2009; Mercer et al. 2009). Denn Variationen in Sequenzen, in denen solche miRNAs oder lncRNAs binden und die Genexpression regulieren, können eine große Auswirkung auf die Proteinexpression haben (Nicoloso et al. 2010).

Mit der Methode der GWAS können einzelne genetische Loci identifiziert werden, in denen SNPs eine unterschiedliche Verteilung in Patienten und gesunden Kontrollpersonen zeigen. Die Assoziationen und die genaue Lage in den Regionen müssen bestätigt und dann eingehender auf Mutationen mit Auswirkungen, die im Kontext mit dem untersuchten Phänotyp stehen können, untersucht werden. Wird dabei ein Gen identifiziert, kann dieser Befund als Grundlage für zellbiologische Untersuchungen der krankheitsbedingenden Mechanismen dienen. Da die technischen Möglichkeiten in der Umsetzung von GWAS-Analysen noch nicht vollständig ausgereizt sind, liefern diese Analysen immer noch wichtige Ansatzpunkte in der genetischen Forschung.

4.3.1 Identifizierung eines neuen Sarkoidose-Risikolocus

Die neu identifizierte Assoziation von Sarkoidose mit der Region *11q13.1* macht weitere Untersuchungen nötig, besonders auf funktioneller Ebene. In diesem Locus liegen neun verschiedene Gene, von denen durch Expressionsanalysen und *in silico*-Analysen die Gene *CCDC88B*, *PRDX5* und *KCNK4* als mögliche Risikovarianten identifiziert werden konnten.

Bisher konnte die Assoziation der Region *11q13.1* mit Sarkoidose in deutschen, schwedischen und tschechischen Populationen nachgewiesen werden, doch genetische Analysen in weiteren Populationen (in Nicht-Europäern) müssen klären, ob die Region unabhängig von dem ethnischen Hintergrund für Sarkoidose prädispositioniert.

In weiteren Experimenten mit relevanten Geweben und Zelltypen muss geklärt werden, welches spezifische Gen bei Trägern der SNPs rs479777 und rs671976 in seiner Expression betroffen ist. Ist das Gen identifiziert, wird es möglich für die Krankheit relevante Signaltransduktionswege und Mechanismen aufzudecken. Hierbei würden sich auch Knockout-Experimente in relevanten Zellen (z.B. mittels RNAi) oder Knockout-Mäusen anbieten, um die betroffenen Signaltransduktionsketten und ihre Auswirkungen in „Funktionsverlust-Phänotypen“ (engl. *loss-of-function phenotype*) weiter zu untersuchen.

Weitere Experimente mit Trägern der SNPs, bei denen die umgebende Region der SNPs sequenziert wird, würden die Möglichkeit bieten expressionsverändernde Varianten und ihre genaue Auswirkung auf z.B. Transkriptionsfaktoren zu identifizieren und damit helfen Genregulationsmechanismen dieser Region besser zu verstehen.

4.3.2 Gemeinsame genetische Faktoren der Krankheiten Tuberkulose und Sarkoidose

In der kombinierten Analyse der Tuberkulose- und Sarkoidose-GWAS konnten mehrere Kandidaten-Risikoloci für weiterführende Untersuchungen identifiziert werden. Durch die Kombination zweier GWAS-Datensätze erhöhte sich die Stichprobenanzahl deutlich und es bestand eine größere statistische Teststärke für die Identifizierung von seltenen Varianten, die vielleicht Faktoren für die Entzündungsreaktionen (oder für weitere Mechanismen) darstellen, die beiden Phänotypen zugrunde liegen.

Für zwei Loci in der Nähe der Gene *IL6* und *IFNG* konnten starke Assoziationssignale detektiert werden. Die beobachteten Assoziationen deuten darauf hin, dass diese Loci mit den Krankheiten in Verbindung stehen und damit Kandidaten für weitere Assoziationsstudien darstellen. Ebenso bieten die identifizierten Regionen interessante Ansatzpunkte in dem funktionellen Kontext von Entzündungskrankheiten. *In silico*-Studien konnten zeigen, dass in den Regionen zum Teil eine erhöhte Histonmethylierung und Histonacetylierung herrscht und in der Nähe der SNPs Transkriptionsfaktorbindestellen liegen. Wobei die assoziierten SNPs sehr wahrscheinlich nur genetische Marker darstellen, in deren Regionen durch Feinkartierungen mögliche kausative Varianten erst identifiziert werden müssen.

Im Zuge der Analysen konnte auch eine neue Assoziation von Sarkoidose mit der Region *10q22.3* festgestellt werden, zusätzlich konnte ein Marker im *MYO1D*-Gen identifiziert werden, der eine Assoziation mit der Krankheit Tuberkulose, nicht aber mit Sarkoidose zeigte.

Die durchgeführten Berechnungen deuten auf eine neue Assoziation der Krankheit Sarkoidose mit der Region *10q22.3* hin. Doch sollten die Ergebnisse in weiteren Sarkoidosestichproben auf Unabhängigkeit von der in unmittelbarer Nachbarschaft liegenden *ANXA11*-Asssoziation überprüft werden. Außerdem sollte eine genaue Feinkartierung der Region durchgeführt werden, um eventuell vorhandene kausative Varianten zu identifizieren.

Der tuberkulosespezifische Marker im *MYO1D*-Gen zeigt in den beiden ghanaischen Stichproben einen vielversprechenden *p*-Wert, auch wenn er keine genomweite Signifikanz erreicht, und stellt damit einen interessanten Kandidaten-Locus für die Krankheit Tuberkulose dar. Eine Replikation dieses Ergebnisses in weiteren Tuberkulosestichproben wird zeigen, ob sich die gefundene Assoziation bestätigt.

Daran anschließend ist nun nötig in Folgeexperimenten die gefundenen Assoziationen in weiteren Stichproben zu replizieren, um die Ergebnisse auf diesem Wege zu bestätigen oder zu widerlegen. Da weltweit die Verfügbarkeit von umfangreichen Tuberkulosestichproben bisher noch sehr begrenzt ist und für eine Replikation sehr wahrscheinlich auf andere Populationen als die ghanaische zurückgegriffen werden muss, wäre es sinnvoll, für eine Replikation eine Art Feinkartierung der

identifizierten Regionen vorzunehmen, da aufgrund der zu erwartenden genetischen Variation zwischen den Populationen eine Verschiebung des Assoziationssignals möglich ist. Des Weiteren sollte sich daran eine „normale“ Feinkartierung der Regionen anschließen, um die den Krankheiten zugrunde liegenden kausativen Varianten zu identifizieren.

Ersatzweise bietet sich auch eine Feinkartierung der Regionen, in den in dieser Arbeit verwendeten Stichproben, an. Auf diesem Wege könnten vielleicht kausative Varianten identifiziert werden, die mit den detektierten *tagging*-SNPs nur in einem schwachen LD stehen und daher deren schwache Assoziation erklären, da die kausativen Varianten von diesen nur unzureichend repräsentiert wurden.

4.3.3 Ausblick

Die Erforschung der genetischen Architektur von Krankheiten ist durch die Entwicklung der GWAS weit vorangeschritten, doch bei der Identifizierung seltener Varianten hat diese Technik auch ihre Grenzen. Während für eine sehr hohe Anzahl an häufigen Varianten eine Assoziation mit Krankheiten oder anderen Phänotypen nachgewiesen werden konnte (zusammengefasst in J. Hardy and Singleton 2009; Manolio et al. 2009), tut sich die GWAS schwer Assoziationen seltener Varianten zu identifizieren. Mit Hilfe großer und umfangreicher Stichproben kann das Problem, dass mittels Assoziationsstudien Varianten mit moderatem Risiko und niederfrequente Varianten bisher schwierig zu detektieren sind, teilweise umgangen werden, da der große Stichprobenumfang ein teilweises Detektieren von Assoziationen der oben genannten Varianten zulässt. Doch erst die Entwicklung neuer Techniken wie Exom- oder Genomsequenzierung, Arrays, mit denen auch auf Insertionen und Deletionen getestet werden kann, und die Weiterentwicklung der Technik der Imputation werden helfen, solche genetischen Varianten zu identifizieren.

5 Zusammenfassung

Es ist bekannt, dass die granulomatösen Lungenkrankheiten Tuberkulose und Sarkoidose neben anderen Auslösern auch genetische Ursachen haben. Die genetische Ätiologie dieser beiden Lungenkrankheiten ist bislang jedoch nicht vollständig geklärt. Mit zwei unabhängige Studien die im Rahmen dieser Arbeit durchgeführt wurden, sollten weitere prädisponierende genetische Faktoren entzündlicher granulomatöser Lungenkrankheiten identifiziert werden:

Identifizierung von Risikoloci der Sarkoidose und ihrer Subphänotypen

In einer genomweiten Assoziationsstudie wurden 1.294.967 SNPs in 564 Sarkoidosepatienten und 1.575 gesunden Kontrollpersonen mit Hilfe der Methode der Imputation genotypisiert. Zusätzlich wurde getestet, ob SNPs subphänotypspezifische Assoziationen zeigen. Von den 30 SNPs mit der höchsten Signifikanz, die in einer größeren unabhängigen Stichprobe genotypisiert wurden, konnte die Assoziation eines SNPs (rs479777) mit Sarkoidose in einer genreichen Region auf Chromosom *11q13.1* validiert und in weiteren unabhängigen europäischen Stichproben repliziert werden. Die durchgeführte Feinkartierung der Region führte zur Identifikation eines weiteren noch stärker assoziierten SNPs (rs671976). *In silico*-Analysen der SNPs mit der höchsten Signifikanz der Region identifizierten die Gene *CCDC88B*, *KCNK4* und *PRDX5* als mögliche Risikofaktoren der Sarkoidose. Expressionsanalysen zeigten eine erhöhte *CCDC88B*-Expression für homozygote Träger des rs671976-Risikoallels.

Mit Hilfe einer imputieren genomweiten Sarkoidose-Assoziationsstudie konnte ein neuer Sarkoidose-Suszeptibilitätslocus auf Chromosom *11q13.1* identifiziert werden. Erste Analysen deuten auf das Gen *CCDC88B* als Sarkoidose-Risikofaktor hin, doch weitere funktionelle Studien der Gene in diesem Locus müssen deren genaue Rolle bei der Pathogenese klären.

Identifizierung gemeinsamer genetischer Faktoren der Krankheiten Tuberkulose und Sarkoidose

Durch die Kombination einer imputierten genomweiten Tuberkulose- und Sarkoidose-Assoziationsstudie konnten 949.988 SNPs in 564 Sarkoidosepatienten, 1.288 Tuberkulosepatienten und 3.365 gesunden Kontrollpersonen analysiert werden. Mit vier Analysemethoden wurden insgesamt 52 SNPs identifiziert, die am stärksten mit den beiden Krankheiten assoziiert sind. In größeren unabhängigen Stichproben wurden diese SNPs erneut genotypisiert und die vier SNPs mit der höchsten Signifikanz als Kandidaten für die zweite Replikationsphase in weiteren unabhängigen Stichproben ausgewählt. Des Weiteren konnte in den Tuberkulose-Daten in dem *MYO1D*-Gen ein SNP als Kandidat für eine tuberkulosespezifische Assoziation identifiziert werden. Auf Grund zu geringer statistischer Teststärken konnten die erste und die zweite Replikationsphase teilweise keine der Assoziationen bestätigen oder widerlegen. Mittels *in silico*-Analysen konnten zwei Kandidaten-

SNPs in den Bereichen der Gene *IL6* und *IFNG* als mögliche Suszeptibilitätsloci für Tuberkulose und Sarkoidose identifiziert werden. Zusätzlich wurde eine potentielle Assoziation mit Sarkoidose im Chromosom *10q22.3* und eine mögliche Assoziation mit Tuberkulose im *MYO1D*-Gen identifiziert.
Weitere Studien in zusätzlichen Stichproben müssen klären, ob die hier beobachteten vielversprechenden Assoziationen Risikoloci für Tuberkulose und Sarkoidose darstellen.

6 Summary

Amongst other causes it is known that genetic factors contribute to the development of the granulomatous lung diseases tuberculosis and sarcoidosis. So far the genetic etiology of these two lung diseases has not been fully elucidated. In order to identify novel susceptibility loci for diseases with granulomatous inflammation, two independent studies were carried out in context of this thesis:

Identification of risk loci for sarcoidosis and its subphenotypes

After carrying out imputation in a genome-wide association study, 1,294,967 SNPs were successfully genotyped in 564 sarcoidosis cases and 1,575 controls. Additionally, all SNPs were screened for subphenotype specific associations. Validation of the top 30 significant SNPs in a large independent panel confirmed association of one SNP (rs479777) with sarcoidosis. This association was successfully replicated in independent European panels further confirming its status. The SNP is located in a gene dense region on chromosome *11q13.1* and fine mapping of the region revealed the strongest association at rs671976. *In silico* analyses of the most significant associated SNPs identified the genes *CCDC88B*, *KCNK4*, and *PRDX5* as potential risk factors for sarcoidosis. Performed expression analyses showed elevated *CCDC88B* expression in patients homozygous for the rs671976 risk allele.

Imputation of a genome-wide sarcoidosis association study identified *11q13.1* as a novel risk locus for sarcoidosis. Initial studies link *CCDC88B* to be the underlying risk factor, but further functional studies of the individual genes at this locus are required to clarify their role in the pathogenesis of sarcoidosis.

Identification of shared genetic factors for tuberculosis and sarcoidosis

By combining an imputed tuberculosis GWAS dataset and an imputed sarcoidosis GWAS dataset, 949,988 SNPs, genotyped in 564 sarcoidosis patients, 1,288 tuberculosis patients, and 3,365 controls, were available for analysis. Four different methods identified a total of 52 top significant SNPs. These SNPs were genotyped in additional large independent panels, identifying four most significant associated SNPs as candidates for a second replication phase in further independent panels. Additionally, in the tuberculosis data a SNP in the *MYO1D* gene was identified as a candidate for a tuberculosis specific association. Due to insufficient statistical power of the second and in part of the first replication phase the observed associations could not be refuted. *In silico* analyses identified two candidate SNPs in the areas of the genes *IL6* and *IFNG* as possible susceptibility loci for tuberculosis and sarcoidosis. Additionally, a potential genetic association between sarcoidosis and *10q22.3* and a possible association of *MYO1D* with tuberculosis were identified.

It is necessary to perform further studies in additional panels to confirm whether the identified associations are true risk loci for tuberculosis and sarcoidosis.

7 Anhang

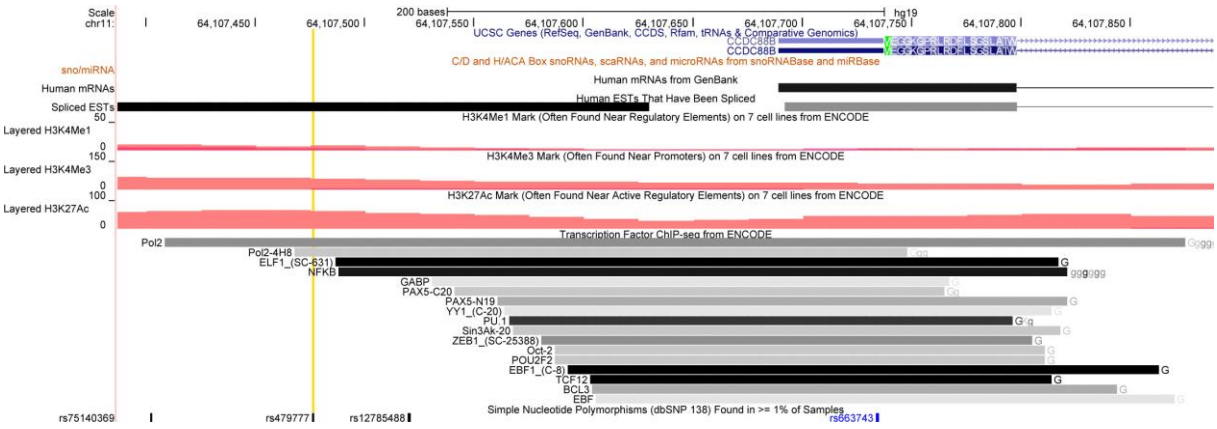

Abb. 7-1: Grafische Übersicht der Region des SNPs rs479777 in der UCSC-Datenbank. In der UCSC-Datenbank sind weitere Datenbanken grafisch integriert, wie z.B. die ENCODE-Datenbank, die für die *in silico*-Analyse der Region auf regulatorische Elemente genutzt wurde. Die Position des SNPs rs479777 ist mit einer gelben Linie markiert. Von oben nach unten sind folgende Elemente zu sehen: Skala und Position des betrachteten Ausschnitts in bp auf dem Chromosom (Die Position bezieht sich auf das *human genome build 19*). „*UCSC Genes*“: Auf verschiedenen Datenbanken basierende Positionen verschiedener Gen-Transkripte. „*snoRNABase and miRBase*“: Datenbank, die vier verschiedene RNA-Typen (pre-miRNA, C/D box snoRNA, H/ACA box snoRNA und scaRNA) und ihre Position im Genom darstellt. „*Human mRNAs from GenBank*“: Diese Datenbank zeigt die Anordnung von mRNAs im Genom. „*Human ESTs That Have Been Spliced*“: Diese Datenbank zeigt transkribierte Bereiche, in denen Spleißen stattfindet. „*ENCODE*“: Diese Datenbank sammelt Informationen, die im Rahmen der Regulation der Transkription relevant sind. Hier wurden ausgewertet: Histonmethylierung und -acetylierung (H3K4Me1, H3K4Me3, H3K27Ac) für eine Lymphoblastoide- und Lungenfibroblasten-Zelllinie; die Transkriptionsfaktor-ChIP-Seq zeigt die Bindung von Transkriptionsfaktoren in verschiedenen Zelllinien an (Lymphoblastoide-Zelllinien: G und g; Lungenfibroblasten-Zelllinien: n; K563-Zelllinien: K); „*Simple Nucleotide Polymorphisms (dbSNP 138)*“: SNP- und Indel-Daten der dbSNP-Datenbank *build 138* (SNP-Farbgebung: synonymer SNP = grün; nicht-synonymer SNP = rot; SNP in einer Spleißstelle = rot; SNP im untranslatierten Bereich = blau; intronische und unbekannte SNPs = schwarz).

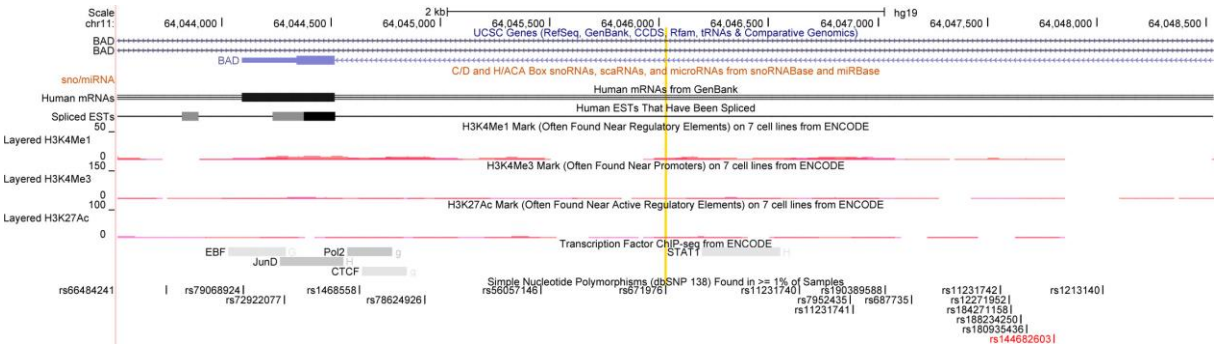

Abb. 7-2: Grafische Übersicht der Region des SNPs rs671976 in der UCSC-Datenbank. In der UCSC-Datenbank sind weitere Datenbanken grafisch integriert, wie z.B. die ENCODE-Datenbank, die für die *in silico*-Analyse der Region auf regulatorische Elemente genutzt wurde. Die Position des SNPs rs671976 ist mit einer gelben Linie markiert. Von oben nach unten sind folgende Elemente zu sehen: Skala und Position des betrachteten Ausschnitts in bp auf dem Chromosom (Die Position bezieht sich auf das *human genome build 19*). „*UCSC Genes*“: Auf verschiedenen Datenbanken basierende Positionen verschiedener Gen-Transkripte. „*snoRNABase and miRBase*“: Datenbank, die vier verschiedene RNA-Typen (pre-miRNA, C/D box snoRNA, H/ACA box snoRNA und scaRNA) und ihre Position im Genom darstellt. „*Human mRNAs from GenBank*“: Diese Datenbank zeigt die Anordnung von mRNAs im Genom. „*Human ESTs That Have Been Spliced*“: Diese Datenbank zeigt transkribierte Bereiche, in denen Spleißen stattfindet. „*ENCODE*“: Diese Datenbank sammelt Informationen, die im Rahmen der Regulation der Transkription relevant sind. Hier wurden ausgewertet: Histonmethylierung und -acetylierung (H3K4Me1, H3K4Me3, H3K27Ac) für eine Lymphoblastoide- und Lungenfibroblasten-Zelllinie; die Transkriptionsfaktor-ChIP-Seq zeigt die Bindung von Transkriptionsfaktoren in verschiedenen Zelllinien an (Lymphoblastoide-Zelllinien: G und g; Lungenfibroblasten-Zelllinien: n; K563-Zelllinien: K); „*Simple Nucleotide Polymorphisms (dbSNP 138)*“: SNP- und Indel-Daten

der dbSNP-Datenbank *build 138* (SNP-Farbgebung: synonymer SNP = grün; nicht-synonymer SNP = rot; SNP in einer Spleißstelle = rot; SNP im untranslatierten Bereich = blau; intronische und unbekannte SNPs = schwarz).

Tab. 7-1: Allelfrequenzen und 95% Konfidenzintervall für die 30 SNPs mit den höchsten Rängen aus den beiden Metaanalysen der GWAS-Daten. Der erste Abschnitt der Tabelle zeigt die 17 Marker, die mit *meta-analysis fixed effects* identifiziert wurden (Meta. f.), der zweite Abschnitt die 13 Marker mit *meta-analysis opposite effects* (Meta. o.). Für die Tuberkulose-SNPs wurden in der Tabelle immer die „originalen" ORs und Allelfrequenzen aus den GWAS-Daten dargestellt. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind in den Kontrollen und Fällen die Allelfrequenzen, das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; OR = Quotenverhältnis (engl. *odds ratio*).

Meta. f.	**Sarkoidose-GWAS**					**Tuberkulose-GWAS**				
dbSNP ID	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**
rs1732581	T	0,3825	0,3177	$6{,}66 \times 10^{-04}$	0,75 [0,63-0,88]	T	0,0351	0,0235	$3{,}36 \times 10^{-04}$	0,45 [0,29-0,69]
rs2621483	A	0,1157	0,1516	$7{,}76 \times 10^{-03}$	1,37 [1,09-1,72]	A	0,1841	0,2103	$5{,}12 \times 10^{-04}$	1,34 [1,14-1,57]
rs11584687	G	0,4578	0,4214	$3{,}59 \times 10^{-03}$	0,79 [0,67-0,93]	G	0,414	0,3754	$1{,}10 \times 10^{-03}$	0,84 [0,76-0,93]
rs11126681	G	0,4346	0,3998	$1{,}77 \times 10^{-03}$	0,77 [0,66-0,91]	G	0,143	0,1165	$3{,}72 \times 10^{-03}$	0,79 [0,67-0,93]
rs1471921	C	0,1838	0,1507	$6{,}28 \times 10^{-03}$	0,74 [0,60-0,92]	T	0,3147	0,3518	$1{,}04 \times 10^{-03}$	0,83 [0,75-0,93]
rs9915508	T	0,5007	0,4638	$4{,}67 \times 10^{-03}$	0,76 [0,63-0,92]	G	0,2652	0,2903	$2{,}26 \times 10^{-03}$	1,28 [1,09-1,49]
rs2505202	C	0,4888	0,4344	$3{,}22 \times 10^{-03}$	0,79 [0,67-0,92]	C	0,4976	0,4619	$2{,}74 \times 10^{-03}$	0,84 [0,75-0,94]
rs3181200	T	0,4676	0,4184	$5{,}61 \times 10^{-03}$	0,80 [0,69-0,94]	G	0,1288	0,1611	$1{,}99 \times 10^{-03}$	1,27 [1,08-1,45]
rs10850905	G	0,3337	0,3036	$5{,}11 \times 10^{-03}$	0,73 [0,59-0,91]	G	0,1554	0,132	$2{,}82 \times 10^{-03}$	0,73 [0,59-0,90]
rs11583306	T	0,254	0,2074	$1{,}54 \times 10^{-03}$	0,74 [0,61-0,89]	T	0,457	0,4169	$3{,}44 \times 10^{-03}$	0,86 [0,77-0,95]
rs758362	A	0,2762	0,322	$4{,}79 \times 10^{-03}$	1,28 [1,08-1,51]	A	0,452	0,4891	$2{,}48 \times 10^{-03}$	1,19 [1,06-1,34]
rs10931252	T	0,2108	0,1782	$5{,}76 \times 10^{-03}$	0,76 [0,62-0,92]	T	0,1464	0,1219	$2{,}65 \times 10^{-03}$	0,79 [0,69-0,93]
rs2300955	A	0,1829	0,148	$3{,}03 \times 10^{-03}$	0,72 [0,58-0,90]	A	0,1247	0,0919	$5{,}64 \times 10^{-03}$	0,80 [0,69-0,93]
rs9527209	A	0,201	0,1711	$3{,}54 \times 10^{-03}$	0,74 [0,60-0,91]	A	0,0221	0,0113	$2{,}61 \times 10^{-03}$	0,51 [0,33-0,79]
rs3181374	G	0,4723	0,4249	$5{,}83 \times 10^{-03}$	0,80 [0,69-0,94]	A	0,1293	0,1584	$5{,}77 \times 10^{-03}$	0,81 [0,70-0,94]
rs1548255	C	0,4384	0,3857	$4{,}96 \times 10^{-03}$	0,79 [0,67-0,93]	C	0,4039	0,3737	$5{,}42 \times 10^{-03}$	0,85 [0,75-0,95]
rs8080480	T	0,3505	0,3928	$1{,}80 \times 10^{-03}$	1,29 [1,10-1,52]	C	0,455	0,4227	$8{,}61 \times 10^{-03}$	1,15 [1,03-1,28]
Meta. o.	**Sarkoidose-GWAS**					**Tuberkulose-GWAS**				
dbSNP ID	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**
rs7194	G	0,1782	0,2108	$1{,}00 \times 10^{-05}$	1,44 [1,22-1,61]	G	0,3208	0,3054	$2{,}83 \times 10^{-03}$	0,85 [0,69-0,99]
rs8084	A	0,44	0,5407	$3{,}44 \times 10^{-05}$	1,41 [1,20-1,66]	A	0,4675	0,4508	$8{,}48 \times 10^{-03}$	0,86 [0,78-0,97]
rs6884494	A	0,105	0,0717	$1{,}69 \times 10^{-04}$	0,57 [0,42-0,76]	A	0,3303	0,3622	$1{,}72 \times 10^{-03}$	1,23 [1,10-1,37]
rs745182	C	0,4151	0,5032	$2{,}01 \times 10^{-04}$	1,36 [1,16-1,60]	T	0,425	0,3978	$9{,}61 \times 10^{-03}$	0,85 [0,74-0,96]
rs11659929	C	0,3267	0,3688	$9{,}68 \times 10^{-04}$	1,32 [1,12-1,56]	C	0,4682	0,4301	$2{,}96 \times 10^{-03}$	0,85 [0,78-0,98]
rs4321884	A	0,4911	0,5231	$5{,}01 \times 10^{-03}$	1,25 [1,07-1,46]	A	0,3503	0,3108	$1{,}92 \times 10^{-03}$	0,84 [0,73-0,93]
rs778199	T	0,3352	0,3626	$1{,}97 \times 10^{-03}$	1,29 [1,10-1,52]	T	0,2589	0,2298	$4{,}54 \times 10^{-03}$	0,84 [0,78-0,91]
rs778193	T	0,4184	0,4468	$4{,}11 \times 10^{-03}$	1,26 [1,07-1,47]	T	0,2827	0,2569	$3{,}21 \times 10^{-03}$	0,82 [0,72-0,86]
rs1390488	T	0,1444	0,1741	$1{,}20 \times 10^{-03}$	1,42 [1,14-1,75]	C	0,4881	0,452	$3{,}33 \times 10^{-03}$	0,85 [0,52-0,98]
rs6444661	T	0,4906	0,4613	$3{,}23 \times 10^{-03}$	0,78 [0,65-0,91]	T	0,1438	0,159	$5{,}75 \times 10^{-03}$	1,35 [1,14-1,47]
rs11948804	T	0,0932	0,064	$2{,}14 \times 10^{-04}$	0,54 [0,39-0,74]	T	0,2315	0,2544	$7{,}08 \times 10^{-03}$	1,23 [1,05-1,39]
rs566353	T	0,4387	0,4828	$3{,}16 \times 10^{-03}$	1,27 [1,08-1,49]	T	0,0816	0,0643	$7{,}16 \times 10^{-03}$	0,74 [0,68-0,90]
rs9480957	A	0,3098	0,3502	$1{,}26 \times 10^{-03}$	1,31 [1,11-1,54]	A	0,2869	0,2601	$1{,}18 \times 10^{-02}$	0,86 [0,79-0,97]

Tab. 7-2: Allelfrequenzen und 95% Konfidenzintervall für die 33 SNPs mit den höchsten Rängen aus der *LD cluster ranking*-Analyse und der *gene ranking*-Analyse. Der erste Abschnitt der Tabelle zeigt die 25 Marker, die mit der *LD cluster ranking*-Analyse identifiziert wurden, der zweite Abschnitt die 8 Marker der *gene ranking*-Analyse. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind in den Kontrollen und Fällen die Allelfrequenzen, das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; OR = Quotenverhältnis (engl. *odds ratio*).

LD cluster ranking											
Sarkoidose						**Tuberkulose**					
dbSNP ID	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**	**dbSNP ID**	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**
rs6919855	C	0,344	0,4089	$1{,}02 \times 10^{-03}$	1,33 [1,12-1,58]	rs9268880	T	0,2786	0,3174	$6{,}32 \times 10^{-05}$	1,21 [1,13-1,43]
rs1732581	T	0,3825	0,3177	$6{,}66 \times 10^{-04}$	0,75 [0,63-0,88]	rs1732581	T	0,0351	0,0235	$3{,}36 \times 10^{-04}$	0,45 [0,29-0,70]
rs11844637	C	0,4883	0,4326	$7{,}53 \times 10^{-04}$	0,77 [0,65-0,89]	rs12431541	A	0,2253	0,2577	$1{,}24 \times 10^{-03}$	1,22 [1,08-1,38]
rs6884494	A	0,105	0,0717	$1{,}69 \times 10^{-04}$	0,57 [0,42-0,72]	rs6884494	A	0,3303	0,3622	$1{,}72 \times 10^{-03}$	1,24 [1,08-1,42]
rs1740736	G	0,4857	0,4323	$2{,}72 \times 10^{-03}$	0,78 [0,67-0,92]	rs2505202	C	0,4976	0,4619	$2{,}74 \times 10^{-03}$	0,84 [0,75-0,94]
rs7194	G	0,1782	0,2108	$1{,}00 \times 10^{-05}$	1,44 [1,22-1,61]	rs7194	G	0,3054	0,3208	$2{,}83 \times 10^{-03}$	1,17 [1,01-1,44]
rs7148429	A	0,3992	0,4402	$1{,}82 \times 10^{-03}$	1,63 [1,20-1,78]	rs8007404	A	0,3203	0,3483	$2{,}89 \times 10^{-03}$	1,21 [1,07-1,37]
rs11659929	C	0,3267	0,3688	$9{,}68 \times 10^{-04}$	1,32 [1,12-1,56]	rs11659929	C	0,4682	0,4301	$2{,}96 \times 10^{-03}$	0,85 [0,77-0,95]
rs778198	T	0,4137	0,4459	$2{,}68 \times 10^{-03}$	1,27 [1,09-1,49]	rs778193	T	0,2827	0,2569	$3{,}21 \times 10^{-03}$	0,81 [0,71-0,93]
rs4845741	A	0,145	0,1761	$7{,}91 \times 10^{-04}$	1,44 [1,16-1,78]	rs1390488	C	0,452	0,4881	$3{,}33 \times 10^{-03}$	1,18 [1,06-1,32]
rs12304940	C	0,1789	0,1463	$1{,}38 \times 10^{-03}$	0,70 [0,57-0,87]	rs2190733	C	0,2857	0,2562	$3{,}39 \times 10^{-03}$	0,84 [0,75-0,94]
rs11580170	T	0,2837	0,2341	$1{,}12 \times 10^{-03}$	0,74 [0,62-0,89]	rs11583306	T	0,457	0,4169	$3{,}44 \times 10^{-03}$	0,86 [0,77-0,95]
rs828270	C	0,4327	0,3972	$1{,}61 \times 10^{-03}$	0,77 [0,66-0,91]	rs11126681	G	0,143	0,1165	$3{,}72 \times 10^{-03}$	0,79 [0,68-0,93]
rs10505536	A	0,3354	0,3896	$9{,}48 \times 10^{-04}$	1,32 [1,12-1,56]	rs2099292	T	0,1051	0,1263	$3{,}83 \times 10^{-03}$	1,27 [1,08-1,49]
rs778199	T	0,3352	0,3626	$1{,}97 \times 10^{-03}$	1,30 [1,10-1,52]	rs778199	T	0,2589	0,2298	$4{,}54 \times 10^{-03}$	0,84 [0,75-0,95]
Gene ranking											
Sarkoidose						**Tuberkulose**					
dbSNP ID	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**	**dbSNP ID**	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**
rs1960574	T	0,427	0,3715	$8{,}05 \times 10^{-04}$	0,76 [0,64-0,89]	rs1029074	A	0,6127	0,5623	$5{,}33 \times 10^{-04}$	0,79 [1,12-1,58]
rs747088	T	0,204	0,1180	$8{,}65 \times 10^{-04}$	0,64 [0,53-0,88]	rs17057784	A	0,3421	0,3956	$5{,}64 \times 10^{-05}$	1,25 [1,12-1,39]
rs1560064	A	0,1378	0,1693	$9{,}06 \times 10^{-04}$	1,45 [1,16-1,80]	rs2913633	T	0,3327	0,3773	$2{,}29 \times 10^{-04}$	1,23 [1,10-1,37]
rs2519866	C	0,3424	0,2939	$3{,}91 \times 10^{-04}$	0,69 [0,57-0,85]	rs225214	T	0,4031	0,4496	$3{,}65 \times 10^{-05}$	1,25 [1,12-1,38]

Tab. 7-3: Allelfrequenzen der Kandidaten-SNPs in den Replikationsstichproben C-I und C-II. Nominell signifikante Ergebnisse (*p*-Wert < 0,05) sind fett hervorgehoben. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind in den Kontrollen und Fällen die Allelfrequenzen, das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; OR = Quotenverhältnis (engl. *odds ratio*); Repl.II = zweite Replikationsphase.

Zweite Replikationsphase Sarkoidose (Repl.II)		Stichprobe C-I				Stichprobe C-II			
dbSNP ID	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**
rs8084	A	0.4211	0.4701	$9{,}87 \times 10^{-02}$	1,22 [0,96-1,54]	**0.4167**	**0.5079**	$\mathbf{2{,}29 \times 10^{-03}}$	**1,45 [1,14-1,83]**
rs4321884	A	0.4946	0.5035	$7{,}65 \times 10^{-01}$	1,04 [0,82-1,31]	0.5033	0.4705	$2{,}74 \times 10^{-01}$	0,88 [0,69-1,11]
rs745182	C	**0.4014**	**0.4613**	$\mathbf{4{,}27 \times 10^{-02}}$	**1,28 [1,01-1,62]**	**0.433**	**0.5394**	$\mathbf{3{,}90 \times 10^{-04}}$	**1,53 [1,21-1,94]**
rs2190733	C	0.1685	0.1972	$2{,}13 \times 10^{-01}$	1,21 [0,90-1,64]	0.1699	0.1476	$3{,}11 \times 10^{-01}$	0,85 [0,61-1,17]

Tab. 7-4: Allelfrequenzen der Kandidaten-SNPs in der Replikationsstichprobe F. A1 bezeichnet das seltenere Allel des SNPs in den Kontrollen. Für das seltenere Allel sind in den Kontrollen und Fällen die Allelfrequenzen, das allelische *odds ratio* und das 95% Konfidenzintervall angegeben. Abkürzungen: AF = Allelfrequenz; CI = Konfidenzintervall (engl. *confidence interval*); Kont. = Kontrollen; OR = Quotenverhältnis (engl. *odds ratio*); Repl.II = zweite Replikationsphase.

Zweite Replikationsphase Tuberkulose (Repl.II)		Stichprobe F			
dbSNP ID	**A1**	**AF Kontr.**	**AF Fälle**	***p*-Wert**	**OR [95% CI]**
rs8084	A	0.4113	0.4223	$6{,}64 \times 10^{-01}$	0,96 [0,85-1,28]
rs4321884	A	0.2164	0.2591	$5{,}12 \times 10^{-02}$	0,79 [1,00-1,61]
rs745182	C	0.4005	0.408	$7{,}66 \times 10^{-01}$	1,03 [0,84-1,27]
rs2190733	C	0.1667	0.1554	$5{,}52 \times 10^{-01}$	0,92 [0,70-1,21]
rs225214	T	0.4086	0.3899	$4{,}57 \times 10^{-01}$	0,93 [0,75-1,14]

Tab. 7-5: Assoziationsergebnisse der Kandidaten-SNPs der kombinierten Metaanalyse aller Stichproben beider Krankheiten. Die *p*-Werte der Kandidaten-SNPs sind nach Krankheiten sortiert für die jeweilige Stichprobe abgebildet. Die Spalte p_{META} zeigt die *p*-Werte der Metaanalyse aller verwendeten Stichproben. Abkürzungen: e = exonisch; i = intronisch; Repl.I = erste Replikationsphase; Repl.II = zweite Replikationsphase; SA = Sarkoidose; TB = Tuberkulose; us = stromaufwärts (engl. *upstream*).

	SA								TB						
dbSNP ID	**Stichprobe A**		**Stichprobe B**		**Stichprobe C-I**		**Stichprobe C-II**		**Stichprobe D**		**Stichproben E-I & E-II**		**Stichprobe F**		p_{META}
	p_{GWAS}	**OR**	$p_{Repl.I}$	**OR**	$p_{Repl.II}$	**OR**	$p_{Repl.II}$	**OR**	p_{GWAS}	**OR**	$p_{Repl.I}$	**OR**	$p_{Repl.II}$	**OR**	
rs8084	$3{,}44 \times 10^{-05}$	1,41	$6{,}90 \times 10^{-15}$	1,46	$9{,}87 \times 10^{-02}$	1,22	$2{,}29 \times 10^{-03}$	1,45	$8{,}48 \times 10^{-03}$	0,87	$6{,}47 \times 10^{-01}$	1,03	$6{,}64 \times 10^{-01}$	0,96	$2{,}52 \times 10^{-16}$
rs4321884	$5{,}01 \times 10^{-03}$	1,25	$6{,}67 \times 10^{-02}$	1,09	$7{,}65 \times 10^{-01}$	1,04	$2{,}74 \times 10^{-01}$	0,88	$1{,}92 \times 10^{-03}$	0,84	$4{,}42 \times 10^{-01}$	0,95	$5{,}12 \times 10^{-02}$	0,79	$1{,}04 \times 10^{-05}$
rs745182	$2{,}01 \times 10^{-04}$	1,36	$2{,}07 \times 10^{-05}$	1,23	$4{,}27 \times 10^{-02}$	1,28	$3{,}90 \times 10^{-04}$	1,53	$9{,}61 \times 10^{-03}$	0,85	$9{,}75 \times 10^{-01}$	0,99	$7{,}66 \times 10^{-01}$	1,03	$1{,}26 \times 10^{-09}$
rs2190733	$3{,}51 \times 10^{-03}$	0,73	$2{,}90 \times 10^{-03}$	0,83	$2{,}13 \times 10^{-01}$	1,21	$3{,}11 \times 10^{-01}$	0,85	$9{,}61 \times 10^{-03}$	0,84	$5{,}37 \times 10^{-01}$	0,94	$5{,}52 \times 10^{-01}$	0,92	$8{,}49 \times 10^{-06}$
rs225214	-	-	-	-	-	-	-	-	$3{,}65 \times 10^{-05}$	1,25	$2{,}00 \times 10^{-03}$	1,14	$4{,}57 \times 10^{-01}$	0,93	$1{,}50 \times 10^{-06}$

Abb. 7-3: Grafische Übersicht der Region des SNPs rs8084 in der UCSC-Datenbank. In der UCSC-Datenbank sind weitere Datenbanken grafisch integriert, wie z.B. die ENCODE-Datenbank, die für die *in silico*-Analyse der Region auf regulatorische Elemente genutzt wurde. Die Position des SNPs rs8084 ist mit einer gelben Linie markiert. Von oben nach unten sind folgende Elemente zu sehen: Skala und Position des betrachteten Ausschnitts in bp auf dem Chromosom (Die Position bezieht sich auf das *human genome build 19*). „*UCSC Genes*“: Auf verschiedenen Datenbanken basierende Positionen verschiedener Gen-Transkripte. „*snoRNABase and miRBase*“: Datenbank, die vier verschiedene RNA-Typen (pre-miRNA, C/D box snoRNA, H/ACA box snoRNA und scaRNA) und ihre Position im Genom darstellt. „*Human mRNAs from GenBank*“: Diese Datenbank zeigt die Anordnung von mRNAs im Genom. „*Human ESTs That Have Been Spliced*“: Diese Datenbank zeigt transkribierte Bereiche, in denen Spleißen stattfindet. „*ENCODE*“: Diese Datenbank sammelt Informationen, die im Rahmen der Regulation der Transkription relevant sind. Hier wurden ausgewertet: Histonmethylierung und -acetylierung (H3K4Me1, H3K4Me3, H3K27Ac) für eine Lymphoblastoide- und Lungenfibroblasten-Zelllinie; die Transkriptionsfaktor-ChIP-Seq zeigt die Bindung von Transkriptionsfaktoren in verschiedenen Zelllinien an (Lymphoblastoide-Zelllinien: G und g; Lungenfibroblasten-Zelllinien: n; K563-Zelllinien: K); „*Simple Nucleotide Polymorphisms (dbSNP 138)*“: SNP- und Indel-Daten der dbSNP-Datenbank *build 138* (SNP-Farbgebung: synonymer SNP = grün; nicht-synonymer SNP = rot; SNP in einer Spleißstelle = rot; SNP im untranslatierten Bereich = blau; intronische und unbekannte SNPs = schwarz).

Abb. 7-4: Grafische Übersicht der Region des SNPs rs8084 in der *„eQTL resources at the Pritchard lab“*-Datenbank. In dieser Datenbank sind eQTL-Daten von mehreren Publikationen gesammelt und grafisch integriert. Von oben nach unten: Ideogramm des Chromosoms und Position des betrachteten Bereichs. Position des betrachteten Ausschnitts in kb auf dem Chromosom (Die Position bezieht sich auf das *human genome build 18*). Gen-Positionen basierend auf der *Entrez genes*-

Datenbank. Aufzählung der verschiedenen Studien, in denen eQTLs identifiziert wurden (farbige Dreiecke geben Anzahl und Position identifizierter eQTLs an). SNPs, für die eine eQTL-Funktion vorhergesagt wird (die farbige Markierung unterhalb der SNPs ordnet die SNPs der jeweiligen Studie zu; leere Dreiecke zeigen eine *cis*-Wirkung, gefüllte Dreiecke eine *trans*-Wirkung der eQTLs an).

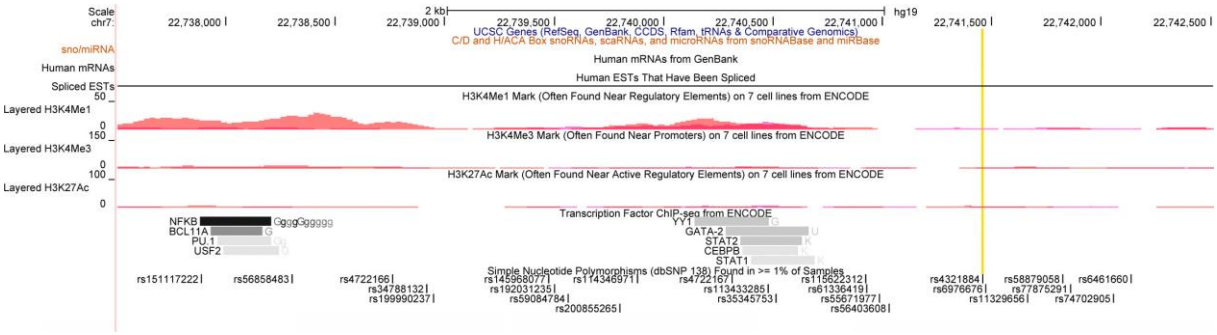

Abb. 7-5: Grafische Übersicht der Region des SNPs rs4321884 in der UCSC-Datenbank. In der UCSC-Datenbank sind weitere Datenbanken grafisch integriert, wie z.B. die ENCODE-Datenbank, die für die *in silico*-Analyse der Region auf regulatorische Elemente genutzt wurde. Die Position des SNPs rs4321884 ist mit einer gelben Linie markiert. Von oben nach unten sind folgende Elemente zu sehen: Skala und Position des betrachteten Ausschnitts in bp auf dem Chromosom (Die Position bezieht sich auf das *human genome build 19*). „*UCSC Genes*“: Auf verschiedenen Datenbanken basierende Positionen verschiedener Gen-Transkripte. „*snoRNABase and miRBase*“: Datenbank, die vier verschiedene RNA-Typen (pre-miRNA, C/D box snoRNA, H/ACA box snoRNA und scaRNA) und ihre Position im Genom darstellt. „*Human mRNAs from GenBank*“: Diese Datenbank zeigt die Anordnung von mRNAs im Genom. „*Human ESTs That Have Been Spliced*“: Diese Datenbank zeigt transkribierte Bereiche, in denen Spleißen stattfindet. „*ENCODE*“: Diese Datenbank sammelt Informationen, die im Rahmen der Regulation der Transkription relevant sind. Hier wurden ausgewertet: Histonmethylierung und -acetylierung (H3K4Me1, H3K4Me3, H3K27Ac) für eine Lymphoblastoide- und Lungenfibroblasten-Zelllinie; die Transkriptionsfaktor-ChIP-Seq zeigt die Bindung von Transkriptionsfaktoren in verschiedenen Zelllinien an (Lymphoblastoide-Zelllinien: G und g; Lungenfibroblasten-Zelllinien: n; K563-Zelllinien: K); „*Simple Nucleotide Polymorphisms (dbSNP 138)*“: SNP- und Indel-Daten der dbSNP-Datenbank *build 138* (SNP-Farbgebung: synonymer SNP = grün; nicht-synonymer SNP = rot; SNP in einer Spleißstelle = rot; SNP im untranslatierten Bereich = blau; intronische und unbekannte SNPs = schwarz).

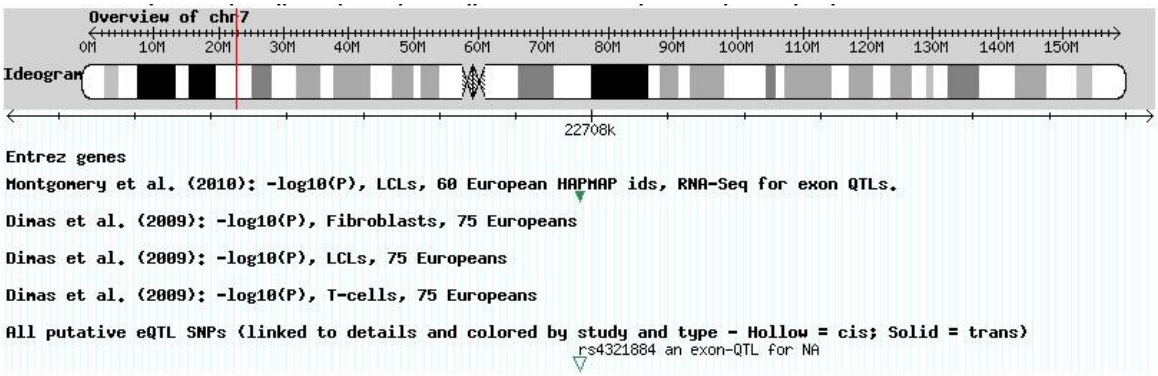

Abb. 7-6: Grafische Übersicht der Region des SNPs rs4321884 in der *„eQTL resources at the Pritchard lab“*-Datenbank. In dieser Datenbank sind eQTL-Daten von mehreren Publikationen gesammelt und grafisch integriert. Von oben nach unten: Ideogramm des Chromosoms und Position des betrachteten Bereichs. Position des betrachteten Ausschnitts in kb auf dem Chromosom (Die Position bezieht sich auf das *human genome build 18*). Gen-Positionen basierend auf der *Entrez genes*-Datenbank. Aufzählung der verschiedenen Studien, in denen eQTLs identifiziert wurden (farbige Dreiecke geben Anzahl und Position identifizierter eQTLs an). SNPs, für die eine eQTL-Funktion vorhergesagt wird (die farbige Markierung unterhalb der SNPs ordnet die SNPs der jeweiligen Studie zu; leere Dreiecke zeigen eine *cis*-Wirkung, gefüllte Dreiecke eine *trans*-Wirkung der eQTLs an).

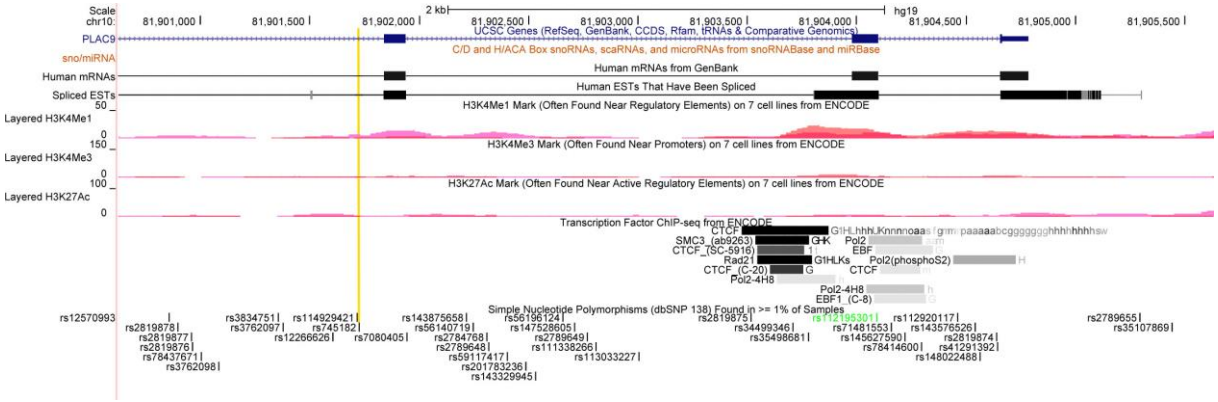

Abb. 7-7: Grafische Übersicht der Region des SNPs rs745182 in der UCSC-Datenbank. In der UCSC-Datenbank sind weitere Datenbanken grafisch integriert, wie z.B. die ENCODE-Datenbank, die für die *in silico*-Analyse der Region auf regulatorische Elemente genutzt wurde. Die Position des SNPs rs745182 ist mit einer gelben Linie markiert. Von oben nach unten sind folgende Elemente zu sehen: Skala und Position des betrachteten Ausschnitts in bp auf dem Chromosom (Die Position bezieht sich auf das *human genome build 19*). „*UCSC Genes*“: Auf verschiedenen Datenbanken basierende Positionen verschiedener Gen-Transkripte. „*snoRNABase and miRBase*“: Datenbank, die vier verschiedene RNA-Typen (pre-miRNA, C/D box snoRNA, H/ACA box snoRNA und scaRNA) und ihre Position im Genom darstellt. „*Human mRNAs from GenBank*“: Diese Datenbank zeigt die Anordnung von mRNAs im Genom. „*Human ESTs That Have Been Spliced*“: Diese Datenbank zeigt transkribierte Bereiche, in denen Spleißen stattfindet. „*ENCODE*“: Diese Datenbank sammelt Informationen, die im Rahmen der Regulation der Transkription relevant sind. Hier wurden ausgewertet: Histonmethylierung und -acetylierung (H3K4Me1, H3K4Me3, H3K27Ac) für eine Lymphoblastoide- und Lungenfibroblasten-Zelllinie; die Transkriptionsfaktor-ChIP-Seq zeigt die Bindung von Transkriptionsfaktoren in verschiedenen Zelllinien an (Lymphoblastoide-Zelllinien: G und g; Lungenfibroblasten-Zelllinien: n; K563-Zelllinien: K); „*Simple Nucleotide Polymorphisms (dbSNP 138)*“: SNP- und Indel-Daten der dbSNP-Datenbank *build 138* (SNP-Farbgebung: synonymer SNP = grün; nicht-synonymer SNP = rot; SNP in einer Spleißstelle = rot; SNP im untranslatierten Bereich = blau; intronische und unbekannte SNPs = schwarz).

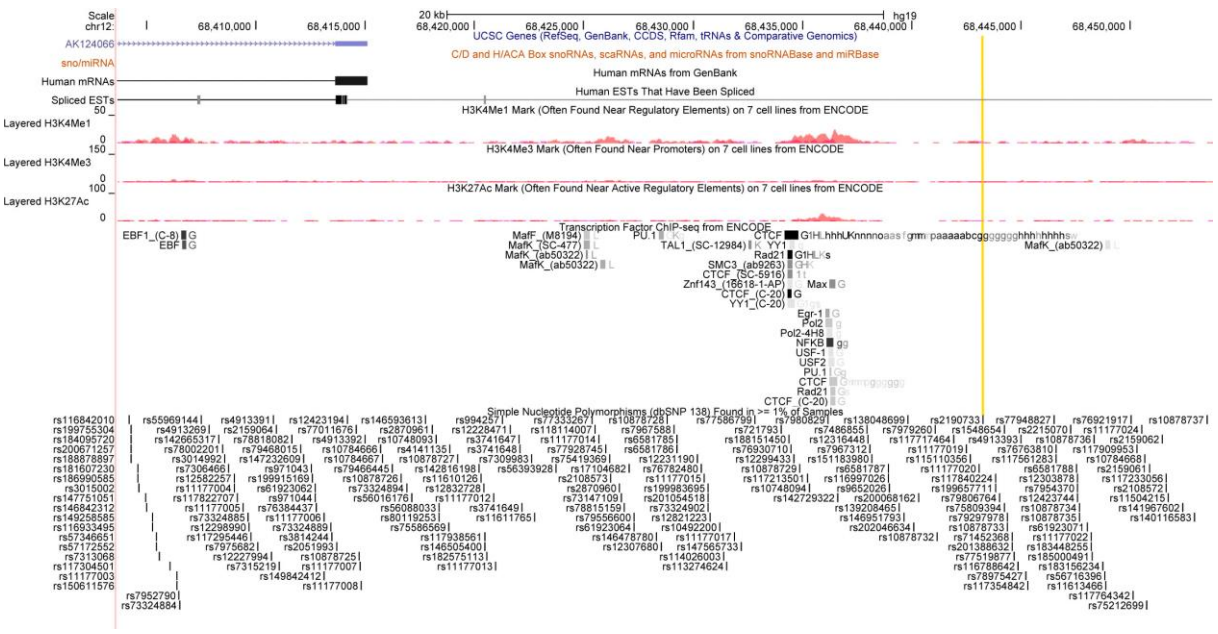

Abb. 7-8: Grafische Übersicht der Region des SNPs rs2190733 in der UCSC-Datenbank. In der UCSC-Datenbank sind weitere Datenbanken grafisch integriert, wie z.B. die ENCODE-Datenbank, die für die *in silico*-Analyse der Region auf regulatorische Elemente genutzt wurde. Die Position des SNPs rs2190733 ist mit einer gelben Linie markiert. Von oben nach unten sind folgende Elemente zu sehen: Skala und Position des betrachteten Ausschnitts in bp auf dem Chromosom (Die Position bezieht sich auf das *human genome build 19*). „*UCSC Genes*“: Auf verschiedenen Datenbanken basierende Positionen verschiedener Gen-Transkripte. „*snoRNABase and miRBase*“: Datenbank, die vier verschiedene RNA-Typen (pre-miRNA, C/D box snoRNA, H/ACA box snoRNA und scaRNA) und ihre Position im Genom darstellt. „*Human mRNAs from GenBank*“: Diese Datenbank zeigt die Anordnung von mRNAs im Genom. „*Human ESTs That Have Been Spliced*“: Diese

Datenbank zeigt transkribierte Bereiche, in denen Spleißen stattfindet. „*ENCODE*“: Diese Datenbank sammelt Informationen, die im Rahmen der Regulation der Transkription relevant sind. Hier wurden ausgewertet: Histonmethylierung und -acetylierung (H3K4Me1, H3K4Me3, H3K27Ac) für eine Lymphoblastoide- und Lungenfibroblasten-Zelllinie; die Transkriptionsfaktor-ChIP-Seq zeigt die Bindung von Transkriptionsfaktoren in verschiedenen Zelllinien an (Lymphoblastoide-Zelllinien: G und g; Lungenfibroblasten-Zelllinien: n; K563-Zelllinien: K); „*Simple Nucleotide Polymorphisms (dbSNP 138)*“: SNP- und Indel-Daten der dbSNP-Datenbank *build 138* (SNP-Farbgebung: synonymer SNP = grün; nicht-synonymer SNP = rot; SNP in einer Spleißstelle = rot; SNP im untranslatierten Bereich = blau; intronische und unbekannte SNPs = schwarz).

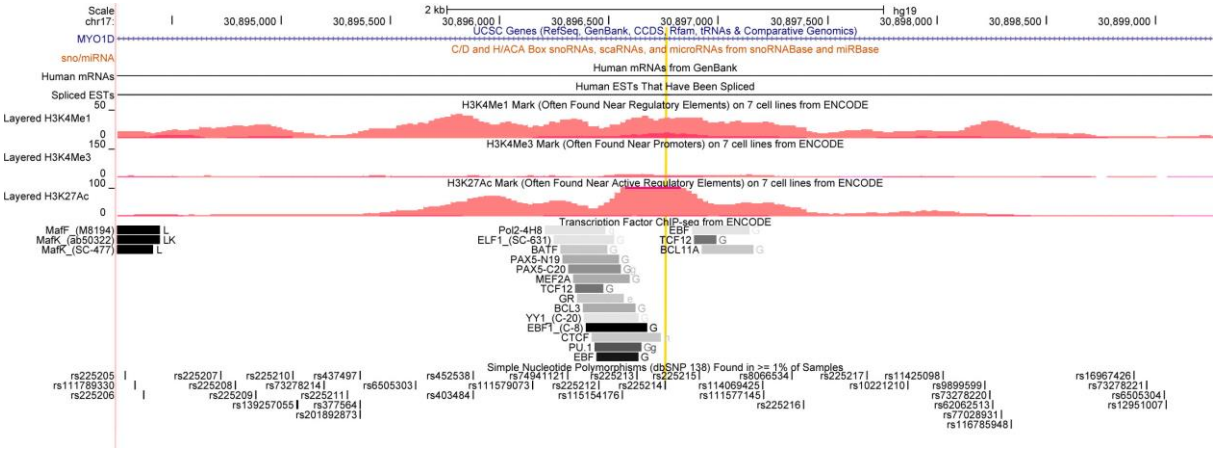

Abb. 7-9: Grafische Übersicht der Region des SNPs rs225214 in der UCSC-Datenbank. In der UCSC-Datenbank sind weitere Datenbanken grafisch integriert, wie z.B. die ENCODE-Datenbank, die für die *in silico*-Analyse der Region auf regulatorische Elemente genutzt wurde. Die Position des SNPs rs225214 ist mit einer gelben Linie markiert. Von oben nach unten sind folgende Elemente zu sehen: Skala und Position des betrachteten Ausschnitts in bp auf dem Chromosom (Die Position bezieht sich auf das *human genome build 19*). „*UCSC Genes*“: Auf verschiedenen Datenbanken basierende Positionen verschiedener Gen-Transkripte. „*snoRNABase and miRBase*“: Datenbank, die vier verschiedene RNA-Typen (pre-miRNA, C/D box snoRNA, H/ACA box snoRNA und scaRNA) und ihre Position im Genom darstellt. „*Human mRNAs from GenBank*“: Diese Datenbank zeigt die Anordnung von mRNAs im Genom. „*Human ESTs That Have Been Spliced*“: Diese Datenbank zeigt transkribierte Bereiche, in denen Spleißen stattfindet. „*ENCODE*“: Diese Datenbank sammelt Informationen, die im Rahmen der Regulation der Transkription relevant sind. Hier wurden ausgewertet: Histonmethylierung und -acetylierung (H3K4Me1, H3K4Me3, H3K27Ac) für eine Lymphoblastoide- und Lungenfibroblasten-Zelllinie; die Transkriptionsfaktor-ChIP-Seq zeigt die Bindung von Transkriptionsfaktoren in verschiedenen Zelllinien an (Lymphoblastoide-Zelllinien: G und g; Lungenfibroblasten-Zelllinien: n; K563-Zelllinien: K); „*Simple Nucleotide Polymorphisms (dbSNP 138)*“: SNP- und Indel-Daten der dbSNP-Datenbank *build 138* (SNP-Farbgebung: synonymer SNP = grün; nicht-synonymer SNP = rot; SNP in einer Spleißstelle = rot; SNP im untranslatierten Bereich = blau; intronische und unbekannte SNPs = schwarz).

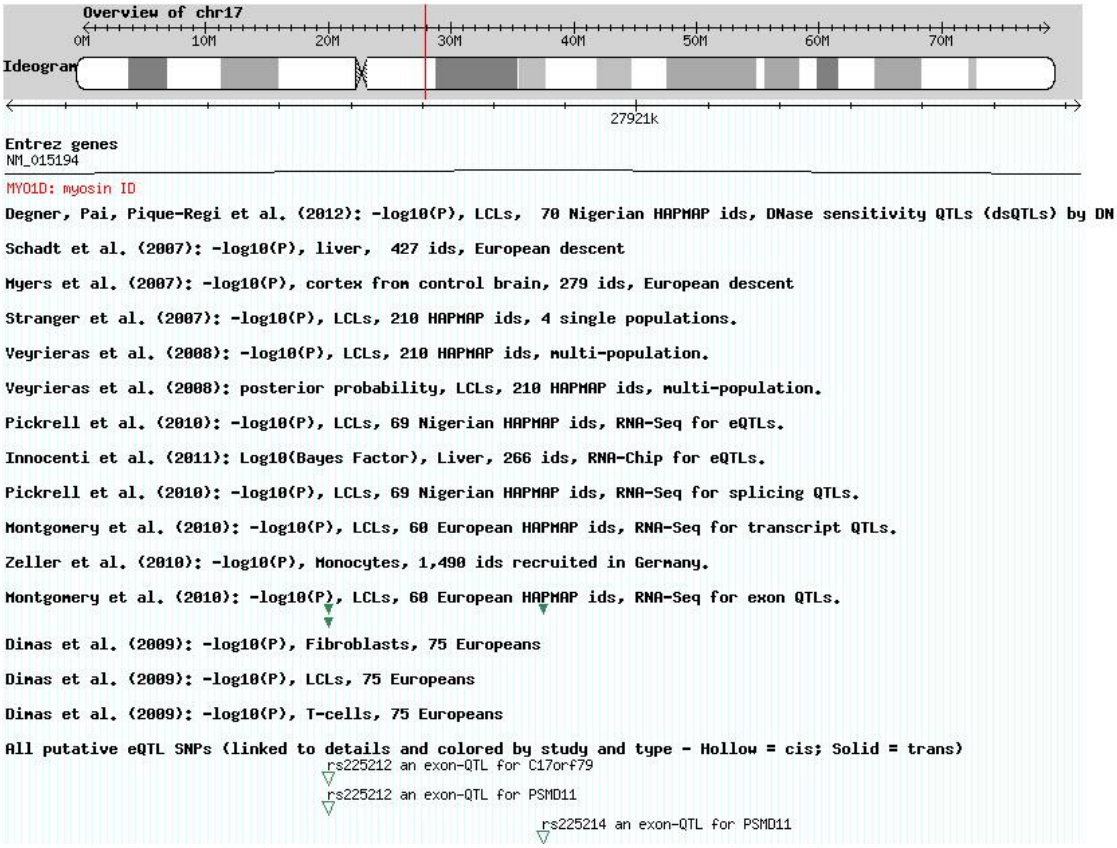

Abb. 7-10: Grafische Übersicht der Region des SNPs rs225214 in der *„eQTL resources at the Pritchard lab“*-Datenbank. In dieser Datenbank sind eQTL-Daten von mehreren Publikationen gesammelt und grafisch integriert. Von oben nach unten: Ideogramm des Chromosoms und Position des betrachteten Bereichs. Position des betrachteten Ausschnitts in kb auf dem Chromosom (Die Position bezieht sich auf das *human genome build 18*). Gen-Positionen basierend auf der *Entrez genes*-Datenbank. Aufzählung der verschiedenen Studien, in denen eQTLs identifiziert wurden (farbige Dreiecke geben Anzahl und Position identifizierter eQTLs an). SNPs, für die eine eQTL-Funktion vorhergesagt wird (die farbige Markierung unterhalb der SNPs ordnet die SNPs der jeweiligen Studie zu; leere Dreiecke zeigen eine *cis*-Wirkung, gefüllte Dreiecke eine *trans*-Wirkung der eQTLs an).

Referenzen

1000 Genomes Project Consortium (2010), 'A map of human genome variation from population-scale sequencing', *Nature,* 467 (7319), 1061-73.

Aaron, L., et al. (2004), 'Tuberculosis in HIV-infected patients: a comprehensive review', *Clin Microbiol Infect,* 10 (5), 388-98.

Abbas, K., et al. (2009), 'Signaling events leading to peroxiredoxin 5 up-regulation in immunostimulated macrophages', *Free Radic Biol Med,* 47 (6), 794-802.

Abecasis, G. R., et al. (2010), 'A map of human genome variation from population-scale sequencing', *Nature,* 467 (7319), 1061-73.

Abecasis, G. R., et al. (2001), 'Extent and distribution of linkage disequilibrium in three genomic regions', *Am J Hum Genet,* 68 (1), 191-97.

Adams, L. A., et al. (2011), 'Polymorphisms in MC3R promoter and CTSZ 3'UTR are associated with tuberculosis susceptibility', *Eur J Hum Genet,* 19 (6), 676-81.

Adrianto, I., et al. (2012), 'Genome-wide association study of African and European Americans implicates multiple shared and ethnic specific loci in sarcoidosis susceptibility', *PLoS One,* 7 (8), e43907.

Affymetrix 'Data Sheet, Genome-Wide Human SNP Array 6.0', <http://www.affymetrix.com/>.

Agostini, C., Adami, F., and Semenzato, G. (2000), 'New pathogenetic insights into the sarcoid granuloma', *Curr Opin Rheumatol,* 12 (1), 71-6.

Agresti, A. and Coull, B. A. (1996), 'Order-restricted tests for stratified comparisons of binomial proportions', *Biometrics,* 52 (3), 1103-11.

Ali, S., et al. (2013), 'IL12B SNPs and copy number variation in IL23R gene associated with susceptibility to leprosy', *J Med Genet,* 50 (1), 34-42.

Almarri, Ajaybe and Batchelor, JR (1994), 'HLA and hepatitis B infection', *The Lancet,* 344 (8931), 1194-95.

Alsmadi, O., et al. (2009), 'Specific and complete human genome amplification with improved yield achieved by phi29 DNA polymerase and a novel primer at elevated temperature', *BMC Res Notes,* 2, 48.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008), 'Genetic mapping in human disease', *Science,* 322 (5903), 881-8.

Altshuler, D. M., et al. (2010), 'Integrating common and rare genetic variation in diverse human populations', *Nature,* 467 (7311), 52-8.

American Journal of Public Health and the Nation's Health (1931), 'The Lubeck Disaster', *Am J Public Health Nations Health,* 21 (3), 282.

American Thoracic Society (1999), 'Statement on sarcoidosis. Joint Statement of the American Thoracic Society (ATS), the European Respiratory Society (ERS) and the World Association of Sarcoidosis and Other Granulomatous Disorders (WASOG) adopted by the ATS Board of Directors and by the ERS Executive Committee, February 1999', *Am J Respir Crit Care Med,* 160 (2), 736-55.

American Thoracic Society (2000), 'Diagnostic Standards and Classification of Tuberculosis in Adults and Children. This official statement of the American Thoracic Society and the Centers for Disease Control and Prevention was adopted by the ATS Board of Directors, July 1999. This statement was endorsed by the Council of the Infectious Disease Society of America, September 1999', *Am J Respir Crit Care Med,* 161 (4 Pt 1), 1376-95.

Amirzargar, A. A., et al. (2004), 'The association of HLA-DRB, DQA1, DQB1 alleles and haplotype frequency in Iranian patients with pulmonary tuberculosis', *Int J Tuberc Lung Dis,* 8 (8), 1017-21.

Anderson, C. A., et al. (2010), 'Data quality control in genetic case-control association studies', *Nat Protoc,* 5 (9), 1564-73.

Ansari, A., et al. (2011), 'Differential combination of cytokine and interferon- gamma +874 T/A polymorphisms determines disease severity in pulmonary tuberculosis', *PLoS One,* 6 (11), e27848.

Applied Biosystems 'Product Bulletin - TaqMan® SNP Genotyping Assays', <http://de-de.invitrogen.com/site/de/de/home/brands/Applied-Biosystems.html>.

Ates, O., et al. (2008), 'Interleukin-10 and tumor necrosis factor-alpha gene polymorphisms in tuberculosis', *J Clin Immunol,* 28 (3), 232-6.

Bachwich, P. R., et al. (1986), 'Tumor necrosis factor production by human sarcoid alveolar macrophages', *Am J Pathol,* 125 (3), 421-5.

Baehner, R. L. and Nathan, D. G. (1967), 'Leukocyte oxidase: defective activity in chronic granulomatous disease', *Science,* 155 (3764), 835-6.

Barrett, J. C., et al. (2008), 'Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease', *Nat Genet,* 40 (8), 955-62.

Bateson, William (1909), 'Mendel's principles of heredity. 1909', *Mendel's principles of heredity*.

Baudat, F., et al. (2010), 'PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice', *Science,* 327 (5967), 836-40.

Baughman, R. P., Lower, E. E., and du Bois, R. M. (2003), 'Sarcoidosis', *Lancet,* 361 (9363), 1111-8.

Baughman, R. P., Culver, D. A., and Judson, M. A. (2011), 'A concise review of pulmonary sarcoidosis', *Am J Respir Crit Care Med,* 183 (5), 573-81.

Baughman, R. P., et al. (1997), 'Predicting respiratory failure in sarcoidosis patients', *Sarcoidosis Vasc Diffuse Lung Dis,* 14 (2), 154-8.

Benjamini, Yoav and Hochberg, Yosef (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Besnier, Ernest (1889), 'Lupus pernio de la face; synovites fongueuses (scrofulo-tuberculeuses) symétriques des extrémités superieures', *Annales de dermatologie et de syphilographie, Paris,* 2nd series, 10: 333-36.

Beyer, A., Bandyopadhyay, S., and Ideker, T. (2007), 'Integrating physical and genetic maps: from genomes to interaction networks', *Nat Rev Genet,* 8 (9), 699-710.

Bhatt, K., Hickman, S. P., and Salgame, P. (2004), 'Cutting edge: a new approach to modeling early lung immunity in murine tuberculosis', *J Immunol,* 172 (5), 2748-51.

Billings, T., et al. (2013), 'DNA binding specificities of the long zinc-finger recombination protein PRDM9', *Genome Biol,* 14 (4), R35.

Bishop, YMM, Fienberg, SE, and Holland, PW (1975), 'Discrete multivariate analysis: theory and practice MIT Press', *Cambridge, MA*.

Blanco, L., et al. (1989), 'Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication', *J Biol Chem,* 264 (15), 8935-40.

Bodmer, Walter and Bonilla, Carolina (2008), 'Common and rare variants in multifactorial susceptibility to common diseases', *Nature genetics,* 40 (6), 695-701.

Boeck, Cæsar Peter Møller (1899), 'Multiple benign sarcoid of the skin', *Journal of Cutaneous and Genitourinary Diseases, Chicago,* (17), 543-50.

Boros, D. L. (1978), 'Granulomatous inflammations', *Prog Allergy,* 24, 183-267.

Brahmajothi, V, et al. (1991), 'Association of pulmonary tuberculosis and HLA in South India', *Tubercle,* 72 (2), 123-32.

Brett, G. Z. (1965), 'Epidemiological trends in tuberculosis and sarcoidosis in a district of London between 1958 and 1963', *Tubercle,* 46 (4), 413-6.

Briken, V., et al. (2004), 'Mycobacterial lipoarabinomannan and related lipoglycans: from biogenesis to modulation of the immune response', *Mol Microbiol,* 53 (2), 391-403.

British Thoracic and Tuberculosis Association (1973), 'Familial associations in sarcoidosis. A report to the research committee of the British Thoracic and Tuberculosis Association', *Tubercle,* 54, 87-98.

Broer, L., et al. (2013), 'Distinguishing true from false positives in genomic studies: p values', *Eur J Epidemiol,* 28 (2), 131-8.

Brown, Christopher D, Mangravite, Lara M, and Engelhardt, Barbara E (2012), 'Integrative modeling of eQTLs and cis-regulatory elements suggest mechanisms underlying cell type specificity of eQTLs', *arXiv preprint arXiv:1210.3294*.

Brownell, I., et al. (2011), 'Evidence for mycobacteria in sarcoidosis', *Am J Respir Cell Mol Biol,* 45 (5), 899-905.

Browning, B. L. and Browning, S. R. (2009), 'A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals', *Am J Hum Genet,* 84 (2), 210-23.

Browning, S. R. (2006), 'Multilocus association mapping using variable-length Markov chains', *Am J Hum Genet,* 78 (6), 903-13.

Browning, S. R. and Browning, B. L. (2007), 'Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering', *Am J Hum Genet,* 81 (5), 1084-97.

Bunn, D. T. and Johnston, R. N. (1972), 'A ten-year study of sarcoidosis', *Br J Dis Chest,* 66 (1), 45-52.

Burke, WMJ, et al. (1990), 'Transmission of sarcoidosis via cardiac transplantation', *The Lancet,* 336 (8730), 1579.

Campbell, M. C. and Tishkoff, S. A. (2008), 'African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping', *Annu Rev Genomics Hum Genet,* 9, 403-33.

Carrington, M., et al. (1999), 'HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage', *Science,* 283 (5408), 1748-52.

Castelli, E. C. 'SNPex version 5.17', <http://immunogen.fmrp.usp.br/snpex/>.

Cervino, A. C., et al. (2000), 'Allelic association between the NRAMP1 gene and susceptibility to tuberculosis in Guinea-Conakry', *Ann Hum Genet,* 64 (Pt 6), 507-12.

Chakravarti, A. (1999), 'Population genetics--making sense out of sequence', *Nat Genet,* 21 (1 Suppl), 56-60.

Chappell, A. G., Cheung, W. Y., and Hutchings, H. A. (2000), 'Sarcoidosis: a long-term follow up study', *Sarcoidosis Vasc Diffuse Lung Dis,* 17 (2), 167-73.

Cheah, F. C., et al. (2009), 'Airway inflammatory cell responses to intra-amniotic lipopolysaccharide in a sheep model of chorioamnionitis', *Am J Physiol Lung Cell Mol Physiol,* 296 (3), L384-93.

Chen, E. S., et al. (2008), 'T cell responses to mycobacterial catalase-peroxidase profile a pathogenic antigen in systemic sarcoidosis', *J Immunol,* 181 (12), 8784-96.

Clark, V. J., et al. (2007), 'Combining sperm typing and linkage disequilibrium analyses reveals differences in selective pressures or recombination rates across human populations', *Genetics,* 175 (2), 795-804.

Clarke, Geraldine M, et al. (2011), 'Basic statistical analysis in genetic case-control studies', *Nature protocols,* 6 (2), 121-33.

Co, Dominic O, et al. (2004), 'Mycobacterial granulomas: keys to a long-lasting host–pathogen relationship', *Clinical Immunology,* 113 (2), 130-36.

Collins, A. (2009), 'Allelic association: linkage disequilibrium structure and gene mapping', *Mol Biotechnol,* 41 (1), 83-9.

Collins, A., Lonjou, C., and Morton, N. E. (1999), 'Genetic epidemiology of single-nucleotide polymorphisms', *Proc Natl Acad Sci U S A,* 96 (26), 15173-7.

Collins, F. S., Guyer, M. S., and Charkravarti, A. (1997), 'Variations on a theme: cataloging human DNA sequence variation', *Science,* 278 (5343), 1580-1.

Comstock, G. W. (1978), 'Tuberculosis in twins: a re-analysis of the Prophit survey', *Am Rev Respir Dis,* 117 (4), 621-4.

Conrad, D. F., et al. (2006), 'A worldwide survey of haplotype variation and linkage disequilibrium in the human genome', *Nat Genet,* 38 (11), 1251-60.

Corbett, E. L., et al. (2003), 'The growing burden of tuberculosis: global trends and interactions with the HIV epidemic', *Arch Intern Med,* 163 (9), 1009-21.

Costabel, U. (2001), 'Sarcoidosis: clinical update', *Eur Respir J Suppl,* 32, 56s-68s.

Cozier, Y. C., et al. (2012), 'Fine-mapping in African-American women confirms the importance of the 10p12 locus to sarcoidosis', *Genes Immun,* 13 (7), 573-8.

Crawford, D. C., et al. (2004), 'Evidence for substantial fine-scale variation in recombination rates across the human genome', *Nat Genet,* 36 (7), 700-6.

Dannenberg, A. M., Jr. (1993), 'Immunopathogenesis of pulmonary tuberculosis', *Hosp Pract (Off Ed),* 28 (1), 51-8.

de Bakker, P. I., et al. (2008), 'Practical aspects of imputation-driven meta-analysis of genome-wide association studies', *Hum Mol Genet,* 17 (R2), R122-8.

De la Vega, F. M., et al. (2005), 'Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System', *Mutat Res,* 573 (1-2), 111-35.

Dean, F. B., et al. (2002), 'Comprehensive human genome amplification using multiple displacement amplification', *Proc Natl Acad Sci U S A,* 99 (8), 5261-6.

Demangel, C., Bertolino, P., and Britton, W. J. (2002), 'Autocrine IL-10 impairs dendritic cell (DC)-derived immune responses to mycobacterial infection by suppressing DC trafficking to draining lymph nodes and local IL-12 production', *Eur J Immunol,* 32 (4), 994-1002.

Dempfle, A., et al. (2008), 'Gene-environment interactions for complex traits: definitions, methodological requirements and challenges', *Eur J Hum Genet,* 16 (10), 1164-72.

Derrien, T., et al. (2012), 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression', *Genome Res,* 22 (9), 1775-89.

Di Rienzo, A. (2006), 'Population genetics models of common diseases', *Curr Opin Genet Dev,* 16 (6), 630-6.

Dickson, S. P., et al. (2010), 'Rare variants create synthetic genome-wide associations', *PLoS Biol,* 8 (1), e1000294.

Diehl, Karl, et al. (1936), *Der Erbeinfluss bei der Tuberkulose* (G. Fischer).

Diet, A., et al. (2007), 'Regulation of peroxiredoxins by nitric oxide in immunostimulated macrophages', *J Biol Chem,* 282 (50), 36199-205.

Ding, S., Li, L., and Zhu, X. (2008), 'Polymorphism of the interferon-gamma gene and risk of tuberculosis in a southeastern Chinese population', *Hum Immunol,* 69 (2), 129-33.

Dinger, M. E., et al. (2009), 'Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications', *Brief Funct Genomic Proteomic,* 8 (6), 407-23.

Doganci, A., et al. (2005), 'The IL-6R alpha chain controls lung CD4+CD25+ Treg development and function during allergic airway inflammation in vivo', *J Clin Invest,* 115 (2), 313-25.

Dubaniewicz, A., et al. (2012), 'Is mycobacterial heat shock protein 16 kDa, a marker of the dormant stage of Mycobacterium tuberculosis, a sarcoid antigen?', *Hum Immunol*.

Dubuisson, M., et al. (2004), 'Human peroxiredoxin 5 is a peroxynitrite reductase', *FEBS Lett,* 571 (1-3), 161-5.

Ducati, R. G., et al. (2006), 'The resumption of consumption -- a review on tuberculosis', *Mem Inst Oswaldo Cruz,* 101 (7), 697-714.

Dudbridge, F. and Gusnanto, A. (2008), 'Estimation of significance thresholds for genomewide association scans', *Genet Epidemiol,* 32 (3), 227-34.

Duerr, R. H., et al. (2006), 'A genome-wide association study identifies IL23R as an inflammatory bowel disease gene', *Science,* 314 (5804), 1461-3.

Duggal, Priya, et al. (2008), 'Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies', *BMC Genomics,* 9 (1), 516.

Dupont, WD and Plummer, WD (1997), 'PS power and sample size program available for free on the Internet', *Controlled Clinical Trials,* 18, 274.

Ellinghaus, D., et al. (2012), 'Combined analysis of genome-wide association studies for Crohn disease and psoriasis identifies seven shared susceptibility loci', *Am J Hum Genet,* 90 (4), 636-47.

Ellinghaus, E., et al. (2010), 'Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2', *Nat Genet,* 42 (11), 991-5.

Enarson, DA and Murray, JF (1996), 'Global epidemiology of tuberculosis', *Tuberculosis. Boston, Little Brown and Company*, 57-76.

Escoubet-Lozach, L., et al. (2011), 'Mechanisms establishing TLR4-responsive activation states of inflammatory response genes', *PLoS Genet,* 7 (12), e1002401.

Evans, Timothy M, et al. (2005), 'Characterization of Rab23, a negative regulator of sonic hedgehog signaling', *Methods in enzymology,* 403, 759-77.

Evseeva, I., et al. (2010), 'Linkage disequilibrium and age of HLA region SNPs in relation to classic HLA gene alleles within Europe', *Eur J Hum Genet,* 18 (8), 924-32.

Ezzie, M. E. and Crouser, E. D. (2007), 'Considering an infectious etiology of sarcoidosis', *Clin Dermatol,* 25 (3), 259-66.

Falk, C. T. and Rubinstein, P. (1987), 'Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations', *Ann Hum Genet,* 51 (Pt 3), 227-33.

Feng, W. X., et al. (2012), 'CCL2-2518 (A/G) polymorphisms and tuberculosis susceptibility: a meta-analysis', *Int J Tuberc Lung Dis,* 16 (2), 150-6.

Fischer, A., et al. (2010), 'A genome-wide linkage analysis in 181 German sarcoidosis families using clustered biallelic markers', *Chest,* 138 (1), 151-7.

Fischer, A., et al. (2011), 'Association of inflammatory bowel disease risk loci with sarcoidosis, and its acute and chronic subphenotypes', *Eur Respir J,* 37 (3), 610-6.

Fischer, A., et al. (2012), 'A Novel Sarcoidosis Risk Locus for Europeans on Chromosome 11q13.1', *Am J Respir Crit Care Med,* 186 (9), 877-85.

Flicek, P., et al. (2010), 'Ensembl's 10th year', *Nucleic Acids Res,* 38 (Database issue), D557-62.

Flynn, J. L. and Chan, J. (2001), 'Immunology of tuberculosis', *Annu Rev Immunol,* 19, 93-129.

Flynn, J. L., et al. (1993), 'An essential role for interferon gamma in resistance to Mycobacterium tuberculosis infection', *J Exp Med,* 178 (6), 2249-54.

Foley, P. J., et al. (2001), 'Human leukocyte antigen-DRB1 position 11 residues are a common protective marker for sarcoidosis', *Am J Respir Cell Mol Biol,* 25 (3), 272-7.

Franke, A., et al. (2010a), 'Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci', *Nature genetics,* 42 (12), 1118-25.

Franke, A., et al. (2008), 'Genome-wide association analysis in sarcoidosis and Crohn's disease unravels a common susceptibility locus on 10p12.2', *Gastroenterology,* 135 (4), 1207-15.

Franke, A., et al. (2010b), 'Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL)', *Nat Genet,* 42 (4), 292-4.

Frazer, K. A., et al. (2007), 'A second generation human haplotype map of over 3.1 million SNPs', *Nature,* 449 (7164), 851-61.

Gabriel, S., Ziaugra, L., and Tabbaa, D. (2009), 'SNP genotyping using the Sequenom MassARRAY iPLEX platform', *Curr Protoc Hum Genet,* Chapter 2, Unit 2 12.

Gabriel, S. B., et al. (2002), 'The structure of haplotype blocks in the human genome', *Science,* 296 (5576), 2225-9.

Gal, A. A. and Koss, M. N. (2002), 'The pathology of sarcoidosis', *Curr Opin Pulm Med,* 8 (5), 445-51.

Gallagher, Sean R and Desjardins, Philippe R (1989), 'Quantitation of DNA and RNA with absorption and fluorescence spectroscopy', *Current Protocols in Human Genetics*, A. 3D. 1-A. 3D. 21.

Gallant, CJ, et al. (2010), 'Impact of age and sex on mycobacterial immunity in an area of high tuberculosis incidence', *The International Journal of Tuberculosis and Lung Disease,* 14 (8), 952-59.

Garmendia, C., et al. (1992), 'The bacteriophage phi 29 DNA polymerase, a proofreading enzyme', *J Biol Chem,* 267 (4), 2594-9.

Gerke, Volker and Moss, Stephen E (2002), 'Annexins: from structure to function', *Physiological reviews,* 82 (2), 331-71.

Gerstein, M. B., et al. (2012), 'Architecture of the human regulatory network derived from ENCODE data', *Nature,* 489 (7414), 91-100.

Girardi, E., et al. (2000), 'Impact of the HIV epidemic on the spread of other diseases: the case of tuberculosis', *AIDS,* 14 Suppl 3, S47-56.

Golden, M. P. and Vikram, H. R. (2005), 'Extrapulmonary tuberculosis: an overview', *Am Fam Physician,* 72 (9), 1761-8.

Gordon, Alexander, et al. (2007), 'Control of the mean number of false discoveries, Bonferroni and stability of multiple testing', *The Annals of Applied Statistics,* 1 (1), 179-90.

GraphPad Software 'GraphPad Prism 6', <http://www.graphpad.com/>.

Greenwood, C. M., et al. (2000), 'Linkage of tuberculosis to chromosome 2q35 loci, including NRAMP1, in a large aboriginal Canadian family', *Am J Hum Genet,* 67 (2), 405-16.

Grunewald, J. and Eklund, A. (2007), 'Role of CD4+ T cells in sarcoidosis', *Proc Am Thorac Soc,* 4 (5), 461-4.

Grunewald, J. and Eklund, A. (2009), 'Lofgren's syndrome: human leukocyte antigen strongly influences the disease course', *Am J Respir Crit Care Med,* 179 (4), 307-12.

Grunewald, J., et al. (2010), 'Major histocompatibility complex class II transactivator gene polymorphism: associations with Lofgren's syndrome', *Tissue Antigens,* 76 (2), 96-101.

Grutters, J. C., et al. (2003), 'The importance of sarcoidosis genotype to lung phenotype', *Am J Respir Cell Mol Biol,* 29 (3 Suppl), S59-62.

Guan, Y. and Stephens, M. (2008), 'Practical issues in imputation-based association mapping', *PLoS Genet,* 4 (12), e1000279.

Guillaudeux, T., et al. (1998), 'The complete genomic sequence of 424,015 bp at the centromeric end of the HLA class I region: gene content and polymorphism', *Proc Natl Acad Sci U S A,* 95 (16), 9494-9.

Gupta, D., et al. (2007), 'Molecular evidence for the role of mycobacteria in sarcoidosis: a meta-analysis', *Eur Respir J,* 30 (3), 508-16.

Guttman, M., et al. (2009), 'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature,* 458 (7235), 223-7.

Haimovic, A., et al. (2012), 'Sarcoidosis: a comprehensive review and update for the dermatologist: part I. Cutaneous disease', *J Am Acad Dermatol,* 66 (5), 699 e1-18; quiz 717-8.

Hampe, J., et al. (2001), 'An integrated system for high throughput TaqMan based SNP genotyping', *Bioinformatics,* 17 (7), 654-5.

Hance, A. J., et al. (1985), 'Characterization of mononuclear phagocyte subpopulations in the human lung by using monoclonal antibodies: changes in alveolar macrophage phenotype associated with pulmonary sarcoidosis', *J Immunol,* 134 (1), 284-92.

Hardy, G. H. (1908), 'Mendelian Proportions in a Mixed Population', *Science,* 28 (706), 49-50.

Hardy, J. and Singleton, A. (2009), 'Genomewide association studies and human disease', *N Engl J Med,* 360 (17), 1759-68.

Harley, J. B., et al. (2008), 'Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci', *Nat Genet,* 40 (2), 204-10.

Hayden, M. S., West, A. P., and Ghosh, S. (2006), 'NF-kappaB and the immune response', *Oncogene,* 25 (51), 6758-80.

Heath, Simon C, et al. (2008), 'Investigation of the fine structure of European populations with applications to disease association studies', *European Journal of Human Genetics,* 16 (12), 1413-29.

Heinrich, P. C., et al. (2003), 'Principles of interleukin (IL)-6-type cytokine signalling and its regulation', *Biochem J,* 374 (Pt 1), 1-20.

Herndon, C. N. and Jennings, R. G. (1951), 'A twin-family study of susceptibility to poliomyelitis', *Am J Hum Genet,* 3 (1), 17-46.

Hershkovitz, I., et al. (2008), 'Detection and molecular characterization of 9,000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean', *PLoS One,* 3 (10), e3426.

Hinch, A. G., et al. (2011), 'The landscape of recombination in African Americans', *Nature,* 476 (7359), 170-5.

Hingorani, Aroon D, et al. (2010), 'Translating genomics into improved healthcare', *BMJ,* 341.

Ho, L. P., et al. (2005), 'Deficiency of a subset of T-cells with immunoregulatory properties in sarcoidosis', *Lancet,* 365 (9464), 1062-72.

Hofmann, S., et al. (2008), 'Genome-wide association study identifies ANXA11 as a new susceptibility locus for sarcoidosis', *Nat Genet,* 40 (9), 1103-6.

Hofmann, S., et al. (2011), 'A genome-wide association study reveals evidence of association with sarcoidosis at 6p12.1', *Eur Respir J,* 38 (5), 1127-35.

Hofmann, S., et al. (2013), 'Genome-wide association analysis reveals 12q13.3-q14.1 as new risk locus for sarcoidosis', *Eur Respir J,* 41 (4), 888-900.

Hoggart, C. J., et al. (2008), 'Genome-wide significance for dense SNP and resequencing data', *Genet Epidemiol,* 32 (2), 179-85.

Holliday, R. and Grigg, G. W. (1993), 'DNA methylation and mutation', *Mutat Res,* 285 (1), 61-7.

Hong, H., Xu, L., and Tong, W. (2010), 'Assessing consistency between versions of genotype-calling algorithm Birdseed for the Genome-Wide Human SNP Array 6.0 using HapMap samples', *Adv Exp Med Biol,* 680, 355-60.

Horton, R., et al. (1998), 'Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC', *J Mol Biol,* 282 (1), 71-97.

Hosoda, Y., Sasagawa, S., and Yasuda, N. (2002), 'Epidemiology of sarcoidosis: new frontiers to explore', *Curr Opin Pulm Med,* 8 (5), 424-8.

Hosoda, Yutaka, Yamaguchi, Momoko, and Hiraga, Yomei (1997), 'Global epidemiology of sarcoidosis: What story do prevalence and incidence tell us?', *Clinics in chest medicine,* 18 (4), 681-94.

International HapMap Consortium (2003), 'The International HapMap Project', *Nature,* 426 (6968), 789-96.

International HapMap Consortium (2005), 'A haplotype map of the human genome', *Nature,* 437 (7063), 1299-320.

Jagielska, Dagny, et al. (2012), 'Follow-up study of the first genome-wide association scan in alopecia areata: IL13 and KIAA0350 as susceptibility loci supported with genome-wide significance', *Journal of Investigative Dermatology,* 132 (9), 2192-97.

Jansen, Bastiaan JH, et al. (2009), 'OS9 interacts with DC-STAMP and modulates its intracellular localization in response to TLR ligation', *Molecular immunology,* 46 (4), 505-15.

Jeffreys, A. J., et al. (2013), 'Recombination regulator PRDM9 influences the instability of its own coding sequence in humans', *Proc Natl Acad Sci U S A,* 110 (2), 600-5.

Jepson, Annette, et al. (2001), 'Genetic Regulation of Acquired Immune Responses to Antigens ofMycobacterium tuberculosis: a Study of Twins in West Africa', *Infection and immunity,* 69 (6), 3989-94.

Jordan, H. T., et al. (2011), 'Sarcoidosis diagnosed after September 11, 2001, among adults exposed to the World Trade Center disaster', *J Occup Environ Med,* 53 (9), 966-74.

Kallmann, Fr J and Reisner, D (1943), 'Twin studies on the significance of genetic factors in tuberculosis', *Am Rev Tuberc,* 47, 549-74.

Kappelman, J., et al. (2008), 'First Homo erectus from Turkey and implications for migrations into temperate Eurasia', *Am J Phys Anthropol,* 135 (1), 110-6.

Kasowski, Maya, et al. (2010), 'Variation in transcription factor binding among humans', *Science,* 328 (5975), 232-35.

Kennedy, G. C., et al. (2003), 'Large-scale genotyping of complex DNA', *Nat Biotechnol,* 21 (10), 1233-7.

Khader, S. A., et al. (2006), 'Interleukin 12p40 is required for dendritic cell migration and T cell priming after Mycobacterium tuberculosis infection', *J Exp Med,* 203 (7), 1805-15.

Kim, H. S., et al. (2011), 'Association of interleukin 23 receptor gene with sarcoidosis', *Dis Markers,* 31 (1), 17-24.

Kim, Mi‐Jeong, et al. (2010), 'Caseation of human tuberculosis granulomas correlates with elevated host lipid metabolism', *EMBO molecular medicine,* 2 (7), 258-74.

Kimman, Tjeerd G., Janssen, Riny, and Hoebee, Barbara (2006), 'Future prospects in respiratory syncytial virus genetics', *Future Virology,* 1 (4), 483-92.

Klareskog, Lars, et al. (2006), 'A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA–DR (shared epitope)–restricted immune reactions to autoantigens modified by citrullination', *Arthritis & Rheumatism,* 54 (1), 38-46.

Klein, R. J., et al. (2005), 'Complement factor H polymorphism in age-related macular degeneration', *Science,* 308 (5720), 385-9.

Knoops, B., et al. (2011), 'Peroxiredoxin 5: structure, mechanism, and function of the mammalian atypical 2-Cys peroxiredoxin', *Antioxid Redox Signal,* 15 (3), 817-29.

Knoops, B., et al. (1999), 'Cloning and characterization of AOEB166, a novel mammalian antioxidant enzyme of the peroxiredoxin family', *J Biol Chem,* 274 (43), 30451-8.

Koboldt, D. C., Miller, R. D., and Kwok, P. Y. (2006), 'Distribution of human SNPs and its effect on high-throughput genotyping', *Hum Mutat,* 27 (3), 249-54.

Korn, J. M., et al. (2008), 'Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs', *Nat Genet,* 40 (10), 1253-60.

Koth, L. L., et al. (2011), 'Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis', *Am J Respir Crit Care Med,* 184 (10), 1153-63.

Krawczak, M., et al. (2006), 'PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships', *Community Genet,* 9 (1), 55-61.

Kruskal, Joseph B (1964), 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika,* 29 (1), 1-27.

Krutilina, R. I., et al. (2006), 'Migrating leukocytes are the source of peroxiredoxin V during inflammation in the airways', *J Inflamm (Lond),* 3, 13.

Kunitake, R., et al. (1999), 'Apoptosis in the course of granulomatous inflammation in pulmonary sarcoidosis', *Eur Respir J,* 13 (6), 1329-37.

Kwok, P. Y. and Gu, Z. (1999), 'Single nucleotide polymorphism libraries: why and how are we building them?', *Mol Med Today,* 5 (12), 538-43.

Lander, E. S. (1996), 'The new genomics: global views of biology', *Science,* 274 (5287), 536-9.

Lander, E. S., et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature,* 409 (6822), 860-921.

Lee, Y. H., et al. (2012a), 'Genome-wide pathway analysis of a genome-wide association study on psoriasis and Behcet's disease', *Mol Biol Rep,* 39 (5), 5953-9.

Lee, Y. H., et al. (2012b), 'Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis', *Mol Biol Rep,* 39 (12), 10627-35.

Lembrechts, R., et al. (2011), 'Expression of mechanogated two-pore domain potassium channels in mouse lungs: special reference to mechanosensory airway receptors', *Histochem Cell Biol,* 136 (4), 371-85.

Lesage, F., Maingret, F., and Lazdunski, M. (2000), 'Cloning and expression of human TRAAK, a polyunsaturated fatty acids-activated and mechano-sensitive K(+) channel', *FEBS Lett,* 471 (2-3), 137-40.

Levin, A. M., et al. (2013), 'Association of ANXA11 genetic variation with sarcoidosis in African Americans and European Americans', *Genes Immun,* 14 (1), 13-8.

Li, Y., et al. (2009), 'Genotype imputation', *Annu Rev Genomics Hum Genet,* 10, 387-406.

Li, Y., et al. (2010), 'First independent replication study confirms the strong genetic association of ANXA11 with sarcoidosis', *Thorax,* 65 (10), 939-40.

Li, Y., et al. (2006), 'BTNL2 gene variant and sarcoidosis', *Thorax,* 61 (3), 273-4.

Lian, Y., et al. (2010), 'Analysis of the association between BTNL2 polymorphism and tuberculosis in Chinese Han population', *Infect Genet Evol,* 10 (4), 517-21.

Lin, T. M., et al. (1989), 'Hepatitis B virus markers in Chinese twins', *Anticancer Res,* 9 (3), 737-41.

Livak, K. J. (1999), 'Allelic discrimination using fluorogenic probes and the 5' nuclease assay', *Genet Anal,* 14 (5-6), 143-9.

Lofgren, S. (1953), 'Primary pulmonary sarcoidosis. I. Early signs and symptoms', *Acta Med Scand,* 145 (6), 424-31.

Lyles, R. H., Lin, H. M., and Williamson, J. M. (2007), 'A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses', *Stat Med,* 26 (7), 1632-48.

Maertzdorf, J., et al. (2012), 'Common patterns and disease-related signatures in tuberculosis and sarcoidosis', *Proc Natl Acad Sci U S A,* 109 (20), 7853-58.

Magira, E. E., et al. (2012), 'HLA-A and HLA-DRB1 amino acid polymorphisms are associated with susceptibility and protection to pulmonary tuberculosis in a Greek population', *Hum Immunol,* 73 (6), 641-6.

Maillard, Ivan, Adler, Scott H, and Pear, Warren S (2003), 'Notch and the immune system', *Immunity,* 19 (6), 781-91.

Mann, Henry B and Whitney, Donald R (1947), 'On a test of whether one of two random variables is stochastically larger than the other', *The annals of mathematical statistics,* 18 (1), 50-60.

Manolio, T. A., et al. (2009), 'Finding the missing heritability of complex diseases', *Nature,* 461 (7265), 747-53.

Marchini, J., et al. (2007), 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nat Genet,* 39 (7), 906-13.

Mariano, M. (1995), 'The experimental granuloma. A hypothesis to explain the persistence of the lesion', *Rev Inst Med Trop Sao Paulo,* 37 (2), 161-76.

Marks, G. B., et al. (2000), 'Incidence of tuberculosis among a cohort of tuberculin-positive refugees in Australia: reappraising the estimates of risk', *Am J Respir Crit Care Med,* 162 (5), 1851-4.

Matsushita, E., et al. (2011), 'Protective role of Gipie, a Girdin family protein, in endoplasmic reticulum stress responses in endothelial cells', *Mol Biol Cell,* 22 (6), 736-47.

Matsuzaki, H., et al. (2004), 'Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays', *Nat Methods,* 1 (2), 109-11.

Mattick, J. S. and Gagen, M. J. (2001), 'The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms', *Mol Biol Evol,* 18 (9), 1611-30.

McCarthy, M. I. (2008), 'Casting a wider net for diabetes susceptibility genes', *Nat Genet,* 40 (9), 1039-40.

McGuire, William, et al. (1999), 'Severe malarial anemia and cerebral malaria are associated with different tumor necrosis factor promoter alleles', *Journal of Infectious Diseases,* 179 (1), 287-90.

Meilang, Q., et al. (2012), 'Polymorphisms in the SLC11A1 gene and tuberculosis risk: a meta-analysis update', *Int J Tuberc Lung Dis,* 16 (4), 437-46.

Mells, George F, et al. (2011), 'Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis', *Nature genetics,* 43 (4), 329-32.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009), 'Long non-coding RNAs: insights into functions', *Nat Rev Genet,* 10 (3), 155-9.

Minshall, E. M., et al. (1997), 'Cytokine mRNA gene expression in active and nonactive pulmonary sarcoidosis', *Eur Respir J,* 10 (9), 2034-9.

Miyara, M., et al. (2006), 'The immune paradox of sarcoidosis and regulatory T cells', *J Exp Med,* 203 (2), 359-70.

Moller, M. and Hoal, E. G. (2010), 'Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis', *Tuberculosis (Edinb),* 90 (2), 71-83.

Montgomery, Stephen B, et al. (2010), 'Transcriptome genetics using second generation sequencing in a Caucasian population', *Nature,* 464 (7289), 773-77.

Morton, N. E. and Collins, A. (1998), 'Tests and estimates of allelic association in complex inheritance', *Proc Natl Acad Sci U S A,* 95 (19), 11389-93.

Moss, Stephen E and Morgan, Reg O (2004), 'The annexins', *Genome biology,* 5 (4), 219.

Mrazek, F., et al. (2011), 'Functional variant ANXA11 R230C: true marker of protection and candidate disease modifier in sarcoidosis', *Genes Immun,* 12 (6), 490-4.

Mullis, K., et al. (1986), 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction', *Cold Spring Harb Symp Quant Biol,* 51 Pt 1, 263-73.

Nachman, M. W., et al. (1998), 'DNA variability and recombination rates at X-linked loci in humans', *Genetics,* 150 (3), 1133-41.

Natoli, G., Ghisletti, S., and Barozzi, I. (2011), 'The genomic landscapes of inflammation', *Genes Dev,* 25 (2), 101-6.

Nemeth, J., et al. (2011), 'Specific cytokine patterns of pulmonary tuberculosis in Central Africa', *Clin Immunol,* 138 (1), 50-9.

Neuman, R. J. and Rice, J. P. (1992), 'Two-locus models of disease', *Genet Epidemiol,* 9 (5), 347-65.

Neville, E., Walker, A. N., and James, D. G. (1983), 'Prognostic factors predicting the outcome of sarcoidosis: an analysis of 818 patients', *Q J Med,* 52 (208), 525-33.

Newman, L. S., Rose, C. S., and Maier, L. A. (1997), 'Sarcoidosis', *N Engl J Med,* 336 (17), 1224-34.

Newman, L. S., et al. (2004), 'A case control etiologic study of sarcoidosis: environmental and occupational risk factors', *Am J Respir Crit Care Med,* 170 (12), 1324-30.

Newman, LS (2005), 'Aetiologies of sarcoidosis', *European Respiratory Monograph,* 32, 23.

Newport, MJ, et al. (2004), 'Genetic regulation of immune responses to vaccines in early life', *Genes and immunity,* 5 (2), 122-29.

Ng, Aylwin CY (2010), 'Integrative systems biology and networks in autophagy', *Seminars in immunopathology* (32: Springer), 355-61.

Nguyen, T., et al. (2006), 'BTNL2, a butyrophilin-like molecule that functions to inhibit T cell activation', *J Immunol,* 176 (12), 7354-60.

NHGRI <http://www.genome.gov/gwastudies/>.

Nica, Alexandra C, et al. (2011), 'The architecture of gene regulatory variation across multiple human tissues: the MuTHER study', *PLoS genetics,* 7 (2), e1002003.

Nicoloso, M. S., et al. (2010), 'Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility', *Cancer Res,* 70 (7), 2789-98.

Oswald-Richter, K. A., et al. (2010), 'Multiple mycobacterial antigens are targets of the adaptive immune response in pulmonary sarcoidosis', *Respir Res,* 11, 161.

Oswald-Richter, K. A., et al. (2012), 'Dual analysis for mycobacteria and propionibacteria in sarcoidosis BAL', *J Clin Immunol,* 32 (5), 1129-40.

Ottenhoff, T. H., et al. (2002), 'Genetics, cytokines and human infectious disease: lessons from weakly pathogenic mycobacteria and salmonellae', *Nat Genet,* 32 (1), 97-105.

Pabst, S., et al. (2011), 'Caspase recruitment domain 15 gene haplotypes in sarcoidosis', *Tissue Antigens,* 77 (4), 333-7.

Pacheco, A. G., Cardoso, C. C., and Moraes, M. O. (2008), 'IFNG +874T/A, IL10 -1082G/A and TNF -308G/A polymorphisms in association with tuberculosis susceptibility: a meta-analysis study', *Hum Genet,* 123 (5), 477-84.

Page, G. P., et al. (2003), '"Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits', *Am J Hum Genet,* 73 (4), 711-9.

Parsons, V. (1960), 'Awareness of family and contact history of tuberculosis in generalized sarcoidosis', *Br Med J,* 2 (5215), 1756-9.

Pe'er, I., et al. (2008), 'Estimation of the multiple testing burden for genomewide association studies of nearly all common variants', *Genet Epidemiol,* 32 (4), 381-5.

Pemberton, T. J., et al. (2010), 'Inference of unexpected genetic relatedness among individuals in HapMap Phase III', *Am J Hum Genet,* 87 (4), 457-64.

Petukhova, Lynn, et al. (2010), 'Genome-wide association study in alopecia areata implicates both innate and adaptive immunity', *Nature,* 466 (7302), 113-17.

Price, A. L., et al. (2006), 'Principal components analysis corrects for stratification in genome-wide association studies', *Nat Genet,* 38 (8), 904-9.

Prior, C. and Haslam, P. L. (1991), 'Increased levels of serum interferon-gamma in pulmonary sarcoidosis and relationship with response to corticosteroid therapy', *Am Rev Respir Dis,* 143 (1), 53-60.

Pritchard, J. K. (2001), 'Are rare variants responsible for susceptibility to complex diseases?', *Am J Hum Genet,* 69 (1), 124-37.

Pritchard, Jonathan 'eQTL resources at the Pritchard lab', <http://eqtl.uchicago.edu/Home.html>.

Prokunina, L., et al. (2002), 'A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans', *Nat Genet,* 32 (4), 666-9.

Purcell, S., et al. (2007), 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *Am J Hum Genet,* 81 (3), 559-75.

R Core Team 'R: A Language and Environment for Statistical Computing', <http://www.r-project.org/>.

Raj, T., et al. (2013), 'Common risk alleles for inflammatory diseases are targets of recent positive selection', *Am J Hum Genet,* 92 (4), 517-29.

Ramakrishnan, L. (2012), 'Revisiting the role of the granuloma in tuberculosis', *Nat Rev Immunol,* 12 (5), 352-66.

Ramesh, V., Misra, R. S., and Jain, R. K. (1987), 'Secondary tuberculosis of the skin. Clinical features and problems in laboratory diagnosis', *Int J Dermatol,* 26 (9), 578-81.

Rava, M., et al. (2013), 'Selection of genes for gene-environment interaction studies: a candidate pathway-based strategy using asthma as an example', *Environ Health,* 12 (1), 56.

Raychaudhuri, S. (2011), 'Mapping rare and common causal alleles for complex human diseases', *Cell,* 147 (1), 57-69.

Reich, D. E. and Lander, E. S. (2001), 'On the allelic spectrum of human disease', *Trends Genet,* 17 (9), 502-10.

Reich, D. E., et al. (2001), 'Linkage disequilibrium in the human genome', *Nature,* 411 (6834), 199-204.

Reich, J. M. (2002), 'Mortality of intrathoracic sarcoidosis in referral vs population-based settings: influence of stage, ethnicity, and corticosteroid therapy', *Chest,* 121 (1), 32-9.

Reich, J. M. (2012), 'On the nature of sarcoidosis', *Eur J Intern Med,* 23 (2), 105-9.

Richmond, B. W. and Drake, W. P. (2010), 'Vitamin D, innate immunity, and sarcoidosis granulomatous inflammation: insights from mycobacterial research', *Curr Opin Pulm Med,* 16 (5), 461-4.

Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science,* 273 (5281), 1516-7.

Risch, N. J. (2000), 'Searching for genetic determinants in the new millennium', *Nature,* 405 (6788), 847-56.

Robinson, B. W., McLemore, T. L., and Crystal, R. G. (1985), 'Gamma interferon is spontaneously released by alveolar macrophages and lung T lymphocytes in patients with pulmonary sarcoidosis', *J Clin Invest,* 75 (5), 1488-95.

Rosen, Y. (2007), 'Pathology of sarcoidosis', *Semin Respir Crit Care Med,* 28 (1), 36-52.

Rosenbloom, K. R., et al. (2012), 'ENCODE whole-genome data in the UCSC Genome Browser: update 2012', *Nucleic Acids Res,* 40 (Database issue), D912-7.

Rossi, G. A., et al. (1984), 'Pulmonary sarcoidosis: excess of helper T lymphocytes and T cell subset imbalance at sites of disease activity', *Thorax,* 39 (2), 143-9.

Rossman, M. D., et al. (2003), 'HLA-DRB1*1101: a significant risk factor for sarcoidosis in blacks and whites', *Am J Hum Genet,* 73 (4), 720-35.

Russell, D. G. (2011), 'Mycobacterium tuberculosis and the intimate discourse of a chronic infection', *Immunol Rev,* 240 (1), 252-68.

Rybicki, B. A., et al. (1997), 'Racial differences in sarcoidosis incidence: a 5-year study in a health maintenance organization', *Am J Epidemiol,* 145 (3), 234-41.

Rybicki, B. A., et al. (2005), 'The BTNL2 gene and sarcoidosis susceptibility in African Americans and Whites', *Am J Hum Genet,* 77 (3), 491-9.

Rybicki, B. A., et al. (2001), 'Familial aggregation of sarcoidosis. A case-control etiologic study of sarcoidosis (ACCESS)', *Am J Respir Crit Care Med,* 164 (11), 2085-91.

Sachidanandam, R., et al. (2001), 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature,* 409 (6822), 928-33.

Sainudiin, R., Clark, A. G., and Durrett, R. T. (2007), 'Simple models of genomic variation in human SNP density', *BMC Genomics,* 8, 146.

Satagopan, J. M., et al. (2002), 'Two-stage designs for gene-disease association studies', *Biometrics,* 58 (1), 163-70.

Sato, H., et al. (2010), 'Sarcoidosis HLA class II genotyping distinguishes differences of clinical phenotype across ethnic groups', *Hum Mol Genet,* 19 (20), 4100-11.

Schoenborn, J. R. and Wilson, C. B. (2007), 'Regulation of interferon-gamma during innate and adaptive immune responses', *Adv Immunol,* 96, 41-101.

Schork, Nicholas J, et al. (2009), 'Common vs. rare allele hypotheses for complex diseases', *Current opinion in genetics & development,* 19 (3), 212-19.

Schurmann, M., et al. (2001), 'Results from a genome-wide search for predisposing genes in sarcoidosis', *Am J Respir Crit Care Med,* 164 (5), 840-6.

Seimon, T. A., et al. (2010), 'Induction of ER stress in macrophages of tuberculosis granulomas', *PLoS One,* 5 (9), e12772.

Sequenom 'iPLEX™ Assay', <http://www.sequenom.com/>.

Shi, G. L., et al. (2011), 'Association of HLA-DRB alleles and pulmonary tuberculosis in North Chinese patients', *Genet Mol Res,* 10 (3), 1331-6.

Silva, E., et al. (2013), 'Quantitative intact proteomics investigations of alveolar macrophages in sarcoidosis', *Eur Respir J,* 41 (6), 1331-9.

Singh, S. P., et al. (1983), 'Human leukocyte antigen (HLA)-linked control of susceptibility to pulmonary tuberculosis and association with HLA-DR types', *J Infect Dis,* 148 (4), 676-81.

Singla, N., et al. (2012), 'Genetic polymorphisms in the P2X7 gene and its association with susceptibility to tuberculosis', *Int J Tuberc Lung Dis,* 16 (2), 224-9.

Slatkin, M. (2008), 'Linkage disequilibrium--understanding the evolutionary past and mapping the medical future', *Nat Rev Genet,* 9 (6), 477-85.

Song, Gwan Gyu, et al. (2013), 'Genome-wide pathway analysis of a genome-wide association study on multiple sclerosis', *Molecular biology reports,* 40 (3), 2557-64.

Song, Z., et al. (2005), 'Mycobacterial catalase-peroxidase is a tissue antigen and target of the adaptive immune response in systemic sarcoidosis', *J Exp Med,* 201 (5), 755-67.

Spagnolo, P., et al. (2003), 'C-C chemokine receptor 2 and sarcoidosis: association with Lofgren's syndrome', *Am J Respir Crit Care Med,* 168 (10), 1162-6.

Spencer, C. C., et al. (2009), 'Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip', *PLoS Genet,* 5 (5), e1000477.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993), 'Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *Am J Hum Genet,* 52 (3), 506-16.

Stamatoyannopoulos, J. A. (2012), 'What does our genome encode?', *Genome Res,* 22 (9), 1602-11.

Stewart, Gareth A, et al. (2003), 'Expression of the developmental Sonic hedgehog (Shh) signalling pathway is up‐regulated in chronic lung fibrosis and the Shh receptor patched 1 is present in circulating T lymphocytes', *The Journal of pathology,* 199 (4), 488-95.

Stranger, Barbara E, et al. (2007), 'Population genomics of human gene expression', *Nature genetics,* 39 (10), 1217-24.

Suzuki, H., et al. (2012), 'Genetic Characterization and Susceptibility for Sarcoidosis in Japanese Patients: Risk Factors of BTNL2 Gene Polymorphisms and HLA Class II Alleles', *Invest Ophthalmol Vis Sci,* 53 (11), 7109-15.

Svendsen, C. B., et al. (2011), 'The continuing search for Mycobacterium tuberculosis involvement in sarcoidosis: a study on archival biopsy specimens', *Clin Respir J,* 5 (2), 99-104.

Sverrild, A., et al. (2008), 'Heredity in sarcoidosis: a registry-based twin study', *Thorax,* 63 (10), 894-6.

Szalai, AJ, et al. (2005), 'Single-nucleotide polymorphisms in the C-reactive protein (CRP) gene promoter that affect transcription factor binding, alter transcriptional activity, and associate with differences in baseline serum CRP level', *Journal of Molecular Medicine,* 83 (6), 440-47.

Taflin, C., et al. (2009), 'FoxP3+ regulatory T cells suppress early stages of granuloma formation but have little impact on sarcoidosis lesions', *Am J Pathol,* 174 (2), 497-508.

Taillon-Miller, P., et al. (2000), 'Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28', *Nat Genet,* 25 (3), 324-8.

Tanaka, T., Narazaki, M., and Kishimoto, T. (2012), 'Therapeutic targeting of the interleukin-6 receptor', *Annu Rev Pharmacol Toxicol,* 52, 199-219.

Tarazona-Santos, E. and Tishkoff, S. A. (2005), 'Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus', *Genes Immun,* 6 (1), 53-65.

Teuber, Markus, et al. (2009), 'GMFilter and SXTestPlate: software tools for improving the SNPlex™ genotyping system', *BMC Bioinformatics,* 10 (1), 81.

Thomas, K. W. and Hunninghake, G. W. (2003), 'Sarcoidosis', *JAMA,* 289 (24), 3300-3.

Thompson, J. R., Attia, J., and Minelli, C. (2011), 'The meta-analysis of genome-wide association studies', *Brief Bioinform,* 12 (3), 259-69.

Thuong, N. T., et al. (2012), 'Epiregulin (EREG) variation is associated with susceptibility to tuberculosis', *Genes Immun,* 13 (3), 275-81.

Thye, T., et al. (2009), 'MCP-1 promoter variant -362C associated with protection from pulmonary tuberculosis in Ghana, West Africa', *Hum Mol Genet,* 18 (2), 381-8.

Thye, T., et al. (2012), 'Common variants at 11p13 are associated with susceptibility to tuberculosis', *Nat Genet,* 44 (3), 257-9.

Thye, T., et al. (2010), 'Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2', *Nat Genet,* 42 (9), 739-41.

Tishkoff, S. A. and Verrelli, B. C. (2003), 'Patterns of human genetic diversity: implications for human evolutionary history and disease', *Annu Rev Genomics Hum Genet,* 4, 293-340.

Tishkoff, S. A., et al. (1996), 'Global patterns of linkage disequilibrium at the CD4 locus and modern human origins', *Science,* 271 (5254), 1380-7.

Tokuhiro, S., et al. (2003), 'An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis', *Nat Genet,* 35 (4), 341-8.

Tso, H. W., et al. (2005), 'Association of interferon gamma and interleukin 10 genes with tuberculosis in Hong Kong Chinese', *Genes Immun,* 6 (4), 358-63.

Ulrichs, T. and Kaufmann, S. H. (2002), 'Mycobacterial persistence and immunity', *Front Biosci,* 7, d458-69.

Ulrichs, T. and Kaufmann, S. H. (2006), 'New insights into the function of granulomas in human tuberculosis', *J Pathol,* 208 (2), 261-9.

Unsal, E., et al. (2005), 'Potential role of interleukin 6 in reactive thrombocytosis and acute phase response in pulmonary tuberculosis', *Postgrad Med J,* 81 (959), 604-7.

Valdar, W., et al. (2006), 'Genome-wide genetic association of complex traits in heterogeneous stock mice', *Nat Genet,* 38 (8), 879-87.

Valentonyte, R., et al. (2005), 'Sarcoidosis is associated with a truncating splice site mutation in BTNL2', *Nat Genet,* 37 (4), 357-64.

Vallinoto, A. C., et al. (2010), 'IFNG +874T/A polymorphism and cytokine plasma levels are associated with susceptibility to Mycobacterium tuberculosis infection and clinical manifestation of tuberculosis', *Hum Immunol,* 71 (7), 692-6.

van de Vosse, E., Hoeve, M. A., and Ottenhoff, T. H. (2004), 'Human genetics of intracellular infectious diseases: molecular and cellular immunity against mycobacteria and salmonellae', *Lancet Infect Dis,* 4 (12), 739-49.

van Soolingen, D., et al. (1997), 'A novel pathogenic taxon of the Mycobacterium tuberculosis complex, Canetti: characterization of an exceptional isolate from Africa', *Int J Syst Bacteriol,* 47 (4), 1236-45.

Varela, M. A. and Amos, W. (2010), 'Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence', *Genomics,* 95 (3), 151-9.

Veyrieras, Jean-Baptiste, et al. (2008), 'High-resolution mapping of expression-QTLs yields insight into human gene regulation', *PLoS genetics,* 4 (10), e1000214.

Vineis, Paolo and Pearce, Neil E (2011), 'Genome-wide association studies may be misinterpreted: genes versus heritability', *Carcinogenesis,* 32 (9), 1295-98.

Visentainer, JEL, et al. (1997), 'Association of leprosy with HLA-DR2 in a Southern Brazilian population', *Brazilian journal of medical and biological research,* 30, 51-59.

Visscher, P. M., et al. (2012), 'Five years of GWAS discovery', *Am J Hum Genet,* 90 (1), 7-24.

Wang, J., Tang, S., and Shen, H. (2010), 'Association of genetic polymorphisms in the IL12-IFNG pathway with susceptibility to and prognosis of pulmonary tuberculosis in a Chinese population', *Eur J Clin Microbiol Infect Dis,* 29 (10), 1291-5.

Wang, Y., et al. (2012), 'Weak binder for MHC molecule is a potent Mycobacterium tuberculosis-specific CTL epitope in the context of HLA-A24 allele', *Microb Pathog,* 53 (3-4), 162-7.

Weinberg, Wilhelm (1908), 'Über vererbungsgesetze beim menschen', *Molecular and General Genetics MGG,* 1 (1), 440-60.

Wellcome Trust Case Control Consortium (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature,* 447 (7145), 661-78.

WHO (2012), 'Global tuberculosis report 2012', *Global tuberculosis report* (World Health Organisation).

Wigginton, J. E., Cutler, D. J., and Abecasis, G. R. (2005), 'A note on exact tests of Hardy-Weinberg equilibrium', *Am J Hum Genet,* 76 (5), 887-93.

Wilcoxon, Frank (1945), 'Individual Comparisons by Ranking Methods', *Biometrics Bulletin,* 1 (6), 80-83.

Xiao, J., et al. (2010), 'Metaanalysis of P2X7 gene polymorphisms and tuberculosis susceptibility', *FEMS Immunol Med Microbiol,* 60 (2), 165-70.

Xu, Z. and Taylor, J. A. (2009), 'SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies', *Nucleic Acids Res,* 37 (Web Server issue), W600-5.

Yeboah, J., et al. (2012), 'Necrotizing sarcoid granulomatosis', *Curr Opin Pulm Med,* 18 (5), 493-8.

Zeggini, E. and Ioannidis, J. P. (2009), 'Meta-analysis in genome-wide association studies', *Pharmacogenomics,* 10 (2), 191-201.

Zhang, F., et al. (2011a), 'Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy', *Nature genetics,* 43 (12), 1247-51.

Zhang, F., et al. (2009), 'Genomewide association study of leprosy', *N Engl J Med,* 361 (27), 2609-18.

Zhang, J., et al. (2011b), 'Interleukin-10 polymorphisms and tuberculosis susceptibility: a meta-analysis', *Int J Tuberc Lung Dis,* 15 (5), 594-601.

Zhang, Y. (2004), 'Persistent and dormant tubercle bacilli and latent tuberculosis', *Front Biosci,* 9, 1136-56.

Zhang, Y., Broser, M., and Rom, W. N. (1994), 'Activation of the interleukin 6 gene by Mycobacterium tuberculosis or lipopolysaccharide is mediated by nuclear factors NF-IL6 and NF-kappa B', *Proc Natl Acad Sci U S A,* 91 (6), 2225-9.

Zhu, K. J., et al. (2012), 'Association of IL23R polymorphisms with psoriasis and psoriatic arthritis: a meta-analysis', *Inflamm Res,* 61 (10), 1149-54.

Ziegler, Andreas, König, Inke R, and Pahlke, Friedrich (2010), *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-learning platform* (Wiley-VCh).

Zissel, G., Prasse, A., and Muller-Quernheim, J. (2010), 'Immunologic response of sarcoidosis', *Semin Respir Crit Care Med,* 31 (4), 390-403.

Lebenslauf

Persönliche Daten

Benjamin Schmid
Geboren am 29. Juni 1981
Hamburg, Deutschland

Promotion

seit 02/2010	**Institut für Klinische Molekularbiologie,** **Christian-Albrechts-Universität zu Kiel** „Identifizierung genetischer Suszeptibilitätsloci für granulomatöse Lungenkrankheiten“ Stipendium über den „Exzellenzcluster Entzündungsforschung“

Ausbildung/Studium

04/2009 – 12/2009	**Diplomarbeit im Fachbereich Genetik in dem Labor von Prof. Dr. Rainer Renkawitz** „Functional characterization of the insulator cofactor CP190”
10/2002 – 12/2009	**Studium der Biologie an der Justus-Liebig-Universität Gießen** Abschluss: Biologie Diplom Prüfungsfächer: • Biochemie • Genetik • Zellbiologie
09/2001 – 07/2002	**Zivildienst,** Blinden- und Sehbehindertenverein Hamburg e.V.
07/1998 – 06/1999	**El Molino High School,** Forestville CA, USA
07/1993 – 06/2001	**Gymnasium Corveystraße,** Hamburg Abschluss: Abitur

Veröffentlichungen

08/2012	Hofmann S, Fischer A, Nothnagel M, Jacobs G, **Schmid B**, Wittig M, Franke A, Gaede KI, Schürmann M, Petrek M, Mrazek F, Pabst S, Grohé C, Grunewald J, Ronninger M, Eklund A, Rosenstiel P, Höhne K, Zissel G, Müller-Quernheim J, Schreiber S. (2012), ‘Genome-wide association analysis reveals 12q13.3-q14.1 as new risk locus for sarcoidosis’, *Eur Respir J*.
07/2012	Fischer A*, **Schmid B***, Ellinghaus D, Nothnagel M, Gaede KI, Schürmann M, Lipinski S, Rosenstiel P, Zissel G, Höhne K, Petrek M, Kolek V, Pabst S, Grohé C, Grunewald J, Ronninger M, Eklund A, Padyukov L, Gieger C, Wichmann HE, Nebel A, Franke A, Müller-Quernheim J, Hofmann S, Schreiber S. (2012), ‘A Novel Sarcoidosis Risk Locus for Europeans on Chromosome 11q13.1’, *Am J Respir Crit Care Med*. * These authors contributed equally to this work.

Präsentationen

09/2012

Poster-Präsentation: Symposium "Chronic Inflammatory Disorders of the Lung", *'Identification of shared genetic susceptibility loci for sarcoidosis and tuberculosis'*. Schmid B, Ellinghaus D, Fischer A, Thye T, Horstmann RD, Meyer CG, Müller-Quernheim J, Franke A, Nebel A, Schreiber S. Universitäres Lungen- und Thoraxzentrum, Freiburg, Deutschland. 28. - 29. Sept., 2012

Danksagung

Zu allererst möchte ich Prof. Dr. Stefan Schreiber dafür danken, dass ich die Möglichkeit hatte am Institut für klinische Molekularbiologie an der Christian-Albrechts-Universität zu Kiel zu promovieren. Danken möchte ich aber vor allem auch meinen drei Betreuerinnen Prof. Dr. Almut Nebel, Prof. Dr. Manuela Dittmar und Dr. Annegret Fischer, die durch ihren fachlichen und konstruktiven Rat mein Promotionsprojekt immer wieder gefördert haben, Dr. Annegret Fischer besonders dafür, dass sie mir während der Zeit der Doktorarbeit als direkte Ansprechpartnerin in fachlichen Fragen immer zur Seite stand. Von Dr. David Ellinghaus habe ich für die Bioinformatik wertvolle Unterstützung erhalten, ihm gilt Dank wie überhaupt allen Mitarbeitern des Instituts für klinische Molekularbiologie, insbesondere den Mitarbeitern in den Laboren. Des Weiteren möchte ich dem 'Exzellenzcluster Entzündungsforschung' und ganz besonders Dr. Helga Andree für den hilfreichen Informationsaustausch während der Doktorarbeit danken. Besonders danke ich Prof. Dr. Rolf Horstmann und Dr. Thorsten Thye für die Bereitstellung des Tuberkulose-GWAS-Datensatzes, der Arbeitsgruppe von Prof. Dr. Paul Van Helden für die gute Zusammenarbeit sowie allen Kooperationspartnern. Alexandra, Lilly, Sanaz, Gregor, Marcel und Thies, ohne euch wären die Kaffeepausen und der Büroalltag nicht so unterhaltsam gewesen: dafür euch allen meinen Dank. Für die Unterstützung, nötige Ablenkung und glückliche Momente danke ich meiner Freundin Sarah. Und nicht zuletzt danke ich meinen Eltern, die in jeglicher Hinsicht die Grundsteine für meinen Weg gelegt haben.

Eidesstattliche Erklärung

Hiermit versichere ich, Benjamin Schmid, an Eides statt, dass ich die vorliegende Dissertation mit dem Titel „Identifizierung genetischer Suszeptibilitätsloci für granulomatöse Lungenkrankheiten“ selbständig und unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft verfasst habe.

Ich habe dabei keine anderen als die angegebenen Hilfsmittel und Quellen verwendet und keine weitere Hilfe, außer der Beratung durch meine wissenschaftlichen Betreuer Dr. Annegret Fischer, Prof. Dr. Almut Nebel und Prof. Dr. Manuela Dittmar, in Anspruch genommen.

Die Arbeit wurde bis jetzt weder vollständig noch in Teilen einer anderen Stelle im Rahmen eines Prüfungsverfahren vorgelegt. Zudem erkläre ich, keine früheren Promotionsversuche unternommen zu haben.

Kiel, _____(Benjamin Schmid)