Identifikation von genetischen Kopienzahl-Variationen in Zusammenhang mit der komplexen Lungenkrankheit Sarkoidose

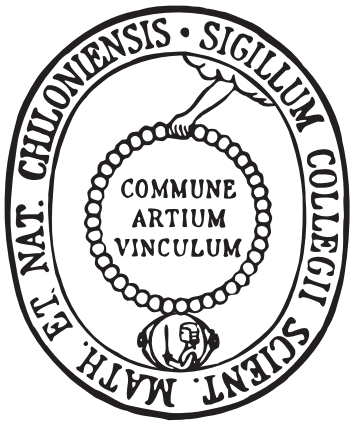
Dissertation zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Dipl.-Math. Marcel E. Nutsua



Kiel, Oktober 2014

Erstgutachter: **Prof. Dr. Hinrich Schulenburg**

Zweitgutachter: **Prof. Dr. Michael Nothnagel**

Tag der mündlichen Prüfung: **08.12.2014**

Zum Druck genehmigt: **08.12.2014**

_____

Dekan Prof. Dr. Wolfgang J. Duschl

„Perfer et obdura“

Ovid

Inhaltsverzeichnis

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

aCGH	*array comparative genomic hybridization*
ACOT1	*acyl-CoA thioesterase 1*
ANXA11	Annexin A11
APT	Affymetrix Power Tools
BAF	relative Häufigkeit des B-Allels (engl. *B-allele frequency*)
BASH	Bourne-Again-Shell (Kommandozeileninterpreter)
BIR	*break induced replikation*
BTNL2	*butyrophilin-like protein 2*
bp	Basenpaar
C_T	Anzahl der PCR-Zyklen bis zur Messung eines Fluoreszenzsignals
C10ORF67	*chromosome 10 open reading frame 67*
CCDC88B	*coiled-coil domain containing 88B*
CEU	in Utah Ansässige Individuen mit nord- und westeuropäischer Abstammung
CGH	*comparative genomic hybridization*
CN	Kopienzahl (engl. *copy number*)
CNV	Kopienzahl-Variation (engl. *copy number variation*)
DDR	Verhältniss von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*)
DGV	Database of Genomic Variation
DNA	Desoxyribonukleinsäure (engl. *deoxyribonucleic acid*)
DSB	Doppelstrangbruch
FAM	FAM™-Fluoreszenzfarbstoff
FISH	Fluoreszenz-in-situ-Hybridisierung
FoSTeS	*fork stalling and template switching*
gDNA	genomische Desoxyribonukleinsäure
GWAS	genomweite Assoziationsstudie (engl. *genome-wide association study*)
HEATR4	*HEAT repeat-containing protein 4*
HLA	humanes Leukozyten-Antigen (engl. *human leukocyte antigen*)
HMM	Hidden-Markov-Modell
HR	homologe Rekombination
IQR	Interquartilsabstand
IKMB	Institut für Klinische Molekularbiologie der Christian-Albrechts-Universität zu Kiel
IL23R	Interleukin-23-Rezeptor
kb	Kilo-Basenpaar
LCR	*low copy repeats*
LD	Kopplungsungleichgewicht (engl. *linkage disequilibrium*)
LOH	Verlust der Heterozygotie (engl. *loss of heterozygosity*)
LRR	Binärer Logarithmus Quotienten der Signalintensitäten (engl. *log2 raw data ratio*)
MAD	Median-Deviation (engl. *median absolute deviation*)

Mb	Mega-Basenpaar
MMBIR	*microhomology-mediated break-induced replication*
NAHR	nicht-allelische homologe Rekombination
NGS	Next-Generation Sequencing
NHEJ	nicht homologe End-zu-End-Verbindung (engl. *non-homologous end joining*)
NOTCH4	*neurogenic locus notch homolog protein 4*
OR	Quotenverhältnis (engl. *odds ratio*)
OS9	*osteosarcoma amplified 9*
PCR	Polymerase-Kettenreaktion (engl. *polymerase chain reaction*)
POPGEN	Biobank des populationsgenetischen Forschungsprojekts des Nationalen Genomforschungsnetzes
QC	Qualitätskontrolle (engl. *quality control*)
RAB23	ras-associated protein 23
SD	Standardabweichung (engl. *standard deviation*)
SDSA	*synthesis-dependent strand annealing*
SNV	Einzelbasen-Variation (engl. *single nucleotide variation*)
SNP	Einzelbasen-Polymorphismus (engl. *single nucleotide polymorphism*)
SSA	Einzelstranganlagerung (engl. *single strand alignment*)
SV	strukturelle Variation
TNF-α	Tumornekrosefaktor-α
V	Cramérs V (Kontingenz-Koeffizient)
VIC	VIC®-Fluoreszenzfarbstoff
WWOX	*WW domain containing oxidoreductase*
YRI	Individuen der Yoruba in Ibadan, Nigeria
χ^2	Chi-Quadrat

Englische Fachtermini

Die Übersetzung vieler Fachtermini aus dem Englischen ins Deutsche ist schwierig und in der molekularbiologischen Literatur nicht üblich, weshalb auch in dieser Arbeit darauf verzichtet wurde die folgenden Begriffe zu übersetzen.

Assay	Standardisierter Reaktionsablauf/Reaktionsansatz
Primer	Kurzes Oligonukleotid, komplementär zu DNA-Sequenzen, dient der gerichteten enzymatischen Amplifikation spezifischer DNA-Abschnitte
Well	Reaktionstasche, -gefäß oder -schale

1 Einleitung

1.1 Sarkoidose – Eine komplexe Erkrankung

Bei der Sarkoidose, auch *Morbus Boeck* oder *Morbus Schaumann-Besnier* genannt, handelt es sich um eine multisystemische Krankheit unbekannter Ätiologie. Sie ist charakterisiert durch die Bildung von Granulomen in verschiedenen Organen, hauptsächlich der Lunge und dem Lymphsystem (Valeyre et al., 2013). Im späten 19. Jahrhundert war die Krankheit nur als dermatologische Kuriosität bekannt und wurde unabhängig 1877 von Jonathan Hutchinson und 1889 von Ernest Henri Besnier beschrieben (Sharma, 2005). Erst im Jahr 1899 wurde durch Cæsar Peter Møller Boeck der Begriff Sarkoidose geprägt, als er eine Ansammlung von Epithelzellen mit großen blassen Nuklei und einigen Riesenzellen („epithelioid cells with large pale nuclei and also a few giant cell“) als multiples gutartiges Sarkom der Haut („multiple benign sarcoid of the skin“) beschrieb (Iannuzzi et al., 2007). In einem 1936 veröffentlichten Artikel vermutete Jörgen Nilsen Schaumann erstmals, dass es sich bei den bis dahin erstellten Befunden um eine systemische Erkrankung verschiedener Organe handelte (The American Thoriacic Society (ATS) et al., 1999). Durch die im Laufe der Zeit immer genauere Beschreibung des Phänotyps der Sarkoidose ist es inzwischen möglich, klinische Diagnosen zu erstellen und verschiedenste Therapien anzuwenden.

Bei zwei Drittel der Patienten tritt nach einer akut auftretenden Entzündungsreaktion ein spontaner Heilungsverlauf innerhalb von 10 Jahren nach der Diagnose ein, bei mehr als der Hälfte bereits nach weniger als drei Jahren (Iannuzzi et al., 2007; The American Thoriacic Society (ATS) et al., 1999). Das Löfgren-Syndrom ist ein spezieller Fall dieser akuten Form der Sarkoidose, charakterisiert durch gleichzeitiges Auftreten von Fieber, Lymphknotenschwellungen (Bihiläre Lymphome) und Hautveränderungen (*Erythema nodosum*) oder Gelenksentzündungen (Polyarthritis) (Iannuzzi et al., 2007; Grunewald & Eklund, 2007). In 10-30 % der Fälle wird ein chronischer Krankheitsverlauf beobachtet, bei dem sich eine zunächst schwache Entzündungsreaktion stetig verschlimmert und zu permanenter Schädigung bis Funktionsverlust der Organe führen kann. In diesen Fällen dauert die Erkrankung länger als zwei Jahre an und kann durch Schädigung der Atemwege, des Herzen oder des zentralen Nervensystems zum Tod führen (Iannuzzi et al., 2007). Neben den Symptomen des Löfgren-Syndroms führen ein anhaltender trockener Husten, Hautauschlag, Augenprobleme (z.B. eine Entzündung der mittleren Augenhaut) und allgemeine Symptome wie Müdigkeit, starker Gewichtsverlust, vermehrter Nachtschweiß und Fieber in 80 % der Fälle zur Diagnose. In weniger als 30 % erfolgt eine Diagnose durch Röntgenaufnahmen des Thorax (Valeyre et al., 2014). Zentrales

Merkmal der Sarkoidose ist die Granulombildung in dem Gewebe der betroffenen Organe, was trotz des heterogenen Krankheitsbildes auf eine gemeinsame Ursache hindeutet.

Ein besseres Verständnis der Ätiologie könnte die Therapiemöglichkeiten weiter verbessern. Es wird vermutet, dass es sich bei der Sarkoidose um eine überhöhte granulomatöse Reaktion auf ein unbekanntes Antigen handelt, welche verstärkt in Individuen auftritt, die eine entsprechende genetische Prädisposition tragen (Valeyre et al., 2013). Auch wenn der Ursprung der Krankheit immer noch unklar ist, konnten die zugrunde liegenden Mechanismen immer detaillierter beschrieben werden (Iannuzzi et al., 2007), und sogar genetische Risikofaktoren und –loci sind mittlerweile bekannt (Valeyre et al., 2013).

1.1.1 Epidemiologie

Die Sarkoidose ist eine global auftretende Erkrankung mit stark variierender Prävalenz und Inzidenz in Abhängigkeit von der geographischen Region, der ethnischen Zugehörigkeit, dem Alter und dem Geschlecht. In Europa reicht die Prävalenz von 0,2 Patienten je 100.000 Einwohner in Portugal bis zu 64 Patienten je 100.000 Einwohner in Schweden. Entsprechend der aktuellen Studienlage kann zwischen diesen beiden Extremen ein grobes Nord-Süd-Gefälle in Europa beobachtet werden (Müller-Quernheim et al., 2012; Müller-Quernheim, 1998). Auch die Inzidenz ist in den nordeuropäischen Ländern mit 5-40 Erkrankungen je 100.000 Einwohner pro Jahr am höchsten, gefolgt von 35,5 Neuerkrankungen bei afroamerikanischen Einwohnern in den USA. Damit ist die Inzidenz in den USA bei Afroamerikanern ungefähr dreimal so groß wie bei der europäisch stämmigen Bevölkerung. Die geringste Inzidenz wurde mit 1-2 Erkrankungen je 100.000 Einwohnern pro Jahr in Japan beobachtet. (Valeyre et al., 2013; Iannuzzi et al., 2007). Die Erkrankung entwickelt sich meist vor einem Alter von 50 Jahren und erreicht die höchste Inzidenz bei europäisch stämmigen Populationen in einem Alter von 20 bis 49 Jahren (Iannuzzi et al., 2007). Über alle geographischen Regionen und ethnischen Gruppen hinweg tritt die Krankheit häufiger bei Frauen als bei Männern und nur selten bei Kindern unter 15 Jahren oder Erwachsenen über 70 Jahren auf (Valeyre et al., 2013; Iannuzzi et al., 2007).

Es wird vermutet, dass die weltweiten Unterschiede in der Inzidenz zusätzlich zu der geographischen Lage von unterschiedlichen Umwelteinflüssen, Erhebungsmethoden und genetischer Prädisposition abhängen (Iannuzzi et al., 2007). Kontakt mit moderigen Umgebungen, Insektiziden oder der metallverarbeitenden Industrie gelten als Risikofaktoren (Valeyre et al., 2013; Deubelbeiss et al., 2010; Newman et al., 2004). Der sozioökonomische Status hingegen ist kein Risikofaktor, dafür aber

ein geringes Einkommen und damit verbunden ein schlechter Zugang zum Gesundheitssystem (Iannuzzi et al., 2007; Rabin et al., 2004).

Sarkoidose tritt üblicherweise vereinzelt auf, kann aber in 3,6-9,6 % der Fälle familiär vererbt sein, wobei Geschwister eines Patienten ein höheres Risiko haben ebenfalls zu erkranken als die Eltern, was auf ein rezessives Vererbungsmuster mit unvollständiger Penetranz schließen lässt (Valeyre et al., 2013; Rybicki et al., 2001). Es wird angenommen, dass genetische Faktoren für zwei Drittel der Suszeptibilität verantwortlich sind, was durch ein 80-mal größeres Risiko einer Erkrankung in monozygoten Zwillingen bestätigt wird (Valeyre et al., 2013; Sverrild et al., 2008).

1.1.2 Pathogenese

Die Erkrankung ist – immunologisch betrachtet – eine überhöhte Immunantwort auf bislang nicht identifizierte Antigene (Valeyre et al., 2013). Der Kontakt mit unbekannten Mikroorganismen sowie organischen und anorganischen Partikeln führt zur Aktivierung von antigen-spezifischen CD4+ T-Zellen durch antigen-präsentierende Zellen (Iannuzzi et al., 2007). Die aktivierten CD4+ T-Zellen sind entscheidend an der Bildung von Granulomen beteiligt (Abbildung 1-1). Anzeichen für einen möglichen Einfluss von Mykobakterien auf die Entwicklung von Sarkoidose werden diskutiert und durch mehrere Studien unterstützt (Ezzie & Crouser, 2007; Song et al., 2005; Eishi et al., 2002), konnten aber bislang nicht nachgewiesen werden. Die nekrosefreien Granulome der Sarkoidose sind kompakte Ansammlungen von Makrophagen, Epitheloidzellen und mehrkernigen Riesenzellen umgeben von Lymphozyten. Sie enthalten die Überreste der nur schwer oder nicht weiter abbaubaren Fremdkörper, welche die Immunantwort ausgelöst haben (Valeyre et al., 2013). Diese Granulome können sich auflösen, bestehen bleiben oder zur Fibrose führen (Zissel et al., 2010; Thomas & Hunninghake, 2003).

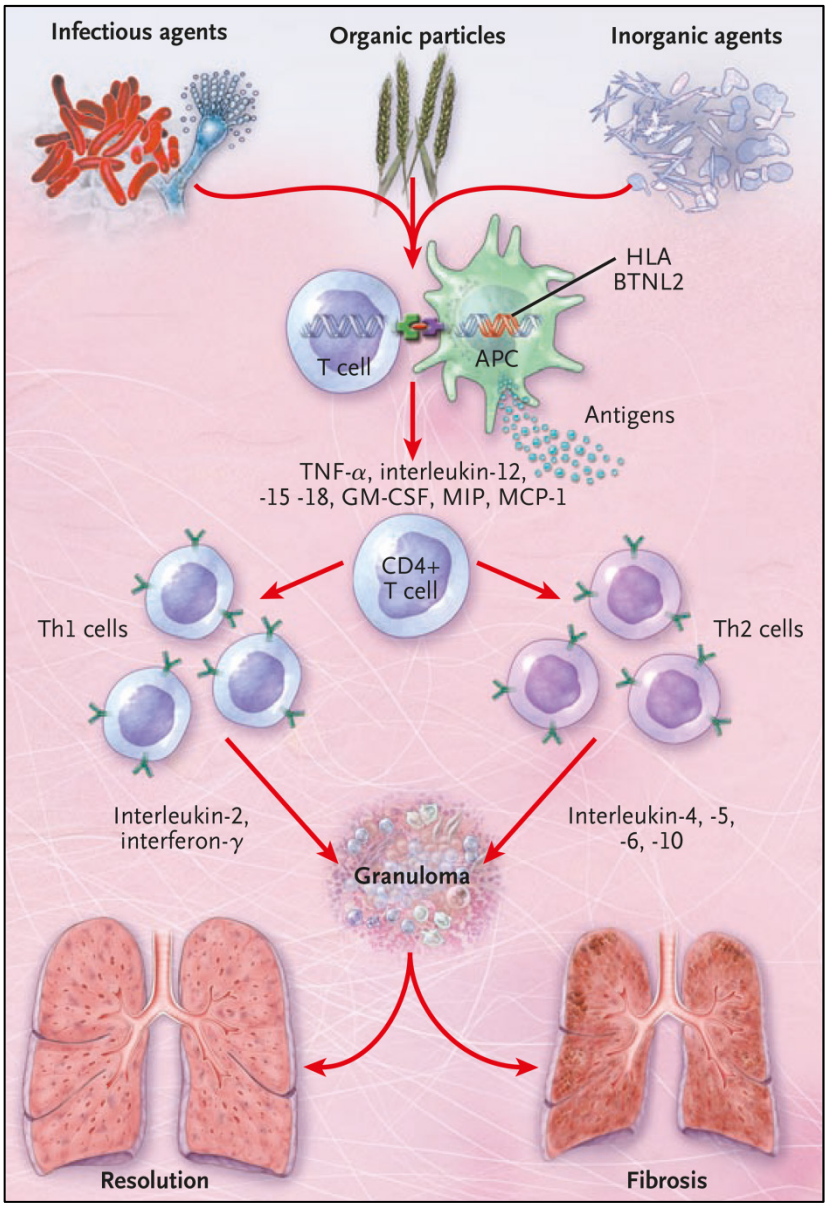


Abbildung 1-1: Postuliertes Modell der Pathogenese der Sarkoidose nach Iannuzzi et al. (2007)

Der Kontakt mit pathogen-assoziierten molekularen Mustern, die als Antigene wirken, lässt die antigen-präsentierenden Zellen TNF-α, Interleukine und Chemokine (Interleukin-12, -15, -18, MIP-1, GM-CSF) produzieren und so CD4+ T-Zellen aktivieren, die wiederum zu Typ 1 (Th1 cells) und Typ 2 (Th2 cells) Helferzellen differenzieren. Diese schütten Interleukine und Inteferon-γ aus, wodurch es zur Granulombildung kommt.

1.1.3 Genetik der Sarkoidose

Die in Abschnitt 1.1.1 beschriebene variable Prävalenz und Inzidenz und das familiär gehäufte Auftreten der Sarkoidose lässt vermuten, dass sowohl unterschiedliche Umwelteinflüsse als auch genetische Prädispositionen und die Wechselwirkung dieser Faktoren zur Entwicklung der Erkrankung beitragen, weshalb sie auch als komplexe Krankheit bezeichnet wird. Genetische Variationen können selbst zu einer Prädisposition beitragen oder als Marker genutzt werden, um durch die Assoziation mit einem Phänotyp die zugrunde liegenden Faktoren zu finden. Als Marker dienen meist Einzelnukleotid-Variationen (engl. *single nucleotide variations*, kurz SNVs) die auch Einzelnukleotid-Polymorphismen (engl. *single nucleotide polymorphisms*, kurz SNPs) genannt werden, wenn die in der Population beobachtete relative Häufigkeit der Variation > 0,01 ist.

Aufgrund ihrer z.T. sehr hohen Dichte und einer begrenzten Rekombinationsrate werden meist mehrere SNPs zusammen vererbt. Geschieht dies öfter als durch Zufall zu erwarten wäre, stehen die SNPs in Kopplungsungleichgewicht (engl. *linkage disequilibrium*, kurz LD).

Basierend auf dem postulierten Modell der Pathogenese beschränkten sich die ersten Bemühungen, genetische Risikofaktoren zu finden, vor allem auf Gene, die aufgrund ihrer möglichen Rolle in der Ätiologie ausgewählt wurden (Kandidaten-Gene). In Anbetracht der Schlüsselrolle des humanen Leukozyten-Antigen-Systems (engl. *human leukocyte antigen system*, kurz HLA-System) bei der Interaktion von antigen-präsentierende Zellen und T-Lymphozyten konzentrierten sich die ersten Studien auf allelische Assoziationen der Gene des HLA-Systems mit Sarkoidose. Die erste Assoziation eines Allels erfolgte 1977 durch Brewerton et al. zwischen dem Klasse I Antigenen HLA-B8 und akuter Sarkoidose (Iannuzzi et al., 2007). Im Zuge weiterer Studien des HLA-Systems konnten neben den Risiko-Allelen HLA-DRB1*03, DRB1*11, DRB1*12, DRB1*14 und DRB1*15, welche die Suszeptibilität, den Phänotyp sowie den Verlauf der Krankheit beeinflussen können, auch die protektiven Allele DRB1*01 und DRB1*04 gefunden werden (Fischer et al., 2014; Valeyre et al., 2013). Im Rahmen einer Kandidaten-Gen-Studie der HLA-Region wurde zusätzlich zu SNPs die Anzahl der Kopien des *C4*-Gens untersucht, der beobachtete Zusammenhang mit dem Phänotyp der Sarkoidose besaß jedoch keine statistische Signifikanz (Wennerström et al., 2012).

Aufgrund ihrer Funktion sind die Gene, die für den Tumornekrosefaktor-α (TNF-α) und den Interleukin-23-Rezeptor (IL23R) codieren, naheliegende Kandidaten für Risikofaktoren. Das *TNF-α*-Gen codiert für ein Zytokin, das bei vielen entzündlichen Prozessen eine wichtige Rolle spielt, insbesondere bei der Bildung von Granulomen (Iannuzzi & Rybicki, 2007). In mehreren Studien konnte eine Assoziation des Allels -308A des SNPs rs1800629 in der Promotorregion mit Sarkoidose detektiert werden (McDougal et al., 2009a; Labunski et al., 2001; Somoskövi et al., 1999; Seitzer et al., 1997). Zusätzlich wurde in einer Studie von Grutters et al. (2002) eine Assoziation des seltenen -857T Allels mit dem Phänotyp der Sarkoidose nachgewiesen. Durch diese Studien konnte nicht geklärt werden, ob das -308A-Allel ein von HLA-DRBI unabhängiges Risiko darstellt, da die beiden Allele in starkem LD stehen (Iannuzzi & Rybicki, 2007). Auch in einer familienbasierten Studie konnte keine Assoziation des *TNF-α*-Gens mit Sarkoidose festgestellt werden (Rybicki et al., 2004). Der Ligand des Interleukin-23-Rezeptors ist das proinflammatorische Zytokin IL23, wodurch der Rezeptor an der Erkennung von körperfremden Substanzen, wie z.B. Mykobakterien, beteiligt ist. Aufgrund der Assoziation von Variationen im *IL23R*-Gen mit der klinisch verwandten Krankheit Morbus Crohn wurde dieses Gen bereits in zwei Studien in Bezug auf Sarkoidose untersucht

(Fischer et al., 2011; Kim et al., 2011). In beiden Studien wurde eine Assoziation des seltenen Allels ARG381Gln des SNPs rs11209026 mit Sarkoidose nachgewiesen. Eine Übersicht weiterer Studien zu Kandidaten-Genen mit zum Teil widersprüchlichen Befunden ist in Tabelle 1-1 dargestellt.

Weitere Assoziationen mit zuvor nicht mit der Erkrankung in Verbindung gebrachten Loci konnten durch hypothesenfreie genomweite Studien entdeckt werden (Tabelle 1-1). Ein Risiko-Allel im Gen für das *butyrophilin-like protein 2* (BTNL2) wurde durch eine Kopplungsstudie und ein anschließendes Eingrenzungsverfahren entdeckt (Valentonyte, Hampe, Huse, et al., 2005; Schürmann et al., 2001). Die Assoziation von Varianten im *BTNL2*-Gen wurde im Folgenden mehrfach und in unterschiedlichen Populationen repliziert (Cozier et al., 2013; Morais et al., 2012; Suzuki et al., 2012; Adrianto et al., 2012; Milman et al., 2011; Wijnen et al., 2011; Li et al., 2006; Rybicki et al., 2005). Das durch das BTNL2-Gen codierte Protein spielt eine wichtige Rolle bei der Aktivierung von T-Zellen (Nguyen et al., 2006) und das Risiko-Allel A des SNPs rs2076530 führt zu einem alternativen Spleißen und so zu einem verkürzten Protein (Valentonyte, Hampe, Huse, et al., 2005). Durch die Anwendung von SNP-basierten genomweiten Assoziationsstudien (GWAS) konnte eine Reihe weiterer Risiko-Allele gefunden werden. Bei der ersten GWAS für Sarkoidose wurde durch Hofmann et al. (2008) eine Assoziation des T Allels des häufigen nicht-synonymen SNPs rs1049550 in dem für *Annexin A11* (ANXA11) codierenden Gen mit Sarkoidose festgestellt. Dieser Befund konnte sowohl in unabhängigen europäischen Populationen (Morais et al., 2013; Adrianto et al., 2012; Li et al., 2010) sowie bei Amerikanern afrikanischer und europäischer Herkunft validiert werden (Levin et al., 2013; Mrazek et al., 2011). Das codierte Protein Annexin A11 spielt bei einer Vielzahl zellulärer Prozesse, wie z.B. der Apoptose von aktivierten Entzündungszellen, eine Rolle (Moss & Morgan, 2004).

Weitere durch GWAS entdeckte Assoziationen betreffen die Genregionen des *ras-associated protein 23* (*RAB23*), des *chromosome 10 open reading frame 67* (*C10ORF67*), des *osteosarcoma amplified 9* (*OS9*), des *coiled-coil domain containing 88B* (*CCDC88B*) und des *neurogenic locus notch homolog protein 4* (*NOTCH4*). Alle in Bezug auf Sarkoidose bisher durchgeführten GWAS nutzten SNPs als Marker, wobei die Assoziation nicht zwangsweise durch den beobachteten SNP sondern, auch durch Varianten in LD verursacht werden können. Aus diesem Grund werden die auf diese Art identifizierten Gene trotz statistischer Assoziation der einzelnen Varianten weiterhin als Kandidaten bezeichnet, wobei die Identifizierung der kausalen Variation noch aussteht.

Tabelle 1-1: Untersuchte Risiko-Loci der Sarkoidose außerhalb der HLA-Region

Locus	Gen	Variation	Allel	Assoziation	Studie(n)
Kandidaten					
17q23.3	ACE	Intron 8 InDel	287 bp Del.	SV (+/-)	(Rybicki et al., 2004; McGrath et al., 2001; Maliarik et al., 1998; Furuya et al., 1996; Arbustini et al., 1996)
3p21.31	CCR2	rs1799864	A	SV (+/-)	(Valentonyte, Hampe, Croucher, et al., 2005; Spagnolo et al., 2003; Petrek et al., 2000; Hizawa et al., 1999)
3p21.31	CCR5	CCR5Δ32	32 bp Del.	SV (-)	(Fischer et al., 2008; Spagnolo et al., 2005; Petrek et al., 2000)
5q31.3	CD14	CD14 Promotor C->T	T	S (+)	(Fridlender et al., 2010; Gazouli et al., 2005)
11q12.3	SCGB1A1	rs3741240	A	SV (-)	(Rob Janssen et al., 2004; Ohchi et al., 2004)
1q32.2	CR1	Pro1827Arg(C507G)	G	S	(Zorzetto et al., 2002)
7q31.2	CFTR	R75Q	Q	V (+/-)	(Makrythanasis et al., 2010; Schürmann et al., 2002; Bombieri et al., 2000)
15q13.3	GREM1	rs1919364	C	S	(Heron et al., 2011)
6p21.33	HSPA1L	HSP70 +2437 C	C	SV (-)	(Bogunia-Kubik et al., 2006; Ishihara et al., 1995)
14q13.2	NFKBI	IκB -297T	T	SV	(Abdallah et al., 2003)
2q13	IL-1α	IL-1α -889C	C	S (-)	(Vasakova et al., 2010; Grutters et al., 2003; Hutyrová et al., 2002)
7p15.3	IL6	IL-6 -174G>C	C	S (-)	(Vasakova et al., 2010; Maver et al., 2007)
5p13.2	IL7R	rs10213865	A	S	(Heron, Grutters, van Moorsel, et al., 2009)
1q32.1	IL10	IL10 -819C & IL10 -592C	C, C	S (-)	(Sakuyama et al., 2012; Vasakova et al., 2010; McDougal et al., 2009b; Muraközy et al., 2001)
11q23.1	IL18	IL18 -607 (A/C)	A	S (+/-)	(Maver et al., 2007; Zhou et al., 2005; R Janssen et al., 2004; Takada et al., 2002)
1p31.3	IL23R	rs11209026	A	S	(Fischer et al., 2011; Kim et al., 2011)
12q15	IFN-γ	T551G (Ile184Arg)	G	S	(Akahoshi et al., 2004)
17p13.2	ITGAE	-1088 A/G	A	S	(Heron, Grutters, Van Moorsel, et al., 2009)
20q13.12	MMP9	T-1702A	T	S	(Piotrowski et al., 2011)
10p12.33	MRC1	111380T/C (rs691005)	C	S	(Hattori et al., 2010)
3p22.2	MyD88	-938A/1944G	A/G	S	(Daniil et al., 2013)
16q12.1	NOD2	2104T (702W)	T	SV (-)	(Pabst, Golebiewski, et al., 2011; Fischer et al., 2011; Sato et al., 2010; Akahoshi et al., 2008; Milman et al., 2007; Zorzetto et al., 2005; Gazouli et al., 2005; Schürmann et al., 2003)
1q31.1	PTGS2	-765C, T8473C, G3050C	C,C,C	S	(Lopez-Campos et al., 2008, 2009; Hill et al., 2006)
2q35	SLC11A1	Promotor $(GT)_n$ Repeat	Allel 3	S (-)	(Dubaniewicz et al., 2005; Maliarik et al., 2000)
19q13.2	TGF-β1	-509C, 10T	C,T	V	(Jonth et al., 2007; Niimi et al., 2002; Muraközy et al., 2001)
1q41	TGF-β2	59941G,		V	(Pabst, Fränken, et al., 2011; Kruit et al., 2006)
14q24.3	TGF-β3	4875A, 17369C,15101G		SV	(Kruit et al., 2006)
9q33.1	TLR4	Asp299Gly & Thre399Ile		V	(Pabst et al., 2006)
3p21.2	TLR9	T1237C	C	V (-)	(Pabst et al., 2013; Veltkamp et al., 2010)
6p21.33	TNF-α	-307 A/G, -857T	A, T	SV (-)	(Rybicki et al., 2004; Grutters et al., 2002; Labunski et al., 2001; Somoskövi et al., 1999; Seitzer et al., 1997; Wilson et al., 1992)
6p21.1	VEGFA	+813T	T	SV	(Pabst et al., 2010; Seyhan et al., 2008; Morohashi et al., 2003)
12q13.11	VDR	*Bsm1 restriction site*	B	S (-)	(Rybicki et al., 2004; Guleva & Seitzer, 2000; Niimi et al., 1999)
GWAS					
6p21.32	BTNL2	rs2076530	A	S (+)	(Cozier et al., 2013; Morais et al., 2012; Suzuki et al., 2012; Milman et al., 2011; Wijnen et al., 2011; Li et al., 2006; Rybicki et al., 2005; Valentonyte, Hampe, Huse, et al., 2005; Schürmann et al., 2001)
10q22.3	ANXA11	rs1049550	T	S (+)	(Levin et al., 2013; Adrianto et al., 2012; Li et al., 2010; Hofmann et al., 2008)
6p11.2	RAB23	rs1040461	A	S (+)	(Adrianto et al., 2012; Hofmann et al., 2011)
10p12.2	C10ORF67	rs1398024	A	S (+)	(Cozier et al., 2012; Franke et al., 2008)
12q13.3	OS9	rs1050045	C	S	(Hofmann et al., 2013)
11q13.1	CCDC88B	rs479777	C	S	(Fischer et al., 2012)
6p22.3	-	rs11966463	C	S	(Rybicki et al., 2011)
6p21.32	NOTCH4	rs715299	G	S	(Adrianto et al., 2012)

Allele: Unabhängig assoziierte Allele am selben Locus sind getrennt durch ein Komma, bei Assoziationen von Haplotypen sind die Allele getrennt durch einen Schrägstrich. **Assoziation:** Suszeptibilität (S), Krankheitsverlauf (V), validiert (+), entkräftet (-)

1.2 Kopienzahl-Variationen

Die Abweichungen eines Humangenoms von einem Referenzgenom werden als genetische Variation bezeichnet und reichen von Variationen, die jeweils nur ein Basenpaar (bp) betreffen, bis zu relativem Zugewinn oder Verlust ganzer Chromosomen (Aneuploidie) (Tabelle 1-2). Neben der Größe der Variationen ist für eine Klassifizierung entscheidend, ob die Struktur des Genoms betroffen ist oder nicht. Strukturelle Variationen (SV) beschreiben dabei Varianten, die zur Umgestaltung des Genoms führen wie Segmente, die im Vergleich zur Referenz ihre Position im Genom verändern (Translokationen), ihre Orientierung umkehren (Inversionen) sowie Segmente die in variabler Kopienzahl auftreten und als Kopienzahl-Variation (engl. *copy number variation*, kurz CNV) bezeichnet werden (Abbildung 1-2). Nach dieser Definition können CNVs im Vergleich zu einem Referenzgenom als Deletionen (d.h. Kopien des Segmentes fehlen) oder Duplikationen (d.h. Kopien des Segmentes sind hinzugekommen) beschrieben werden. Es ist Konvention bei der Beschreibung von CNVs eine Größe von mindestens einem Kilo-Basenpaar (kb) anzunehmen (Conrad et al., 2010; Redon et al., 2006), auch wenn der Übergang zu anderen Formen von Variation fließend ist, und in neueren Studien unter Verwendung moderner Methoden der Sequenzierung mit hoher Auflösung strukturelle Variationen ab einer Größe von 50 bp detektiert wurden (Mills et al., 2011).

Tabelle 1-2: Klassifizierung genomischer Variationen nach Sharp et al. (2006)

Klasse	Definition	Größe	Referenz
Einzelbasenveränderung	Einzelnukleotidpolimorphismen, Punktmutationen (SNVs, SNPs)	1bp	The International HapMap Consortium (2005)
Kleine Insertionen und Deletionen	Diallelische Insertionen/Deletionen (InDels)	1-50bp	Weber et al. (2002)
Short Tandem Repeats	Mikrosatelliten und andere einfache repetitiven Sequenzen	1-500bp	Dib et al. (1996)
Kleine strukturelle Variationen	Duplikationen, Deletionen, Tandemwiederholungen, Inversionen	50bp – 5kb	(Conrad et al., 2010; McCarroll et al., 2006; Tuzun et al., 2005)
Mittlere strukturelle Variationen	Duplikationen, Deletionen, Tandemwiederholungen, Inversionen	300bp – 10kb	(Conrad et al., 2010; McCarroll et al., 2006)
Große strukturelle Variationen	Duplikationen, Deletionen, große Tandemwiederholungen, Inversionen	5kb – 50kb	(Conrad et al., 2010; McCarroll et al., 2006; Iafrate et al., 2004)
Chromosomale Variationen	Zytogenetisch sichtbare Deletionen, Duplikationen, Translokationen und Inversionen sowie Aneuploidie	~5Mb bis ganze Chromosomen	(Shaffer & Lupski, 2000; Jacobs et al., 1992)

Die angegebene Größe dient der groben Einordnung und definiert nicht die Klasse, Übergänge zwischen den Klassen sind fließend.

Viele Mechanismen, die zu strukturelle Veränderungen am Genom und damit auch zu CNVs führen, beruhen auf der homologen Rekombination (HR) und der nicht-homologen Rekombination. Die am

besten untersuchten Mechanismen basieren auf Fehlern in der HR. Andere wichtige Mechanismen, z.B. die nicht homologe End-zu-End-Verbindung, führen zu sehr kleinen strukturellen Veränderungen. Die resultierenden Variationen werden aufgrund ihrer Größe häufig nicht als CNVs klassifiziert. Weitere Mechanismen sind bislang nur wenig untersucht, wie z.B. der Wechsel der Matrize bei einem Abbruch der Replikationsgabel (Hastings, Lupski, et al., 2009). In diesem Abschnitt werden verschiedene Reparaturmechanismen vorgestellt, die z.T. auf HR basieren, sowie Mechanismen der Replikation, die zu CNVs führen können.

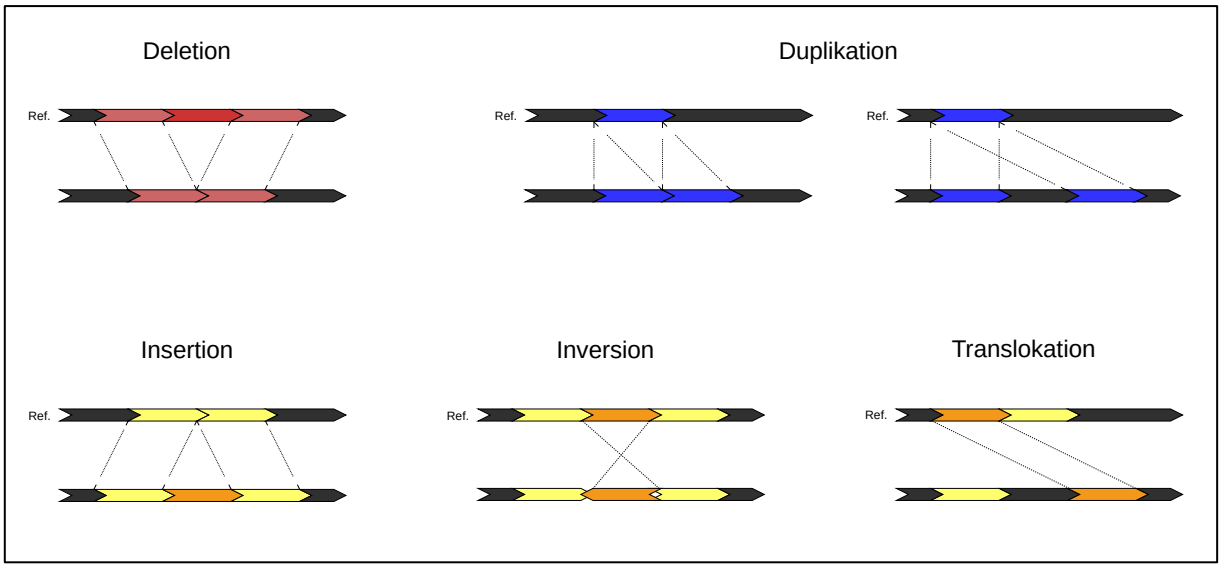


Abbildung 1-2: Strukturelle Variationen

In der oberen Reihe sind schematisch die verschiedenen Arten von CNVs dargestellt, Deletionen (rot) und Duplikationen (blau). Liegen die duplizierten Segmente unmittelbar nebeneinander spricht man von einer Tandem-Duplikation. Die untere Reihe zeigt weitere Formen häufig auftretender struktureller Variationen (gelb).

1.2.1 Reparaturmechanismen basierend auf homologer Rekombination

Die HR ist ein Mechanismus zur Reparatur von Doppelstrangbrüchen (DSB), bei dem es zum Austausch von genetischen Informationen zwischen homologen Regionen zweier DNA-Moleküle kommt. Nach einem DSB in einem DNA-Molekül werden die 5‘-Enden auf beiden Seiten des DSB durch eine Exonuklease entfernt, und ein freies 3‘-Ende verbindet sich unter Einwirkung des Proteins Rad51 mit dem komplementären Strang des homologen Moleküls (Abbildung 1-3A). Bei der Anlagerung und Verlängerung des gebrochenen Moleküls entlang des homologen Einzelstranges des anderen Moleküls wird dessen zweiter Einzelstrang verdrängt. Ist der verlängerte Einzelstrang lang genug, um die durch den DSB entstandenen Lücken zu überbrücken, kann es zur Auflösung der Struktur kommen. Dabei lagert sich die andere Seite des gebrochenen Moleküls an das verlängerte 3‘-Ende an, und die Lücken werden geschlossen. Dieser Vorgang wird *synthesis-dependent strand*

annealing (SDSA) genannt. Es kann aber auch zu einer Anlagerung des zweiten 3‘-Endes des gebrochenen Moleküls an den verdrängten Einzelstrang des homologen Moleküls kommen, wodurch die Einzelstränge beider Moleküle lokal überkreuzen und an Einzelstränge des jeweils anderen Moleküls gebunden sind (Holliday-Strukturen). Bei der Auflösung dieser Strukturen durch eine Endonuklease kann es dann zu einer chromosomalen Rekombination kommen (Hastings, Lupski, et al., 2009).

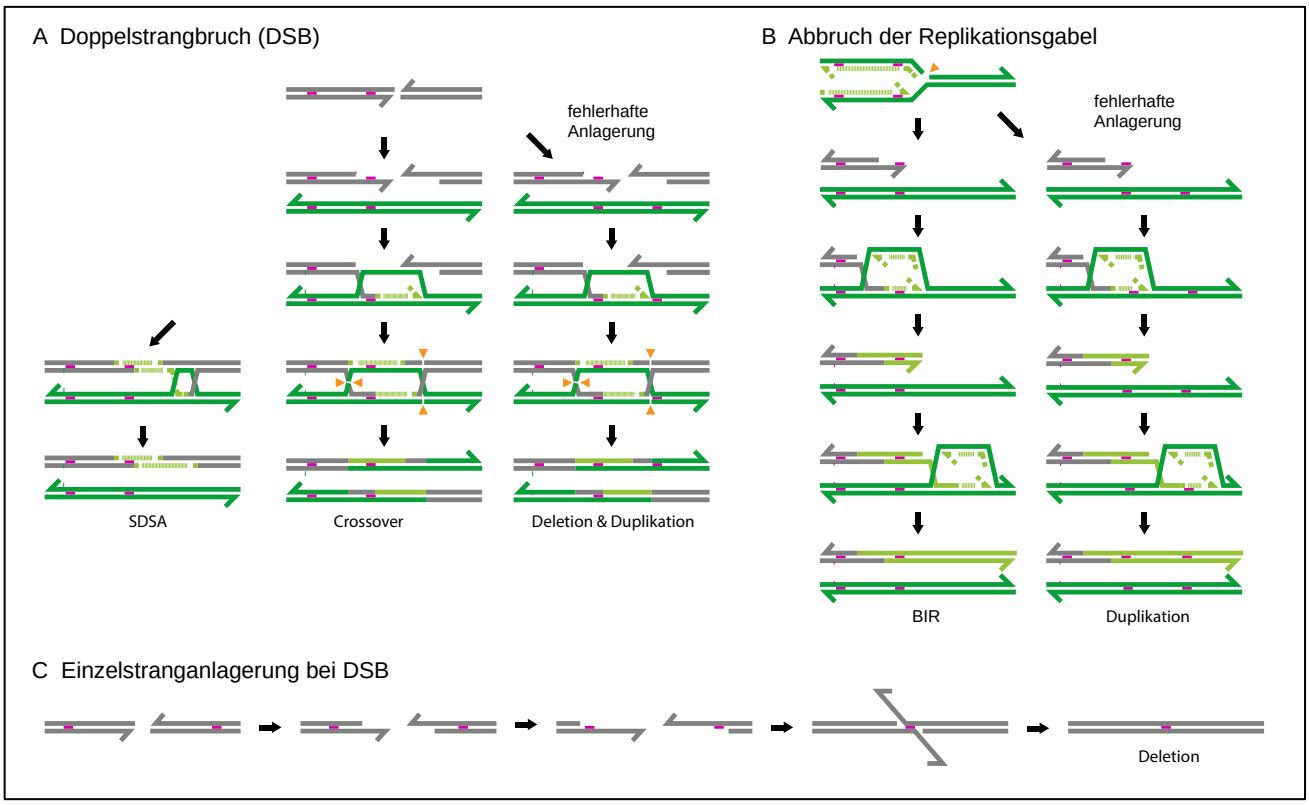


Abbildung 1-3: Fehler bei Reparaturmechanismen basierend auf homologer Rekombination

A) und B) Bei Reparaturmechanismen basierend auf homologer Rekombination wird ein Doppelstrangbruch eines DNA-Moleküls (grau) repariert, indem ein homologes DNA-Molekül (grün) als Vorlage genutzt wird, um verlorengegangene Sequenzen neu zu synthetisieren (hellgrün). **C)** Liegt auf beiden Seiten des Doppelstrangbruchs Mikrohomologie vor, z.B. durch eine repetitive Sequenz (violett), so kann eine fehlerhafte Reparatur auch ohne ein anderes Molekül erfolgen. Schematische Darstellung nach Hastings, Lupski, et al. (2009).

HR Mechanismen kommen auch bei einem Bruch der Replikationsgabel zum Tragen, wenn der Leitstrang beschädigt ist und die Replikation abbricht (Abbildung 1-3B). In diesem Fall gibt es nur ein freies 3‘-Ende, welches eine Bindung mit dem komplementären Strang des homologen Moleküls eingeht und so eine Holliday-Struktur bildet. Da kein zweites freies 3‘-Ende verfügbar ist, löst sich die Struktur wie bei dem SDSA auf. Da der neu synthetisierte Strang immer noch ein offenes Ende

besitzt, wird dieser Vorgang so lange wiederholt, bis sich erneut eine stabile Replikationsgabel bilden kann. Dieser Mechanismus wird auch als *break induced replikation* (BIR) bezeichnet.

Fehler in diesen Reparaturmechanismen können zu CNVs führen. Lagert sich das freie 3‘-Ende eines gebrochen DNA-Moleküls an einer nicht-allelischen homologen Region eines anderen Moleküls an (z.B. einer repetitiven Sequenz), kann es zur nicht-allelischen homologen Rekombination (NAHR) kommen (Abbildung 1-3A). Ist dies der Fall bei einem DSB, entstehen in den resultierenden Doppelsträngen sowohl eine Deletion als auch eine Duplikation der Sequenz zwischen den repetitiven Sequenzen (Hastings, Lupski, et al., 2009). NAHR kann auch während einer BIR auftreten (Smith et al., 2007), wobei lediglich eine Deletion oder Duplikation entsteht (Abbildung 1-3B).

1.2.2 Nicht homologe Reparaturmechanismen

Wenn nach einem DSB das freie 3‘-Ende nicht an eine komplementäre Sequenz binden kann, wird das 5‘-Ende immer weiter abgebaut. Wenn durch den Abbau der 5‘-Enden auf beiden Seiten des DSB komplementäre Einzelstränge, z.B. einer repetitiven Sequenz, freigelegt werden, so können sich diese an einander anlagern (Abbildung 1-3C). Das Entfernen der überhängenden 3‘-Enden und Auffüllen der noch bestehenden Lücken auf den Einzelsträngen vervollständigt diesen Reparaturprozess, der auch Einzelstranganlagerung (engl. *single strand alignment*, kurz SSA) genannt wird. Dabei geht das Segment zwischen den repetitiven Sequenzen verloren (Haber, 1992).

Häufig wird auch die nicht homologe End-zu-End-Verbindung (engl. *non-homologous end joining*, kurz NHEJ) als Mechanismus angegeben, der zur Entstehung von CNVs führen kann. Dabei werden ggf. beschädigte oder überstehende Basen des DSBs entfernt und die Enden anschließend durch eine Ligase neu verbunden (Lieber et al., 2003). Die durch diesen Mechanismus entstehenden Deletionen sind mit 1-4 bp jedoch so klein, dass sie nicht in die obige Definition von CNVs fallen und an dieser Stelle nicht weiter betrachtet werden.

1.2.3 Mechanismen basierend auf DNA-Replikation

Neben Fehlern bei Reparaturmechanismen kann es auch bei der DNA-Replikation zu Fehlern kommen, die zu CNVs führen, allerdings sind diese Mechanismen noch nicht so gut untersucht wie die bereits vorgestellten Reparaturmechanismen. Es gibt aber zunehmend Belege dafür, dass Replikationsmechanismen bei der Entstehung zumindest einen Teil der strukturellen Variation involviert sind (Hastings, Ira, et al., 2009). Insbesondere das Verlangsamen oder Abbrechen der Replikationsgabel (Replikationsstress) an chromosomalen fragilen Stellen wird seit langem mit der

Entstehung von CNVs in Verbindung gebracht (Carr & Lambert, 2013; Arlt et al., 2009, 2012; Coquelle et al., 1997, 2002; Kuo et al., 1994).

Wenn während des Replikationsvorgangs kurze Abschnitte identischer Sequenzen auftreten, kann es bei der Bildung einer Sekundärstruktur im Folgestrang zur Deletion des betroffenen Segments kommen (Abbildung 1-4A). Dieser Vorgang, der nur auf der Länge einer Replikationsgabel operieren kann, wird *replication slippage* genannt (Chen et al., 2005) und erzeugt Deletionen von der Länge eines Okazaki-Fragments. Aufgrund ihrer Größe fallen diese nicht in die obige Definition von CNVs, können aber durch moderne Methoden der Sequenzierung detektiert werden.

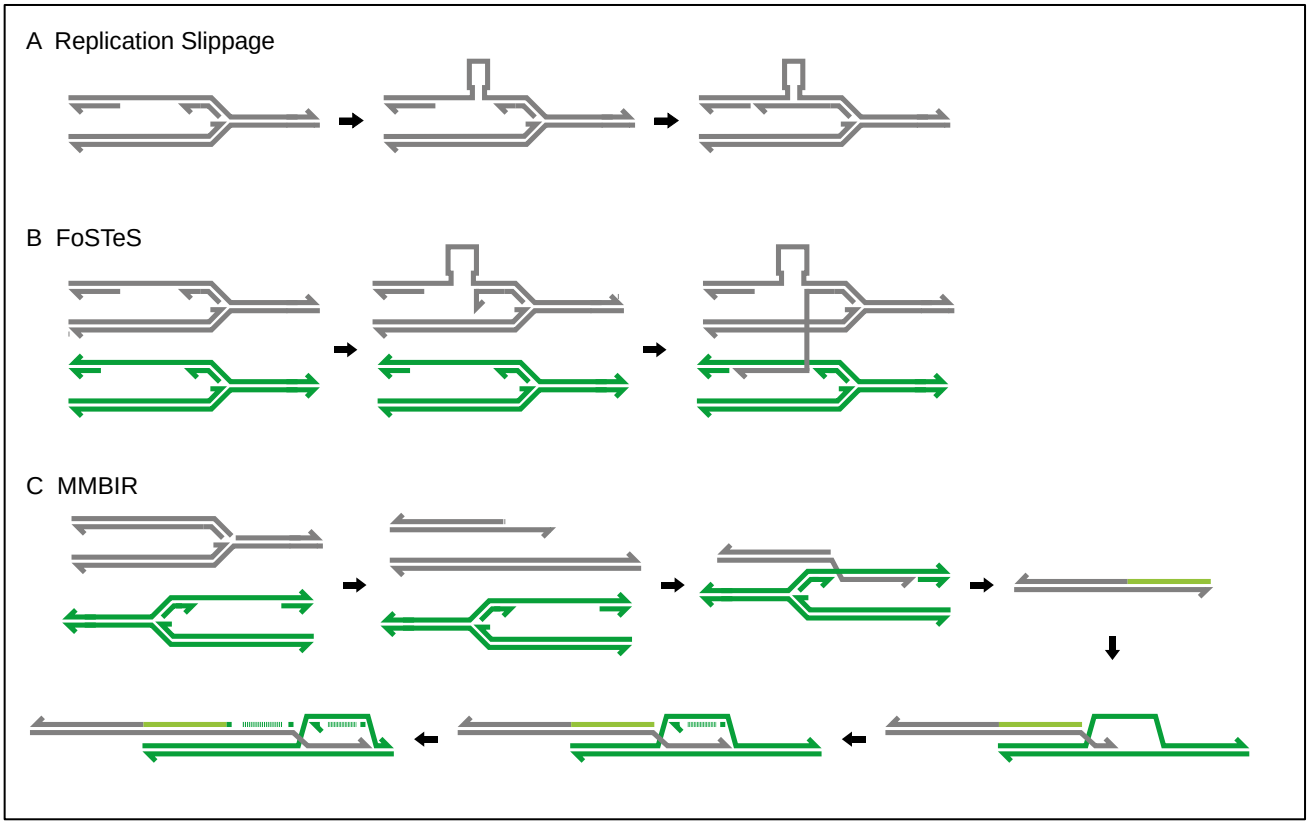


Abbildung 1-4: Entstehung von CNVs durch Fehler in der DNA-Replikation

A) und B) Die Entstehung von Sekundärstrukturen während der DNA-Replikation können zu Deletionen und Duplikationen führen, wobei auch die Informationen eines anderen DNA-Moleküls (grün) integriert werden können. **C)** Nach Abbruch der Replikation entsteht ein DNA-Molekül mit offenem Ende (grau), welches sich an den Folgestrang einer anderen Replikationsgabel (grün) anlagern kann, sofern Mikrohomologie besteht. Schematische Darstellung nach Hastings, Lupski, et al. (2009).

Entsteht bei der Replikation eine Sekundärstruktur im Folgestrang, die keine repetitive Sequenz betrifft, kann die Replikationsgabel blockiert werden (Abbildung 1-4B). Das 3‘-Ende des neu synthetisierten Moleküls löst sich von der Matrize und kann dann in eine andere Replikationsgabel integriert werden, sofern Mikrohomologie zum dortigen Folgestrang vorhanden ist. Dies kann zu

Deletionen, Duplikationen oder anderen strukturellen Variationen führen (Zhang et al., 2009; Lee et al., 2007) und wird *fork stalling and template switching* (FoSTeS) genannt.

Wenn wie bei der BIR ein Arm der Replikationsgabel abbricht. aber nicht genügende Mengen des Proteins Rad51 zur DNA-Anlagerung vorhanden sind, um den Fehler durch homologe Rekombination zu reparieren, kann der freie 3‘-Einzelstrang an einen freien Leitstrang einer anderen Replikationsgabel binden, falls Mikrohomologie vorliegt (Abbildung 1-4C). Da nur wenig Homologie erforderlich ist, kann die Anlagerung am Schwesterchromatid entweder vor oder hinter der ursprünglichen Position der Replikationsgabel stattfinden, was zu Deletionen oder Duplikationen führt. Andere Arten von struktureller Variation können ebenfalls auftreten, wenn die Anlagerung in umgekehrter Ausrichtung (Inversion) oder an einem anderen Chromosom (Translokation) erfolgt. Eine Anlagerung am homologen Chromosom anstelle des Schwesterchromatids führt zu dem Verlust der Heterozygotie (engl. *loss of heterozygosity*, kurz LOH). Dieser Vorgang wird *microhomology-mediated break-induced replication* (MMBIR) genannt (Hastings, Lupski, et al., 2009).

1.2.4 Einfluss von CNVs auf Phänotypen

Strukturelle Variationen beeinflussen durch ihre Auswirkungen auf Zellen, Organismen und Populationen alle Aspekte der Genetik und Populationsgenetik (Hurles et al., 2008) und sind für einen Großteil der Variation des Humangenoms verantwortlich (Conrad et al., 2010). In einer auf Mikro-Array-Methoden basierenden Studie von Conrad et al. (2010) betrafen die in 41 Individuen beobachteten CNVs insgesamt 3,7 % des Genoms. Die Autoren beobachten darüber hinaus, dass im Mittel 40,5 % aller in einer Probe detektierten CNVs Genregionen betrafen, die wiederum 3,1 % aller zu der Zeit bekannten Gene (RefSeq) ausmachten (Abbildung 1-5).

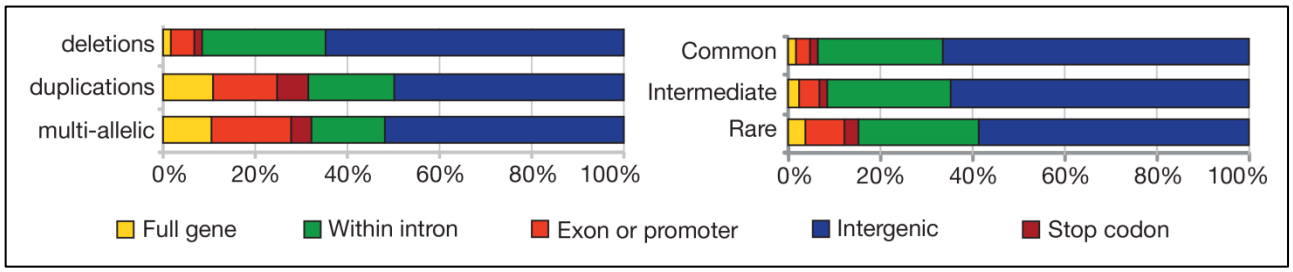


Abbildung 1-5: Von CNVs betroffene Regionen des Genoms nach Conrad et al. (2010)

Anteil der validierten CNVs in unterschiedlichen Bereichen des Genoms, aufgeteilt nach CN-Klasse (links) und relativer Häufigkeit h (rechts). Common: $h \geq 0{,}1$. Intermediate: $0{,}1 \geq h \geq 0{,}01$. Rare: $h \leq 0{,}01$.

Funktionelle Auswirkungen reichen von zellulären Vorgängen wie Genexpression (Stranger et al., 2007) bis zu den genetischen Grundlagen komplexer Erkrankungen (Buchanan & Scherer, 2008).

Strukturelle Variationen, insbesondere CNVs, können einerseits keinen oder nur wenig Einfluss auf einen Phänotyp haben, oder andererseits einen Risikofaktor für eine Erkrankung darstellen (Abbildung 1-6; Feuk et al., 2006). Der direkteste Einfluss von CNVs auf einen Phänotyp entsteht bei einer Veränderung der für den Phänotyp relevanten codierenden Sequenzen. Ist ein ganzes Gen von einer CNV betroffen, so kann dieses im Vergleich zu einem Referenzgenom nicht mehr vorhanden sein oder in höherer Kopienzahl vorliegen, was sich direkt auf das Expressionslevel des Gens auswirken kann. Deletionen auf nur einem der homologen Chromosomen können zur Demaskierung rezessiver Allele führen. Sind genregulatorische Regionen in ganzem Umfang von einer CNV betroffen, kann dies ebenfalls die Regulation und Expression der Gene verändern. Sind codierende Regionen nur teilweise von CNVs betroffen, kann auch das schwerwiegende Folgen haben, da schon kleine Veränderungen der Sequenz die Funktion der ganzen Region verändern oder zerstören können.

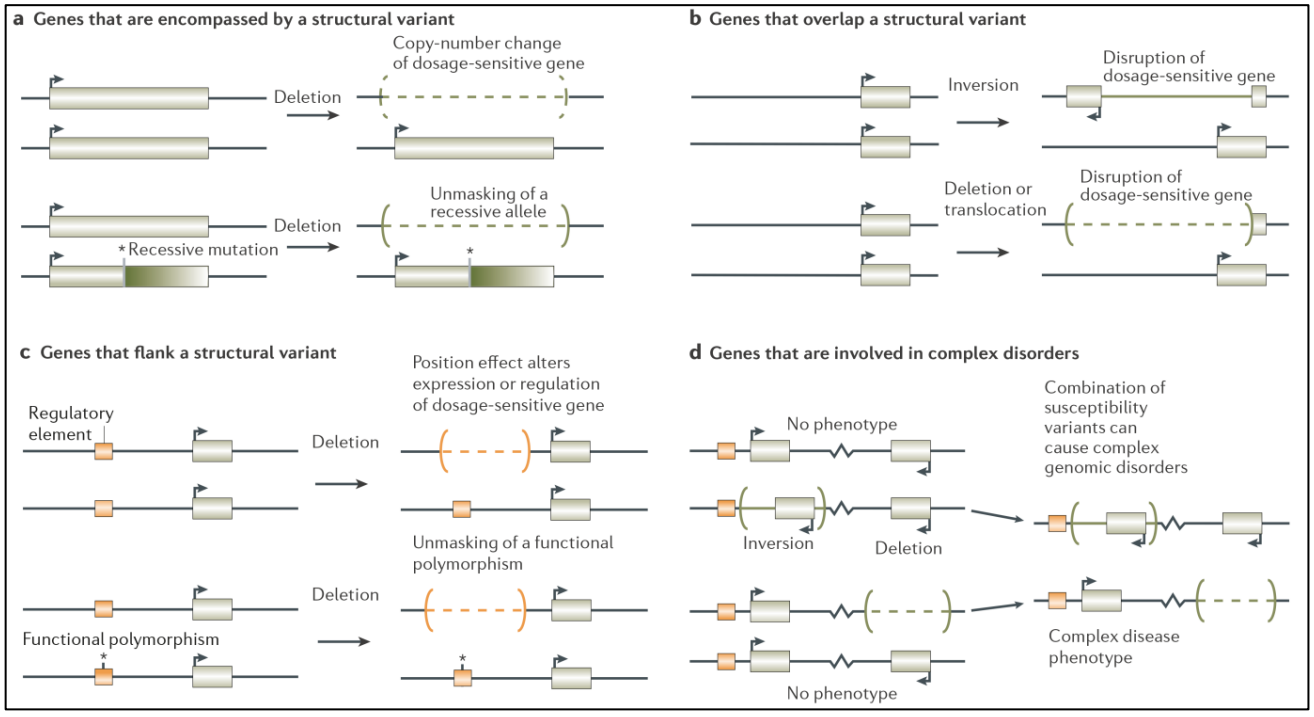


Abbildung 1-6: Der Einfluss von CNVs auf Phänotypen nach Feuk et al. (2006)

Sind Gene oder regulatorische Elemente ganz oder teilweise von strukturellen Variationen (SVs) betroffen, kann dies ihre Funktion ändern, einschränken oder unmöglich machen. Dargestellt sind **A)** Gene, die durch eine SV nicht mehr vorhanden sind, **B)** Gene, deren Sequenz teilweise durch SVs verändert werden, **C)** Gene, deren regulatorische Elemente durch SVs verändert werden und **D)** Gene, die durch SVs Auswirkungen auf komplexe Erkrankungen haben.

CNVs können auch erst in Kombination mit anderen genetisch Einflüssen (z.B. SNPs) oder Umwelteinflüssen in Rahmen komplexer Erkrankungen eine Auswirkung auf den Phänotyp entwickeln. Darüber hinaus können CNVs indirekt Einfluss auf einen Phänotypen haben, wenn sie

für weitere strukturelle Veränderungen prädisponieren, zum Beispiel als Duplikationen in geringer Kopienzahl (engl. *low copy repeats*, kurz LCR) auftreten und wie in den vorigen Abschnitten beschrieben, zur Entstehung weiteren strukturellen Veränderungen beitragen (Feuk et al., 2006).

Die Tatsache, dass nur ein Bruchteil der Heritabilität der häufigen komplexen Krankheiten durch SNPs erklärt werden kann, bekräftigt die Vermutung, dass andere Formen genetischer Variationen, wie CNVs, eine entscheidende Rolle in der Ätiologie dieser Erkrankungen spielen (Visscher et al., 2012; Manolio et al., 2009; Maher, 2008). In Übereinstimmung mit dieser Annahme wurden Assoziationen von CNVs mit verschiedenen häufigen Erkrankungen wie Morbus Crohn (McCarroll, Huett, et al., 2008), rheumatoide Arthritis und Asthma (The Wellcome Trust Case Control Consortium, 2010), Psoriasis (de Cid et al., 2009), geistigen Behinderungen (Coe et al., 2012), Adipositas (Bochukova et al., 2010), Myokardinfarkt (Kathiresan et al., 2009), Schizophrenie (Stefansson et al., 2008) und Autismus (Sebat et al., 2007) beschrieben. Eine Zusammenfassung dieser und weiterer bekannter Assoziationen findet sich in Almal und Padh (2012).

1.3 Bestimmung und Analyse von CNVs

Die Methode zur Analyse struktureller Variationen ist abhängig von der Größe der CNVs. Mikroskopische Variationen bis zu einer Größe von ca. 5 Mega-Basenpaaren (Mb) können noch direkt beobachtet und durch zytogenetische Methoden wie z.B. der Fluoreszenz-in-situ-Hybridisierung (FISH) analysiert werden (Feuk et al., 2006). Durch die in den frühen 1990er Jahren entwickelte Methode der *comparative genomic hybridization* (CGH) war es erstmals möglich submikroskopische strukturelle Variationen zu detektieren (Kallioniemi et al., 1992). Die Mitte der 1990er entwickelte *array CGH* (aCGH) stellt eine Weiterentwicklung der CGH dar und nutzt als Hochdurchsatzverfahren die DNA-Chip Technologie. Die Methode der aCGH wird vorwiegend eingesetzt, um bereits bekannte CNVs zu detektieren, z.B. in der klinischen Diagnostik. Im Jahre 1999 wurde von Pollack erstmals vorgeschlagen, bereits bestehenden Datensätze basierend auf SNP-Chips für eine CNV-Genotypisierung zu nutzen (Pollack et al., 1999). Die Detektion von submikroskopischen Variationen basierte in der Vergangenheit ausschließlich auf der Analyse von Signalintensitäten der aCGH oder SNP-Chip Technologie (Carter, 2007). Die Entwicklung moderner Methoden der DNA-Sequenzierung (Next-Generation Sequencing, kurz NGS) macht es mittlerweile teilweise möglich, auch diese sehr kleinen Varianten direkt zu analysieren (Mills et al., 2011). Die dafür notwendigen Algorithmen werden derzeit noch weiterentwickelt und beschränken sich gegenwärtig hauptsächlich auf eine Detektion von Deletionen. Eine Zusammenfassung dieser Methoden findet sich in Zhao et al., (2013), ein Vergleich von verfügbarer Software in Duan et al.,

(2013). Trotz dieser Fortschritte bleiben Mikro-Array-Methoden auch weiterhin ein wichtiger Ansatz für die CNV-Analyse, nicht zuletzt wegen der häufig bereits verfügbaren Daten aus früheren GWAS. Die Grundlagen dieser Methoden werden im Folgenden beschrieben.

1.3.1 Mikro-Array-Methoden

Die DNA-Chip-Technologie verwendet spezielle Methoden der DNA-Hybridisierung und DNA-Markierung zur Hochdurchsatzanalyse von Proben. Dabei wird unter Verwendung von kurzen DNA-Fragmenten mit bekannten Sequenzen (DNA-Sonden) die Menge der teilweise oder vollständig komplementären genomischen DNA-Abschnitte in einer Probe untersucht. Dazu werden die Sonden vor der Reaktion mit Fluoreszenzfarbstoffen markiert und für die spätere Analyse auf einem festen Träger fixiert. Ein Mikro-Array ist ein solcher Träger, genauer eine kleine Glasplatte, auf der viele hunderttausend DNA-Sonden angebracht sind. Die Mikro-Arrays werden in DNA-Chips verarbeitet, auf denen zunächst der Reaktionsablauf stattfindet, wobei die Sonden spezifisch an bestimmte Sequenzen in der Probe binden. Danach werden die DNA-Chips in speziell angefertigten Scannern ausgelesen, und die Menge der Verbindungen aus Sonden mit DNA-Sequenzen der Probe über die Intensität des Fluoreszenzsignals gemessen. Die DNA-Chip Technologie wird hauptsächlich bei der Genexpressionsanalyse, der SNP-Analyse und der Analyse von strukturellen Variationen eingesetzt.

1.3.2 CNV-Analyse

Bei der Analyse von CNVs wird der Verlust oder Zugewinn von genetischem Material an vorgegebenen Loci detektiert, sodass die Resultate in die Copy Number (CN)-Klassen „Deletion“ „Normalzustand“ und „Duplikation“ eingeteilt werden können. Moderne Verfahren geben sogar die Anzahl der betroffenen Allele an, also ob es sich um eine einfache oder mehrfache Deletion bzw. Duplikation handelt (CN-Genotyp). Die Möglichkeit CN-Klassen oder CN-Genotypen vorherzusagen (Auflösung der Vorhersage), hängt von der Methode der CNV-Analyse und dem Algorithmus der zur Vorhersage verwendeten Software ab. Die noch immer am häufigsten eingesetzten Methoden zur Analyse von CNVs sind die Mikro-Array-Methoden der aCGH und SNP-Chips. Bei der aCGH dient der binäre Logarithmus des Quotienten aus der Signalintensität der Probe und der einer Referenzprobe (engl. *log2 raw data ratio*, kurz LRR) als Kenngröße zur Ermittlung der Kopienzahl (Alkan et al., 2011). Bei der Verwendung von SNP-Chips wird die gleiche Kennzahl verwendet, wobei für den Quotienten ebenfalls eine Referenzprobe oder der Erwartungswert über alle Proben der Analyse verwendet werden kann. Da auf SNP-Chips beide Allele (A und B) von diallelischen SNPs untersucht werden, kann zusätzlich die relative Häufigkeit des B-Allels (engl. *B-allele frequency*, kurz BAF) als weitere Kenngröße herangezogen werden, was die

Auflösung der Vorhersagen gegenüber aCGH-Methoden deutlich verbessert. Da es sich um indirekte Methoden der Analyse handelt, bei der über die Auswertung der Kenngrößen Rückschlüsse auf die CNV-Genotypen gemacht werden, benötigen beide Ansätze Algorithmen zur Auswertung der Rohdaten. Die ersten Algorithmen, die für diesen Zweck entwickelt wurden, basierten auf der Segmentierung der DNA-Sequenz der Proben anhand der ermittelten Kenngrößen. Im Zuge immer komplexerer Ansätze wurden später auch Hidden-Markov Modelle verwendet, um den komplexen Zusammenhang zwischen der beobachtbaren Signalintensität und dem zugrundeliegenden CN-Genotyp besser zu modellieren. Beide Ansätze werden in Abschnitt 2.1 im Detail beschrieben.

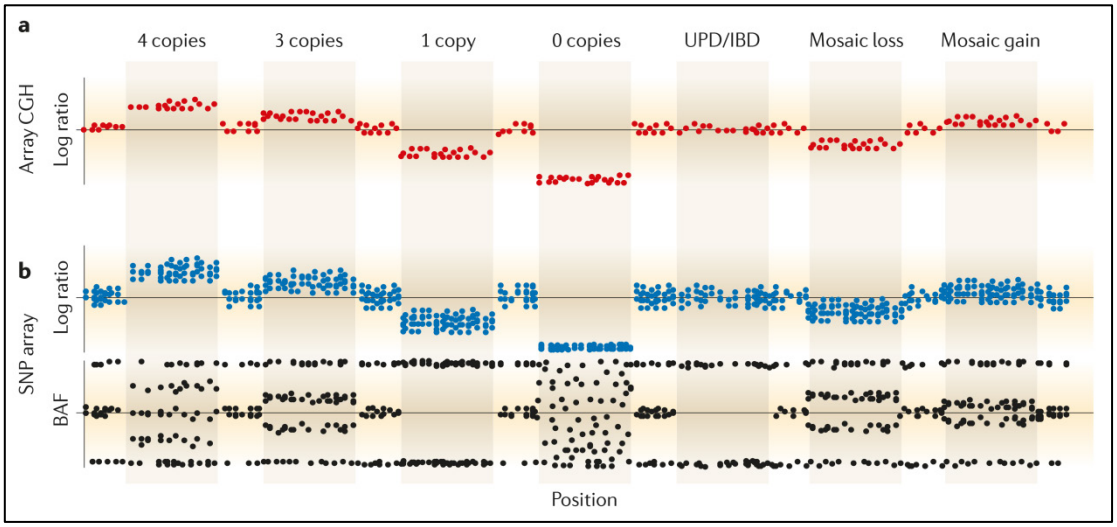


Abbildung 1-7: Signalintensitäten von aCGH- und SNP-Chips nach Alkan et al. (2011)

A) aCGH-Methoden verwenden ausschließlich die Signalintensität (rot) in der Analyse von CNVs. **B)** SNP-Chip-Methoden verwenden zusätzlich zur Signalintensität (blau) den Anteil des gesamten Signals der durch ein Allel erklärt werden kann (schwarz) in der Analyse, was den im Vergleich zu aCGH-Methoden hohen Einfluss von Störgrößen ausgleicht und eine detaillierte Analyse der CN-Genotypen zulässt.

Es wurden in mehreren Studien bereits unterschiedliche Ansätze, einen sogenannten Benchmark, also eine vergleichende Analyse der Leistungsfähigkeit und Zuverlässigkeit der Softwares zur CNV-Vorhersage, durchzuführen, verfolgt. Unter Verwendung von Daten basierend auf dem frühen *Affymetrix 100K* SNP-Chip konnten Baross et al. (2007) erhebliche Falsch-Positiv-Raten bei den Softwares CNAG (Copy Number Analyzer for GeneChip (Nannya et al., 2005)), dChip (DNA-Chip Analyzer (Zhao et al., 2004)) und GLAD (Gain and Loss of DNA (Hupé et al., 2004)) feststellen. Die Autoren berichten zusätzlich über eine große Variabilität der Softwares in Bezug auf die Anzahl der vorhergesagten CNVs. Winchester et al. (2009) bewerteten die Präzision der CNV-Vorhersagen von fünf weiteren Softwares unter Verwendung des neueren Affymetrix® Genome-Wide Human SNP Array 6.0 und des Illumina® 1M-Duo BeadChip Chips. Die Studie umfasste einen Vergleich der

auf SNP-Chip-Daten basierten Vorhersagen mit den Resultaten aus zuvor veröffentlichten Studien basierend auf Sequenzierung (Kidd et al., 2008; Korbel, Urban, Affourtit, et al., 2007; Redon et al., 2006), allerdings lediglich für eine HapMap-Probe. Trotzdem konnten Winchester und Kollegen zeigen, dass eine große Anzahl der vorhergesagten CNVs nicht durch die vorangegangen Studien bestätigt werden konnten (bis zu 80 %, in Abhängigkeit von der verwendeten Software) und dass es große Unterschiede sowohl zwischen den Softwares als auch zwischen den zur Bestätigung verwendeten Studien gibt. In einem ähnlichen Ansatz wendeten Zhang et al. (2011) Birdsuite (Korn et al., 2008), Partek (Partek Inc, St. Loius, MO), HelixTree (Golden Helix, Inc) und PennCNV (Wang et al., 2007) auf drei verschiedene Datensätze an und beobachteten eine positive Korrelation zwischen der Anzahl der Marker in einer CNV-Vorhersage und dem Anteil der Vorhersagen, die durch bereits publizierte CNVs die als validiert gelten bestätigt wurde. Bemerkenswert ist, dass dieser Anteil in einer negativen Korrelation zur relativen Häufigkeit des CNVs stand. Darüber hinaus wurde nur eine geringe Übereinstimmung der CNV-Vorhersagen für acht bereits durch Kidd et al. (2008) und Conrad et al. (2010) analysierten Proben mit den bereits publizierten Ergebnissen gefunden. Eine jüngst durch Eckel-Passow et al. (2011) durchgeführte Studie berichtete über eine erhebliche Variabilität der paarweisen Konkordanz der Vorhersage durch PennCNV (Wang et al., 2007), Affymetrix Power Tools (APT (Affymetrix Inc., 2011)), Aroma.Affymetrix (Bengtsson et al., 2008) und CRLMM (Corrected Robust Linear Model with Maximum Likelihood Distance (Scharpf et al., 2011)). Eine detaillierte Betrachtung von PennCNV und CRLMM zeigte einen Median der Konkordanz von 52 % bei Deletionen und 48 % bei Duplikationen. Von beiden Softwares wurden mehr Deletionen als Duplikationen vorhergesagt und die empirische Falsch-Positiv-Rate lag bei 26 % für CRLMM und 24 % für PennCNV. In einer umfangreichen Studie analysierten Pinto et al. (2011) sechs Proben mit 11 unterschiedlichen Mikro-Array-Methoden und nutzten ebenso viele Softwares zur CNV-Vorhersage, einschließlich PennCNV und QuantiSNP. Die durch die unterschiedlichen Mikro-Array-Plattformen generierten Daten wurden mit jeweils einer bis fünf der Softwares analysiert. Alle Experimente erfolgten in dreifacher Ausführung, und die Autoren konnten eine Reproduzierbarkeit der Ergebnisse in < 70 % der Fälle und eine Konkordanz zwischen den Softwares von < 50 % beobachten.

1.4 Ziele der Studie

Die genomweite Assoziationsanalyse von CNVs ist eine umfangreiche Aufgabe, für die es derzeit keinen allgemein anerkannten Ansatz oder standardisierte Verfahren gibt. Der Umfang der genetischen Variation, die auf CNVs zurückzuführen sind, ist sehr groß (Conrad et al., 2010) und

auch wenn in den letzten Jahren große Fortschritte in der Erforschung der Zusammenhänge mit Phänotypen gemacht worden sind, bleibt das ganze Ausmaß des Einflusses auf komplexe Krankheiten unbekannt (Almal & Padh, 2012). Sarkoidose ist eine solche komplexe Erkrankung, deren Ursache ebenfalls noch größtenteils unbekannt ist. Hier spielt genetische Suszeptibilität eine Rolle, und schon mehrere Risiko-Loci sind gefunden worden. Eine Analyse von CNVs in Zusammenhang mit dieser Krankheit soll dazu beitragen, neue genetische Risikofaktoren zu identifizieren und die Ätiologie der Krankheit besser zu verstehen. In dieser Studie wurde dieser Ansatz in zwei Schritten verfolgt:

1) Ein Benchmark von vorhandenen CNV-Analysesoftwares einschließlich der Entwicklung eines Verfahrens zur systematischen genomweiten *in-silico* CNV-Analyse (Pipeline),
2) eine Assoziationsanalyse von CNVs in Zusammenhang mit Sarkoidose.

Das Benchmark der CNV-Software diente der Auffindung der am besten geeigneten Software, auf der die CNV-Analyse-Pipeline aufgebaut wurde. Ziel der Entwicklung dieser Pipeline ist eine Assoziationsanalyse basierend auf CNV-Vorhersagen, bei der ganze Regionen auf Assoziation getestet werden. Durch die *in-silico* Analyse sollen Kandidaten-Regionen identifiziert werden, in welchen mit Sarkoidose assoziierten CNVs liegen können. Diese Analyse soll im Rahmen einer Fall-Kontroll-Studie basierend auf den SNP-Chip Daten von 1654 deutschen Proben (564 Fälle und 1090 Kontrollen) unter Verwendung des *Affymetrix Human SNP Array 6.0* durchgeführt werden. Zur Validierung der vielversprechendsten Kandidaten-Regionen sollen in einer unabhängigen deutschen Stichprobe CNVs mittels TaqMan® genotypisiert werden.

2 Material & Methoden

2.1 Algorithmen zur Vorhersage von CNVs

Zur Analyse von CNVs basierend auf SNP-Chip Daten wurden bisher diverse Softwares entwickelt, die meist eigene Algorithmen zur Vorhersage von CNVs verwenden. Die in dieser Studie verwendeten Softwares nutzen alle die Rohdaten von Affymetrix® SNP-Chips für ihre Vorhersage von CNVs. SNP-Chips messen die Menge an genetischem Material in einer Probe an verschiedenen Loci (Abschnitt 1.3.1). Die SNP-Chips der Firma Affymetrix® verwenden dazu mehrere Sonden (3-4) je Allel für jeden SNP der auf dem Mikro-Array analysiert wird (McCarroll, Kuruvilla, et al., 2008). Alle Sonden für ein Allel werden als Kanal (A und B) bezeichnet. Die gemessenen Signalintensitäten beider Kanäle R_A und R_B korrelieren mit der Menge an hybridisiertem genetischem Material und dienen zusammen als Marker zur Detektion von genetischen Variationen. Der SNP-Genotyp lässt sich über das Verhältnis der beiden Signalintensitäten ermitteln. Die Signalintensitäten lassen des Weiteren Rückschlüsse auf die gesamte Menge an genetischem Material zu, das in der Probe an dieser Stelle vorhanden ist. Der SNP-Chip *Affymetrix Human SNP Array 6.0* enthält zusätzlich CNV-Sonden. Diese Sonden decken Regionen ab, welche möglicherweise CNVs enthalten und messen ausschließlich die gesamte Menge an genetischem Material. Alle im Folgenden dargestellten Algorithmen nutzen hauptsächlich oder ausschließlich das Verhältnis der gemessenen Signalintensität R_{obs} zu einer erwarteten Intensität R_{exp}, um auf den CN-Status an einem bestimmten Lokus zu schließen, wobei $R_{obs} = R_A + R_B$ gilt. Die erwartete Intensität R_{exp} unterscheidet sich je nach Ansatz des Algorithmus. Der binäre Logarithmus des Quotienten aus beobachteter und erwarteter Signalintensität (engl. *log2 raw data ratio*, kurz LRR) ist definiert als $LRR = R_{obs} + R_{exp}$ und wird von allen Algorithmen als Eingangsgröße verwendet. Manche Algorithmen verwenden zusätzlich die relative Häufigkeit des B-Allels (engl. *B-allele frequency*, kurz BAF), welche den Anteil der gesamten Signalstärke darstellt, der durch die Kopien eines Allels erklärt wird. Die BAF nimmt Werte zwischen 0 und 1 an, wobei $BAF = 0$ bedeutet, dass in der Probe keine Kopien des B-Allels vorliegen (Genotyp AA); $BAF = 0{,}5$ bedeutet, dass beide Allele in gleicher Kopienzahl vorliegen (Genotyp AB), und $BAF = 1$ bedeutet, dass nur Kopien des B-Allels vorliegen (Genotyp BB). CNVs haben häufig spezielle Genotypen, z.B. ABB, wodurch die BAF Werte zwischen 0 und 0,5, sowie zwischen 0,5 und 1 annimmt. Unter Laborbedingungen können verschiedene Störgrößen Einfluss auf die gemessene Signalintensität der einzelnen Allele haben, was ein direktes Messen der wahren BAF unmöglich macht. Zur Bestimmung der BAF an dem Lokus j

wird zunächst das Verhältnis θ der allelischen Signalintensitäten für jede Person i berechnet. Die θ-Werte der einzelnen Personen können dann mit den Medianen der θ-Werte der kanonischen Genotypen AA, AB und BB in der Stichprobe, θ^{AA}, θ^{AB}und θ^{BB} verglichen werden. Diese Werte entsprechen aufgrund der bereits erwähnten Störgrößen nur selten den theoretischen Werten $\theta^{AA} = 0$, $\theta^{AB} = 0{,}5$ und $\theta^{BB} = 1$. Die aus den θ-Werten abgeleitete BAF ist dann definiert als:

$$BAF = \begin{cases} 0, \text{ wenn } \theta < \theta^{AA} \\ 0{,}5(\theta - \theta^{AA})/(\theta^{AB} - \theta^{AA}), \text{ wenn } \theta^{AA} \leq \theta < \theta^{AB} \\ 0{,}5 + 0{,}5(\theta - \theta^{AB})/(\theta^{BB} - \theta^{AB}), \text{ wenn } \theta^{AB} \leq \theta < \theta^{BB} \\ 1, \text{ wenn } \theta > \theta^{BB} \end{cases}$$

Die in dieser Studie verwendeten Algorithmen zur Vorhersage von CNVs können, basierend auf dem mathematischen Modell welches sie verwenden, in zwei Klassen unterteilt werden: Zum einen die Hidden-Markov-Model-Algorithmen (HMM-Algorithmen) und zum anderen die Segmentierungs-Algorithmen. Die Vertreter der Klasse der HMM-Algorithmen ähneln sich in ihrem Ansatz stark, da sie alle auf Hidden-Markov Modellen beruhen und unterscheiden sich hauptsächlich in den Parametern des Modells. Die heterogene Klasse der Segmentierungs-Algorithmen umfasst alle Algorithmen, die auf verschiedene Art versuchen die Chromosomen in Segmente zu unterteilen. Beide Klassen von Algorithmen wurden in diversen Softwareprogrammen implementiert. In dieser Studie wurden die dreiauf HMM-Algorithmen beruhenden Softwares Affymetrix Power Tools (APT) (Affymetrix Inc., 2012), PennCNV (Wang et al., 2007) und QuantiSNP (Colella et al., 2007), sowie die drei auf Segmentierungs-Algorithmen beruhenden Softwares R-gada (Pique-Regi et al., 2010), GLAD (Hupé et al., 2004) und VEGA (Morganella et al., 2010) verwendet.

2.1.1 HMM-Algorithmen

Hidden-Markov Modelle sind komplexe stochastische Modelle, die ein System anhand einer Markow-Kette mit unbeobachteten Zuständen beschreiben. Die HMM-Algorithmen nutzen dieses Modell, um anhand der beobachteten LRR-Werte die nicht beobachtbaren CN-Genotypen an gegebenen Positionen vorherzusagen (Abbildung 2-1). In den Algorithmen werden die nicht beobachtbaren CN-Genotypen als Markow-Kette und die LRR-Werte (und ggf. die BAF-Werte) als emittierte Beobachtungen modelliert. Die Übergangs- und Emissionswahrscheinlichkeiten werden durch einen erwartungsmaximierenden Algorithmus (EM-Algorithmus), wie dem Baum-Welch-Algorithmus, aus den Daten geschätzt. Anschließend wird der Viterbi-Algorithmus verwendet um die wahrscheinlichste Sequenz der verborgenen CN-Genotypen entlang der Marker des SNP-Chips zu ermitteln. Trotz der Ähnlichkeiten im Ansatz verwenden die HMM-Algorithmen unterschiedliche

Modelle, deren Parameter teilweise aus den Daten geschätzt werden, durch die Software definiert sind oder durch den Anwender angegeben werden müssen.

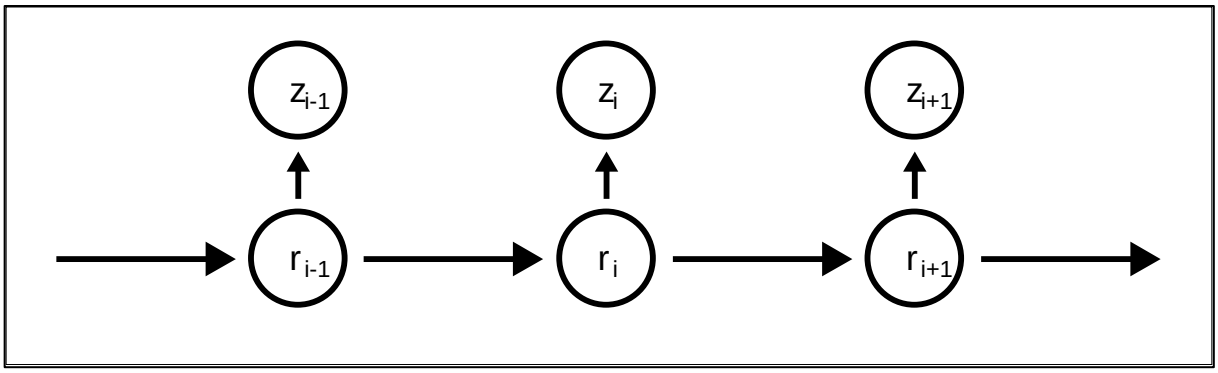


Abbildung 2-1: Schematische Darstellung eines Hidden-Markov Modells

i: genomische Position, r: verdeckter Zustand des Systems an Position i (CN-Genotyp); z: mögliche Beobachtung (LRR & BAF) an Position i; vertikale Pfeile: Übergangswahrscheinlichkeiten, horizontale Pfeile: Emissionswahrscheinlichkeiten

2.1.2 Segmentierungs-Algorithmen

In der heterogenen Gruppe der Segmentierungs-Algorithmen existieren viele verschiedene Ansätze die Chromosomen anhand der gemessenen Intensitäten in Segmente zu unterteilen. Den ersten Algorithmus dieser Klasse, den *Circular Segmentation* Algorithmus, entwickelten Olshen et al., (2004). Dabei wird das LRR als Kenngröße für die Signalintensität verwendet um entlang des Genoms nach Unstetigkeitsstellen zu suchen (Abbildung 2-2).

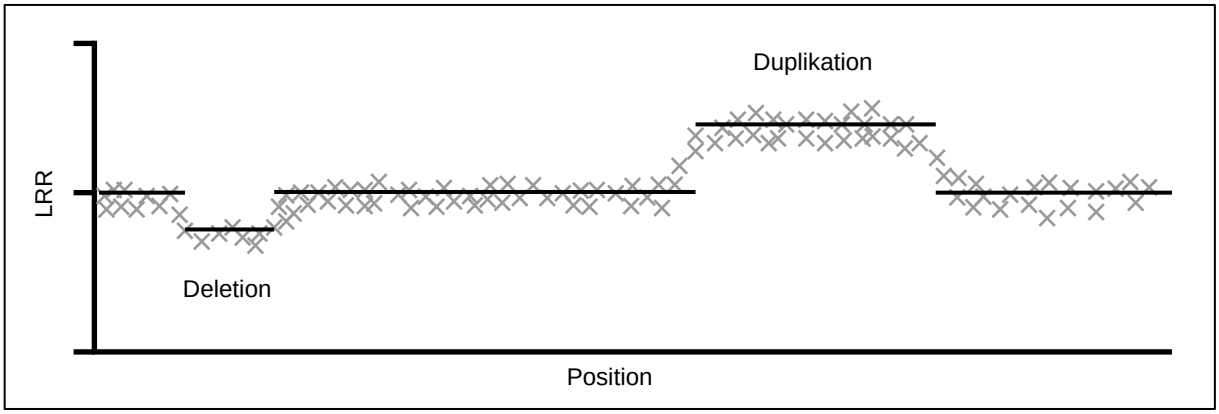


Abbildung 2-2: Segmentierung einer DNA-Sequenz anhand von LRR-Werten

Den resultierenden Segmenten wird in einem zweiten Schritt eine CN-Klasse zugewiesen, wobei die LRR-Werte eines Segments mit einem Referenzwert oder den restlichen Werten des Chromosoms verglichen wird. Die Bestimmung eines CN-Genotyps ist mit diesem Ansatz nicht möglich. Die in dieser Studie verwendeten Softwares nutzen bei der Suche nach Unstetigkeitsstellen Sparse Byesian

Learning, einen Adaptive Weights Smoothing Algorithmus (Polzehl & Spokoiny, 2000) und einen Algorithmus basierend auf dem Mumford und Shah Modell (Mumford & Shah, 1989).

2.2 Software-Benchmark

In Übereinstimmung mit früheren Studien (Conrad et al., 2010; Redon et al., 2006) wurde in dem Benchmark die obige Definition von CNVs als DNA-Segmente >1 kb, die im Vergleich zu einer Referenz entweder in erhöhter oder verminderter Kopienzahl vorliegen, verwendet. Die Referenz für die CNV-Vorhersage war abhängig von der verwendeten Software, genetische Loci beziehen sich auf die Referenz der UCSC Genome Assembly Version hg19.

2.2.1 Datensatz

In dem Benchmark wurden die Rohdaten (Signalintensitäten) von 60 Trios (180 Individuen) aus dem *Affymetrix® Sample Data Set* verwendet. Diese Individuen sind Teil des *Release 21a* (Veröffentlicht am 01.11.2007) der Phase 1 und 2 des *HapMap*-Projektes (The International HapMap Consortium, 2005). Die Hälfte der Trios ist von afrikanischer Abstammung (Yoruba in Ibadan, Nigeria, YRI), während die andere Hälfte von europäischer Abstammung (in Utah Ansässige mit nord- und westeuropäischer Abstammung, CEU) ist. Alle Proben wurden auf dem *Affymetrix Genome-Wide Human SNP Array 6.0* genotypisiert. Dieser SNP-Chip enthält Sonden für 906.600 SNPs und zusätzlich 945.826 CNV-Sonden (Affymetrix Inc., 2011). Den Markern des Chips wurden mit Hilfe von NetAffx™ Annotationsdateien (Release 32, UCSC hg19) Loci im menschlichen Genom zugewiesen. Die durchschnittliche Rate der SNP-Genotypisierung (engl. *call rate*) im vollständigen Datensatzes des Release 21a (270 Proben) betrug 99,83% (technische Dokumentation) und die Konkordanz zu HapMap Genotypen (Release 21a) betrug 99,84%. Da es keine allgemeinen CNV-spezifischen Kriterien zur Qualitätskontrolle gibt, wurden alle Proben verwendet und softwarespezifische Qualitätskriterien angewandt soweit vorhanden. Es wurden daher alle 60 unabhängige Proben der unverwandten Nachkommen in der CNV-Vorhersage und die 120 Proben der zugehörigen Eltern in der anschließenden Validierung verwendet. Die Analyse wurde auf Autosomen beschränkt.

2.2.2 Software

Es wurden sechs häufig verwendete Softwares für die Detektion von CNVs in Affymetrix® SNP-Chip-Daten analysiert, namentlich APT (Affymetrix Inc., 2012), QuantiSNP (Colella et al., 2007) und PennCNV (Wang et al., 2007), welche einen HMM Algorithmus verwenden, sowie die auf Segmentierung basierenden Softwares R-gada (Pique-Regi et al., 2010), GLAD (Hupé et al., 2004)

und VEGA (Morganella et al., 2010). Alle Softwares wurde mit Standardeinstellungen verwendet sofern nicht anders angegeben.

2.2.2.1 APT

Die Affymetrix Power Tools (APT) (Affymetrix Inc., 2012) sind eine Sammlung von Kommandozeilen-Programmen zur Analyse von Daten welche durch Affymetrix® Arrays erzeugt wurden. Das Programm `apt-copynumber-workflow` aus den APT Version 1.14.2 dient der CNV-Detektion. Der implementierte Algorithmus setzt die erwartete Signalintensität an einem gegebenen Locus *i* R_i^{exp} gleich dem Median aller probenspezifischen Signalintensitäten an diesem Locus oder verwendet eine zuvor berechnete Referenz (Bengtsson et al., 2009), um die markerspezifischen LRR-Werte zu berechnen. Die Sequenz der LRR-Werte wird dann in einem Hidden Markov Model verwendet, um die nicht direkt beobachtbaren CN-Genotypen zu ermitteln. Das Programm `apt-copynumber-workflow` wurde mit Standarteinstellungen im single-sample Modus und der Option `--text-output 'true'` verwendet. Als Referenz wurde die von Affymetrix® zur Verfügung gestellte *Copy Number Analysis HapMap Reference* Datei (Release 31) eingesetzt, und zur Annotation wurde die *NetAffx Annotation File* (Release 31) verwendet (offen zugänglich auf der Affymetrix® Website http://www.affymetrix.com).

2.2.2.2 PennCNV

Bei der Software PennCNV handelt es sich um eine Sammlung von Perl-Skripten zur Detektion von CNVs in SNP-Chip-Daten. Zur Analyse von Affymetrix® Daten wurde das *PennCNV-Affy* Protokoll (http://www.openbioinformatics.org/penncnv/) angewendet. Die LRR und BAF werden nach dem Protokoll aus kanonischen Genotyp-Clustern (Peiffer et al., 2006) mittels linearer Interpolation abgeleitet. Die Genotyp-Cluster werden aus den durch die APT Software (siehe oben) ermittelten Genotypen generiert. Die Sequenz der LRR- und BAF-Werte wird in einem HMM-Algorithmus verwendet, um die nicht direkt beobachtbaren CN-Genotypen abzuleiten. In dieser Studie wurde die Version 2011Jun16 der PennCNV Software verwendet. Zunächst wurde entsprechend dem Protokoll mit Hilfe der Programme `apt-probeset-genotype` und `apt-probeset-summarize` der APT Software (Version 1.14.2) die Genotypen ermittelt sowie die allelspezifischen Signale extrahiert. In einem zweiten Schritt wurden die kanonischen Genotyp-Cluster (Peiffer et al., 2006) mit Hilfe des PennCNV-Programmes `generate_affy_geno_cluster.pl` erzeugt. Diese Cluster werden in dem PennCNV-Programm `normalize_affy_geno_cluster.pl` verwendet, um durch lineare Interpolation die LRR und BAF-Werte zu berechnen. Die Sequenz der LRR und BAF-Werte wurde mit dem Programm `detect_cnv.pl` unter Verwendung der Standardparameter analysiert.

2.2.2.3 QuantiSNP

Die Software QuantiSNP nutzt separat erstellte LRR- und BAF-Werte (z.B. nach dem PennCNV-Affy Protokoll), welche in einem eigenen HMM Algorithmus verwendet werden, um auf die verborgenen CN-Zustände entlang des Genoms zu schließen. QuantiSNP (Version 2) wurde gemäß den Anweisungen der QuantiSNP Projekt-Website (https://sites.google.com/site/quantisnp/) angewendet. Es wurden LRR- und BAF-Werte verwendet, welche zuvor mit der Software PennCNV erstellt wurden.

2.2.2.4 R-gada

Die Software R-gada (Pique-Regi et al., 2010) implementiert einen Algorithmus zur Segmentierung von Chromosomen in einem R-Paket. Der Algorithmus sucht durch Sparse Byesian Learning nach Unstetigkeitsstellen entlang zuvor berechneter LRR-Werte. Nach der Segmentierung wird zunächst der LRR-Referenzwert ermittelt, der einer normalen Kopienzahl entsprich, indem der Median der LRR-Werte auf allen Autosomen berechnet wird. Liegt der Mittelwert aller LRR-Werte eines Segments signifikant unter (Deletion) oder über (Duplikation) dem Referenzwert, wird dem Segment die entsprechende CN-Klasse zugeordnet. Die Software R-gada (Version 0.8-5) wurde als R-Paket in R 2.15 mit durch dem APT Programm `apt-copynumber-workflow` erstellten LRR-Werten (siehe oben) verwendet.

2.2.2.5 GLAD

Bei der Software GLAD (Hupé et al., 2004) handelt es sich um ein R-Paket, das ursprünglich entwickelt wurde, um aCGH Daten zu analysieren. Da die Software aber Signalintensitäten für die Segmentierung verwendet, kann der Algorithmus auch auf SNP-Chip-Daten angewendet werden. Die Software GLAD verwendet separat berechnete LRR-Werte in dem Adaptive-Weights-Smoothing Algorithmus, um Unstetigkeitsstellen entlang des Genoms zu finden und jedes Chromosom anhand dieser in Segmente zu unterteilen. Nach Abschluss der Segmentierung repräsentiert das Segment, dessen Median der LRR-Werte sich am nächsten an Null, befindet den normalen CN-Zustand. Allen anderen Segmenten werden dann abhängig von dem Unterschied zu diesem Wert CN-Klassen zugewiesen. Die Software GLAD (Version 2.20.0) wurde als R-Paket in R 2.15 mit durch APT erstellten LRR-Werten (siehe oben) verwendet.

2.2.2.6 VEGA

Das R-Paket VEGA verwendet separat berechnete LRR-Werte in einem Algorithmus zur Segmentierung basierend auf dem *Mumford-Shah-Variational* Modell (Morganella et al., 2010).

Nach der Segmentierung werden den Segmenten auf Grundlage des mittleren LRR-Wertes μ eines Segmentes CN-Klassen zugewiesen. Ein Segment wird als Deletion deklariert, wenn $\mu < -0{,}2$ gilt, und als Duplikation für $\mu > 0{,}2$. Gilt $-0{,}2 \leq \mu \leq 0{,}2$, wird angenommen, dass das entsprechende Segment in zu erwartender Kopienzahl vorliegt. Die Software VEGA (Version 1.7.0) wurde als R-Paket in R 2.15 mit durch APT erstellten LRR-Werten (siehe oben) verwendet.

2.2.3 Standardisierung der Software-Ergebnisse

Das Ausgabeformat der verschiedenen Softwares unterscheidet sich stark und ist nicht immer direkt vergleichbar. Während die Vorhersagen der HMM-Algorithmen eine höhere Auflösung besitzen und bis zu sechs verschiedene CN-Genotypen bestimmen, geben die Segmentierungs-Algorithmen lediglich die CN-Klasse an. Aus diesem Grund mussten für einen Vergleich die CN-Genotypen der HMM-Algorithmen in CN-Klassen zusammengefasst werden. Des Weiteren geben die Softwares PennCNV, QuantiSNP, R-gada und VEGA ihre Resultate jeweils als Listen von Segmenten aus. APT und GLAD geben jeden analysierten Marker und den zugewiesenen CN-Genotyp bzw. die CN-Klasse aus. Um einen Vergleich unter den Softwares zu ermöglichen, wurden die Resultate von APT und GLAD in Listen von Segmenten umgewandelt, indem alle aufeinander folgenden Marker der gleichen CN-Klasse zu einem Segment zusammengefasst wurden. Da manche Algorithmen (z.B. PennCNV) standartmäßig nicht die Analyse der Gonosomen unterstützen, wurden ausschließlich Autosomen betrachtet.

2.2.4 Durchführung des Benchmarkings

Es wurden sechs Softwares in Bezug auf die Charakteristiken der vorhergesagten CNVs (z.B. ihre Anzahl, Länge und Typ) sowie auf die Validität der gemachten Vorhersagen untersucht. Zusätzlich wurde die Anzahl und Dichte der Marker innerhalb der in den 60 unverwandten Nachkommen vorhergesagten CNVs verglichen. Zur Validierung der CNVs in den Nachkommen wurde das Auftreten von CNVs bei den Eltern analysiert. Eine CNV in einem Nachkommen wurde als validiert betrachtet, wenn ein oder mehrere Segmente derselben CN-Klasse, welche die CNV zu mehr als 90 % abdecken, in mindestens einem Elternteil durch die gleiche Software detektiert wurden (Abbildung 2-3). Es wurden auch weitere Grenzwerte der Abdeckung als Kriterium für eine Validierung verwendet. Der mögliche Einfluss des zugrunde liegenden mathematischen Modells auf die CNV-Vorhersage wurde analysiert, indem für jede Zielgröße der Median über die Ergebnisse der auf HMM basierenden Softwares (APT, PennCNV, QuantiSNP) und der auf Segmentierung basierenden Softwares (GLAD, R-gada, VEGA) verglichen wurde. Abschließend wurde die paarweise Konkordanz der Softwares in Bezug auf die CNV-Detektion untersucht indem der Anteil

der softwarespezifischen CNV-Vorhersagen ermittelt wurde, der auch durch eine andere Software vorhergesagt wurde. Eine durch eine Software vorgesagte CNV galt als durch eine andere Software verifiziert, wenn diese mehr als 90 % des Segments ebenfalls als CNV derselben CN-Klasse deklariert hatte (Abbildung 2-3). Auch hier wurden weitere Grenzwerte der Abdeckung als Kriterium der Verifikation verwendet.

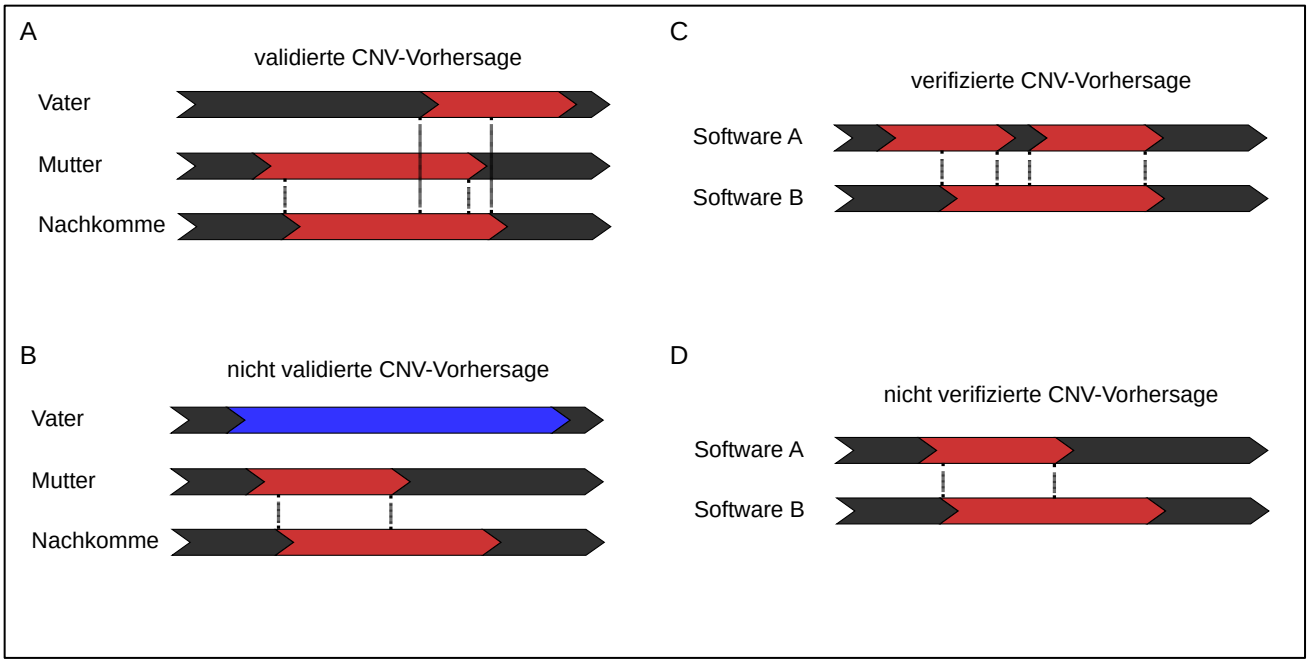


Abbildung 2-3: Validierungskriterien des Software-Benchmarks

Die dargestellten Kriterien sind Beispiele und stellen nicht jedes mögliche Szenario dar. Die Kriterien gelten sowohl für die familienbasierte Validierung als auch die paarweise Verifizierung durch andere Software. **A)** Die Deletion in dem Nachkommen wird zu mehr als 90 % von einer Deletion in der Mutter abgedeckt und gilt als validiert. **B)** Die Deletion (rot) in dem Nachkommen wird vollständig von einer CNV-Vorhersage für den Vater überdeckt, da es sich dabei aber um eine Duplikation (blau) handelt, gilt die Deletion als nicht validiert. **C)** Eine softwarespezifische CNV-Vorhersage (Software B) wird zu mehr als 90% von CNV-Vorhersagen der gleichen CN-Klasse einer anderen Software überdeckt und gilt als verifiziert. **D)** Die softwarespezifische Vorhersage wird zu weniger als 90% durch eine Vorhersage einer anderen Software abgedeckt und gilt als nicht verifiziert.

Alle Analysen wurden separat für die afrikanischen (YRI) und die europäischen (CEU) Trios wiederholt, um mögliche populationsspezifische Unterschiede in Bezug auf die CNV-Detektion zu erkennen. Zusätzlich wurde die Zuordnung der Eltern zu den Trios zehn Mal permutiert um die Wahrscheinlichkeit einer zufälligen Validierung zu erheben. Dieses Verfahren wurde sowohl auf alle Trios (YRI und CEU) zusammen angewendet, sowie beschränkt auf die Trios jeweiligen Populationen.

Alle statistischen Analysen wurden in der Programmiersprache R Version 2.15.2 (R Core Team 2013, http://www.R-project.org) durchgeführt. Unterschiede in den Ergebnissen der Softwares wurden unter Verwendung des Wilcoxon-Vorzeichen-Rang-Tests auf statistische Signifikanz überprüft.

2.3 Sarkoidose-Stichproben

Zur Analyse von CNVs in Zusammenhang mit Sarkoidose wurden zwei unabhängige Datensätze verwendet.

2.3.1 Sarkoidose-GWAS-Datensatz

Für die genomweite *in-silico* Vorhersage von CNVs und die Bestimmung von Kandidaten-Regionen wurde ein Datensatz aus den zuvor durchgeführten Sarkoidose-GWAS von Fischer et al., (2012) und Hofmann et al., (2013) verwendet. Die SNP-Chip Rohdaten, basierend auf dem *Affymetrix Human SNP Array 6.0*, wurden von dem Institut für Klinische Molekularbiologie (IKMB) der Christian-Albrechts-Universität zu Kiel (CAU) zur Verfügung gestellt. An der Rekrutierung der deutschen Sarkoidosepatienten dieses Datensatzes waren die Deutsche Sarkoidose-Vereinigung e.V., die Krankenkassen und spezialisierte Krankenhäuser und Ärzte beteiligt. Die Diagnose erfolge bei allen Patienten durch eine histologische Untersuchung entsprechend internationaler Standards (The American Thoriacic Society (ATS) et al., 1999). Patienten mit unsicherer Diagnose wurden von der Studie ausgeschlossen. Die Kontrollpersonen des GWAS-Datensatzes wurden durch die POPGEN Biobank (Krawczak et al., 2006) rekrutiert, und die entsprechenden Proben wurden für die Studien zur Verfügung gestellt. Die Probenkollektion, die experimentelle Vorgehensweise und die Einhaltung des Datenschutzes wurde von der Ethikkommission des CAU und dem lokalen Datenschutzbeauftragen gemäß den datenschutzrechtlichen Bestimmungen geprüft und beaufsichtigt. Der verwendete Sarkoidose-GWAS-Datensatz umfasste insgesamt 564 Proben von Sarkoidosepatienten (Fälle) und 1090 Proben von gesunden Kontrollpersonen (Kontrollen), die den allgemein anerkannten und etablierten Kriterien zur Qualitätskontrolle für SNP-Genotypisierung entsprachen (Hofmann et al., 2013).

2.3.2 Sarkoidose-Proben zur Validierung

In der *in-vitro* Analyse zur Validierung möglicher Assoziationen der detektierten Kandidaten-Regionen wurden Proben eines unabhängigen Datensatzes verwendet, der bereits in den Sarkoidose-GWAS von Fischer et al., (2012) und Hofmann et al., (2013) zur Validierung der Ergebnisse diente. Die DNA-Proben für die CNV-Analyse durch TaqMan®-Genotypisierung wurden vom IKMB zur

Verfügung gestellt und durch die dortigen technischen Assistenten aufbereitet. Die Patienten und Kontrollpersonen wurden nach den beschriebenen Kriterien der GWAS-Studien (von den oben angegebenen Institutionen) rekrutiert. Für jede Probe wurde mittels Gelelektrophorese eine Qualitätskontrolle durchgeführt, und Proben, bei denen nur wenig oder stark fragmentierte DNA vorhanden war, wurden nicht in der CNV-Analyse verwendet. Der Datensatz zur Validierung umfasste insgesamt DNA-Proben von 552 Fällen und 552 Kontrollen.

2.4 CNV-Analyse Pipeline

Aufbauend auf den Ergebnissen des Software Benchmarks wurde ein Verfahren zur systematischen genomweiten *in-silico* Analyse von CNVs basierend auf SNP-Chip Daten (Pipeline) entwickelt. Im Rahmen einer Fall-Kontroll-Studie diente diese Pipeline der Durchführung einer, auf CNV-Vorhersagen basierenden, Assoziationsanalyse zur Bestimmung von Kandidaten-Regionen (Abbildung 2-4). Die Pipeline ist für die Verwendung von PennCNV konzipiert, ließe sich aber auch mit anderer CNV-Analyse Software verwenden, deren Ausgabe für jede Probe CNV-Vorhersagen enthält.

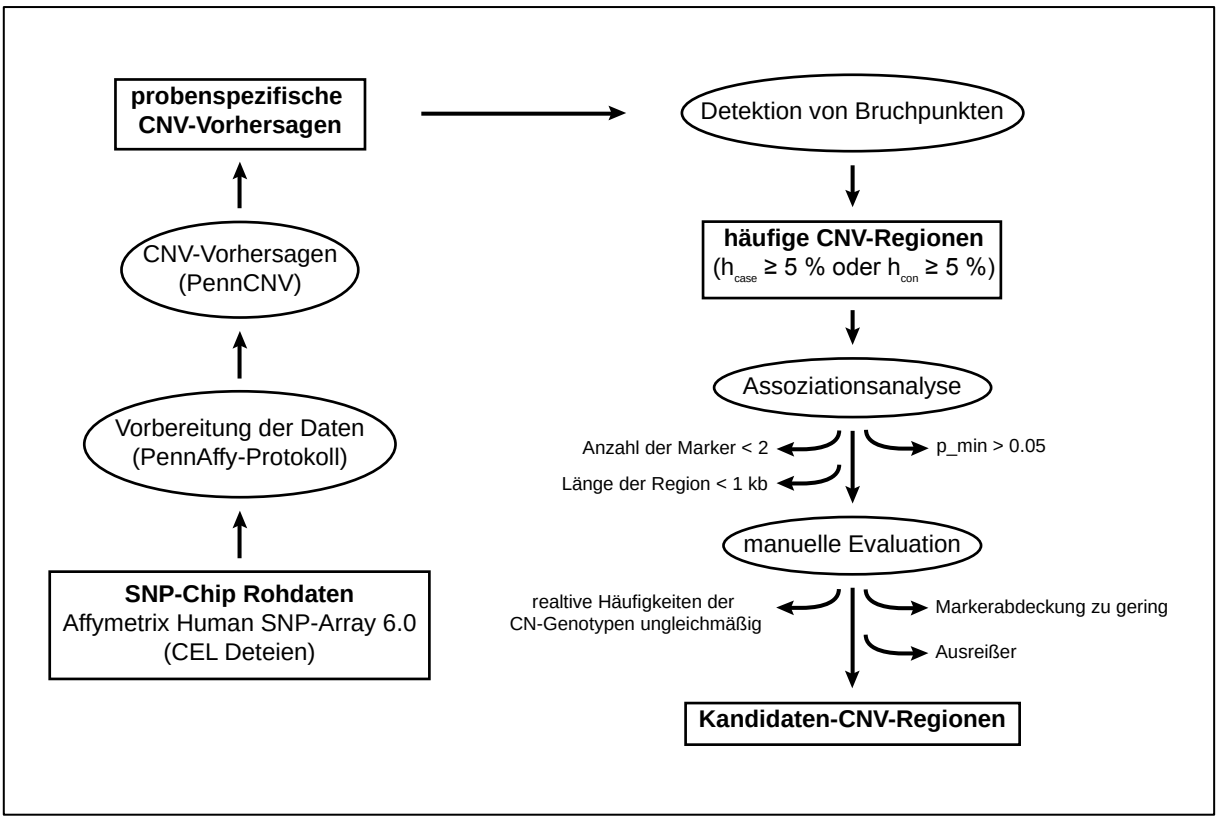


Abbildung 2-4: Aufbau der CNV-Analyse-Pipeline

Schematische Darstellung der Kontrollpunkte (Ovale), Operationen (Rechtecke) und Filterkriterien (freier Text) der zur Analyse verwendeten Skripte.

Die einzelnen Operationen der Pipeline wurden durch kleine Programme (Skripte) für den Kommandozeileninterpreter *Bourne-Again-Shell* (BASH) gesteuert. Zuerst wurden auf diese Weise die Rohdaten (CEL-Dateiformat), welche die gemessenen Signalintensitäten jedes Markers auf dem Chip enthalten, für die Analyse durch PennCNV vorbereitet. Die Vorbereitung der Daten und die CNV-Vorhersage erfolgte, wie in Abschnitt 2.2.2 beschrieben, unter Verwendung der Standardeinstellungen der Software gemäß des *PennCNV-Affy* Protokolls der Entwickler (http://www.openbioinformatics.org/penncnv/). Die Assoziationsanalyse auf Grundlage der vorhergesagten CNVs erfolgte unter Verwendung von Skripten in der Programmiersprache R und wird in Abschnitt 2.4.2 im Detail erklärt. Die Ausgabe der R-Skripte zur Assoziationsanalyse beinhaltet eine Liste von häufigen CNV-Regionen (relative Häufigkeit >5 % in Fällen oder Kontrollen), sowie die p-Werte aller durchgeführten Tests. Die Ausgabe sowie ihre Auswertung wird in Abschnitt 2.4.3 im Detail beschrieben. Alle verwendeten Skripte befinden sich auf einem dieser Arbeit beiliegenden Datenträger.

2.4.1 Detektion von Bruchpunkten

Die Ungenauigkeit der probenspezifischen CNV-Vorhersagen, bei der Bestimmung der Grenzpositionen, führt zu Überschneidungen in den Anfangs- und Endbereich der Vorhersagen unterschiedlicher Proben am gleichen Locus (Abbildung 2-5).

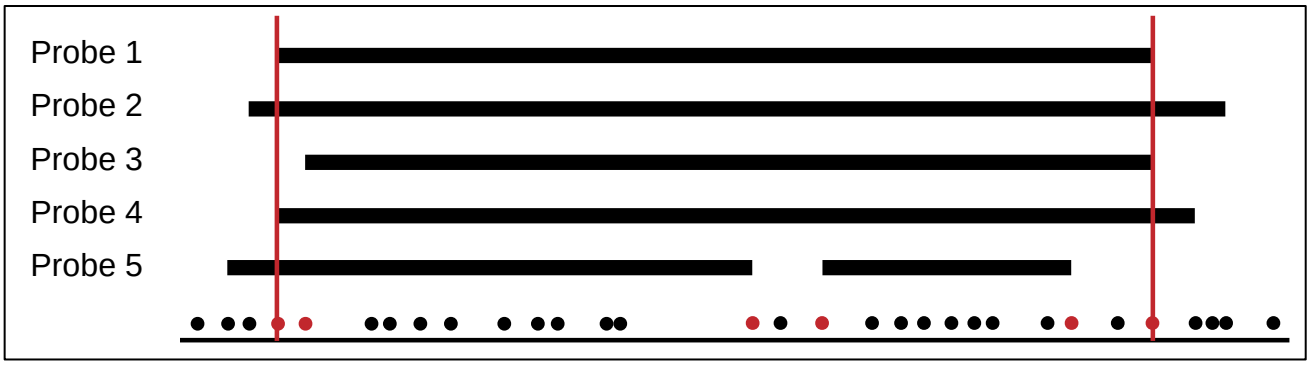


Abbildung 2-5: Schematische Darstellung einer CNV-Region

Eine CNV-Region beginnt, wenn der Anteil der Proben, für die eine CNV vorhergesagt wurde größer, ist als ein vordefinierter Grenzwert und endet wenn dieser Wert wieder unterschritten wird (rote vertikale Linien). Jeder Marker (Punkte), an dem eine CNV-Vorhersage beginnt oder endet, dient als Bruchpunkt (rot) der Analyse der Region.

Um einheitliche Grenzpositionen für alle Proben festzulegen, wurden in Regionen mit einer hohen relativen Häufigkeit an CNV-Vorhersagen Schnittmengen der vorhergesagten CNVs gebildet. Dazu wurde die relative Häufigkeit der CNV-Vorhersagen in den Fällen (h_{case}) und in den Kontrollen (h_{con}) an allen Markern ausgewertet, an denen eine probenspezifische CNV-Vorhersage beginnt

oder endet (Bruchpunkte). Eine CNV-Region wurde dann definiert als ein DNA-Segment, in der für alle Marker $h_{case} > 0{,}05$ oder $h_{con} > 0{,}05$ gilt. An allen relevanten Markern einer Region wurden statistische Tests durchgeführt, um eine Assoziation der CNV-Region mit dem Phänotyp zu prüfen.

2.4.2 Assoziationsanalyse

Bei der gleichzeitigen Betrachtung eines diskreten Merkmals (CN-Klassen oder CN-Genotypen) in zwei Gruppen (Fällen und Kontrollen) wird die Verteilung der absoluten Häufigkeiten, sowie der Ausprägungen des Merkmals in den Gruppen in einer Kontigenztafel verglichen (Tabelle 2-1). Um festzustellen, ob die beiden Gruppen bezüglich des betrachteten Merkmals unterschiedlich verteilt sind, wird ein χ^2-Test durchgeführt. Dabei wird geprüft, ob die empirische Verteilung der Häufigkeiten statistisch signifikant von der Verteilung abweicht, die erwartet wird, wenn beide Gruppen gleich verteilt sind.

Tabelle 2-1: Aufbau einer 2×5 Kontingeztafel

CN-Klasse	**Deletionen**		**Keine CNV**	**Duplikationen**		**Summe**
CN-Genotyp	**0**	**1**	**2**	**3**	**4**	
Fälle	m_{10}	m_{11}	m_{12}	m_{13}	m_{14}	$m_{1\cdot}$
Kontrollen	m_{20}	m_{21}	m_{22}	m_{23}	m_{24}	$m_{2\cdot}$
Summe	$m_{\cdot 0}$	$m_{\cdot 1}$	$m_{\cdot 2}$	$m_{\cdot 3}$	$m_{\cdot 4}$	n

$\boldsymbol{m_{ij}}$**:** absolute Häufigkeit der Ausprägung des Merkmals, wobei i die Gruppe und j den CN-Genotyp angibt.

Die Wahl der zu verwendenden Kontingenztabelle hängt maßgeblich von der zu prüfenden Hypothese ab. Wird ein Unterschied in der Verteilung der einzelnen CN-Genotypen erwartet, so muss eine 2×5-Kontingenztafel untersucht werden, um diesen Unterschied auch erfassen zu können. Der resultierende 2×5-χ^2-Test ist allerdings weniger gut geeignet Unterschiede zu detektieren, die sich nur auf CN-Klassen zurückführen lassen, weshalb für diesen Fall ein 2×3-χ^2-Test die bessere Wahl ist. Steht nur eine CN-Klasse in Verdacht unterschiedlich in Fällen und Kontrollen verteilt zu sein, reicht sogar ein 2×2-χ^2-Test aus, bei dem ausschließliche die Verteilung zweier CN-Klassen (keine CNVs und Deletionen oder Duplikationen) untersucht wird. Da die Pipeline einen hypothesenfreien Ansatz verfolgt, um mögliche Kandidaten-Regionen zu bestimmen, wurde jeder der vier genannten Tests an allen relevanten Marker der CNV-Regionen durchgeführt. Da die Resultate ausschließlich zur Priorisierung verwendet werden, ist eine Korrektur für multiples Testen nicht erforderlich. Für die Priorisierung wurden die CNV-Regionen nach dem jeweiligen Minimum der p-Werte aller 2×5-χ^2-Tests der jeweiligen Region geordnet. Die CNV-Regionen mussten

anschließend manuell gefiltert werden, um die vielversprechendsten Kandidaten für eine anschließende statistische Validierung zu bestimmen.

2.4.3 Kandidaten-Regionen

Die nach den minimalen p-Werten der 2×5-χ^2-Tests priorisierten CNV-Regionen wurden visuell überprüft, da keine allgemein gültigen Kriterien zur Qualitätskontrolle definiert werden konnten. Aufgrund der sich überschneidenden Grenzpositionen der CNV-Vorhersagen und der geringen Markerdichte des SNP-Chips ist es möglich, dass eine durch die Pipeline definierte CNV-Region die Vorhersagen von mehreren, nahe bei einander liegenden, CNVs beinhaltet. Die probenspezifischen CNV-Grenzpositionen müssen also manuell angepasst und die Region ggf. aufgeteilt werden. Die resultierenden Segmente stellen die bestmöglichste Annäherung an die zugrunde liegenden und nicht direkt beobachtbaren CNVs da. Die durch die Pipeline detektierten Regionen sind durch mindestens zwei Marker abgedeckt und haben eine Länge von mindestens 1 kb. Alle Regionen mit einem nicht signifikanten minimalen P-Wert werden automatisch herausgefiltert, da sie als Kandidaten nicht in Frage kommen. Weitere Kriterien zur Qualitätskontrolle müssen manuell geprüft werden. Dazu wird für jede automatisch detektierte Region ein Plot erstellt, der die p-Werte der einzelnen Tests sowie die relativen Häufigkeiten der CN-Genotypen für Fälle und Kontrollen darstellt (Abbildung 2-6).

Die Hauptkriterien bei der manuellen Auswahl der Kandidaten-Regionen war eine gleichmäßige Abdeckung durch Marker über die gesamte Länge des relevanten Segments sowie gleichmäßig hohe relative Häufigkeiten von CNV-Vorhersagen an allen Bruchpunkten innerhalb des Segments. Regionen die nach manueller Neudefinition der Grenzen relevante Segmente besaßen, die kleiner als 1 kb waren oder deren signifikante p-Werte außerhalb relevante Segmente lagen (Ausreißer), wurden von der weiteren Untersuchung ausgeschlossen. Die sehr kleinen p-Werte bei Ausreißer-Regionen sind auf die nicht am selben Marker beginnenden oder endenden CNV-Vorhersagen und die sich dadurch ergebenden, lokalen, großen Unterschiede in der relativen Häufigkeit zurückzuführen (Abbildung 2-6). Dadurch werden zunächst auch Segmente, die sonst keine signifikanten p-Werte aufweisen, als hoch priorisiert eingestuft. Um diese Ausreißer-Regionen zu identifizieren, wird eine gleichmäßig hohe relative Häufigkeiten der CN-Genotypen, an allen Markern des relevanten CNV-Segments, visuell überprüft. Entscheidend ist auch eine gleichmäßig hohe Abdeckung des Segments mit Markern, um sicherzustellen, dass es entlang des Segments keine großen Veränderungen der relativen Häufigkeiten gibt. Im Extremfall könnte eine gleichmäßig hohe relative Häufigkeit beobachtet werden, weil die Region nur durch zwei Marken abgedeckt ist, jeweils an der Start- und Endposition der Region und über das Segment dazwischen keine Informationen vorliegen. Die

schlussendliche Deklarierung eines Segments einer CNV-Region als Kandidaten-Region, die eine möglicherweise mit dem Phänotyp assoziierte CNV enthält, erfolgt nach manueller Evaluation.

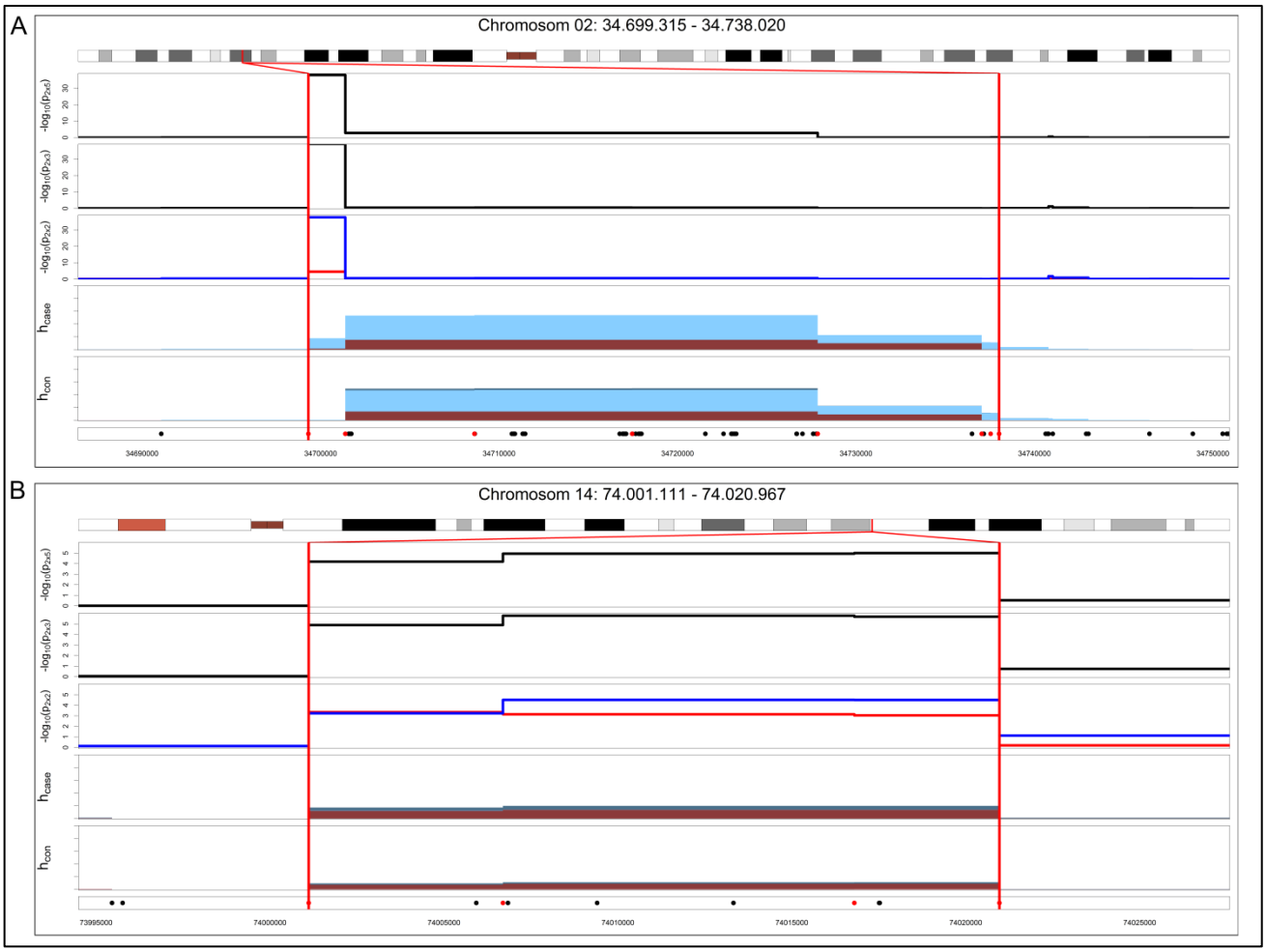


Abbildung 2-6: Plots der CNV-Regionen

A) Beispiel für eine Region, deren relevantes Segment mit einer hohen relativen Häufigkeit der CNV-Vorhersagen deutlich kleiner ist als die automatisch detektierte CNV-Region. Der minimale p-Wert befindet sich jedoch außerhalb dieses Segmentes (Ausreißer), weshalb es trotz guter Abdeckung nicht als Kandidaten-Region in Frage kommt. **B)** Beispiel für eine CNV-Region mit gleichmäßig hohen relativen Häufigkeiten über das gesamte Segment und eine gute Abdeckung durch Marker. $-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).

2.5 CNV-TaqMan® Experiment

Zur experimentellen Validierung vorhergesagter CNVs wurde die auf der Polymerase-Kettenreaktion (engl. *polymerase chain reaction*, kurz PCR, Mullis *et al.*, 1986) beruhende TaqMan®-Technologie verwendet. Die Vorbereitung der DNA Proben sowie die Ausführung der CNV-TaqMan®-Experimente wurde vollständig von den technischen Assistenten des IKMB durchgeführt.

2.5.1 TaqMan® Assays

Zur Bestimmung des CN-Genotyps wurden spezielle TaqMan® Copy Number Assays (CN-Assays) zusammen mit TaqMan®-Referenzassays in einer duplex real-time PCR verwendet (Abbildung 2-7). Diese TaqMan®-Assays enthalten, neben den für die Zielsequenz spezifischen, nicht markierten Forward- und Reverse-Primern, die mit einem Fluoreszenzfarbstoff (FAM™ oder VIC®) markierten Sonden (Applied Biosystems, 2010). Diese Sonden bestehen aus für die Zielsequenz komplementären Oligonukleotiden, an deren 5‘-Ende der Fluoreszenzfarbstoff und am 3‘-Ende ein nichtfluoreszierender Quencher (NFQ) und ein *Minor Groove Binder* (MGB) angebracht sind. Der Quencher verhindert durch seine räumliche Nähe die Floreszenz des an die Sonde gebundenen Farbstoffs. Die Sonden der CN-Assays sind FAM™-markiert und bestimmen die auf CNVs zu untersuchende Zielsequenz. Die VIC®-markierten Sonden der Referenzassays binden spezifisch an eine Referenzsequenz, von der bekannt ist, dass sie in zweifacher Kopienzahl in einem diploiden Genom vorkommt (Applied Biosystems, 2010).

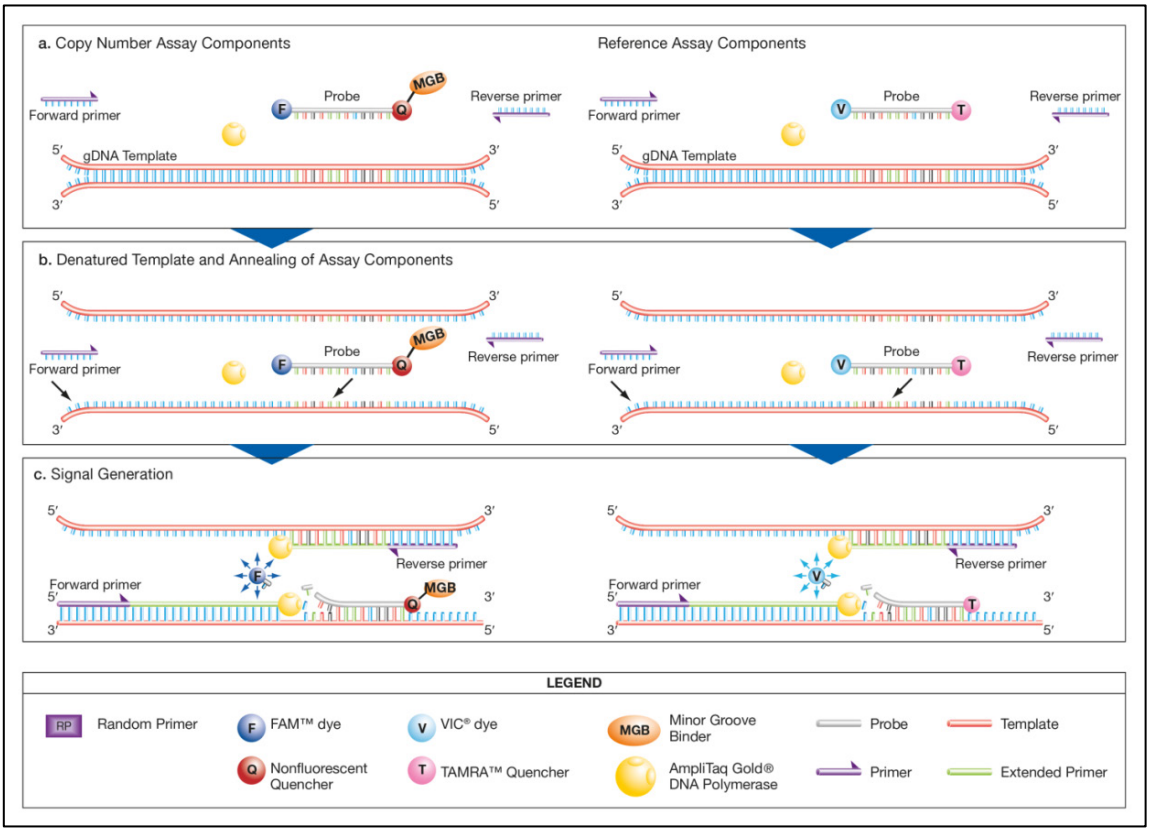


Abbildung 2-7: Funktionsweise eines TaqMan®-CNV-Assays nach Applied Biosystems (2010)

Die aufgereinigte genomische DNA (gDNA) der Probe wird zusammen mit dem CN-Assay, dem Referenz Assay sowie dem *TaqMan® Genotyping Master Mix*, welcher die *AmpliTaq Gold DNA Polymerase* (Taq-Polymerase) enthält, gemischt. Während eines PCR-Zyklus wird die gDNA

denaturiert, wodurch Primer und Sonden komplementär an die Einzelstränge der DNA binden können. Solange die Oligonukleotide der Sonden intakt sind, wird das Fluoreszenzsignal des Reporterfarbstoffes durch den Quencher unterdrückt. Während der Amplifikation der Einzelstränge durch die Taq-Polymerase werden Nukleotide abgetrennt, die entlang des Amplikons bereits hybridisiert sind, wodurch Reporterfarbstoff und Quencher getrennt werden, und ein auslesbares Fluoreszenzsignal entsteht (Abbildung 2-7).

2.5.2 PCR-Experiment

In Vorbereitung für das PCR-Experiment wurden die DNA-Proben mit einer Konzentration von 20 ng/µl durch die technischen Assistenten des IKMB auf 96-Deepwell-PCR-Platten (Master-Platten) aufgetragen. Die photometrischen Messungen der DNA-Konzentration erfolgen mittels Qbit (Life Technologies) oder DropSense96 (Trinean). Die Bravo Automated Liquid Handling Platform (Agilent Technologies) wurde verwendet, um die Proben von den 96-Well Master-Platten in vierfacher Replikation zusammen mit einem TaqMan® Copy Numer Assay, einem TaqMan® Referenz Assay und dem TaqMan® Genotyping Master Mix auf 384-Well-PCR-Platten aufzutragen, welche anschließend in dem PCR-Experiment verwendet wurde (Abbildung 2-8). Das PCR-Experiment wurde entsprechend dem offiziellen Protokoll des Herstellers der verwendeten TaqMan® Assays (Applied Biosystems, 2010), durchgeführt. Sowohl die PCR als auch die Detektion der Fluoreszenzsignale erfolgte auf einem *7900HT Fast Real-Time System* (Applied Biosystems®).

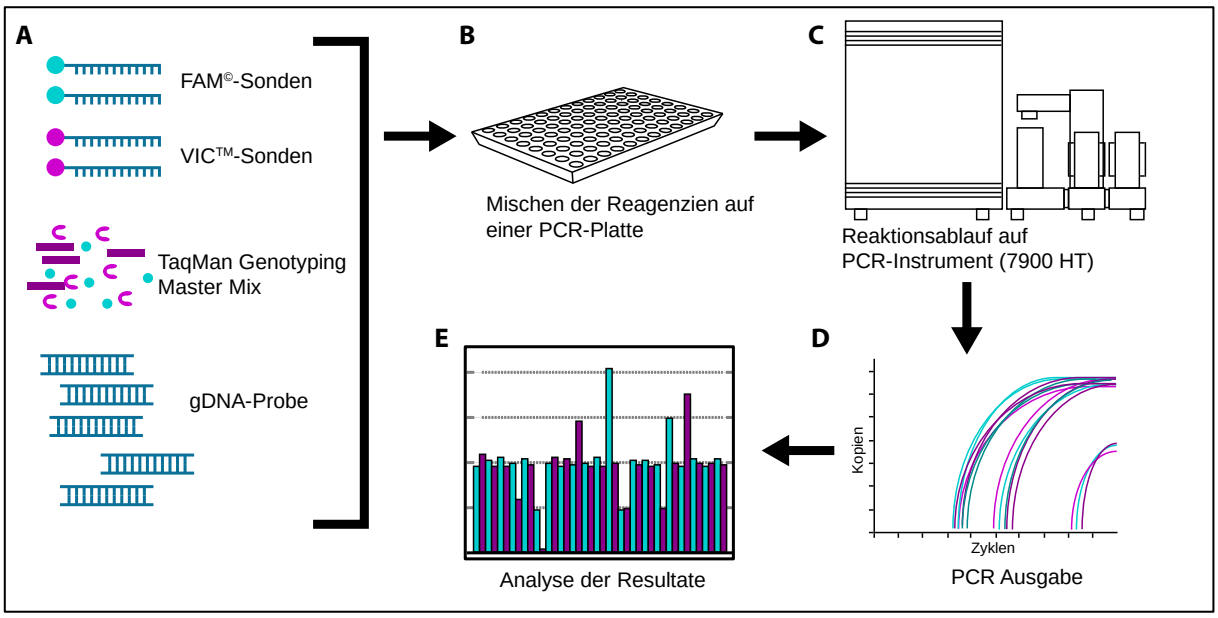


Abbildung 2-8: Schematische Darstellung des Ablaufs der TaqMan® CNV-Analyse

Die gDNA-Proben werden mit den für die PCR benötigten Reagenzien (A) gemischt und in vierfacher Replikation auf eine PCR-Platte aufgetragen (B). Nach durchführen des PCR-Experiments (C) werden die Ergebnisse (D) mit Hilfe der CopyCaller®-Software ausgewertet (E).

2.6 Auswertung der CNV-Genotypisierungsdaten

Zur Auswertung der in dem PCR-Experiment erhobenen Daten wurde zunächst die Software CopyCaller® von Applied Biosystems (www.appliedbiosystems.com/support/software/copycaller/) verwendet, welche aus den gemessenen Signalen CN-Genotypen für jede Probe errechnet. Anschließend wurde eine Assoziationsanalyse unter Verwendung von Skripten in der Programmiersprache R Version 2.15.2 (R Core Team 2013, http://www.R-project.org) durchgeführt.

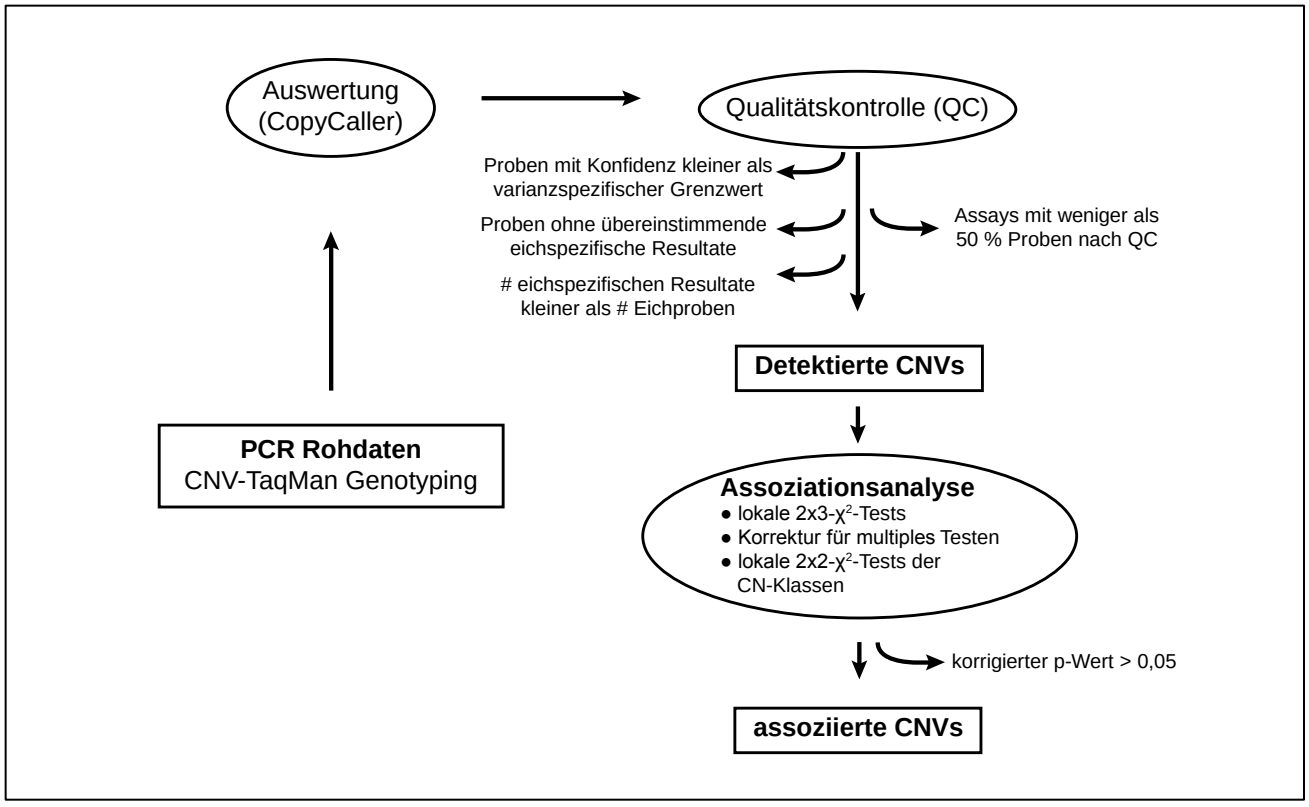


Abbildung 2-9: Schematische Darstellung der *in-vitro* Assoziationsanalyse

Schematische Darstellung der Kontrollpunkte (Ovale), Operationen (Rechtecke) und Filterkriterien (freier Text) der zur Analyse verwendeten Skripte.

2.6.1 Algorithmus der CopyCaller® Software

Der Algorithmus der CopyCaller® Software (Applied Biosystems, 2011) bestimmt CN-Genotypen basierend auf einem theoretischen Modell, welches die biochemischen Vorgänge abbildet, die dem PCR-Experiment zugrunde liegen und dabei mögliche Störgrößen statistisch einbezieht. Basierend auf diesem Modell wird für jede Probe eine Kopienzahl sowie zugehörige Qualitätsmetriken bestimmt.

Während der CN-Genotypisierung wird für jede Probe die Anzahl der PCR-Zyklen C_T gemessen, die nötig sind, bis genügend freier Reporterfarbstoff vorhanden ist, um ein Fluoreszenzsignal zu

erfassen. Diese Messungen erfolgen separat für beide Reporterfarbstoffe (C_T^{FAM} und C_T^{VIC}). Die Referenzsequenz unterliegt keiner Änderung in der Kopienzahl, somit repräsentiert der C_T^{VIC}-Wert die Anzahl der Zyklen, die unter gegebenen Bedingungen bei zwei Kopien der Zielsequenz benötigt werden, bis ein Fluoreszenzsignal gemessen wird. Die PCR verdoppelt unter optimalen Bedingungen die Menge an genetischem Material mit jedem Zyklus, somit bedeutet ein kleiner C_T-Wert, dass die entsprechende Sequenz häufig in der Probe vorkommt und zu Beginn der PCR bereits viel genetisches Material zur Verfügung stand. Ein C_T^{FAM}-Wert der um einen Zyklus größer ist als der C_T^{VIC}-Wert der Probe bedeutet, dass die Zielsequenz halb so häufig vorkommt wie die Referenzsequenz. Ist der C_T^{FAM}-Wert um einen Zyklus geringer als der C_T^{VIC}-Wert, würde das bedeuten, dass die Zielsequenz doppelt so häufig vorkommt wie die Referenzsequenz. In der Praxis weicht das Verhältnis der Differenz der C_T-Werte zu der Kopienzahl von dem beschriebenen Modell ab, da die PCR die Menge an genetischem Material mit jedem Zyklus nicht exakt verdoppelt. Weitere mögliche Störgrößen sind z.B. unterschiedliche Genauigkeiten bei der Messung der beiden Reporterfarbstoffe, ein unterschiedlicher Einfluss der verschiedenen Ziel- und Referenzsequenzen auf die PCR sowie Mutationen in der Referenzsequenz. Der Einfluss dieser Störgrößen wird in einem statistischen Modell berücksichtig.

Da die PCR das genomische Material in jedem Zyklus verdoppelt, verringert sich ΔC_T exponentiell für steigende Kopienzahl der Zielsequenz (sinkender C_T^{FAM}-Wert) bei konstanter Kopienzahl der Referenzsequenz (fester C_T^{VIC}-Wert). Dieser Zusammenhang wird von der Software durch die Gleichung

$$\Delta C_T = K - \log_{1+E}(cn)$$

modelliert, wobei K eine Konstante ist, E die Effizienz der PCR und cn die Kopienzahl der Zielsequenz, die Werte in $[1, \infty)$ annahemen kann. Unter optimalen Bedingungen gilt $K = 0$ und $E = 1$, allerdings gilt unter Einfluss der oben erwähnten Störgrößen in der Praxis $K \neq 0$ und $E < 1$. Dieser Umstand bedeutet, dass bei gleicher Kopienzahl der Zielsequenz in unterschiedlichen Proben die entsprechenden Differenzen ΔC_T der C_T-Werte verschieden seinen können. Um diese Variation zu modellieren, wird angenommen, dass die ΔC_T-Werte unterschiedlicher Proben mit gleicher Kopienzahl normalverteilt sind.

2.6.2 Bestimmung des CN-Genotyps

Zur Bestimmung der CN-Genotypen werden alle Parameter des in Abschnitt 2.6.1 beschriebenen Modells aus den gemessenen Daten durch den Algorithmus der Software CopyCaller® geschätzt, wobei alle analysierten Proben einer Platte einbezogen werden (Applied Biosystems, 2011). Kann

für ein Replikat einer Probe kein FAM-Signal gemessen werden, bleibt der zugehörige C_T^{FAM}-Wert unbestimmt. Wird zusätzlich auch kein VIC-Signal oder nur ein sehr schwaches ($C_T^{VIC} > 32$) gemessen, gilt die PCR-Reaktion (vermutlich aufgrund einer zu geringen gDNA-Konzentration) als fehlgeschlagen. Kann für die entsprechende Reaktion ein VIC-Signal gemessen werden ($C_T^{VIC} \leq 32$), so wird dem Replikat automatisch die Kopienzahl $cn = 0$ zugewiesen. Nachdem die Daten fehlgeschlagener Replikate herausgefiltert wurden, berechnet die CopyCaller®-Software die reaktionsspezifischen Differenzen der C_T-Werte als $\Delta C_T = C_T^{FAM} - C_T^{VIC}$. Ist das FAM-Signal im Vergleich zu dem VIC-Signal sehr schwach ($\Delta C_T > 4$) wird angenommen, dass das schwache Signal durch nichtspezifische Bindung der Sonde entsteht, und dem Replikat wird die Kopienzahl $cn = 0$ zugewiesen.

Um die Variabilität der ΔC_T-Werte pro Platte wird zu berechnen, wird davon ausgegangen, dass die ΔC_T-Werte aller Replikate mit demselben CN-Genotyp normalverteilt sind, wobei der Mittelwert der ΔC_T-Werte den Erwartungswert der Verteilung darstellt. Weiter wird vereinfachend angenommen, dass die Varianz für alle genotypspezifischen Verteilungen gleich ist. Die Verteilungen werden aus den Daten berechnet, indem die Parameter ermittelt werden, die unter den gemachten Annahmen die Daten am besten erklären. Die Standardabweichung dieser Verteilungen der plattenspezifischen ΔC_T-Werte gilt dann als Kenngröße für die Variabilität der gemessenen Daten. Dabei spricht eine hohe Variabilität für einen großen Einfluss von Störgrößen, was die genaue Ermittlung der CN-Genotypen erschwert. Zur Qualitätskontrolle werden Replikate, deren ΔC_T-Werte um mehr als das Vierfache der plattenspezifischen Standardabweichung von dem Mittelwert der ΔC_T-Werte aller Replikate der jeweiligen Probe abweichen, von der weiteren Auswertung ausgeschlossen. Sind für eine Probe die Daten von mehr als einem Replikat verfügbar, so wird anschließend der probenspezifische Mittelwert $\mu(\Delta C_T) = \frac{1}{n}\sum_{w=1}^{n}(\Delta C_T)_w$ über die Messungen der n Wells der Replikate berechnet.

Unter optimalen Bedingungen repräsentiert $\Delta C_T = 0$ einen unveränderten CN-Genotyp mit $cn = 2$, da die FAM™- und VIC®-Sonden gleich viel PCR-Zyklen gebraucht haben, um ein messbares Fluoreszenzsignal zu produzieren. Die in Abschnitt 2.6.1 beschriebenen Störgrößen führen aber dazu, dass in der Praxis auch für Proben ohne Kopienzahl-Variation ($cn = 2$) unterschiedliche C_T^{FAM}- und C_T^{VIC}-Werte gemessen werden und damit $\Delta C_T \neq 0$ gilt. Aus diesem Grund ist eine Eich-Probe nötig, um einen ΔC_T-Wert zu ermitteln, der $cn = 2$ repräsentiert. Dazu muss der CN-Genotyp der Zielsequenz in der der Eich-Probe cn_c bekannt sein, wobei idealerweise $cn_c = 2$ gilt.

Zur Bestimmung des CN-Genotyps einer Probe s wird die Differenz des probenspezifischen Mittels der ΔC_T-Werte von dem der Eich-Probe c als $\Delta\Delta C_T = \mu(\Delta C_T)_s - \mu(\Delta C_T)_c$berechnet. Der CN-Genotyp der Zielsequenz in der zu testenden Probe s ergibt sich dann als $cn_s^{calc} = cn_c 2^{-\Delta\Delta C_T}$. Die auf diese Weise berechnete Kopienzahl ist in der Regel nicht ganzzahlig und wird anschließend auf die nächste ganze Zahl cn_s^{pred} gerundet.

2.6.3 Verwendete Eich-Proben

Auf jeder analysierten PCR-Platte wurden Proben von drei unverwandte HapMap Individuen zur Qualitätskontrolle des PCR-Experimentes verwendet. Diese Proben wurden während der Auswertung als Eich-Proben genutzt, wobei die CN-Genotypen an den entsprechenden Zielsequenzen aus den Ergebnissen der Studie von Conrad et al., (2010) sowie den Ergebnissen des HapMap Projects (The International HapMap Consortium, 2005) und der ersten Phase des 1000 Genomes Projects (The 1000 Genomes Project Consortium, 2012) entnommen wurde. Diese Studien geben die CN-Genotypen oder –Klassen detektierter CNVs an. Lag für eine gegebene Probe keine Detektion vor, wurde ein unveränderter CN-Genotyp ($cn = 2$) angenommen. Die Resultate der verschiedenen Studien sind zum Teil nicht konkordant, sodass für Zielsequenzen mit nicht eindeutiger Studienlage der am häufigsten detektierte CN-Genotyp für die entsprechende Probe angenommen wurde.

2.6.4 Auswertung und Qualitätskontrolle

Durch die im PCR-Experiment verwendeten HapMap Proben standen für die Auswertungen der Daten jedes Assays mit CopyCaller® bis zu drei Eich-Proben zur Verfügung, deren CN-Genotypen bereits in mehreren Studien untersucht wurden (Abschnitt 2.6.3). Die HapMap-Proben deren PCR-Reaktionen fehlschlugen, oder für die eine vollständige Deletion ($cn = 0$) bestimmt wurde, konnten nicht als Eich-Proben verwendet werden. Die assay- und plattenspezifische Auswertung durch die CopyCaller® Software wurde mit jeder verfügbaren Eich-Probe wiederholt. HapMap Proben die nach der CN-Genotypisierung und der Auswertung unter Verwendung der anderen Eich-Proben einen von der Literatur abweichenden CN-Genotypen aufwiesen, wurden nicht als Eich-Probe für das entsprechende Assay verwendet. Dies war nur möglich für Assays, bei denen alle drei HapMap Proben als Eich-Proben verwendet werden konnten. Standen keine Eich-Proben für die Auswertung zur Verfügung so wurde angenommen, dass der Median aller plattenspezifischen ΔC_T-Werte am besten den unveränderten CN-Genotyop ($cn = 2$) repräsentiert und als Referenz verwendet.

Zur Qualitätskontrolle der mit CopyCaller® ermittelten CN-Genotypen wurden drei Qualitätsmetriken verwendet: (i) Die Konkordanz der eichspezifischen Resultate, (ii) die durch die Software berechnete Konfidenz der ermittelten CN-Genotypen, sowie (iii) die relative Anzahl der eichspezifischen CN-Genotypen.

Bei einer hohen Variabilität der Daten kann es vorkommen, dass die eichspezifischen CN-Genotypen einzelner Proben nicht konkordant sind. Diese Proben wurden für die weitere Analyse des Assays entfernt. Somit wurden nur Proben verwendet, deren CN-Genotypen in den Auswertungen unter Verwendung der verfügbaren Eich-Proben gleich waren.

Die durch die Software berechnete Konfidenz gibt die Wahrscheinlichkeit an, dass eine berechnete Kopienzahl nach dem statistischen Modell des CopyCaller® Algorithmus korrekt ist. Die Genauigkeit, mit der das von der Software verwendete Modell zwischen zwei aufeinander folgenden CN-Genotypen unterscheiden kann, nimmt mit ansteigender Kopienzahl ab. Zusätzlich hängt die Genauigkeit, mit der die CN-Genotypen bestimmt werden können, von der Variabilität der ΔC_T-Werte der analysierten Platte ab. Das statistische Modell funktioniert optimal bei einer geringen Variabilität ($SD(\Delta C_T) \leq 0{,}05$), normal bei einer mittleren Variabilität ($0{,}05 < SD(\Delta C_T) \leq 0{,}09$) und schlecht bei einer hohen Variabilität ($SD(\Delta C_T) > 0{,}09$). Für alle drei zu erwartenden Funktionsweisen wurden von den Entwicklern der Software Grenzwerte für die Konfidenz angegeben, bei denen zu erwarten ist, dass 95 % der Proben der jeweiligen Kopienzahl diesen Konfidenzwert erreichen (Tabelle 2-2). Proben deren Konfidenzwerte der ermittelten CN-Genotypen nicht mindestens den plattenspezifischen Grenzwerten entsprachen, wurden aus der weiteren Analyse des entsprechenden Assays entfernt.

Tabelle 2-2: Grenzwerte der Konfidenz nach Applied Biosystems (2011)

Variabilität	Kopienzahl									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
gering	1,0000	1,0000	1,0000	1,0000	0,9998	0,9903	0,8976	0,6619	0,4324	0,2934
normal	1,0000	1,0000	0,9916	0,7468	0,2833	0,1938	0,1477	0,1299	0,1196	0,0856
hoch	0,9570	0,6251	0,1092	0,0897	0,0769	0,0760	0,0747	0,0682	0,0527	0,0306

Proben für die eine oder mehrere eichspezifische CN-Genotypen nicht den Kriterien der Qualitätskontrolle entsprachen, wurden aus der Analyse des Assays entfernt. Assays mit einer zu hohen Rate an fehlenden CN-Genotypdaten (> 50 %) wurden von der Assoziationsanalyse ausgeschlossen.

2.7 Validierungen

Die Validierung der CNV-Vorhersagen mittels CNV-TaqMan®-Genotypisierung erfolgte in zwei Schritten. Zunächst wurden die Kandidaten-Regionen in einer Teilstichprobe des GWAS-Datensatzes technisch validiert, um abzuschätzen, wieviel der Vorhersagen sich experimentell bestätigen lassen. Anschließend wurde durch eine Assoziationsanalyse, basierend auf CNV-TaqMan®-Genotypisierungsdaten einer unabhängigen Stichprobe, untersucht ob der vorhergesagte Zusammenhang der Kandidaten-Regionen mit dem Phänotyp der Sarkoidose statistisch validiert werden kann. Die CNV-TaqMan®-Genotypisierung wurde in beiden Fällen wie in den Abschnitten 2.5 und 2.6 beschrieben durchgeführt.

2.7.1 Vergleich mit *in-silico* CNV-Vorhersagen (technische Validierung)

Zur technischen Validierung der Kandidaten-Regionen wurden entsprechende CNV-TaqMan®-Assays ausgewählt (Abschnitt 3.3), und das PCR-Experiment mit 92 zufällig ausgewählten Proben des Datensatzes zur CNV-Vorhersage durchgeführt. Die Ergebnisse der Auswertung durch die CopyCaller®-Software wurden mit den Vorhersagen durch PennCNV verglichen. CNV-Regionen deren experimentell ermittelte CN-Klassen in mehr als 60 % der verwendeten Proben den Vorhersagen entsprachen, galten als technisch Validiert, da diese Rate nach dem Benchmark zu erwarten war.

2.7.2 Assoziationsanalyse (statistische Validierung)

Zur statistischen Validierung wurde das PCR-Experiment mit den entsprechenden CNV-TaqMan®-Assays in einer unabhängigen Stichprobe durchgeführt (Abschnitt 3.3.2). Aufgrund des beträchtlichen Einflusses von Störgrößen auf das PCR-Experiment und der damit verbundenen Ungenauigkeit bei der Bestimmung der CN-Genotypen wurde nur die Verteilung der CN-Klassen untersucht. Der dafür nötige 2×3-χ^2-Test wurde für jedes Assays separat durchgeführt. Die daraus resultierenden lokalen p-Werte wurden anschließend durch ein Permutationsverfahren (Westfall & Young, 1993) mit 100.000 Wiederholungen für multiples Testen korrigiert.

3 Ergebnisse

3.1 CNV-Software Benchmark

Ziel des Benchmarks war es, eine vergleichende Analyse in Bezug auf die Charakteristiken der vorhergesagten CNVs sowie auf die Validität der gemachten Vorhersagen von sechs CNV-Analyse-Softwares durchzuführen. Dazu wurden die Vorhersagen für 30 unverwandten HapMap Individuen eingehend verglichen und ein familienbasierter Ansatz zur Validierung angewendet.

3.1.1 CNV-Vorhersage

Die von den Softwares in den Proben der Kinder der 30 europäischen (CEU) und 30 afrikanischen (YRI) Trios getroffenen Vorhersagen der CNVs unterschieden sich stark in der Anzahl pro Individuum, der Länge und der Markerdichte. Die Anzahl reichte im Median von 75 CNVs pro Individuum bei Vorhersagen durch PennCNV bis 211 CNVs pro Individuum bei Vorhersagen durch R-gada (Abbildung 3-1 und Tabelle 3-1). Softwares basierend auf Segmentierungs-Algorithmen sagten im Median eine signifikant höhere Anzahl von CNVs voraus als Softwares basierend auf HMM-Algorithmen (Median:182 bzw. 98, p-Wert des Wilcoxon-Rangsummentests: $p = 1{,}6 \times 10^{-11}$). Darüber hinaus zeigten die Segmentierungs-Algorithmen einen nicht signifikanten Trend zu einer größeren Variabilität zwischen den Softwares in der Anzahl der vorhergesagten CNVs (Median-Deviation (engl. *median absolute deviation*, kurz MAD): 42,3 bzw. 19,3, $p = 0{,}12$ bei 10.000 Permutationen der zugeordneten Algorithmus-Klassen). Alle Softwares bis auf PennCNV sagten weniger CNVs in den europäischen Proben (CEU) als in den afrikanischen Proben (YRI) voraus ($p < 0{,}05$ für alle Softwares; Appendix Tabelle 1 und Appendix Tabelle 2).

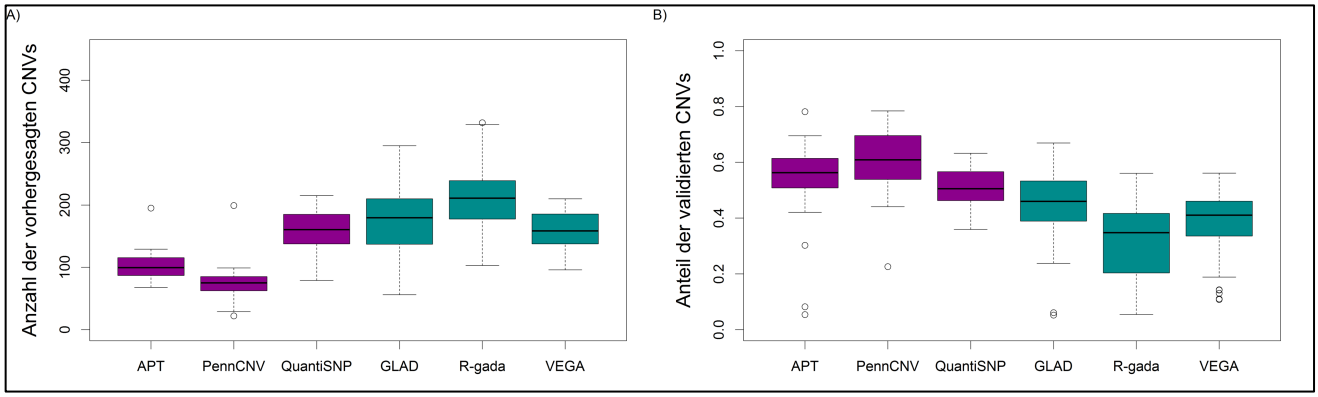


Abbildung 3-1: CNV-Vorhersagen und familienbasierte Validierung

A: Anzahl der vorhergesagten CNVs pro Probe. **B:** Anteil der durch Familieninformationen validierten CNVs pro Probe.

Tabelle 3-1: Probenspezifische Eigenschaften der CNV-Vorhersagen

Software	Anzahl	Median der Länge [kb]	Median der kumulativen Länge [kb]	Median der Anzahl der Marker	Median der Distanz der Marker [kb]	DDR
APT	99,5 (87,0 - 115,2)	8,9 (8,1 - 10,3)	4,6 (3,7 - 5,7)	10,2 (8,0 - 14,0)	0,22 (0,19 - 0,26)	4,3 (3,2 - 5,1)
GLAD	179,5 (139,5 - 210,0)	7,1 (6,6 - 8,5)	6,2 (4,4 - 8,3)	6,0 (5,0 - 8,0)	0,20 (0,17 - 0,26)	2,8 (2,3 - 3,4)
PennCNV	75,0 (63,2 - 84,5)	21,7 (17,3 - 25,8)	5,1 (4,0 - 6,5)	25,0 (23,0 - 29,2)	0,18 (0,14 - 0,21)	5,5 (4,8 - 6,6)
QuantiSNP	160,5 (137,8 - 184,5)	9,0 (8,3 - 10,0)	8,2 (5,7 - 23,2)	6,0 (5,9 - 7,0)	0,23 (0,20 - 0,26)	3,1 (2,6 - 3,6)
R-gada	211,0 (177,8 - 236,5)	8,0 (7,1 - 9,6)	121,0 (18,9 - 281,4)	7,0 (6,8 - 10,0)	0,28 (0,23 - 0,33)	4,4 (3,6 - 5,4)
VEGA	158,5 (137,8 - 185,2)	7,0 (6,3 - 7,6)	6,2 (4,7 - 7,9)	7,0 (5,0 - 8,0)	0,28 (0,23 - 0,31)	3,6 (3,1 - 4,6)
Algorithmus						
HMM	98,0 (87,0-111,5)	9,7 (8,9-11,2)	5,2 (4,1-6,6)	10,8 (8,0-14,0)	0,21 (0,19-0,24)	4,3 (3,3-5,0)
Segmentierung	182,0 (150,8-210,0)	7,4 (6,6-8,1)	7,0 (5,0-8,7)	7,0 (6,0-8,0)	0,26 (0,22-0,30)	3,6 (3,1-4,4)

Angegeben ist der Median und in Klammern der Interquartilsabstand (IQR). **DDR:** Verhältnis von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*).

Für alle Softwares wurden linksschiefe Verteilungen des probenspezifischen Medians der CNV-Länge sowie Proben mit einem besonders großen Median der CNV-Länge (Ausreißer) beobachtet (Abbildung 3-2). Insbesondere R-gada wies CNV-Vorhersagen mit einer Länge von bis zu 1,9 Mb und einer kumulativen Länge aller Vorhersagen einer Probe von bis zu 126 Mb auf. Der mittlere probenspezifische Median der CNV-Länge, berechnet über alle vorhergesagten CNVs, ähnelte sich bei allen Softwares, ausgenommen PennCNV, welche einen Trend zu längeren CNVs aufwiesen. Im Allgemeinen tendierten die HMM-Softwares dazu im Median je Probe längere CNVs (Median: 9,7 kb) als die Segmentierungs-Softwares vorherzusagen (Median: 7,4 kb, $p=1{,}6\times10^{-11}$; Tabelle 3-1). Auch die kumulative Länge der CNVs je Probe variierte zwischen den Softwares und reichte von einem Median von 4,6 Mb (Interquartilsabstand (IQR): 3,7-5,7 Mb) bei APT über 8,7 Mb (5,7-23,2 Mb) bei QuantiSNP bis zu 121,0 Mb (18,9-281,4 Mb) bei R-gada. Der Median über die kumulative Länge je Probe war bei europäischen Proben durchgehend größer als bei afrikanischen Proben ($p < 0{,}05$ für alle Softwares; Appendix Tabelle 1 und Appendix Tabelle 2).

Der Median je Probe über die Anzahl der Marker pro CNV war in den Vorhersagen für alle Softwares außer PennCNV ähnlich groß. In den Vorhersagen von PennCNV war der Median über die probenspezifischen Mediane bis zu vier Mal so hoch wie in den anderen Softwares, was trotz der größeren Länge der CNV-Vorhersagen zu der geringsten Distanz zwischen den Markern führte (Tabelle 3-1). Alle Softwares wiesen probenspezifische Mediane der Distanz auf, die im Mittel über dem Median der paarweisen Distanzen aller Marker auf dem *Affymetrix Human SNP Array 6.0* (684 bp) liegen. Dies spiegelt eine bevorzugte CNV-Vorhersage in Regionen mit hoher Markerdichte

wieder. Die Distanz zwischen den Markern wies keine signifikanten Unterschiede zwischen europäischen und afrikanischen Proben auf (Appendix Tabelle 1 und Appendix Tabelle 2).

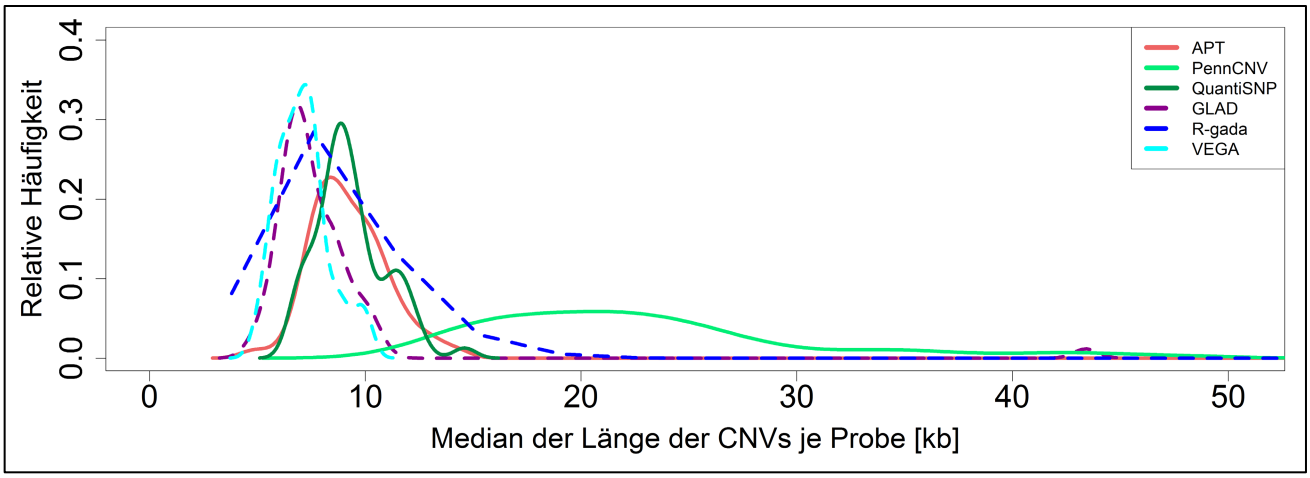


Abbildung 3-2: Median der probenspezifischen Länge der CNV-Vorhersagen

Kerndichteschätzung des Medians der CNV-Länge je Probe. Ausreißer mit einer Länge > 50 kb sind nicht dargestellt. **Durchgezogene Linien:** Software basierend auf HMM-Algorithmen. **Gestrichelte Linien:** Software basierend auf Segmentierungs-Algorithmen.

Alle sechs Softwares sagten deutlich mehr Deletionen als Duplikationen voraus. Der Median des Verhältnisses von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*, kurz DDR) je Probe reichte von 2,8 bei GLAD bis 5,5 bei PennCNV (Tabelle 3-1). Für Softwares basierend auf HMM-Algorithmen ergab sich im Vergleich zu Softwares basierend auf Segmentierung ein höheres DDR (4,3 vs. 3,6, $p = 6{,}9 \times 10^{-4}$; Tabelle 3-1). Im Vergleich der DDR-Werte bei europäischen und afrikanischen Proben konnte kein Unterschied festgestellt werden (Appendix Tabelle 1 und Appendix Tabelle 2).

3.1.2 *In-silico* Validierung

Im Hinblick auf die in Abschnitt 3.1.1 beschriebenen Unterschiede zwischen den Softwares bei der Vorhersage von CNVs wurde ein familienbasierter Ansatz (Abschnitt 2.2.4) verwendet, um die Vorhersagen *in-silico* zu validieren. Als bestätigt galten dabei CNV-Vorhersagen, deren Sequenz zu mindestens 90 % in einem der beiden Eltern ebenfalls als CNV derselben CN-Klasse vorhergesagt wurden. Der Anteil der CNVs, die validiert werden konnten, unterschied sich stark zwischen den verschiedenen Softwares und reichte von 41,1 % bei R-gada bis zu 60,9 % bei PennCNV (Tabelle 3-2). Alle Softwares, die auf HMM-Algorithmen basieren, hatten höhere Validierungsraten als die auf Segmentierung basierende, was zu einem deutlichen Unterschied zwischen den Software-Gruppen führte (55,9 % vs. 41,4 %, Tabelle 3-2). Dieser Trend blieb auch bei getrennter Betrachtung der Validierungsraten der Deletionen und Duplikation bestehen (Tabelle 3-2). PennCNV, QuantiSNP

und R-gada wiesen geringfügig höhere Validierungsraten für Duplikationen auf (Median DDR<1), wohingegen APT, GLAD und VEGA leicht erhöhte Validierungsraten bei Duplikationen zeigten (Median DDR>1). Trotz dieser Unterschiede schloss der Interquartilsabstand bei allen sechs Softwares die Eins ein.

Tabelle 3-2: CNV-Validierungsraten

Software	Validierung CNVs [%]	Validierung Deletionen [%]	Validierung Duplicationen [%]	DDR, Validierungs-ratens	Validierung kumulierte CVN Sequenz [%]
APT	56,3 (50,8-61,3)	55,7 (50,9-61,1)	60,0 (48,0-67,2)	0,9 (0,8-1,2)	55,7 (42,2-66,9)
GLAD	46,0 (39,1-53,2)	46,3 (35,9-52,5)	54,8 (41,0-60,9)	1,3 (1,0-1,5)	45,1 (31,7-58,2)
PennCNV	60,9 (53,9-69,5)	64,4 (57,4-74,1)	52,2 (44,3-59,0)	1,2 (0,9-1,4)	56,0 (43,4-65,1)
QuantiSNP	50,5 (46,3-56,6)	52,2 (47,2-58,2)	46,4 (37,9-55,7)	1,0 (0,8-1,2)	53,8 (32,3-77,5)
R-gada	34,8 (20,4-41,6)	34,6 (20,8-43,1)	32,3 (22,3-44,0)	0,8 (0,6-1,0)	5,2 (1,4-16,4)
VEGA	41,1 (33,6-45,9)	39,2 (31,1-45,4)	47,3 (36,6-54,9)	0,9 (0,7-1,1)	36,0 (21,2-53,0)
Algorithmus					
HMM	55,9 (49,5-61,3)	57,5 (50,6-60,6)	51,9 (46,1-60,0)	1,1 (0,9-1,3)	55,3 (43,4-65,0)
Segmentation	41,4 (35,1-47,6)	40,5 (33,9-46,1)	45,0 (36,0-53,5)	0,9 (0,7-1,1)	34,9 (19,9-47,2)

Angegeben ist der Median und in Klammern der Interquartilsabstand (IQR). **DDR:** Verhältnis von Deletionen zu Duplikationen.

Es konnten keine relevanten Unterschiede der prozentualen Anteile validierter CNVs für die Softwares bei Betrachtung unterschiedlicher Größenordnungen festgestellt werden, was für eine von der Länge der vorhergesagten CNVs größtenteils unabhängige Validierungsrate spricht (Abbildung 3-3). Die erweiterte Validierung, bei der elterliche CNV-Vorhersagen aller Softwares verwendet wurden, führte zu einer Erhöhung der Validierungsraten um ~10-20 %, wobei die relativen Verhältnisse der Softwares zueinander größtenteils unverändert blieben (Appendix Tabelle 6).

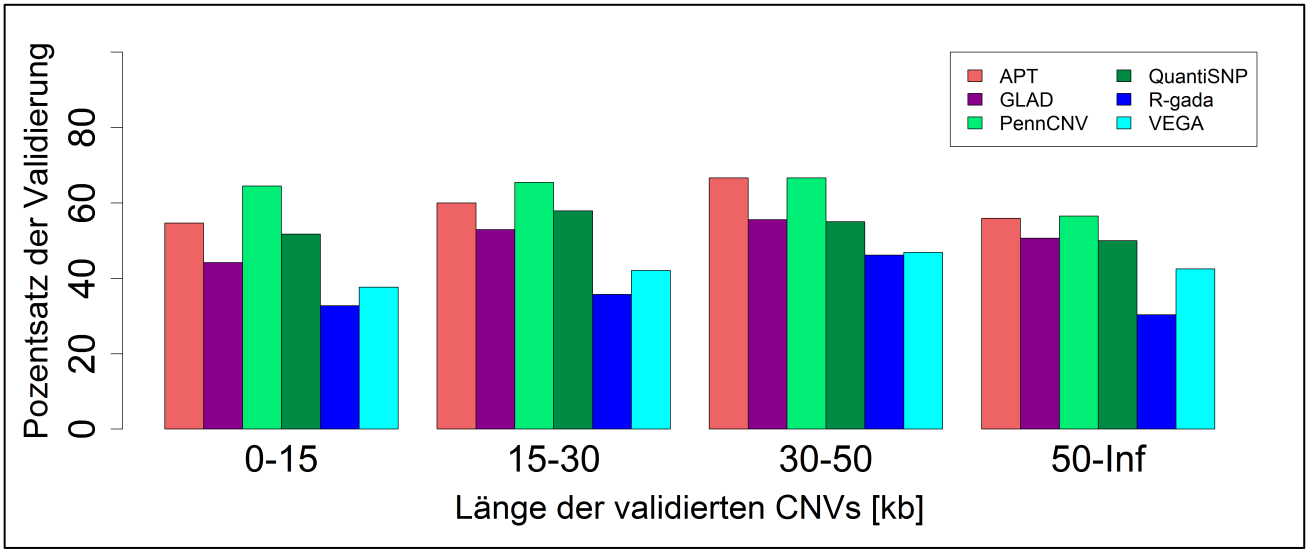


Abbildung 3-3: Validierungsraten aufgeteilt nach Länge der CNVs

Für eine zusätzliche Verifizierung der validierten CNVs durch externe Daten wurden sie mit zwei Datensätzen aus der *Database of Genomic Variation* (DGV) verglichen (je ein Datensatz basierend auf Array- und Sequenzierungstechnologie). Vollständige CNV Daten waren lediglich für insgesamt sechs der Trio-Kinder verfügbar, vier CEU-Proben (NA07048, NA10847, NA108 51, NA12878) und zwei YRI Proben (NA19129, NA19240). Daten aller Individuen eines Trios waren für keines der 60 Trios dieser Studie verfügbar. Die Verifizierung durch den DGV-Array-Datensatz war mit 40,6 % für R-gada bis 63,4 % für PennCNV deutlich höher als die durch den DGV-Sequencing-Datensatz mit 25,6 % für VEGA bis 36,2 % für PennCNV (Abbildung 3-4).

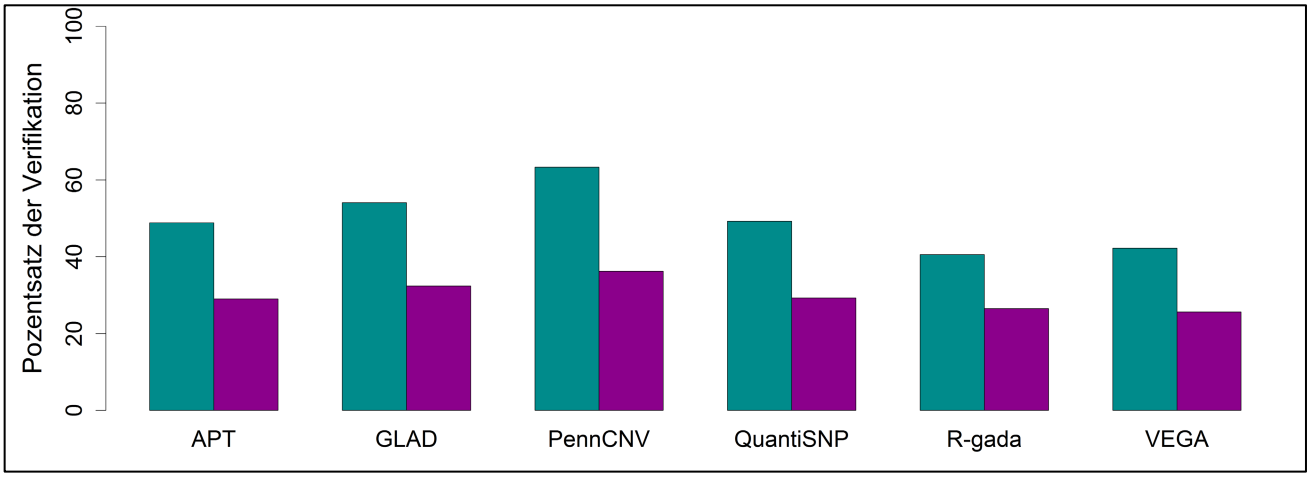


Abbildung 3-4: Verifizierung der CNV-Vorhersagen durch DGV-Datensätze

Anteil der Validierung durch zwei Datensätze aus der *Database of Genomic Variants* (DGV). Ein Datensatz basiert auf CNV-Detektionen durch Sequenzierungs-Technologie (rot), der andere Datensatz basiert auf CNVs, die durch Array-Methoden detektiert wurden (blau).

Der Anteil der kumulierten Sequenz der CNV-Vorhersagen, der validiert werden konnte, betrug bei den HMM-Softwares im Median 55,3 % und lag damit deutlich über den 34,9 % bei den Segmentierungs-Softwares (Tabelle 3-2). Der geringste Anteil von lediglich 5,2 % wurde bei R-gada festgestellt. Dieser extrem geringe Wert kam durch die Vorhersagen einer Reihe sehr langer CNVs zu Stande, die nicht validiert werden konnten. Populationsspezifische Unterschiede konnten weder bei den Validierungsraten noch bei dem Anteil der kumulierten CNV-Sequenzen festgestellt werden (Appendix Tabelle 3 und Appendix Tabelle 4).

Die validierten und nicht-validierten CNVs unterschieden sich in Bezug auf die Anzahl der CNVs, dem Median der Länge, dem Median über die Anzahl der Marker pro CNV und dem Median über die Distanz der Marker pro CNV. Validierte CNVs tendierten dazu länger zu sein und eine höhere Markerdichte zu besitzen als die nicht validierten CNVs (vergleiche Tabelle 3-3 mit Appendix Tabelle 5). Die Unterschiede von validierten zu nicht-validierten CNVs in Bezug auf diese

Kenngrößen waren für jede Software statistisch signifikant (Wilcoxon-Rangsummentest: $p < 0{,}05$ für alle Softwares).

Tabelle 3-3: Probenspezifische Eigenschaften der validierten CNVs

Software	Anzahl	Median der Länge [kb]	Median der kumulativen Länge [kb]	Median der Anzahl der Marker	Median der Distanz der Marker [kb]	DDR
APT	55,0 (48,8-60,0)	10,3 (9,1-12,0)	2,3 (1,9-3,1)	16,0 (14,0-18,0)	0,19 (0,15-0,21)	3,7 (3,1-4,6)
GLAD	79,0 (62,0-94,2)	8,7 (7,3-9,5)	2,7 (1,8-3,5)	10,0 (9,0-12,0)	0,15 (0,11-0,16)	3,6 (2,8-4,1)
PennCNV	42,5 (37,8-51,0)	21,2 (16,1-26,0)	2,4 (2,0-3,3)	26,5 (23,9-31,1)	0,15 (0,12-0,19)	6,2 (5,3-8,9)
QuantiSNP	83,5 (69,5-95,0)	9,7 (8,9-11,0)	3,2 (2,2-11,0)	9,0 (7,5-9,6)	0,19 (0,13-0,21)	3,0 (2,5-3,7)
R-gada	66,0 (37,0-87,0)	8,9 (7,7-10,3)	3,4 (1,4-5,1)	12,0 (10,0-15,2)	0,14 (0,12-0,19)	3,6 (2,9-4,2)
VEGA	62,0 (48,5-74,0)	8,2 (7,2-9,2)	2,3 (1,4-3,6)	11,0 (9,0-14,5)	0,16 (0,12-0,20)	3,0 (2,2-4,1)
Algorithmus						
HMM	56,0 (49,5-61,0)	10,7 (9,8-12,2)	2,4 (2,0-3,2)	16,0 (14,4-18,0)	0,16 (0,13-0,20)	3,9 (3,3-5,2)
Segmentierung	70,5 (57,0-79,2)	8,6 (7,6-9,3)	2,5 (1,6-3,5)	11,0 (9,8-13,0)	0,14 (0,12-0,18)	3,2 (2,9-3,8)

Angegeben ist der Median und in Klammern der Interquartilsabstand (IQR). **DDR:** Verhältnis von Deletionen zu Duplikationen.

Der Median über die Anzahl der validierten CNVs reichte von 42,5 (PennCNV) bis zu 83,5 (QuantiSNP). Das Verhältnis von Deletionen zu Duplikation der validierten CNVs war mit DDR-Werten von 3,0 (QuantiSNP und VEGA) bis 6,2 (PennCNV) ähnlich wie bei den nicht validierten CNVs (Tabelle 3-3, Appendix Tabelle 5 und Abbildung 3-5).

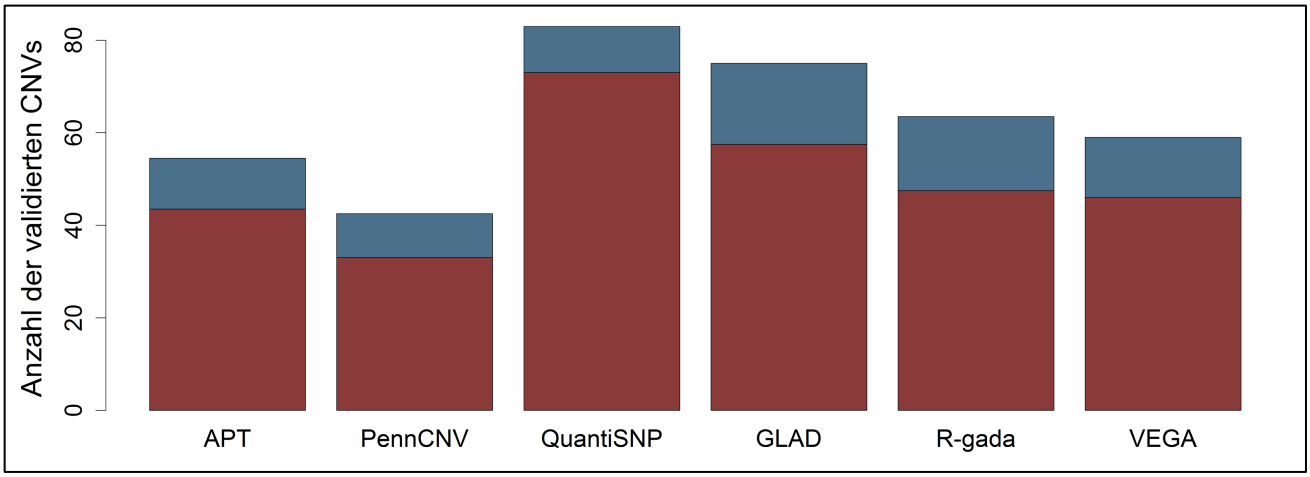


Abbildung 3-5: Median der probenspezifischen Anzahl der validierten CNVs

Blau: Duplikationen; **Rot:** Deletionen.

3.1.3 Pseudo-Validierung

In einigen Fällen kann die familienbasierte Validierung rein zufällig und nicht aufgrund einer Vererbung der CNVs auf den Nachkommen erfolgen. Um die Häufigkeit einer solchen „Pseudo-

Validierung“ abzuschätzen, wurde die Zuordnung der Eltern zu den Nachkommen wiederholt permutiert, und die so neu entstandenen Trios wurden wie zuvor untersucht. Es wurden zehn solcher Permutationen durchgeführt und jedes Mal der Median der Validierungsraten je Software erhoben. Der Anteil an CNVs, die auf diese Weise pseudo-validiert wurden, war unerwartet hoch und reichte, im Median über alle Permutationen, von 13,6 % (R-gada) bis 20,3 % (APT) (Tabelle 3-4). Bei Beschränkung der Permutation der Eltern auf die ursprüngliche Population des Trios waren die Raten der Pseudo-Validierung höher als bei einer Permutation über beide Populationen. Der Median der softwarespezifischen Validierungsraten reichte von 16,8 % (R-gada) bis 29,2 % (APT) bei Proben der europäischen Population und von 14,3 % (VEGA) bis 24,1 % (APT) bei Proben aus der afrikanischen Population (Tabelle 3-4). Die Raten der Pseudo-Validierung geben dabei Aufschluss über die Häufigkeit des Auftretens von CNVs an populationsspezifischen Loci und dienen nicht zur Beurteilung der relativen Genauigkeit der CNV-Vorhersagen der einzelnen Softwares.

Tabelle 3-4: Zufällige Validierung bei permutierter Zuordnung der Eltern

Software	CEU und YRI Proben [%]	Ausschließlich CEU Proben [%]	Ausschließlich YRI Proben [%]
APT	20,3 (20,1-20,7)	29,2 (29,0-30,6)	24,1 (23,9-24,5)
GLAD	16,5 (15,4-16,8)	24,2 (23,9-24,6)	19,9 (18,8-19,9)
PennCNV	18,6 (18,3-18,9)	23,5 (23,1-24,0)	16,3 (16,0-17,2)
QuantiSNP	16,8 (16,4-17,1)	18,5 (18,1-18,9)	15,7 (14,9-16,2)
R-gada	13,6 (13,0-13,9)	16,8 (16,4-18,0)	16,6 (15,9-17,0)
VEGA	13,7 (13,1-13,9)	22,5 (22,3-23,2)	14,3 (14,2-14,9)
Algorithmus			
HMM	18,1 (17,7-18,4)	22,9 (22,7-24,0)	17,6 (16,9-18,2)
Segmentation	14,8 (14,2-15,5)	21,8 (21,4-22,9)	16,5 (15,8-16,9)

3.1.4 Paarweise Konkordanz der Software

Ein weit verbreiteter, wenngleich heuristischer Ansatz zur Steigerung der Validität der CNV-Vorhersagen ist es, zwei oder mehr Softwares simultan zu verwenden und nur Varianten zu untersuchen, die von unterschiedlichen Algorithmen vorhergesagt wurden. Die paarweise Konkordanz zweier Softwares in der CNV-Vorhersage wurde für jeden der 60 Nachkommen der HapMap Trios bestimmt, indem der Anteil der von einer Software als CNVs vorhergesagten Sequenzen ermittelt wurde (Vorhersage), der auch durch eine andere Software als CNVs vorhergesagt wurde (Verifikation). Diese Konkordanz ist per Definition nicht immer symmetrisch. Berechnet wurde (i) der Median über den Anteil der konkordanten Sequenz pro CNV je Probe und (ii) der Median über die kumulierte konkordante CNV-Sequenz je Probe. Eine CNV wurde als durch

einen anderen Algorithmus verifiziert betrachtet, wenn der Anteil der konkordanten Sequenz mehr als 90 % betrug.

Tabelle 3-5: Paarweise Konkordanz der Softwares in der Vorhersage von CNVs

	Verifikation						
Vorhersage	**APT**	**GLAD**	**PennCNV**	**QuantiSNP**	**R-gada**	**VEGA**	**3+**
Median der prozentualen konkordanten Sequenz pro CNV							
APT	-	58,3	48,7	62,1	49,6	71,3	54,0
GLAD	63,7	-	52,9	60,6	58,3	62,6	59,1
PennCNV	61,0	58,1	-	73,2	50,2	53,2	66,6
QuantiSNP	40,4	40,0	40,1	-	40,9	41,4	41,0
R-gada	40,4	46,1	35,6	50,5	-	52,1	42,6
VEGA	61,9	52,3	37,4	52,5	63,3	-	50,8
3+	49,9	51,6	44,7	67,5	46,9	55,9	-
Median der konkordanten kumulierten CNV Sequenz							
APT	-	65,8	71,7	46,6	51,4	62,9	57,3
GLAD	66,7	-	60,5	38,1	43,9	55,4	54,9
PennCNV	70,7	58,6	-	50,2	46,5	55,4	58,3
QuantiSNP	31,0	29,1	33,9	-	16,6	23,3	34,2
R-gada	42,8	36,5	34,6	24,7	-	47,2	36,0
VEGA	66,7	64,1	55,7	39,6	70,6	-	53,8
3+	47,7	50,6	52,0	47,0	41,9	50,0	-

Konkordanz wurde definiert als übereinstimmende Vorhersage der CN-Klasse an einem gegebenen Locus. Ausschließlich validierte CNV-Vorhersagen wurden berücksichtigt. **3+:** Konsensus-Regionen die durch mindestens drei Softwares als CNV vorhergesagt wurden.

Der Anteil übereinstimmender Vorhersagen zeigte große Unterschiede zwischen den Paaren der Softwares (Tabelle 3-5). Der im Mittel größte Anteil konkordanter Sequenz pro CNV von 73,2 % wurde für Vorhersagen von PennCNV beobachtet, die durch QuantiSNP verifiziert wurden. Umgekehrt war die Konkordanz bei Vorhersagen durch QuantiSNP und Verifizierung durch PennCNV deutlich geringer (40,1 %). Generell zeige GLAD den höchsten Gad an Verifikation durch andere Software (52,9-63,7 %) und QuantiSNP den geringsten (40,0-41,4 %). Eine softwarespezifische CNV-Vorhersage galt als durch eine andere Software verifiziert, wenn der Anteil der Sequenz, die ebenfalls durch die andere Software vorhergesagt wurde, mindestens 90 % betrug. Der Anteil verifizierter CNVs je Probe zeigte einen ähnlichen Trend und reichte von 32,5 % (Vorhersage durch R-gada, Verifikation durch PennCNV) bis zu 68,3 % (PennCNV, QuantiSNP; Appendix Tabelle 7). Die Verwendung anderer Grenzwerte zur Verifizierung führte zu ähnlichen Ergebnissen. Der Median über die kumulierte konkordante CNV-Sequenz je Probe reichte von 14,2 % (Vorhersage durch QuantiSNP, Verifikation durch R-gada) bis zu 67,8 % (VEGA, R-gada).

Keine Software zeigte einen generellen Trend zu einem gleichmäßig hohen Grad an Verifikation (Tabelle 3-5). Aufgrund der großen Unterschiede in der Anzahl der vorhergesagten CNVs eignen sich die paarweisen Vergleiche nur schlecht als Indikatoren für die Sensitivität oder Validität der Vorhersagen. Aus diesem Grund wurden alle CNVs die durch mindestens drei Softwares konkordant vorhergesagt wurden in einem Datensatz (Konsensus-Regionen) zusammengefasst, der mit den softwarespezifischen Vorhersagen verglichen wurde. Die Verifikation der Konsensus-Regionen durch die einzelnen Softwares diente als Indikator für die Sensitivität des jeweiligen Algorithmus. Der Anteil der im Mittel konkordanten Sequenz je CNV reichte von 44,7 % für PennCNV bis 67,5 % für QuantiSNP. Diese Kenngröße berücksichtigt insbesondere die Anzahl der Konsensus-Regionen. Der Verifikation der softwarespezifischen Vorhersagen durch die Konsensus-Regionen diente als Indikator für die Validität der Vorhersagen. Der Anteil der im Mittel konkordanten kumulierten CNV-Sequenz berücksichtigt insbesondere die Länge der Vorhersagen und reichte von 40,9 % für R-gada bis 52,0 % für PennCNV.

Für die bisher beschriebenen Ergebnisse wurde parallel zu der vorliegenden Arbeit das Manuskript „Family-based benchmarking of copy number variation detection software“ angefertigt und zur Veröffentlichung eingereicht.

3.1.5 Auswahl einer Software für die Pipeline zur CNV-Analyse

Der Anteil von CNV-Vorhersagen die durch andere Software verifiziert werden konnte war für PennCNV am höchsten, was ein Indikator für ein hohes Maß an Validität der Vorhersagen ist. Auch die Validierungsraten waren für PennCNV am höchsten, was diese Software zur besten Wahl für eine initiale Analyse im Rahmen einer Pipeline macht. Die im Vergleich zu den anderen Softwares geringe Anzahl an CNV-Vorhersagen je Probe lässt jedoch vermuten, dass die durch PennCNV vorhergesagten CNVs nur einen Teil der CNVs ausmachen, die auf Grundlage von SNP-Chip-Daten identifiziert werden könnten.

3.2 Assoziationsanalyse

Ziel der Assoziationsanalyse basierend auf den *in-silico* Vorhersagen war eine Priorisierung der ermittelten CNV-Region in Bezug auf eine mögliche Assoziation mit Sarkoidose. Für diesen Ansatz ist eine hohe Validität der CNV-Vorhersagen wünschenswert, da die anschließend notwendige experimentelle Validierung mit hohen Kosten und großem zeitlichen Aufwand verbunden ist. Daher wurde für die Vorhersage der CNVs, basierend auf den oben gezeigten Ergebnissen, die Software PennCNV verwendet.

3.2.1 CNV-Vorhersagen durch PennCNV

Insgesamt wurden CNV Vorhersagen für 1654 Individuen, davon 564 Sarkoidose-Patienten (Fälle) und 1090 Kontrollpersonen (Kontrollen), getroffen. Dabei wurden die Daten der Fall-Kohorte (SARC) und zweier Kontroll-Kohorten, (CON1 (465 Proben) und CON2 (625 Proben)) verwendet. Die Anzahl der durch PennCNV vorhergesagten CNVs je Probe lag im Median über alle Proben bei 50 und reichte in den einzelnen Kohorten im Median von 48 (CON1) bis 52 (SARC) (Tabelle 3-6 und Abbildung 3-6). In den Kontroll-Kohorten wurden tendenziell weniger CNVs je Probe vorhergesagt als in der Fall-Kohorte (Median: 49 vs 52, p-Wert des Wilcoxon-Rangsummentests: $p = 5{,}2 \times 10^{-7}$). Auch bei einer getrennten Betrachtung der CNV-Vorhersagen unterschiedlicher Länge bleiben die Unterschiede signifikant, aber gering (Abbildung 3-6). Die separate Auswertung der Kohorten ergab, dass in den Kontroll-Kohorten vereinzelt Proben mit einer sehr hohen Zahl an vorhergesagten CNVs auftraten (bis zu 891 in CON1).

Tabelle 3-6: Probenspezifische Eigenschaften der vorhergesagten CNVs

	Anzahl	**Länge [kb]**	**Kumulative Länge**	**Anzahl der Marker**	**Abstand der Marker**	**DDR**
Stichprobe	50 (40-60)	21,7 (15,8 - 26,8)	3,5 (2,6 - 4,7)	27,0 (24,0 - 31,5)	0,19 (0,15 - 0,23)	2,0 (1,5 - 2,7)
Kohorten						
SARC	52 (44 - 60)	19,9 (14,3 - 24,7)	3,4 (2,6 - 4,5)	26,0 (23,0 - 29,0)	0,19 (0,15 - 0,23)	1,9 (1,5 - 2,5)
CON1	48 (38 - 57)	21,7 (16,4 - 27,1)	3,5 (2,5 - 4,5)	28,0 (24,5 - 32,0)	0,19 (0,14 - 0,22)	1,9 (1,4 - 2,5)
CON2	50 (37 - 62)	22,9 (17,5 - 28,6)	3,7 (2,7 - 5,0)	28,0 (24,0 - 33,0)	0,20 (0,15 - 0,24)	2,1 (1,5 - 2,9)

Angegeben ist der Median und in Klammern der Interquartilsabstand. **DDR:** Verhältnis von Deletionen zu Duplikationen.

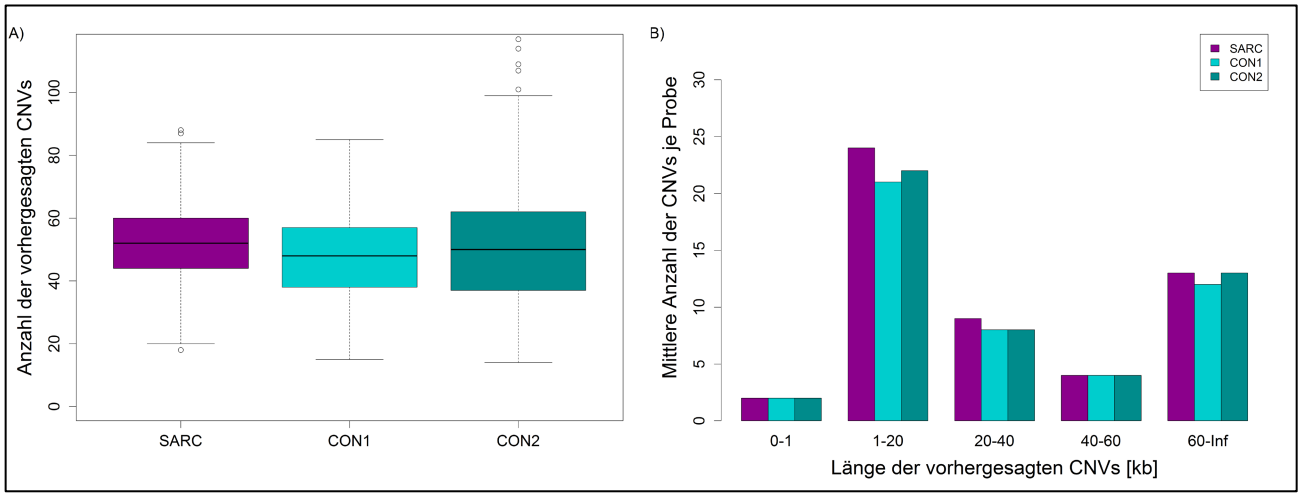


Abbildung 3-6: Anzahl der vorhergesagten CNVs

A: Anzahl der vorhergesagten CNVs. **B:** Probenspezifischer Median der Anzahl der vorhergesagten CNVs aufgeteilt nach Länge.

Der probenspezifische Median der CNV-Länge besaß in allen Kohorten eine leicht linksschiefe Verteilung, deren Mediane stets kleiner waren als das jeweilige arithmetische Mittel (Abbildung 3-7). In beiden Kontroll-Kohorten wurden tendenziell längere CNVs vorhergesagt als in der Fall-Kohorte (19,9 kb vs 21,7 kb (SARC vs CON1), $p = 2{,}7 \times 10^{-4}$; 19,9 kb vs 22,9 kb (SARC vs CON2), $p = 1{,}0 \times 10^{-6}$).

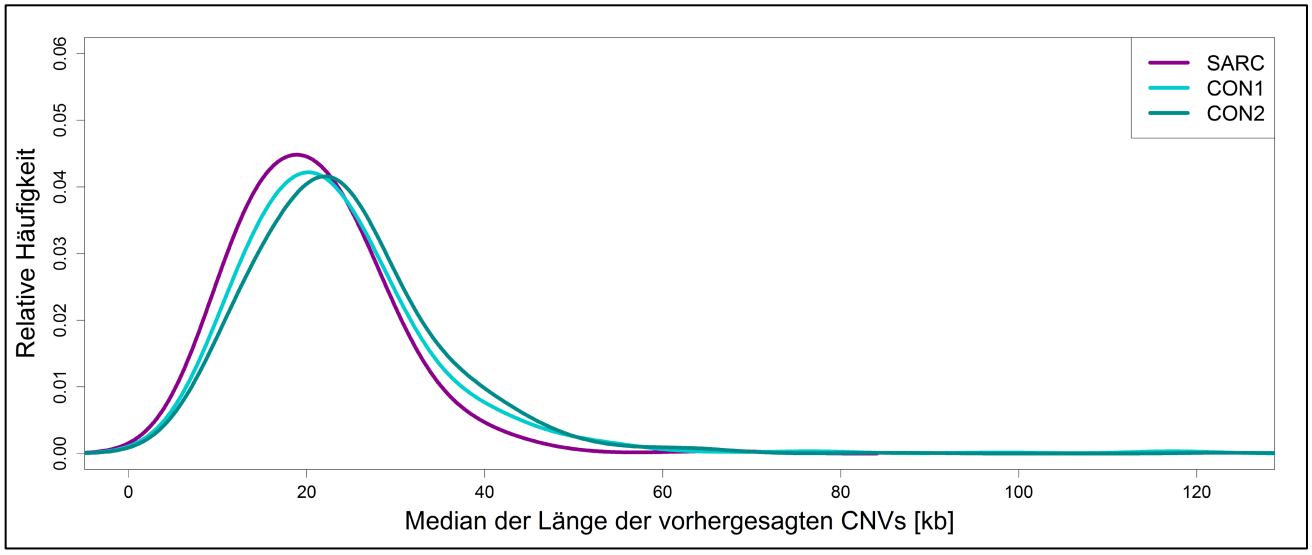


Abbildung 3-7: Verteilung der Länge der vorhergesagten CNVs

Kerndichteschätzung der Verteilung des probenspezifischen Medians der CNV-Länge. Verwendet wurden Gaußkerne mit einer Bandbreite von 5 kb.

In allen Kohorten wurden doppelt so viele Deletionen wie Duplikationen vorhergesagt (Tabelle 3-6).

3.2.2 CNV-Regionen

Insgesamt wurden in der verwendeten Studienpopulation 194 häufige CNV-Regionen (rel. Häufigkeit > 5 %) detektiert. In den regionsspezifischen 2×5-χ^2-Tests zeigten 65 Regionen mindestens an einem Marker einen lokal signifikanten p-Wert ($p_{2x5} < 0.05$) und wurden nach dem minimalen p-Wert der Region priorisiert (Appendix Tabelle 8). Im Rahmen der Auswahl der CNVs zur technischen Validierung wurden alle lokal signifikanten Regionen visuell überprüft und die entsprechend den in Abschnitt 2.4.3 beschriebenen Kriterien manuell gefiltert. Insgesamt 17 Regionen enthielten Segmente > 1 kb (CNVs) mit gleichmäßig hohen relativen Häufigkeiten der vorhergesagten CN-Genotypen bei ausreichender Abdeckung durch Marker (Tabelle 3-7; Appendix Abbildung 1 bis Appendix Abbildung 17). Diese Regionen wurden als Kandidaten deklariert und die relevanten Segmente weitergehend untersucht. Bei der Identifikation der wahrscheinlichsten CNV-Start- und Endpositionen durch Auswertung der relativen Häufigkeiten ergaben sich für Region #12 zwei Segmente von Interesse (Region #12a und #12b). Die relative Häufigkeit der vorhergesagten

CNVs reichte im Median über alle Marker des jeweiligen Segments von 11,7% in Region #34 bis 77,3% in Region #1. Auch die Länge der relevanten Segmente variierte Stark und reichte von 1,6 kb (Region #12a) bis 1,1 Mb (Region #10).

Tabelle 3-7: Liste der 17 CNV-Kandidaten-Regionen

#	Assay	Chr.	Position	CNV-Start	CNV-Ende	CNV-Länge	p_{min}^{2x5}	$\tilde{p}_{2x3}$	$\tilde{p}_{2x2}^{dup}$	$\tilde{p}_{2x2}^{del}$	h_n
1	Hs03239014_cn	4	10,228,766	10,223,404	10,235,256	11,852	3.45E-60	4.59E-05	4.26E-03	2.25E-05	77.3%
3	Hs03930310_cn	16	78,377,478	78,372,428	78,379,663	7,235	8.32E-43	2.10E-04	1.03E-01	5.27E-05	34.3%
10	Hs03228413_cn	3	162,556,898	162,512,645	162,623,885	111,240	1.65E-09	5.43E-05	5.53E-02	1.68E-03	44.0%
12a	Hs04367713_cn	9	44,259,399	44,241,805	44,273,229	1,617	1.69E-08	5.17E-03	2.25E-03	1.19E-01	23.3%
12b	Hs04367763_cn	9	44,825,092	44,727,847	44,840,906	109,668	1.69E-08	5.81E-02	3.70E-02	5.85E-01	38.7%
13	Hs04309608_cn	6	32,505,546	32,474,929	32,570,353	86,354	2.52E-08	4.57E-06	8.55E-04	1.30E-05	28.5%
14	Hs04389783_cn	11	55,405,125	55,374,020	55,442,306	68,286	4.35E-08	4.05E-06	5.90E-04	6.35E-03	28.5%
16	Hs03240113_cn	4	34,814,645	34,786,488	34,824,713	37,683	1.19E-07	1.49E-03	3.13E-04	1.88E-01	48.0%
17	Hs03260288_cn	6	79,000,327	78,969,053	79,032,580	63,527	3.31E-07	4.17E-05	5.70E-01	5.66E-05	61.0%
20	Hs03301139_cn	13	38,078,493	38,072,024	38,084,745	4,231	1.91E-06	6.11E-07	7.32E-01	8.99E-08	15.8%
21	Hs04989338_cn	7	133,791,351	133,785,165	133,799,185	14,020	2.46E-06	2.12E-02	8.28E-01	7.52E-03	24.4%
24	Hs03873630_cn	14	74,002,801	74,001,111	74,020,967	19,856	1.00E-05	2.11E-06	7.17E-04	3.20E-05	14.3%
27	Hs03114102_cn	7	142,481,429	142,476,707	142,486,033	9,326	3.68E-05	1.36E-05	1.00E-03	2.91E-05	40.2%
28	NA	3	NA	129,762,847	129,808,799	45,952	7.52E-05	8.26E-06	NA	NA	39,9%
34	Hs04028005_cn	19	35,856,313	35,855,341	35,861,836	6,495	9.76E-04	9.11E-03	6.76E-03	1.33E-01	11.7%
37	Hs03571282_cn	5	180,409,283	180,378,698	180,418,091	39,393	2.08E-03	6.74E-03	6.68E-01	1.57E-03	18.4%
40	Hs03315651_cn	17	39,425,034	39,423,002	39,430,519	7,517	3.64E-03	1.09E-01	8.16E-01	4.13E-02	50.9%
43	Hs03286888_cn	10	67,312,170	67,307,911	67,314,434	6,523	6.05E-03	1.46E-01	2.42E-01	1.47E-01	12.8%

#: Region, priorisiert über den minimalen p-Wert. $\boldsymbol{p_{min}^{2x5}}$**:** Minimaler p-Wert der Region. $\boldsymbol{\tilde{p}_{2x3}}$**:** Median der p-Werte aller durchgeführten 2x3-χ^2-Tests eines Segments. $\boldsymbol{h_n}$**:** relative Häufigkeit von detektierten CNVs in der gesamten Stichprobe. **NA:** kein CNV-TaqMan®-Assay verfügbar.

In Vorbereitung für die technische Validierung wurde für jede Region ein vorgefertigter CNV-TaqMan®-Assay ausgewählt, dessen Zielsequenz innerhalb der relevanten Region liegt. Für Region #28 stand kein Assay zur Verfügung, sodass die Region von der Validierung ausgeschlossen wurde. Im Rahmen der Hypothesenbildung für die Assoziationsanalyse wurde für jede Region anhand der Verteilung der relativen Häufigkeiten der CN-Klassen geprüft, ob eine Assoziation aufgrund der Unterschiede in der Häufigkeit der Deletionen, der Duplikationen oder beider CN-Klassen zwischen Fällen und Kontrollen zu erwarten war.

3.3 Experimentelle Validierung

Die durch die CNV-Analyse-Pipeline identifizierten Kandidaten-Regionen basieren auf CNV-Vorhersagen und müssen experimentell validiert werden. Dazu werden zunächst die Vorhersagen

einer kleinen Teilstichprobe für jede Region mittels TaqMan®-Genotypisierung auf ihre Validität untersucht (technische Validierung). Regionen deren Validierungsraten über dem zu erwarteten Grenzwert von 60 % lagen wurden in einer unabhängigen Stichprobe erneut untersucht. Dabei wurden die Ergebnisse der TaqMan®-Genotypisierung in einer Assoziationsanalyse ausgewertet (statistische Validierung).

3.3.1 Technische Validierung

Für die technische Validierung der 17 ausgewählten Segmente wurden zwei PCR-Master-Platten (SAR01 und SAR02) erstellt, die in Kombination mit neun bzw. acht CNV-TaqMan®-Assays verwendet wurden (Tabelle 3-8). Jede Platte enthielt die Proben von 92 Individuen, die so ausgewählt wurden, dass die relativen Häufigkeiten aller vorhergesagten CN-Genotypen aus der *in-silico* Analyse für die jeweiligen CNVs ähnlich waren. Das CNV-TaqMan®-Experiment wurde wie in Abschnitt 2.5.2 beschrieben durchgeführt.

Tabelle 3-8: CNV-TaqMan® Assays und Eichproben für die technischen Validierung

SAR01						
#	**Assay**	**Chr.**	**Position**	**Verwendete CN-Genotypen der Eichproben**		
				283504	**283516**	**283526**
16	Hs03240113_cn	4	34,814,645	2	2	2
43	Hs03286888_cn	10	67,312,170	2	2	NA
20	Hs03301139_cn	13	38,078,493	NA	NA	2
34	Hs04028005_cn	19	35,856,313	2	2	2
13	Hs04309608_cn	6	32,505,546	2	2	amb_exp
12a	Hs04367713_cn	9	44,259,399	NA	amb_exp	amb_exp
12b	Hs04367763_cn	9	44,825,092	amb_exp	2	2
21	Hs04989338_cn	7	133,791,351	2	NA	NA
24	Hs03873630_cn	14	74,002,801	amb_exp	2	2
SAR02						
#	**Assay**	**Chr.**	**Position**	**Verwendete CN-Genotypen der Eichproben**		
				283516	**283504**	**283526**
27	Hs03114102_cn	7	142,481,429	2	NA	2
10	Hs03228413_cn	3	162,556,898	amb_exp	2	2
1	Hs03239014_cn	4	10,228,766	2	NA	2
17	Hs03260288_cn	6	79,000,327	2	2	2
40	Hs03315651_cn	17	39,425,034	2	amb_exp	2
37	Hs03571282_cn	5	180,409,283	NA	NA	NA
3	Hs03930310_cn	16	78,377,478	amb_exp	2	2
14	Hs04389783_cn	11	55,405,125	2	2	amb_exp

NA: Die in der Auswertung des TaqMan®-Experiments ermittelte Kopienzahl war $cn = 0$. **amb_exp**: Der CN-Genotyp aus der Literatur stimmt nicht mit dem Wert der TaqMan®-Genotypisierung überein.

Zusätzlich zu den 92 Patienten- bzw. Kontrollproben befanden sich auf beiden Master-Platten jeweils drei HapMap-Proben. Diese Proben wurden als Eich-Proben verwendet, da sie bereits in mehreren Studien auf CNVs untersucht worden sind und validierte CN-Genotypdaten zur Verfügung stehen

(The 1000 Genomes Project Consortium, 2012; Conrad et al., 2010; The International HapMap Consortium, 2005). Ergab die Analyse durch die Software CopyCaller®, dass eine HapMap Probe an dem Locus des CNV-Assays eine Deletion besaß so konnte sie nicht als Eich-Probe verwendet werden (Tabelle 3-8). Gleiches galt für HapMap Proben, deren Auswertung mit CopyCaller® einen von der Literatur abweichenden CN-Genotyp ergab.

Wie in Abschnitt 2.6.2 beschrieben, gibt die Standardabweichung der ΔC_T-Werte $SD(\Delta C_T)$, als Kenngröße für die plattenspezifische Variabilität der Messungen, Auskunft darüber wie gut die Software CopyCaller® in der Auswertung der PCR-Ergebnisse die CN-Genotypen voneinander unterscheiden kann. Die Standardabweichung der ΔC_T-Werte war für fast alle Assays auf beiden Platten hoch ($0{,}09 < SD(\Delta C_T) \leq 0{,}2$) bis sehr hoch ($0{,}2 < SD(\Delta C_T)$) (Tabelle 3-9). Entsprechend der plattenspezifischen Variabilität der ΔC_T-Werte, wurde bei der Auswertung für jeden Assay die genotypspezifischen Grenzwerte für die Konfidenz der berechneten CN-Genotypen verwendet (Abschnitt 2.6.4).

Tabelle 3-9: Variabilität der ΔC_T-Werte währen der technischen Validierung

Assay	SAR01	Assay	SAR02
Hs03240113_cn	hoch	**Hs03114102_cn**	sehr hoch
Hs03286888_cn	hoch	**Hs03228413_cn**	hoch
Hs03301139_cn	hoch	**Hs03239014_cn**	normal
Hs04028005_cn	normal	**Hs03260288_cn**	normal
Hs04309608_cn	sehr hoch	**Hs03315651_cn**	normal
Hs04367713_cn	hoch	**Hs03571282_cn**	normal
Hs04367763_cn	sehr hoch	**Hs03930310_cn**	normal
Hs04989338_cn	hoch	**Hs04389783_cn**	normal
Hs03873630_cn	hoch	-	-

gering: $SD(\Delta C_T) \leq 0{,}05$. **normal**: $0{,}05 < SD(\Delta C_T) \leq 0{,}09$. **hoch**: $0{,}09 < SD(\Delta C_T) \leq 0{,}2$. **sehr hoch**: $0{,}2 < SD(\Delta C_T)$.

Im Vergleich mit den *in-silico* Vorhersagen wurden nach den in Abschnitt 2.7.1 beschriebenen Kriterien acht Regionen (#1, #12a, #12b, #13, #14, #21, #37 und #40) von weiteren Analysen ausgeschlossen, da sie die zu geringe Validierungsraten (< 60%) für die CN-Klassen aufwiesen. Die verbleibenden neun Regionen zeigten Validierungsraten von 61,5% - 100% (Tabelle 3-10). Die Regionen #10 und #17 wurden ebenfalls von weiteren Untersuchungen ausgeschlossen, da keine Vorhersagen in den für die vermutete Assoziation relevanten CN-Klassen validiert werden konnten. Die verbleibenden sieben Regionen (#3, #16, #20, #24, #27, #34 und #43) wurden als Kandidaten-Regionen für die statistische Validierung ausgewählt. Als Kenngröße für die beobachtete

Effektstärke wurde der Kontingenz-Koeffizient *Cramérs V* berechnet. Dieser reichte bei der Betrachtung von 2×3-Kontingenztabellen (Test der CN-Klassen) von 0,065 (Region #34) bis 0,225 (Region #3). Für die statistische Validierung standen 1104 Proben (552 Fälle, 552 Kontrollen) zur Verfügung. Die Teststärke im 2×3-χ^2-Test ist mit 0,709 – 0,999 ausreichend hoch für die Regionen #3, #16, #20, #24 und #2. Für Region #34 und Region #43 werden mit dieser Stichprobengröße geringere Teststärken erreicht (0,471 und 0,151).

Tabelle 3-10: Ergebnisse der technischen Validierung

#	QC [%] (n_{QC}/n_s)	CN-Klasse [%] (n_v/n_{QCs})	Deletionen [%] (n_v/n_{QCs})	Duplikationen [%] (n_v/n_{QCs})	Normal [%] (n_v/n_{QCs})	V_{2x5}	$(1-\beta)_{2x5}$	V_{2x3}	$(1-\beta)_{2x3}$
1	98,9% (91/92)	59,3% (54/91)	100,0% (38/38)	16,7% (6/36)	58,8% (10/17)	-	-	-	-
3	94,6% (87/92)	93,1% (81/87)	100,0% (25/25)	100,0% (11/11)	88,2% (45/51)	0,241	0,999	0,225	0,999
10	39,1% (36/92)	77,8% (28/36)	52,9% (9/17)	0,00% (0/ 0)	100,0% (19/19)	-	-	-	-
12a	78,0% (71/91)	39,4% (28/71)	100,0% (14/14)	0,0% (0/10)	29,8% (14/47)	-	-	-	-
12b	60,4% (55/91)	10,9% (6/55)	100,0% (6/ 6)	0,0% (0/ 0)	00,0% (0/49)	-	-	-	-
13	89,0% (81/91)	19,8% (16/81)	100,0% (16/16)	0,0% (0/33)	00,0% (0/32)	-	-	-	-
14	98,9% (91/92)	61,5% (56/91)	100,0% (26/26)	0,0% (0/16)	61,2% (30/49)	-	-	-	-
16	31,9% (29/91)	69,0% (20/29)	60,9% (14/23)	0,0% (0/ 0)	100,0% (6/ 6)	0,145	0,983	0,090	0,771
17	100,0% (92/92)	71,7% (66/92)	100,0% (30/30)	0,0% (0/21)	87,8% (36/41)	-	-	-	-
20	41,8% (38/91)	100,0% (38/38)	100,0% (24/24)	0,0% (0/ 0)	100,0% (14/14)	0,086	0,612	0,084	0,709
21	61,5% (56/91)	48,2% (27/56)	100,0% (3/ 3)	100,0% (7/ 7)	37,0% (17/46)	-	-	-	-
24	69,2% (63/91)	92,1% (58/63)	100,0% (31/31)	0,0% (0/ 0)	84,4% (27/32)	0,126	0,936	0,122	0,960
27	54,3% (50/92)	100,0% (50/50)	100,0% (16/16)	0,0% (0/ 0)	100,0% (34/34)	0,123	0,923	0,115	0,940
34	97,8% (89/91)	79,8% (71/89)	100,0% (38/38)	66,7% (2/ 3)	64,6% (31/48)	0,084	0,588	0,065	0,471
37	100,0% (92/92)	48,9% (45/92)	100,0% (31/31)	00,0% (0/20)	34,1% (14/41)	-	-	-	-
40	97,8% (90/92)	28,9% (26/90)	100,0% (19/19)	08,0% (2/25)	10,9% (5/46)	-	-	-	-
43	100,0% (91/91)	72,5% (66/91)	100,0% (33/33)	00,0% (0/ 4)	61,1% (33/54)	0,055	0,266	0,033	0,151

n_s: Anzahl aller durch einen Assay analysierten Proben. **n_{QC}:** Anzahl der Proben nach Qualitätskontrolle. **n_v:** Anzahl der validierten Proben. **V:** Cramérs V, **$(1-\beta)$:** Teststärke bei einer Stichprobengröße von 1104 Proben

3.3.2 Statistische Validierung

Die sieben erfolgreich technisch validierten Kandidaten-Regionen wurden in einem unabhängigen Datensatz bestehend aus den Proben von insgesamt 1104 Individuen, davon 552 Sarkoidose-Patienten (Fälle) und 552 Kontrollpersonen (Kontrollen), mittels CNV-TaqMan®-Assays untersucht (*in-vitro* Analyse). Das Experiment wurde wie in Abschnitt 2.5.2 beschrieben durchgeführt.

Das Layout aller zwölf verwendeten PCR-Master-Platten (SAR52 – SAR63) enthielt jeweils dieselben drei HapMap-Proben. Diese Proben wurden nach den bereits in der technischen Validierung (Abschnitt 3.3.1) verwendeten Kriterien als Eich-Proben verwendet. Wieder wurden HapMap-Proben, für die eine Auswertung durch die Software CopyCaller® an dem Locus des CNV-

Assays eine Deletion ergab oder deren CN-Genotypen von denen der Literatur abwichen, nicht als Eich-Probe verwendet (Tabelle 3-11).

Tabelle 3-11: CNV-TaqMan®-Assays und Eichproben in der statistischen Validierung

#	Assay	Chr.	Position	283516	283504	283526
3	Hs03930310_cn	16	78,377,478	amb_exp	2	2
16	Hs03240113_cn	4	34,814,645	amb_exp	2	2
20	Hs03301139_cn	13	38,078,493	2	2	amb_exp
24	Hs03873630_cn	14	74,002,801	2	2	2
27	Hs03114102_cn	7	142,481,429	2	NA	2
34	Hs04028005_cn	19	35,856,313	2	amb_exp	2
43	Hs03286888_cn	10	67,312,170	amb_exp	2	2

NA: Die in der Auswertung des TaqMan®-Experiments der ermittelten Kopienzahl war $cn = 0$. **amb_exp**: Der CN-Genotyp aus der Literatur stimmt nicht mit dem Wert der TaqMan®-Genotypisierung überein.

Die Variabilität der der ΔC_T-Werte war überwiegend als normal einzustufen, nur für Assay Hs03114102_cn ergabt sich eine hohe Variabilität der Daten (Tabelle 3-12). Die Assays Hs03240113_cn und Hs03930310_cn zeigten im Vergleich zu den restlichen Assays eine relativ geringe Variabilität.

Tabelle 3-12: Variabilität der ΔC_T-Werte währen der statistischen Validierung

Assay	SAR57	SAR58	SAR59	SAR60	SAR53	SAR56	SAR54	SAR55	SAR63	SAR61	SAR62	SAR52
Hs03930310_cn	normal	gering	normal	normal	normal	normal	hoch	normal	gering	normal	normal	gering
Hs03240113_cn	normal	normal	normal	gering	normal	normal	normal	normal	normal	normal	normal	normal
Hs03301139_cn	hoch	normal	normal	hoch	hoch	hoch	hoch	hoch	normal	normal	normal	hoch
Hs03873630_cn	hoch	hoch	hoch	normal	hoch	hoch	normal	hoch	hoch	hoch	hoch	normal
Hs03114102_cn	sehr hoch	sehr hoch	sehr hoch	hoch	hoch	sehr hoch	sehr hoch	hoch	hoch	hoch	hoch	sehr hoch
Hs04028005_cn	normal	normal	normal	hoch	normal	hoch	hoch	normal	normal	hoch	hoch	hoch
Hs03286888_cn	normal	hoch	normal	hoch	normal	normal	normal	hoch	normal	normal	hoch	hoch

gering: $var(\Delta C_T) \leq 0{,}05$. **normal**: $0{,}05 < var(\Delta C_T) \leq 0{,}09$. **hoch**: $0{,}09 < var(\Delta C_T) \leq 0{,}2$. **sehr hoch**: $0{,}2 < var(\Delta C_T)$.

Nach Überprüfung der Kriterien zur Qualitätskontrolle (QC-Kriterien) ergaben sich unter den analysierten Regionen große Differenzen in dem Anteil der Proben, welche die QC-Kriterien erfüllten (Tabelle 3-13). Die Region #27 wurde aus weiteren Analysen ausgeschlossen, da der Anteil der Proben die den QC-Kriterien genügten lediglich 27,2 % betrug. Für alle anderen Regionen reichte der Anteil QC-gefilterter Proben von 68,3 % bis 95,1 % (Tabelle 3-13). Für jeden Assay

ergaben sich Unterschiede in der Rate zwischen Fällen und Kontrollen, es konnte jedoch kein allgemeiner Trend festgestellt werden.

Tabelle 3-13: Datensatz nach Qualitätskontrolle

#	Assay	Stichprobe [%] (1104)	Fälle [%] (552)	Kontrollen [%] (552)	Eich-Proben	Plattenspezifische Variabilität
3	Hs03930310_cn	87,2% (963)	91,1% (503)	83,3% (460)	283504, 283526	gering: 3, normal: 8, hoch: 1, sehr hoch: 0
16	Hs03240113_cn	82,6% (912)	88,4% (488)	76,8% (424)	283504, 283526	gering: 1, normal: 11, hoch: 0, sehr hoch: 0
20	Hs03301139_cn	95,8% (1058)	98,2% (542)	93,5% (516)	283516, 283504	gering: 0, normal: 5, hoch: 7, sehr hoch: 0
24	Hs03873630_cn	65,6% (724)	65,6% (362)	65,6% (362)	283516, 283504, 283526	gering: 0, normal: 3, hoch: 9, sehr hoch: 0
27	Hs03114102_cn	24,5% (270)	31,7% (175)	17,2% (95)	283516, 283526	gering: 0, normal: 0, hoch: 6, sehr hoch: 6
43	Hs03286888_cn	91,3% (1008)	86,2% (476)	96,4% (532)	283504, 283526	gering: 0, normal: 7, hoch: 5, sehr hoch: 0
34	Hs04028005_cn	92,4% (1020)	87,3% (482)	97,5% (538)	283516, 283526	gering: 0, normal: 6, hoch: 6, sehr hoch: 0

Angegeben ist der Anteil der Proben, welche die Kriterien zur Qualitätskontrolle erfüllten sowie die Anzahl der Proben in Klammern. **Plattenspezifische Variabilität:** Anzahl der Platten mit dem jeweiligen Grad an Variabilität.

Die relative Häufigkeit detektierter CNVs ($cn \neq 2$) reichte von 0,36 bis 0,56 und war damit in der *in-vitro* Analyse bei allen Assays höher als in der auf CNV-Vorhersagen in einer unabhängigen Stichprobe basierenden *in-silico* Analyse (Tabelle 3-14).

Tabelle 3-14: Relative Häufigkeit der CNVs (*in-vitro* Analyse & *in-silico* Vorhersagen)

#	Assay	Analyse	h_n	h_{case}	h_{con}	h_n^{del}	h_{case}^{del}	h_{con}^{del}	h_n^{dup}	h_{case}^{dup}	h_{con}^{dup}
3	Hs03930310_cn	*in-silico*	0,27	0,18	0,09	0,40	0,24	0,15	0,21	0,15	0,06
		in-vitro	0,42	0,46	0,38	0,28	0,27	0,28	0,14	0,19	0,10
16	Hs03240113_cn	*in-silico*	0,48	0,15	0,33	0,52	0,19	0,33	0,45	0,13	0,33
		in-vitro	0,56	0,59	0,53	0,08	0,08	0,09	0,48	0,51	0,44
20	Hs03301139_cn	*in-silico*	0,08	0,06	0,02	0,10	0,06	0,05	0,07	0,06	0,01
		in-vitro	0,55	0,57	0,53	0,55	0,57	0,53	0,00	0,00	0,00
24	Hs03873630_cn	*in-silico*	0,13	0,10	0,03	0,19	0,13	0,06	0,10	0,08	0,02
		in-vitro	0,36	0,40	0,31	0,35	0,40	0,30	0,00	0,00	0,01
27	Hs03114102_cn	*in-silico*	0,40	0,18	0,23	0,48	0,20	0,28	0,36	0,16	0,20
		in-vitro	0,51	0,43	0,66	0,51	0,43	0,66	0,00	0,00	0,00
34	Hs04028005_cn	*in-silico*	0,12	0,11	0,01	0,09	0,09	0,00	0,14	0,13	0,01
		in-vitro	0,45	0,45	0,45	0,45	0,45	0,45	0,00	0,00	0,01
43	Hs03286888_cn	*in-silico*	0,07	0,06	0,01	0,07	0,07	0,01	0,07	0,06	0,01
		in-vitro	0,50	0,54	0,48	0,50	0,54	0,48	0,00	0,00	0,00

Angegeben sind die relativen Häufigkeiten der CNVs nach der Vorhersage durch PennCNV (*in-silico*) und nach der Validierung durch TaqMan®-Genotypisierung in einer unabhängigen Kohorte (*in-vitro*). $\boldsymbol{h}$: relative Häufigkeit. $\boldsymbol{n}$: Gesamte Stichprobe. $\boldsymbol{case}$: Fälle. $\boldsymbol{con}$: Kontrollen. $\boldsymbol{del}$: Deletionen. $\boldsymbol{dup}$: Duplikationen

Zur statistischen Validierung der in den Kandidaten-Regionen beobachteten lokalen Assoziation mit Sarkoidose wurde für jeden Assay ein 2×3-χ^2-Test durchgeführt. Die daraus resultierenden lokalen p-Werte wurden anschließend durch ein Permutationsverfahren nach Westfall und Young (1993) mit 100.000 Wiederholungen für multiples Testen korrigiert. Zwei Regionen zeigten einen auch nach der Korrektur für multiples Testen signifikante p-Werte von $p_{corr} = 1{,}01 \times 10^{-3}$ (Region #3) bzw. $p_{corr} = 2{,}26 \times 10^{-2}$ (Region #24).

Tabelle 3-15: Ergebnisse des Assoziationstests

#	Assay	p_{2x3}^{corr}	p_{2x3}	V_{2x3}	p_{2x2}^{dup}	OR_{2x2}^{dup}	p_{2x2}^{del}	OR_{2x2}^{del}
3	Hs03930310_cn	$1{,}01 \times 10^{-3}$	$3{,}20 \times 10^{-4}$	0,13	$9{,}14 \times 10^{-5}$	2,204	$4{,}59 \times 10^{-1}$	-
16	Hs03240113_cn	$2{,}79 \times 10^{-1}$	$1{,}28 \times 10^{-1}$	0,07	-	-	-	-
20	Hs03301139_cn	$4{,}21 \times 10^{-1}$	$3{,}43 \times 10^{-1}$	0,04	-	-	-	-
24	Hs03873630_cn	$2{,}26 \times 10^{-2}$	$9{,}91 \times 10^{-3}$	0,11	$3{,}91 \times 10^{-1}$	-	$1{,}03 \times 10^{-2}$	1,513
34	Hs04028005_cn	$7{,}64 \times 10^{-1}$	$6{,}70 \times 10^{-1}$	0,03	-	-	-	-
43	Hs03286888_cn	$2{,}34 \times 10^{-1}$	$1{,}63 \times 10^{-1}$	0,06	-	-	-	-

Angegeben ist der lokale p-Wert ($\boldsymbol{p_{2x3}}$), der p-Wert nach Korrektur für Multiples Testen ($\boldsymbol{p_{2x3}^{corr}}$), der Kontingenz-Koeffizient Cramérs V ($\boldsymbol{V_{2x3}}$) als Kenngröße für die Effektstärke und die p-Werte nachgelagerter lokaler Assoziationstest beschränkt auf Deletionen ($\boldsymbol{p_{2x3}^{del}}$) und Duplikationen ($\boldsymbol{p_{2x3}^{dup}}$) sowie das jeweilige Quotenverhältnis ($\boldsymbol{OR}$).

In einer nachfolgenden Assoziationsanalyse wurden die einzelnen CN-Klassen der bereits signifikant assoziierten Regionen separat untersucht. Ein Vergleich von Proben nur einer CN-Klasse mit CN-neutralen Proben ($cn = 2$) kann genaueren Aufschluss darüber geben, ob die Assoziation auf Unterschiede in der relativen Häufigkeit von Deletionen oder Duplikationen zurückzuführen ist. Die durchgeführten 2×2-χ^2-Tests ergaben lokal signifikante p-Werte bei der Analyse von Duplikationen in Region #3 ($p_{2x2}^{dup} = 9{,}14 \times 10^{-5}$) und der Analyse von Deletionen in Region #24 ($p_{2x2}^{del} = 1{,}03 \times 10^{-2}$) (Tabelle 3-15). Dabei trat die Duplikation in Region #3 häufiger in den Fällen als in den Kontrollen auf ($h_{cas}^{dup} = 0{,}19$ vs $h_{con}^{dup} = 0{,}10$) und auch die Deletion in Region #24 wurde häufiger in den Fällen als in den Kontrollen beobachtet ($h_{cas}^{dup} = 0{,}40$ vs $h_{con}^{dup} = 0{,}30$). Der Vergleich der Duplikationen mit den CN-neutralen Proben in Region #3 ergab ein Quotenverhältnis von 2,204, was für Träger der Duplikation ein mehr als doppelt so hohes Risiko bedeutet an Sarkoidose zu erkranken. Bei dem Vergleich der Deletionen mit den CN-neutralen Proben in Region #24 ergab sich ein Quotenverhältnis von 1,513.

3.3.3 Signifikant assoziierte CNV-Regionen

Eine Auswertung der signifikant mit Sarkoidose assoziierten Regionen im *UCSC Genome Browser* (Kent et al., 2002) ergab, dass beide codierende Sequenzen betreffen. Die Region #3 liegt auf Chromosom 16 (chr16q23.1 – q23.2) und erstreckt sich über 7,2 kb, von Position 78.372.428 bis 78.379.663. Damit liegt sie in der intronischen Region des Gens für die *WW domain containing oxidoreductase* (WWOX), zwischen dem fünften und sechsten Exon (Abbildung 3-8).

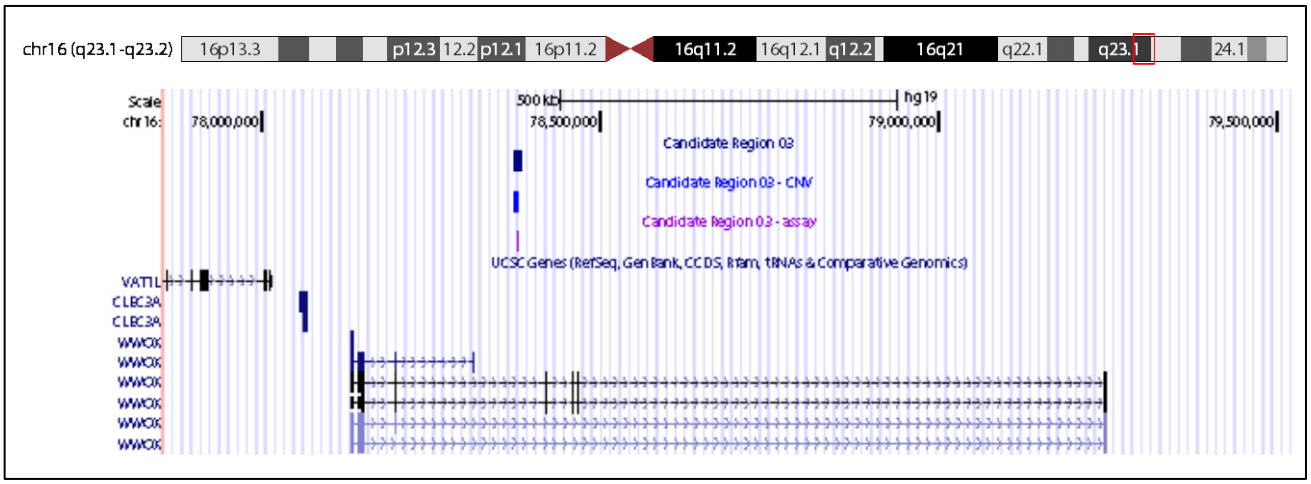


Abbildung 3-8: Genomische Position der CNV-Region #3

Aus der Website des *UCSC Genome Browser* (http://genome.ucsc.edu).

Region #24 liegt auf Chromosom 14 (chr14q24.3) und betrifft eine 19,8 kb lange Sequenz an der Position 74.001.111 bis 74.020.967. Dieser Locus codiert für zwei Genprodukte: Der Plusstrang codiert für das *HEAT repeat-containing protein 4* (HEATR4), auf dem Minusstrang befindet sich die codierende Sequenz des Gens für die *acyl-CoA thioesterase 1* (ACOT1) (Abbildung 3-9).

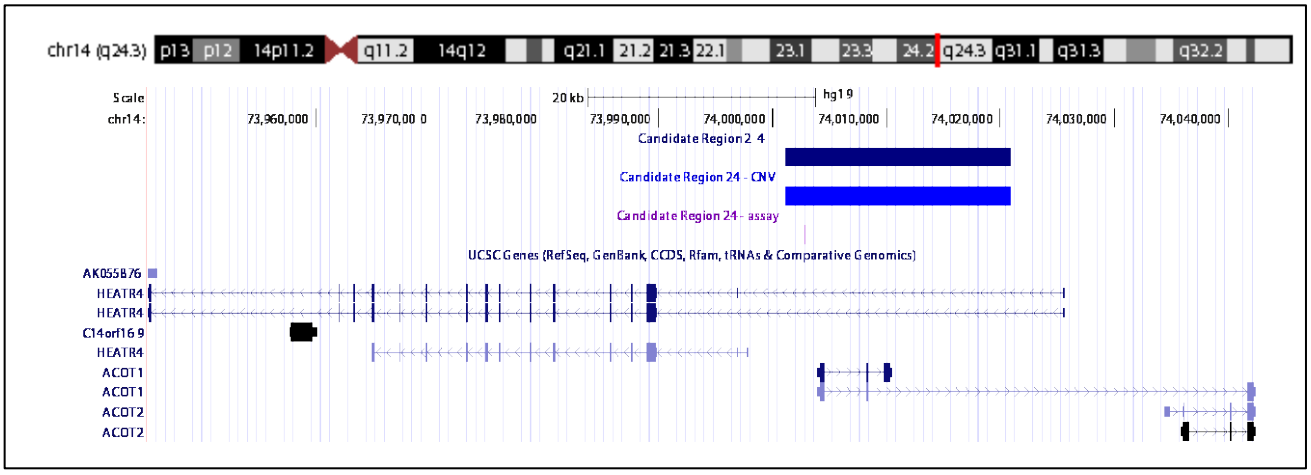


Abbildung 3-9: Genomische Position der CNV-Region #24

Aus der Website des *UCSC Genome Browser* (http://genome.ucsc.edu).

3.3 Experimentelle Validierung

Durch eine systematische genomweite *in-silico* CNV-Analyse wurden mehrere Kandidaten-Regionen identifiziert werden, die möglicherweise CNVs beinhalten, die mit dem Phänotyp der Sarkoidose assoziiert sind. Die in zwei dieser Regionen lokalisierten CNVs konnten durch CNV-TaqMan®-Genotypisierung in einer unabhängigen Stichprobe validiert werden und stellen somit neue potentielle Risikofaktoren für Sarkoidose dar.

4 Diskussion

Komplexe Erkrankungen, bei denen mehrere Gene sowie Umweltfaktoren für die Ausprägung eines Merkmals verantwortlich sind, werden intensiv untersucht, um die zugrunde liegenden molekularen Mechanismen besser zu verstehen. Für viele häufige Erkrankungen, wie Typ I Diabetes oder koronare Herzkrankheiten, konnte in der Vergangenheit nachgewiesen werden, dass genetische Varianten die Suszeptibilität erhöhen. Im Mittelpunkt der Untersuchung genetischer Variationen, die als Risikofaktoren einen Einfluss auf die Ätiologie komplexer Erkrankungen haben, standen bislang vor allem SNPs als genetische Marker. Im Rahmen von Chip-basierten GWAS konnten viele dieser Polymorphismen als Risikofaktoren detektiert werden. Dies trug entscheidend zur Identifizierung von Genen bei, die bei der Entstehung von häufigen und komplexen Krankheiten eine Rolle spielen (Visscher et al., 2012). Dennoch bleibt ein großer Teil der Heritabilität dieser Krankheiten bislang unbekannt (Zuk et al., 2012; Eichler et al., 2010; Manolio et al., 2009; Maher, 2008). Dies legt die Vermutung nahe, dass auch andere Formen von genetischer Variation eine wichtige Rolle in der Entwicklung dieser Krankheiten spielen könnten. Genetische Kopienzahl-Variationen wurden in der Vergangenheit immer wieder mit der Ätiologie komplexer Krankheiten in Verbindung gebracht (Almal & Padh, 2012) und können möglicherweise einen Teil der unbekannten Heritabilität erklären. Bereits für mehrere komplexe Erkrankungen, wie z.B. Morbus Crohn, wurden CNVs als Risikofaktoren identifiziert werden (The Wellcome Trust Case Control Consortium, 2010).

Die Sarkoidose ist eine komplexe Erkrankung mit unbekannter Ätiologie. Die Identifikation genetischer Risikofaktoren für diese Krankheit ist daher Gegenstand intensiver Forschung. Durch Kandidaten-Gen-Studien konnten bislang neben verschiedenen HLA-Haplotypen auch die Gene, die für den *Tumornekrosefaktor-α* (TNF-α) und den *Interleukin-23-Rezeptor* (IL23R) codieren, als Suszeptibilitätsloci identifiziert werden (zusammengefasst in Fischer et al. 2014). Durch genomweite Studien wurden zusätzlich *BTNL2* (Valentonyte, Hampe, Huse, et al., 2005), *ANXA11* (Hofmann et al., 2008), *RAB23* (Hofmann et al., 2011), *C10ORF67* (Franke et al., 2008), *OS9* (Hofmann et al., 2013), *CCDC88B* (Fischer et al., 2012) und *NOTCH4* (Adrianto et al., 2012) als Risiko-Loci mit der Erkrankung in Verbindung gebracht. Jedoch untersuchte keine der bisher durchgeführten Studien dabei CNVs in einem genomweiten Rahmen auf einen Zusammenhang mit Sarkoidose.

Die Analyse von CNVs ist methodisch äußerst anspruchsvoll, da eine direkte Beobachtung der Variationen meist nicht möglich ist. Eine Vielzahl von Softwares zur Identifikation von CNVs basierend auf SNP-Chip-Daten steht zur Verfügung. Diese verwenden unterschiedliche Ansätze und

Algorithmen, um über die Menge der auf den SNP-Chips hybridisierten DNA-Fragmente Rückschlüsse auf den zugrunde liegenden CN-Genotyp zu ziehen. Wie in dieser Arbeit gezeigt, unterscheiden sich die Resultate dieser Softwares stark voneinander, sowohl in der Anzahl der vorhergesagten CNVs als auch in den Charakteristika der detektierten Variationen. Die Auswahl der Software für eine CNV-Analyse erfordert unter diesen Umständen einen sorgfältigen Vergleich der zur Verfügung stehenden Methoden.

In der vorliegenden Studie wurde eine vergleichende Analyse der Leistungsfähigkeit und Zuverlässigkeit von sechs Softwares zur CNV-Vorhersage durchgeführt (Benchmark). Dabei zeigte sich, dass die Vorhersagen der Software PennCNV die größte Validität besaßen. Basierend auf den Ergebnissen des Benchmarks wurde ein Verfahren zur systematischen genomweiten *in-silico* CNV-Analyse (CNV-Analyse-Pipeline) entwickelt und auf einen vorhandenen Datensatz von Sarkoidose-Patienten und Kontrollen angewendet. Die Analyse identifizierte siebzehn Kandidaten-Regionen, die möglicherweise CNVs enthalten, die in Zusammenhang mit dem Phänotyp der Sarkoidose stehen. Nach der experimentellen Validierung in einer unabhängigen Stichprobe konnte für zwei der CNV-Segmente in der *WWOX*- bzw. der *HEATR4/ACOT1*-Genregion eine Assoziation mit Sarkoidose nachgewiesen werden. Diese Ergebnisse belegen, dass CNVs auch bei der Sarkoidose zur genetischen Suszeptibilität beitragen.

4.1 Software zur Vorhersage von CNVs

In vielen Fällen stehen SNP-Chip-Daten durch bereits durchgeführte GWAS zur Verfügung, und eine zusätzliche CNV-Analyse verursacht keine weiteren Kosten. Es ist daher sinnvoll, CNV-Analysen anhand von SNP-Chip-Daten durchzuführen. In dieser Studie wurden sechs verschiedene Softwares zur CNV-Vorhersage basierend auf SNP-Chip-Daten verglichen und die Vorhersagen durch familienbasierte Analysen validiert. Verwendet wurden dafür die Daten von 60 HapMap-Trios, die zur Hälfte afrikanischer (YRI) und zur anderen Hälfte europäischer Abstammung (CEU) waren. Des Weiteren wurde ein möglicher Einfluss von Populationen auf die CNV-Vorhersage untersucht, indem die Auswertung nicht nur für alle Trios zusammen (CEU und YRI) durchgeführt wurde, sondern zusätzlich beschränkt auf Trios der jeweiligen Population. Abschließend wurde die Konkordanz der Softwares in Bezug auf ihre Vorhersagen untersucht.

4.1.1 Software-Benchmark

Im Benchmark wurde primär die Validität der durch die Softwares getroffenen Vorhersagen untersucht. Sie stellt die wichtigste Kenngröße zur Beurteilung der Leistungsfähigkeit der Softwares

zur Analyse von CNVs mittels SNP-Chip-Daten dar. Darüber hinaus wurde über die Charakteristika der Vorhersagen, wie Länge oder Markerdichte, festgehalten, welche Bandbreite an struktureller Variation mit der jeweiligen Methode analysiert werden kann. Über die paarweise Konkordanz wurde festgestellt, ob zwei Softwares komplementäre Vorhersagen treffen oder dazu neigen, dieselben CNVs vorherzusagen. Ein Vergleich der Vorhersagen mit Konsensus-Regionen, die CNVs beinhalten, welche von der Mehrzahl der Softwares übereinstimmend vorhergesagt wurden, gab Aufschluss über die Sensitivität der jeweiligen Methode. Die Validität der Vorhersagen ließ sich damit nur bedingt überprüfen, da eine ausbleibende Verifikation durch Konsensus-Regionen auch ein Indikator für die fehlende Sensitivität anderer Softwares sein kann.

Die sechs Softwares wiesen große Unterschiede in der Anzahl und Länge der vorhergesagten CNVs auf. Die in dieser Studie betrachteten HMM-Softwares tendierten dazu, weniger CNVs als die Segmentierungssoftwares vorherzusagen, die allerdings konsistent valider waren. Die auf HMM basierende Software PennCNV sagte die geringste Anzahl an CNVs je Probe voraus und zeigte die höchste Validität von allen betrachteten Softwares. Auch bei den validierten CNVs zeigten sich große Diskrepanzen unter den Softwares. Die Unterschiede zwischen den Softwares ähnelten denen, die auch bei den nicht validierten CNVs festgestellt wurden. Diese Studie bestätigt also bereits publizierte Befunde über eine generell geringe Validität und eine hohe Falsch-Positiv-Rate bei Vorhersagen auf Grundlage von SNP-Chip-Daten (Pinto et al., 2011; Eckel-Passow et al., 2011; Winchester et al., 2009). Die Verifikation der validierten CNVs durch externe Datensätze der *Database of Genomic Variation* (DGV) ergab, ähnlich wie in einer Studie durch Pinto et al. (2011), gemischte Ergebnisse. Der Anteil an CNV-Vorhersagen, die durch Array-Daten verifiziert wurden, lag deutlich über dem Anteil der durch einen NGS-Datensatz verifizierten Vorhersagen. Dies lässt sich durch die Ähnlichkeit der Methoden zur Detektion von CNVs und das Fehlen von Duplikationen in dem NGS-Datensatz erklären. Trotz der beschriebenen Herausforderungen bei der CNV-Analyse basierend auf SNP-Chip Daten birgt dieser Ansatz noch immer ein großes Potential, bislang unbekannte Assoziationen von strukturellen Variationen mit diversen Phänotypen zu identifizieren. Das Ausmaß der bereits verfügbaren Daten aus vorangegangenen SNP-basierten GWAS und der vergleichsweise geringe Aufwand einer *in-silico* Analyse sind gute Voraussetzungen für umfangreiche Studien, die ein neues Licht auf Regionen werfen könnten, die bisher nicht als Kandidaten für Risikofaktoren in Betracht gezogen worden sind.

Der Einfluss von Populationen auf die Vorhersage der CNVs war gering. Trotzdem konnte beobachtet werden, dass alle Softwares bis auf PennCNV weniger und kürzere CNVs in den

europäischen als in den afrikanischen Proben vorhersagten. Dieser Befund kann durch den generell höheren Grad an genetischer Heterogenität in afrikanischen Populationen als nicht-afrikanischen Populationen erklärt werden (zusammengefasst in Campbell & Tishkoff, 2008). Darüber hinaus lässt sich daraus ableiten, dass PennCNV kaum in der Lage ist, die für afrikanische Populationen spezifischen Variationen zu detektieren.

Trotz dieser Unterschiede waren die Validierungsraten in beiden Populationen nahezu gleich. Die Wahrscheinlichkeit einer „Pseudo-Validierung“ war unerwartet hoch und wurde weiter verstärkt, wenn die Permutation der Zuordnung der Eltern zu den Nachkommen auf die Proben einer Population beschränkt wurde. Diese Beobachtung deutet auf eine mögliche populationsspezifische Verteilung der CNVs hin, was im Einklang zu anderen Studien steht (Chen et al., 2011; Kato et al., 2010; Li et al., 2009; Redon et al., 2006). Dieser Umstand könnte in zukünftigen Ansätzen der CNV-Analyse genutzt werden, indem eine *a-priori* Wahrscheinlichkeit in das Modell aufgenommen wird. An dieser Stelle soll darauf hingewiesen werden, dass die CNV-Vorhersage basierend auf SNP-Chip-Daten stark von der Verteilung der Sonden des jeweiligen Chips abhängen. Dadurch kann es zu populationsspezifischen Unterschieden in den Vorhersagen kommen, wenn der gleiche Chip in unterschiedlichen Populationen verwendet wird. Der in dieser Studie verwendete SNP-Chip *Affymetrix Human SNP Array 6.0* wurde so designt, dass die in den Proben der ersten und zweiten Phase des HapMap Projekts häufigen SNPs optimal abgedeckt sind (McCarroll, Kuruvilla, et al., 2008). Dennoch ist nicht auszuschließen, dass die Proben auf dem Chip die Varianten bestimmter Populationen tendenziell besser abdecken.

Die paarweise Konkordanz der Softwares war häufig hoch, jedoch nicht immer symmetrisch. Die Verifikation der softwarespezifischen Vorhersagen durch andere Softwares und einen Datensatz bestehend aus CNVs, die von mindestens drei Softwares vorhergesagt wurden (Konsensus-Regionen), diente als Indikator für die Validität der Vorhersagen. Bei der Verifizierung der Vorhersagen durch die Konsensus-Regionen waren der Anteil der im Mittel konkordanten Sequenz je CNV sowie der Anteil der im Mittel konkordanten kumulierten CNV-Sequenz für PennCNV am größten. Dies macht PennCNV in Kombination mit der höchsten beobachteten Validierungsrate zur besten Wahl für eine initiale CNV-Analyse, wenn eine hohe Validität gefordert ist. QuantiSNP zeigte die zweithöchste Validierungsrate und nur eine geringe Konkordanz zu den Vorhersagen anderer Software. Damit eignet sich QuantiSNP ideal für ergänzende Analysen, um in Kombination mit PennCNV die Sensitivität zu erhöhen. Ist die Validität der Vorhersagen nicht das wichtigste Kriterium bei der Auswahl der Software für eine CNV-Analyse, so stellt QuantiSNP auch für sich

genommen eine sehr gute Wahl dar. Die Software besitzt eine hohe Validität und die im Mittel höchste Anzahl vorhergesagter CNVs von allen HMM-Softwares. Konkordante Vorhersagen von mehreren Softwares erhöhen die Sicherheit, können aber aufgrund der generell niedrigen Validität eine experimentelle Validierung nicht ersetzen. Darüber hinaus bedeutet eine geringe oder fehlende Konkordanz nicht zwingend, dass die Vorhersagen von schlechter Qualität sind. Aufgrund der unterschiedlichen Algorithmen kann es vorkommen, dass die Softwares verschiedene CNVs vorhersagen.

Die hohe Falsch-Positiv-Rate, die hohe Wahrscheinlichkeit für eine „Pseudo-Validierung“ und ein unzureichendes Maß an Konkordanz bei der CNV-Vorhersage durch unterschiedliche Software, wie es in dieser Studie beobachtet wurde, lassen zwei Implikationen für die CNV-Analyse zu. Erstens ist es unwahrscheinlich, alle CNVs in einem Genom durch Analysen von SNP-Chip-Daten zu finden, auch wenn unterschiedliche Softwares ergänzend verwendet werden. Dies schließt eine rein quantitative Untersuchung der CNV-Vorhersagen aus. Zweitens benötigen CNV-Vorhersagen eine unabhängige und experimentelle Validierung, auch wenn die Vorhersage durch mehrere Algorithmen verifiziert wurde, wie bereits durch Winchester et al., (2009) vorgeschlagen. Auf quantitativer PCR basierende Ansätze, wie die in dieser Studie verwendete TaqMan®-CNV-Genotypisierung, stellen dafür geeignete und etablierte Methoden dar (Alkan et al., 2011). Darüber hinaus ist die Verteilung der Sonden auf dem Chip entscheidend für die Möglichkeit, CNVs vorherzusagen. Aus diesem Grund sind SNP-Chips mit einer hohen Markerdichte in jedem Fall für eine CNV-Analyse zu bevorzugen. Eine exakte Bestimmung der Start- und Endposition eines CNVs ist generell durch die hohe paarweise Distanz der Marker äußerst schwierig, in schlecht abgedeckten Regionen ist unter Umständen eine Vorhersage gar nicht möglich. CNV-Vorhersagen sollten daher eher als Vorhersagen von Region aufgefasst werden, welche die wahren CNVs überschneiden oder sie sogar ganz umfassen. Bei Assoziationsstudien muss daher genau darauf geachtet werden, dass die Positionen, an denen eine mögliche Assoziation untersucht wird, möglichst in einem Segment liegt, das gut durch Marker abgedeckt ist und in dem eine möglichst konstante relative Häufigkeit von CNV-Vorhersagen zu beobachten ist.

4.1.2 CNV-Analyse Pipeline

Das in dieser Studie entwickelte Verfahren zur systematischen genomweiten *in-silico* CNV-Analyse (CNV-Analyse-Pipeline) dient der Identifikation von Kandidaten-Regionen, die CNVs beinhalten, welche möglicherweise mit dem Phänotyp der Sarkoidose assoziiert sind. Dazu musste eine Assoziationsanalyse auf Basis von CNV-Vorhersagen durchgeführt werden. Da in jedem Fall eine

experimentelle Validierung der Ergebnisse erforderlich ist, war eine möglichst hohe Validität der CNV-Vorhersagen wünschenswert. Die Software PennCNV zeigte im Benchmark die höchste Validität und wurde aus diesem Grund für die Pipeline verwendet. Die geringe Anzahl an Vorhersagen durch PennCNV, im Vergleich mit den anderen Softwares im Benchmark, lässt vermuten, dass die Vorhersagen in der Pipeline zwar die höchstmöglichste Validität besitzen, aber nicht alle CNVs ausmachen, die auf Grundlage der Daten zu finden gewesen wären. Um die Rate an falsch-positiven Vorhersagen so gering wie möglich zu halten, wurde darauf verzichtet, weitere Softwares komplementär zu verwenden. Die für einen solchen Ansatz notwendige geringe Konkordanz der Softwares würde bedeuten, dass viele der falsch-positiven Vorhersagen einer weiteren Software zusätzlich in die Analyse aufgenommen würden.

Die Pipeline wurde entwickelt mit dem Ziel, CNV-Regionen zu identifizierten, in denen für mindestens 5 % der Fälle oder Kontrollen eine CNV vorhergesagt wurde. Dies erlaubt eine gezielte Analyse häufiger CNVs in Zusammenhang mit einem Phänotyp. Für die Analyse von seltenen CNVs wäre die Bestimmung von CNV-Regionen ohne eine minimale relative Häufigkeit an CNV-Vorhersagen nötig. Die geringe Validität der Vorhersagen würde in einer solchen Analyse in Kombination mit der niedrigen relativen Häufigkeit der vorhergesagten CNVs jedoch zu einer erhöhten Rate an falsch-positiven Kandidaten-Regionen führen. Die geringe Auflösung der Daten ermöglicht keine genaue Definition der Grenzpositionen. Dadurch ließ sich nicht bestimmen, ob die probenspezifischen Unterschiede der Start- und Endpositionen aufgrund ungenauer Vorhersagen oder echter molekularbiologischer Gegebenheiten bestehen. Die weitere Analyse der detektierten häufigen CNV-Regionen beschränkte sich auf Segmente mit einer gleichmäßigen relativen Häufigkeit der vorhergesagten CN-Genotypen. Diese Segmente stellen die beste Annäherung an die zugrundeliegenden CNVs dar, aber beinhalteten häufig nicht die Sequenzen am Anfang und Ende der CNV-Regionen. In diesen Bereichen wurden aufgrund der Unterschiede in den probenspezifischen Start- und Endpositionen häufig große Differenzen in den relativen Häufigkeiten der vorhergesagten CN-Genotypen beobachtet. Für eine verlässliche Aussage über die Verteilung der relativen Häufigkeiten entlang des Segments ist eine hohe Markerdichte erforderlich. Die geringe Auflösung der SNP-Chip-Daten führt dazu, dass trotz einer Länge der CNV-Regionen von mindestens 1 kb die Markerdichte der relevanten Segmente manuell überprüft werden musste. Diese Limitationen der in der Pipeline angewandten Methoden führen dazu, dass wahrscheinlich viele potentielle CNV-Regionen nicht detektiert wurden oder aufgrund einer zu geringen Markerdichte nicht weiter analysiert wurden. Die gemachten Vorhersagen liefern jedoch eine erste Abschätzung

darüber, welche Regionen als Kandidaten für eine mögliche Assoziation in Frage kommen könnten. Die schlussendlich als Kandidaten ausgewählten CNV-Regionen zeichneten sich demnach durch eine ausreichende Markerdichte in Kombination mit nur wenigen Bruchpunkten aus. Unter diesen Umständen ist es dann möglich, die gesamte Region mit nur einem Assay zu testen. Die detektierten lokalen Assoziationen mit Sarkoidose wurden anschließend in einem unabhängigen Datensatz experimentell überprüft.

4.2 Identifizierung von CNVs in Zusammenhang mit Sarkoidose

Genomweite Assoziationsanalysen von CNVs sind noch immer selten, haben jedoch schon mehrfach neue Risikofaktoren für diverse Erkrankungen identifiziert (zusammengefasst in Almal & Padh (2012)). Diese Studien basieren bislang auf Daten, die durch Mikro-Array-Technologie gewonnen wurden. In Zusammenhang mit Sarkoidose wurde eine solche CNV-GWAS noch nicht durchgeführt.

4.2.1 CNV-Vorhersagen

Für die Vorhersage von CNVs in einem Sarkoidose-Datensatz wurde die Software PennCNV im Rahmen der in dieser Studie entwickelten Pipeline verwendet. Der Datensatz beinhaltete die SNP-Chip-Rohdaten des *Affymetrix Human SNP Array 6.0* von 564 Sarkoidose-Patienten (Fälle) und 1090 gesunden Kontroll-Personen (Kontrollen). Für den Sarkoidose-GWAS-Datensatz wurden keine großen Unterschiede in Bezug auf die Anzahl oder Länge der CNV-Vorhersagen zwischen den Fall- und Kontrollkohorten beobachtet. Die geringen Unterschiede in der Anzahl der CNVs beruhen möglicherweise auf der geringen Sensitivität der genutzten Methode. Eine quantitative Beurteilung wäre aufgrund der in dem Benchmark beobachteten geringen Validität der Vorhersagen rein spekulativ.

Eine exakte Bestimmung der Grenzpositionen ist wegen der geringen Auflösung der SNP-Chip-Daten nicht möglich. Aufgrund der probenspezifischen Unterschiede der Start- und Endpositionen der CNV-Vorhersagen war für die stichprobenübergreifende Definition von CNV-Regionen daher eine manuelle Definition der Grenzpositionen des zu untersuchenden Segments nötig. Dies lässt nur eine grobe Abschätzung der CNV-Grenzen zu. Um diese verlässlich zu bestimmen, wäre eine detaillierte Analyse der CNV-Grenzpositionen nötig. Diese Analysen sind sehr umfangreich und beruhen auf speziell dafür entwickelter Software, wie zuerst durch (Korbel, Urban, Grubert, et al., 2007) beschrieben. Die Auswahl der Kandidaten-Regionen erfolgte manuell, da eine gleichmäßige relative Häufigkeit der vorhergesagten CN-Genotypen und eine hohe Abdeckung des relevanten Segments durch Marker visuell überprüft werden musste. Bei einem Großteil der CNV-Regionen mit

lokal signifikanten p-Werten stellte sich heraus, dass die Assoziationen aus Bereichen stammten, die nicht zu dem relevanten Segment gehörten, was auf die ungenaue Bestimmung der Grenzpositionen zurückzuführen ist.

Trotz der Unterschiede in den Start- und Endpositionen der probenspezifischen CNV-Vorhersagen waren die relativen Häufigkeiten innerhalb der relevanten Segmente der Kandidaten-Regionen meist weitestgehend konsistent. Die Vorhersagen eignen sich gut für eine CNV-Analyse im Rahmen einer Fall-Kontroll-Studie, wobei berücksichtigt werden muss, dass die Grenzpositionen nicht eindeutig bestimmt sind und nicht alle häufigen CNVs detektiert werden können. Bei sorgfältiger Auswahl der Positionierung des CNV-TaqMan®-Assays innerhalb der Region ist es möglich, durch eine punktuelle Assoziationsanalyse die gesamte Region auszuwerten.

4.2.2 Technische Validierung und Assoziationsanalyse

Die Strategie, CNVs und deren Assoziation mit Sarkoidose zunächst aus SNP-Chip-Daten vorherzusagen und anschließende technische und statistische Validierungen der Befunde durchzuführen, resultierte in dieser Studie in der Identifikation von zwei Regionen, in denen sich Fälle und Kontrollen signifikant in der Kopienzahl unterscheiden. Dabei wurde die Hypothese getestet, dass die relativen Häufigkeiten der CN-Klassen in den Fällen und Kontrollen gleich verteilt sind. Die ermittelten CN-Genotypen wurden zu CN-Klassen zusammengefasst. Verwendet wurde ein χ^2-Test basierend auf einer 2×3-Kontingenztafel. Dieser Ansatz lässt nur Aussagen darüber zu, ob der Verlust oder Zugewinn von genetischem Material (CN-Klassen) an einem bestimmten Locus mit dem Phänotyp der Sarkoidose assoziiert ist. Auf eine Analyse der einzelnen CN-Genotypen wurde verzichtet, da auch die TaqMan®-Genotypisierung mit einiger Ungenauigkeit behaftet ist. Aus diesem Grund können keine Aussagen darüber gemacht werden, ob die Unterschiede in der relativen Häufigkeit zwischen Fällen und Kontrollen mit steigender oder abnehmender Kopienzahl zunehmen oder abnehmen.

Die zum Teil sehr geringen Validierungsraten während der technischen Validierung entsprachen im Mittel den Ergebnissen, die nach dem Benchmark erwartet wurden. Für eine Validierung mussten dabei die experimentellen Ergebnisse zu mindestens 60 % mit den Vorhersagen übereinstimmen. Interessanterweise wurden für manche CNV-Regionen alle Vorhersagen validiert, in anderen gar keine. Da sich bisher durchgeführte CNV-GWAS auf die Verwendung von *in-silico* Methoden zur Detektion von CNVs beschränken (Kawamura et al., 2011), ist ein Vergleich der in dieser Studie beobachteten Validierungsraten mit der Literatur nicht möglich. Insgesamt konnten sieben der siebzehn Kandidaten-Regionen technisch validiert werden.

Bei einer CNV-Assoziationsanalyse wird durch punktuelle statistische Tests der Zusammenhang einer ganzen Region mit einem Phänotyp überprüft. Die Information von aufeinander folgenden Markern ist dabei redundant, solange sich die relative Häufigkeit nicht ändert. Bei der Auswahl der für die Assoziationsanalyse relevanten Segmente wurde auf eine gleichmäßig hohe relative Häufigkeit von CNV-Vorhersagen geachtet. Dadurch reduzierte sich in der *in-silico* Analyse die Anzahl der pro Region durchzuführenden Tests stark. Da konstante relative Häufigkeiten der vorhergesagten CN-Genotypen bedeutet, dass in dem Segment keine probenspezifischen CNV-Vorhersagen beginnen oder enden, war dies auch für die experimentelle Validierung entscheidend. Durch die Auswahl von vorgefertigten CNV-TaqMan®-Assays, deren Primer innerhalb dieser CNV-Segmente binden, war es möglich, durch nur jeweils einen statistischen Test, eine Aussagen für das gesamte Segment zu treffen. Die großen Unterschiede in den Validierungsraten der Kandidaten-Regionen während der technischen Validierung legen nahe, dass sich die relativen Häufigkeiten auch in technisch validierten Kandidaten-Regionen deutlich von den Vorhersagen unterscheiden können. In zwei der technisch validierten Kandidaten-Regionen konnte die vorhergesagte Assoziation dennoch statistisch bestätigt werden.

4.2.3 CNV-Region im WWOX Locus

Das von einer CNV in Region #3 betroffene Gen für die *WW domain containing oxidoreductase* (*WWOX*) liegt in der chromosomalen Region chr16q23.1-23.2. Das Gen enthält 9 Exone, die für ein 414 Aminosäuren umfassendes Protein codieren (Ried et al., 2000; Bednarek et al., 2000). Die WW-Domäne, benannt nach zwei hoch konservierten Tryptophan-Bausteinen (W), ist mit ca. 38 Aminosäuren eines der kleinsten Proteinmodule und bindet an prolinreichen Motive (Macias et al., 1996). Die Region #3 liegt in einem intronischen Abschnitt des *WWOX*-Gens und die assoziierte CNV dupliziert einen regulatorischen Bereich, der eine Transkriptionsfaktor-Bindestelle für SETDB1 enthält. SETDB1 ist eine Methyltransferase, die unter anderem in durch T-Zell-Rezeptoren vermittelten Signalwegen eine Rolle spielt (Pereira et al., 2014). Durch die zusätzlichen Transkriptionsfaktor-Bindestellen besteht die Möglichkeit, dass Transkriptionsfaktoren vermehrt in der duplizierten Sequenz binden können, und somit der Effekt auf die Genexpression verstärkt wird. Die Duplikation muss jedoch keine Tandemduplikation sein und kann daher auch an anderer Stelle im Genom die Genregulation verändern. Des Weiteren können Transkriptionsfaktoren sowohl in *cis* als auch in *trans* wirken. Zusätzlich kann eine strukturelle Variation im Intron Auswirkungen auf posttranskriptionale Modifikation, wie z.B. dem Spleißen, haben. Diese molekularbiologischen Gegebenheiten machen Hypothesen zur funktionellen Auswirkung der Duplikation bzw. zu einem

potentiellen Kandidaten-Gen höchst spekulativ. Nur im Falle einer Tandem-Duplikation sowie einer Wirkung in *cis* wäre das WWOX Protein von diesem genetischen Risikofaktor betroffen.

Das *WWOX*-Gen liegt in der zweithäufigsten fragilen Region, FRA16D, einem bekannten Brennpunkt für strukturelle Variation (Yunis & Soreng, 1984; Glover et al., 1984). Fragile Regionen wurden als Loci identifiziert, an denen häufig Lücken und Brüche eines Chromosoms zu beobachten sind (Durkin & Glover, 2007). Diese Regionen sind besonders anfällig für Replikationsstress, wodurch es zum Abbruch der Replikationsgabel und zur Entstehung von strukturellen Variationen kommen kann (Carr & Lambert, 2013; Arlt et al., 2012). Entsprechend war bereits vor der Erstbeschreibung des Gens bekannt, dass Deletionen in der fragilen Region FRA16D gehäuft bei Brustkrebs (Chen et al., 1996; Aldaz et al., 1995; Tsuda et al., 1994; Sato et al., 1990) und anderen Tumoren wie Prostatakrebs auftreten (Carter et al., 1990). In jüngeren Studien wurden Veränderungen in der Struktur und der Expression des *WWOX*-Gens bei Lungenkrebs beobachtet (Donati et al., 2007; Yendamuri et al., 2003). Darüber hinaus wurde gezeigt, dass WWOX als Tumorsuppressor wirkt (Fabbri et al., 2005), was die Vermutung unterstützt, dass Variationen auf der genetischen oder transkriptionellen Ebene eine Rolle in der generellen Karzinogenese spielen könnten (Bednarek et al., 2001).

Im *WWOX*-Gen wurde kürzlich ein SNP identifiziert, der mit der sogenannten forcierten Vitalkapazität (engl. *forced vital capacity*, kurz FVC) assoziiert ist (Loth et al., 2014), einem Parameter, der Aufschluss über das Lungenvolumen gibt. Eine verminderte Lungenfunktion ist ein der Sarkoidose verwandter Phänotyp. Der identifizierte SNP rs1079572 liegt stromaufwärts der Region #3. In der Studie durch Loth et al. (2014) wurde keine CNV-Analyse durchgeführt, es ist aber denkbar, dass dieser Assoziation eine CNV zugrunde liegt und der SNP diesen markiert (engl. *tagging SNP*). In den SNP-basierten Sarkoidose-GWAS, auf deren Datensätze die vorliegende Studie basiert, wurden in dieser Genregion keine genomweit signifikant assoziierte Marker identifiziert, die *tagging* SNPs darstellen könnten (Hofmann et al., 2013; Fischer et al., 2012).

Das *WWOX*-Gen wird hauptsächlich in sekretorischen Epithelzellen der Fortpflanzungsorgane, endokrinischer und exokrinischer Organe sowie in Epithelzellen der ableitenden Harnwege exprimiert (Nunez et al., 2006). Die Autoren dieser Studie konnten darüber hinaus eine hohe Expression des WWOX Proteins in verschiedenen Zellen des Nervensystems, den Nervenzellen, Ependymzellen und Astrozyten belegen. Eine Expression in Fett-, Binde-, und lymphatischem Gewebe, myelinisierte Strukturen und Blutgefäßen konnte nicht beobachtet werden. Eine jüngere Studie belegte darüber hinaus die Expression in Lungengewebe (Loth et al., 2014). In einer weiteren

Studie konnte gezeigt werden, dass eine Expression des Gens Apoptose induzieren kann (Chang et al., 2003). WW Domänen spielen eine Rolle in der Signaltransduktion und vermitteln Wechselwirkungen zwischen Proteinen. Die Charakterisierung ihrer Interaktionen deckte eine Verbindung des *WWOX*-Gens mit verschiedenen Proteinkomplexen auf (Abu-Odeh et al., 2014). Viele der beteiligten Proteine sind in molekulare Prozesse einschließlich der Transkription, der RNA-Prozessierung, der Bildung von Tight Junctions und dem Metabolismus involviert.

4.2.4 CNV-Region im HEATR4/ACOT1 Locus

Die Assoziationsanalyse ergab für den Assay Hs03873630_cn in der Region #24 eine signifikante Assoziation mit Sarkoidose. An dem untersuchten Locus wurden fast ausschließlich Deletionen beobachtet, die häufiger in den Fällen als in den Kontrollen auftraten. Die Region #24 liegt auf Chromosom 14 (chr14q24.3) und betrifft die codierenden Sequenzen für das *HEAT repeat-containing protein 4* (HEATR4) und die Acetyl-Coenzym A Thioesterase 1 (engl. *acyl-CoA thioesterase 1*, kurz ACOT1). Die Deletionen liegen in der intronischen Region des *HEATR4*-Gens, das auf dem Plusstrang codiert wird. Das Transkrip 1 des *ACOT1*-Gens, welches auf dem Minusstrang codiert, wird komplett von der Region #24 umfasst. Im Bereich des ersten Exons des *ACOT1*-Gens befinden sich mehrere Transkriptionsfaktorbindestellen, die durch ihre unmittelbare Nähe zu der codierenden Sequenz des *HEATR4*-Gens auch Einfluss auf dessen Expression haben können. Eine CNV in der Region #24 kann also das komplette ACOT1-Gen deletieren und direkten Einfluss auf die Regulation des *HEATR4*-Gens haben. Das Fehlen von Transkriptionsfaktor-Bindestellen kann sowohl einen *cis*- als auch als *trans*-wirkenden Effekt auf die Genregulation haben. Zusätzlich kann eine strukturelle Variation im Intron Auswirkungen auf posttranskriptionale Modifikationen, wie z.B. dem Spleißen, haben.

Das *HEATR4*-Gen ist nach der *HEAT repeat* Domäne des codierten Proteins benannt und spielt eine Rolle bei der Organisation der Zellstruktur und dem zellulären Transport (Girirajan et al., 2009). Die durch das *ACOT1*-Gen codierte Acetyl-Coenzym A Thioesterase 1 gehört zu einer Gruppe von Enzymen, welche die Hydrolyse von Acetyl-Coenzyms A zu freien Fettsäuren und Coenzym A katalysieren (Hunt et al., 2005). Das Acetyl-Coenzym A ist am Kohlenhydrat- und Eiweißstoffwechsel beteiligt. Bislang sind keine Assoziationen des *HEATR4*-Gens oder des *ACOT1*-Gens mit einer Krankheit oder einem Phänotyp bekannt. Beide Gene werden in dem von der Sarkoidose primär betroffenen Gewebe des Lymphsystems, der Lunge und der Leber exprimiert (Wu et al., 2009).

4.2.5 Neue Risikofaktoren in Zusammenhang mit Sarkoidose

In dieser Studie konnten zwei neue Risikofaktoren für Sarkoidose identifiziert werden, eine Duplikation im *WWOX*-Gen sowie eine Deletion in der *HEATR4*/*ACOT1*-Genregion. Die Interpretation der Duplikation in dem *WWOX*-Gen ist äußerst schwierig, da regulatorische Elemente des duplizierten Segments das *WWOX*-Gen selbst betreffen, aber auch an einem anderen Locus wirken können. Darüber hinaus ist es möglich, dass die betroffene Sequenz an einen anderen Locus dupliziert wurde. Eine detaillierte Analyse der Grenzpositionen sowie des duplizierten Segments und seiner angrenzenden Sequenzen durch Sequenzierung wären nötig, um weitere Aussagen zu treffen. Obwohl die Duplikation einer Transkriptionsfaktor-Bindestelle Auswirkungen auf die Genexpression haben kann, ist es also aufgrund der Datenlage nicht möglich zu bestimmen, welches Gen betroffen ist. Das *WWOX*-Gen selbst kann deshalb nicht als Kandidaten-Gen betrachtet werden. Die Duplikation eines Segments der *WWOX*-Genregion kann jedoch als Risikofaktor eingestuft werden. Die Wirkung der Deletion in der *HEATR4*/*ACOT1*-Genregion ist besser einzuschätzen, da die Variation den Locus direkt betrifft. Allerdings sind für diese Genregion keine Assoziationen mit anderen Phänotypen bekannt und die Funktionen der betroffenen Gene spielen keine bekannte Rolle in der Pathogenese der Sarkoidose. Die *HEATR4*/*ACOT1*-Genregion kann somit als neu identifizierter Kandidat für einen Risiko-Locus vorgeschlagen werden.

Das für die Duplikation im *WWOX*-Gen beobachtete Quotienten-Verhältnis dient als Indikator für die Effektgröße der Variation und liegt mit $OR = 2{,}20$ im Bereich der für Varianten im *BTNL2*-Gen beobachteten Effektgröße von $OR = 2{,}75$ (Valentonyte, Hampe, Huse, et al., 2005). Für die Deletion in der *HEATR4*/*ACOT1*-Genregion wurde ein Quotienten-Verhältnis von $OR = 1{,}51$ beobachtet, womit der Effekt in der Größenordnung des für eine Variation im *ANXA11*-Gen beobachteten Effekts von $OR = 0{,}62$ liegt (Hofmann et al., 2008). Da für häufige CNVs ein relativ geringer Effekt erwartet wird (The Wellcome Trust Case Control Consortium, 2010) waren die in dieser Studie beobachteten Effektgrößen der mit Sarkoidose assoziierten CNVs unerwartet hoch. Somit besteht anscheinend die Möglichkeit diese Befunde auch in kleinen unabhängigen Populationen zu replizieren. Dabei ist zu bedenken, dass sich die vorhergesagten Effekte der *in-silico* Analyse stark von den in der experimentellen Validierung beobachteten Effekten unterscheiden. Bei einer experimentellen Replikation mittels TaqMan®-Genotypisierung wäre auf eine ausreichend hohe Zahl an probenspezifischen Replikaten zu achten, um die Validität der Ergebnisse zu erhöhen und die Effektgröße genauer zu bestimmen.

Die einzige dieser Arbeit vorausgegangene Analyse von CNVs in Zusammenhang mit Sarkoidose erfolgte im Rahmen einer Kandidaten-Gen-Studie der HLA-Region (Wennerström et al., 2012). In dieser Studie wurde ein nicht signifikanter Unterschied der Defizienz des C4-Komplementfaktors bei Sarkoidosepatienten und gesunden Kontrollpersonen beobachtet. Die Defizienz eines Komplementfaktors stellt die verminderte Expression oder das Fehlen von Proteinen des Komplementsystems dar, was durch eine Deletion des entsprechenden Gens verursacht wird.

4.3 Schlussfolgerung und Ausblick

Um die kausalen CN-Genotypen zu identifizieren, wäre eine erneute CNV-TaqMan®-Genotypisierung mit einer erheblich größeren Anzahl an Replikaten für jede Probe nötig, wodurch sich die Validität der experimentellen CNV-Detektion erhöhen würde. Da es sich bei CNVs häufig nicht um genetische Marker, sondern wahrscheinlich um kausale Variationen handelt (Schlattl et al., 2011; Feuk et al., 2006), sind für eine bessere Beurteilung und Interpretation der Assoziation detaillierte Analysen der Start- und Endpositionen der CNVs nötig. Die genaue Bestimmung der Grenzpositionen ist eine wichtige Voraussetzung für funktionelle Studien, die nötig sind, um die Mechanismen zu beschreiben, durch die sich die CNVs auf den Phänotyp auswirken. Darüber hinaus wären Replikationsstudien sinnvoll, um zu untersuchen, ob es sich bei den identifizierten CNVs um populationsspezifische Variationen handelt.

Die erwartete Effektgröße der Variationen in zwei der sieben technisch validierten Kandidaten-Regionen (#34 und #43) war sehr gering. Aufgrund der begrenzten Stichprobengröße in der statistischen Validierung war die resultierende Teststärke zu gering, um auszuschließen, dass eine Assoziation existiert, obwohl sie nicht experimentell validiert werden konnte. Dazu müssten diese Regionen in einer erneuten Analyse mit einer größeren Stichprobe überprüft werden.

Der Einfluss von Störgrößen bei den SNP-Chip Experimenten kann bei Algorithmen, die ausschließlich die Signalintensität analysieren eine wichtige Fehlerquelle darstellen. Dieser Einfluss kann zum Teil kompensiert werden, indem durch den Algorithmus zusätzliche Informationen in Form der BAF-Werte verwendet werden, wie es bei den HMM-Algorithmen der Fall ist. Aus diesem Grund sollten für eine *in-silico* CNV-Analyse Software verwendet werden, die auf HMM-Algorithmen basieren. Es ist zweifelhaft, ob neue Algorithmen die Analysen verbessern können, da die bereits etablierten HMM-Softwares alle relevanten Informationen der Rohdaten auswerten.

Für zukünftige genomweite Analysen häufiger CNVs anhand von SNP-Chip-Daten ist die im Rahmen dieser Studie entwickelte Pipeline sehr gut geeignet. Ist eine höhere Sensitivität erwünscht,

müssen mehrere Softwares zur CNV-Vorhersage komplementär genutzt werden, wodurch sich jedoch auch die Falsch-Positiv-Rate erhöht. Bei einer CNV-GWAS sollte in jedem Fall ein zweistufiges Studiendesign verfolgt werden, basierend auf einer genomweiten Assoziationsanalyse zur Selektion von Kandidaten-Regionen sowie einer anschließenden experimentellen Validierung, z.B. durch CNV-TaqMan®-Genotypisierung. Dieser Ansatz hat sich in dieser Arbeit als erfolgreiche Strategie zur CNV-Analyse und Identifikation bislang unbekannter Risikofaktoren erwiesen.

Die Detektion von CNVs ist heute auch auf Basis von NGS-Daten möglich, die dafür entwickelten Softwares sind jedoch sehr stark auf bestimmte Formen von struktureller Variation spezialisiert (Zhao et al., 2013). Für eine umfangreiche NGS-basierte CNV-Analyse müssen deshalb mehrere Softwares komplementär angewendet werden, wobei unterschiedliche Fehlerquellen die verschiedenen Algorithmen beeinflussen. Vor allem die ungleichmäßige Abdeckung der Referenz-Sequenz durch die sequenzierten DNA-Fragmente stellt ein großes Problem bei der Analyse dar. Die resultierenden falsch-positiven CNV-Detektionen erfordern umfangreiche Validierung der Ergebnisse, wie auch bei der CNV-Vorhersage durch Array-Daten (Abel & Duncavage, 2013). Die im Vergleich zu SNP-Chips immer noch sehr hohen Kosten der DNA-Sequenzierung machen groß angelegte Fall-Kontroll-Studien basierend auf dieser Technologie sehr kostspielig. Die höhere Auflösung von NGS-basierten Methoden der CNV-Analyse ermöglicht eine Detektion von CNVs ab einer Größe von 50 bp (The 1000 Genomes Project Consortium, 2010). Dies führt zu einer genaueren Bestimmung der Grenzbereiche, die präzise Definition der Grenzpositionen bleibt jedoch auch dann noch eine große Herausforderung.

5 Zusammenfassung

Die Analyse von strukturellen Variationen, insbesondere von Kopienzahl-Variationen (CNVs), hat sich als wichtiger Ansatz bei der Untersuchung von genetischen Grundlagen komplexer Erkrankungen erwiesen. Im Rahmen dieser Arbeit wurde eine vergleichende Analyse von Softwares zur Vorhersage von CNVs basierend auf SNP-Chip Daten durchgeführt sowie ein Verfahren zur systematischen genomweiten *in-silico* CNV-Analyse (CNV-Analyse-Pipeline) entwickelt. Diese Pipeline wurde anschließend verwendet, um CNVs in Zusammenhang mit der komplexen Lungenkrankheit Sarkoidose zu identifizieren.

In dieser Studie wurden sechs verschiedene, häufig angewandte Softwares zur CNV-Vorhersage verglichen: Affymetrix Power Tools, QuantiSNP, PennCNV, GLAD, R-gada und VEGA. Unter der Verwendung der SNP-Chip-Rohdaten von 60 europäischen und afrikanischen HapMap-Trios wurden die Vorhersagen durch familienbasierte Analysen validiert und die Konkordanz der Softwares in Bezug auf ihre Vorhersagen untersucht. Diese intra-familiäre Validierung zeigte eine durchgängig geringe Genauigkeit der Vorhersagen mit Validierungsraten von 35-61 %. Die höchste Validität wurde für PennCNV beobachtet (60,9 %). Software basierend auf Hidden Markov Modellen traf weniger Vorhersagen für CNVs als Segmentierungs-Software, dafür mit einer höheren Validität. Die Auswertung der paarweisen Konkordanz zeigte, dass die verwendeten Softwares häufig unterschiedliche CNVs detektieren und sich somit nur eingeschränkt eignen, die Vorhersagen anderer Softwares zu verifizieren. Zur Vorhersage der CNVs wurde in der Pipeline ausschließlich die verlässlichste Software, PennCNV, verwendet. Durch die Pipeline wurden CNV-Regionen identifiziert, in denen für mehr als 5 % der Fälle oder Kontrollen eine CNV vorhergesagt wurde. Eine *in-silico* Assoziationsanalyse erlaubte die Selektion von Kandidaten-Regionen, die häufige CNVs beinhalten, welche möglicherweise in Zusammenhang mit dem Phänotyp der Sarkoidose stehen.

Der zur Vorhersage von CNVs und deren Assoziation mit Sarkoidose genutzte Datensatz beinhaltete die SNP-Chip-Rohdaten des *Affymetrix Human SNP Array 6.0* von 564 deutschen Sarkoidose-Patienten (Fällen) und 1090 gesunden deutschen Kontroll-Personen (Kontrollen). Insgesamt 17 Kandidaten-Regionen wurden durch die CNV-Analyse-Pipeline bestimmt, sieben von ihnen konnten durch CNV-TaqMan®-Genotypisierung technisch validiert werden. Die Assoziation mit Sarkoidose wurde in einer unabhängigen deutschen Stichprobe (552 Fällen und 552 Kontrollen) experimentell überprüft. In zwei der sieben technisch validierten Regionen konnte die vorhergesagte

Assoziation statistisch bestätigt werden. Dies führte zur Identifikation von zwei neuen möglichen Risikofaktoren für Sarkoidose: Einer Duplikation im *WWOX*-Gen ($p = 1{,}01 \times 10^{-3}$, $OR = 2{,}20$) sowie einer Deletion in der *HEATR4*/*ACOT1*-Genregion ($p = 2{,}26 \times 10^{-2}$, $OR = 1{,}51$). Die Duplikation im *WWOX*-Gen trat häufiger in den Fällen (19 %) als in den Kontrollen (10 %) auf. Die duplizierten regulatorischen Elemente können das *WWOX*-Gen selbst betreffen, aber auch an einem anderen Locus wirken, insbesondere wenn die Sequenz an eine anderen Stelle im Genom dupliziert wurde. Aufgrund der Datenlage ist es nicht möglich zu bestimmen, welches Gen beeinflusst wird, die beobachtete Duplikation eines Segments der *WWOX*-Genregion kann jedoch als Risikofaktor eingestuft werden. Die Deletion in der *HEATR4*/*ACOT1*-Genregion hat eine direkte Auswirkung auf den Locus und wurde ebenfalls häufiger in den Fällen (40 %) als in den Kontrollen (30 %) beobachtet. Die Funktionen der betroffenen Gene spielen keine bekannte Rolle in der Pathogenese der Sarkoidose, weshalb die *HEATR4*/*ACOT1*-Genregion als neuer Kandidat für einen Risiko-Locus vorgeschlagen wird.

6 Summary

The analysis of structural variation, in particular of copy number variations (CNVs), has proven a valuable approach to unravel the genetic basis of complex diseases. In this study, a comparative evaluation of six commonly used CNV detection software tools was performed and a framework for the systematic genome-wide CNV analysis based on SNP chip data (CNV analysis pipeline) was developed. This pipeline was subsequently used to identify CNVs associated with the complex lung disorder sarcoidosis.

The comparison of the CNV detection software tools, namely Affymetrix Power Tools, QuantiSNP, PennCNV, GLAD, R-gada and VEGA, was performed using the SNP chip raw data from 60 European and African HapMap trios. The tool-specific accuracy in CNV prediction was assessed *in silico* by means of family-based validation. Additionally, the level of pairwise prediction concordance was evaluated. Intra-familial validation revealed consistently low levels of prediction accuracy with regard to the number of validated CNVs (35-61%). The highest validity of predicted CNVs was observed for PennCNV, yielding the highest validation rate (60.9 %). Software using Hidden Markov models showed a trend to predict fewer CNVs than segmentation-based algorithms, albeit with greater validity. Finally, the pairwise concordance of CNV prediction was found to vary widely across software tools, indicating different sets of predicted CNVs and allowing only limited use of predictions from one software tool to verify the predictions of other tools. The most reliable software, namely PennCNV, was used for CNV prediction within the analysis framework that was developed to perform a genome-wide screening for regions containing overlapping CNV predictions in more than 5% of either cases or controls. The subsequent association analysis allowed the selection of candidate regions containing common CNVs that may be associated with sarcoidosis.

The dataset used for the prediction of CNVs and their association with sarcoidosis contained SNP chip raw data from the *Affymetrix Human SNP Array 6.0* from 564 German sarcoidosis patients (cases) and 1090 healthy German control individuals (controls). A total of 17 candidate regions were identified using the CNV analysis pipeline and seven of these could be technically validated using CNV TaqMan® genotyping. The predicted association of these seven regions with sarcoidosis was experimentally verified using an independent German sample (552 cases and 552 controls). The association analysis based on CNV TaqMan® genotyping verified two of the seven technically validated CNVs and thus identified two new possible genetic risk factors for sarcoidosis: a duplication in the *WWOX* gene ($p = 1{,}01 \times 10^{-3}, OR = 2{,}20$) and a deletion in the

HEATR4/*ACOT1* gene region ($p = 2{,}26 \times 10^{-2}$, $OR = 1{,}51$). The duplication in the *WWOX* gene is more common in the cases (19 %) than in the controls (10 %). The duplicated regulatory elements could affect the gene itself or a different region, especially if the sequence is duplicated to another locus. Based on the data it is not possible to assess which gene is affected, but the observed duplication of a segment of the *WWOX* gene can be classified as a new risk factor for sarcoidosis. The deletion in the *HEATR4*/*ACOT1* gene region alters the structure at this specific locus and was also observed more frequently in the cases (40 %) than in the controls (30 %). The functions of the affected genes are not known to play a role in the pathogenesis of sarcoidosis. Therefore, the *HEATR4*/*ACOT1* gene region can be proposed as a new risk locus for the disease.

A Literatur

Abdallah, a, Sato, H., Grutters, J. C., Veeraraghavan, S., Lympany, P. a, Ruven, H. J. T., et al. (2003). Inhibitor Kappa B-Alpha (IκB-A) Promoter Polymorphisms in UK and Dutch Sarcoidosis. *Genes and Immunity*, 4(6), 450–4.

Abel, H. J. & Duncavage, E. J. (2013). Detection of Structural DNA Variation from next Generation Sequencing Data: A Review of Informatic Approaches. *Cancer Genetics*, 206(12), 432–40.

Abu-Odeh, M., Bar-Mag, T., Huang, H., Kim, T., Salah, Z., Abdeen, S. K., et al. (2014). Characterizing WW Domain Interactions of Tumor Suppressor WWOX Reveals Its Association with Multiprotein Networks. *The Journal of biological chemistry*, 289(13), 8865–80.

Adrianto, I., Lin, C. P., Hale, J. J., Levin, A. M., Datta, I., Parker, R., 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 Inc. (2011). *Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 User Guide*.

Affymetrix Inc. (2012). Affymetrix Power Tools Manual (1.14.2). Retrieved June 6, 2012, from http://media.affymetrix.com/support/developer/powertools/changelog/index.html

Akahoshi, M., Ishihara, M., Namba, K., Kitaichi, N., Ando, Y., Takenaka, S., et al. (2008). Mutation Screening of the CARD15 Gene in Sarcoidosis. *Tissue Antigens*, 71(6), 564–7.

Akahoshi, M., Ishihara, M., Remus, N., Uno, K., Miyake, K., Hirota, T., et al. (2004). Association between IFNA Genotype and the Risk of Sarcoidosis. *Human Genetics*, 114(5), 503–9.

Aldaz, C. M., Chen, T., Sahin, A., Cunningham, J. & Bondy, M. (1995). Comparative Allelotype of in Situ and Invasive Human Breast Cancer: High Frequency of Microsatellite Instability in Lobular Breast Carcinomas. *Cancer research*, 55(18), 3976–81.

Alkan, C., Coe, B. P. & Eichler, E. E. (2011). Genome Structural Variation Discovery and Genotyping. *Nature Reviews Genetics*, 12(5), 363–76.

Almal, S. H. & Padh, H. (2012). Implications of Gene Copy-Number Variation in Health and Diseases. *Journal of Human Genetics*, 57(1), 6–13.

Applied Biosystems. (2010). *TaqMan ® Copy Number Assays Protocol*.

Applied Biosystems. (2011). CopyCaller® Software User Guide.

Arbustini, E., Grasso, M., Leo, G., Tinelli, C., Fasani, R., Diegoli, M., et al. (1996). Polymorphism of Angiotensin-Converting Enzyme Gene in Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 153(2), 851–4.

Arlt, M. F., Mulle, J. G., Schaibley, V. M., Ragland, R. L., Durkin, S. G., Warren, S. T., et al. (2009). Replication Stress Induces Genome-Wide Copy Number Changes in Human Cells That Resemble Polymorphic and Pathogenic Variants. *American Journal of Human Genetics*, 84(3), 339–50.

Arlt, M. F., Wilson, T. E. & Glover, T. W. (2012). Replication Stress and Mechanisms of CNV Formation. *Current Opinion in Genetics & Development*, 22(3), 204–10.

Baross, A., Delaney, A. D., Li, H. I., Nayar, T., Flibotte, S., Qian, H., et al. (2007). Assessment of Algorithms for High Throughput Detection of Genomic Copy Number Variation in Oligonucleotide Microarray Data. *BMC Bioinformatics*, 8, 368.

Bednarek, A. K., Keck-Waggoner, C. L., Daniel, R. L., Laflin, K. J., Bergsagel, P. L., Kiguchi, K., et al. (2001). WWOX, the FRA16D Gene, Behaves as a Suppressor of Tumor Growth. *Cancer Research*, 61(22), 8068–73.

Bednarek, A. K., Laflin, K. J., Daniel, R. L., Liao, Q., Hawkins, K. A. & Aldaz, C. M. (2000). WWOX, a Novel WW Domain-Containing Protein Mapping to Human Chromosome 16q23.3-24.1, a Region Frequently Affected in Breast Cancer. *Cancer Research*, 60(8), 2140–5.

Bengtsson, H., Wirapati, P. & Speed, T. P. (2009). A Single-Array Preprocessing Method for Estimating Full-Resolution Raw Copy Numbers from All Affymetrix Genotyping Arrays Including GenomeWideSNP 5 & 6. *Bioinformatics*, 25(17), 2149–56.

Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., et al. (2010). Large, Rare Chromosomal Deletions Associated with Severe Early-Onset Obesity. *Nature*, 463(7281), 666–70.

Bogunia-Kubik, K., Koscinska, K., Suchnicki, K. & Lange, A. (2006). HSP70-Hom Gene Single Nucleotide (+2763 G/A and +2437 C/T) Polymorphisms in Sarcoidosis. *International Journal of Immunogenetics*, 33(2), 135–40.

Bombieri, C., Luisetti, M., Belpinati, F., Zuliani, E., Beretta, A., Baccheschi, J., et al. (2000). Increased Frequency of CFTR Gene Mutations in Sarcoidosis: A Case/control Association Study. *European Journal of Human Genetics*, 8(9), 717–20.

Brewerton, D. a, Cockburn, C., James, D. C., James, D. G. & Neville, E. (1977). HLA Antigens in Sarcoidosis. *Clinical and Experimental Immunology*, 27(2), 227–9.

Buchanan, J. A. & Scherer, S. W. (2008). Contemplating Effects of Genomic Structural Variation. *Genetics in Medicine*, 10(9), 639–47.

Campbell, M. C. & Tishkoff, S. a. (2008). African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual Review of Genomics and Human Genetics*, 9, 403–33.

Carr, A. M. & Lambert, S. (2013). Replication Stress-Induced Genome Instability: The Dark Side of Replication Maintenance by Homologous Recombination. *Journal of Molecular Biology*, 425(23), 4733–44.

Carter, B. S., Ewing, C. M., Ward, W. S., Treiger, B. F., Aalders, T. W., Schalken, J. a, et al. (1990). Allelic Loss of Chromosomes 16q and 10q in Human Prostate Cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 87(22), 8751–5.

Carter, N. P. (2007). Methods and Strategies for Analyzing Copy Number Variation Using DNA Microarrays. *Nature Genetics*, 39(7 Suppl), S16–21.

Chang, N.-S., Doherty, J., Ensign, A., Lewis, J., Heath, J., Schultz, L., et al. (2003). Molecular Mechanisms Underlying WOX1 Activation during Apoptotic and Stress Responses. *Biochemical Pharmacology*, 66(8), 1347–54.

Chen, J.-M., Chuzhanova, N., Stenson, P. D., Férec, C. & Cooper, D. N. (2005). Meta-Analysis of Gross Insertions Causing Human Genetic Disease: Novel Mutational Mechanisms and the Role of Replication Slippage. *Human Mutation*, 25(2), 207–21.

Chen, T., Sahin, A. & Aldaz, C. M. (1996). Deletion Map of Chromosome 16q in Ductal Carcinoma in Situ of the Breast: Refining a Putative Tumor Suppressor Gene Region. *Cancer research*, 56(24), 5605–9.

Chen, W., Hayward, C., Wright, A. F., Hicks, A. a, Vitart, V., Knott, S., et al. (2011). Copy Number Variation across European Populations. *PLoS ONE*, 6(8), e23087.

De Cid, R., Riveira-Munoz, E., Zeeuwen, P. L. J. M., Robarge, J., Liao, W., Dannhauser, E. N., et al. (2009). Deletion of the Late Cornified Envelope LCE3B and LCE3C Genes as a Susceptibility Factor for Psoriasis. *Nature Genetics*, 41(2), 211–5.

Coe, B. P., Girirajan, S. & Eichler, E. E. (2012). The Genetic Variability and Commonality of Neurodevelopmental Disease. *American Journal of Human Genetics*, 160C(2), 118–29.

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., et al. (2007). QuantiSNP: An Objective Bayes Hidden-Markov Model to Detect and Accurately Map Copy Number Variation Using SNP Genotyping Data. *Nucleic Acids Research*, 35(6), 2013–25.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., et al. (2010). Origins and Functional Impact of Copy Number Variation in the Human Genome. *Nature*, 464(7289), 704–12.

Coquelle, A., Pipiras, E., Toledo, F., Buttin, G. & Debatisse, M. (1997). Expression of Fragile Sites Triggers Intrachromosomal Mammalian Gene Amplification and Sets Boundaries to Early Amplicons. *Cell*, 89(2), 215–25.

Coquelle, A., Rozier, L., Dutrillaux, B. & Debatisse, M. (2002). Induction of Multiple Double-Strand Breaks within an Hsr by meganucleaseI-SceI Expression or Fragile Site Activation Leads to Formation of Double Minutes and Other Chromosomal Rearrangements. *Oncogene*, 21(50), 7671–9.

Cozier, Y., Ruiz-Narvaez, E. A., McKinnon, C. J., Berman, J. S., Rosenberg, L. & Palmer, J. R. (2012). Fine-Mapping in African-American Women Confirms the Importance of the 10p12 Locus to Sarcoidosis. *Genes and Immunity*, 13(7), 573–8.

Cozier, Y., Ruiz-Narvaez, E., McKinnon, C., Berman, J., Rosenberg, L. & Palmer, J. (2013). Replication of Genetic Loci for Sarcoidosis in US Black Women: Data from the Black Women's Health Study. *Human Genetics*, 132(7), 803–10.

Daniil, Z., Mollaki, V., Malli, F., Koutsokera, A., Antoniou, K. M., Rodopoulou, P., et al. (2013). Polymorphisms and Haplotypes in MyD88 Are Associated with the Development of Sarcoidosis: A Candidate-Gene Association Study. *Molecular Biology Reports*, 40(7), 4281–6.

Deubelbeiss, U., Gemperli, A., Schindler, C., Baty, F. & Brutsche, M. H. (2010). Prevalence of Sarcoidosis in Switzerland Is Associated with Environmental Factors. *The European Respiratory Journal*, 35(5), 1088–97.

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., et al. (1996). A Comprehensive Genetic Map of the Human Genome Based on 5,264 Microsatellites. *Nature*, 380(6570), 152–4.

Donati, V., Fontanini, G., Dell'Omodarme, M., Prati, M. C., Nuti, S., Lucchi, M., et al. (2007). WWOX Expression in Different Histologic Types and Subtypes of Non-Small Cell Lung Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 13(3), 884–91.

Duan, J., Zhang, J.-G., Deng, H.-W. & Wang, Y.-P. (2013). Comparative Studies of Copy Number Variation Detection Methods for next-Generation Sequencing Technologies. *PLoS ONE*, 8(3), e59128.

Dubaniewicz, A., Jamieson, S. E., Dubaniewicz-Wybieralska, M., Fakiola, M., Nancy Miller, E. & Blackwell, J. M. (2005). Association between SLC11A1 (formerly NRAMP1) and the Risk of Sarcoidosis in Poland. *European Journal of Human Genetics*, 13(7), 829–34.

Durkin, S. G. & Glover, T. W. (2007). Chromosome Fragile Sites. *Annual Review of Genetics*, 41, 169–92.

Eckel-Passow, J. E., Atkinson, E. J., Maharjan, S., Kardia, S. L. R. & de Andrade, M. (2011). Software Comparison for Evaluating Genomic Copy Number Variation for Affymetrix 6.0 SNP Array Platform. *BMC Bioinformatics*, 12(1), 220.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease. *Nature Reviews Genetics*, 11(6), 446–50.

Eishi, Y., Suga, M., Ishige, I., Kobayashi, D., Yamada, T., Takemura, T., et al. (2002). Quantitative Analysis of Mycobacterial and Propionibacterial DNA in Lymph Nodes of Japanese and European Patients with Sarcoidosis. *Journal of Clinical Microbiology*, 40(1), 198–204.

Ezzie, M. E. & Crouser, E. D. (2007). Considering an Infectious Etiology of Sarcoidosis. *Clinics in Dermatology*, 25(3), 259–66.

Fabbri, M., Iliopoulos, D., Trapasso, F., Aqeilan, R. I., Cimmino, A., Zanesi, N., et al. (2005). WWOX Gene Restoration Prevents Lung Cancer Growth in Vitro and in Vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15611–6.

Feuk, L., Carson, A. R. & Scherer, S. W. (2006). Structural Variation in the Human Genome. *Nature Reviews Genetics*, 7(2), 85–97.

Fischer, A., Grunewald, J., Spagnolo, P., Nebel, A., Schreiber, S. & Müller-Quernheim, J. (2014). Genetics of Sarcoidosis. *Seminars in Respiratory and Critical Care Medicine*, 35(3), 296–306.

Fischer, A., Nothnagel, M., Franke, A., Jacobs, G., Saadati, H. R., Gaede, K. I., et al. (2011). Association of Inflammatory Bowel Disease Risk Loci with Sarcoidosis, and Its Acute and Chronic Subphenotypes. *The European Respiratory Journal*, 37(3), 610–6.

Fischer, A., Schmid, B., Ellinghaus, D., Nothnagel, M., Gaede, K. I., Schürmann, M., et al. (2012). A Novel Sarcoidosis Risk Locus for Europeans on Chromosome 11q13.1. *American Journal of Respiratory and Critical Care Medicine*, (C), 1–35.

Fischer, A., Valentonyte, R., Nebel, A., Nothnagel, M., Müller-Quernheim, J., Schürmann, M., et al. (2008). Female-Specific Association of C-C Chemokine Receptor 5 Gene Polymorphisms with Löfgren's Syndrome. *Journal of Molecular Medicine*, 86(5), 553–61.

Franke, A., Fischer, A., Nothnagel, M., Becker, C., Grabe, N., Till, 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.

Fridlender, Z. G., Schwartz, A., Kohan, M., Amir, G., Glazer, M. & Berkman, N. (2010). Association between CD14 Gene Polymorphisms and Disease Phenotype in Sarcoidosis. *Respiratory Medicine*, 104(9), 1336–43.

Furuya, K., Yamaguchi, E., Itoh, A., Hizawa, N., Ohnuma, N., Kojima, J., et al. (1996). Deletion Polymorphism in the Angiotensin I Converting Enzyme (ACE) Gene as a Genetic Risk Factor for Sarcoidosis. *Thorax*, 51(8), 777–80.

Gazouli, M., Mantzaris, G., Kotsinas, A., Zacharatos, P., Papalambros, E., Archimandritis, A., et al. (2005). Association between Polymorphisms in the Toll-like Receptor 4, CD14, and CARD15/NOD2 and Inflammatory Bowel Disease in the Greek Population. *World Journal of Gastroenterology*, 11(5), 681–5.

Girirajan, S., Chen, L., Graves, T., Marques-Bonet, T., Ventura, M., Fronick, C., et al. (2009). Sequencing Human-Gibbon Breakpoints of Synteny Reveals Mosaic New Insertions at Rearrangement Sites. *Genome research*, 19(2), 178–90.

Glover, T. W., Berger, C., Coyle, J. & Echo, B. (1984). DNA Polymerase Alpha Inhibition by Aphidicolin Induces Gaps and Breaks at Common Fragile Sites in Human Chromosomes. *Human Genetics*, 67(2), 136–42.

Grunewald, J. & Eklund, A. (2007). Sex-Specific Manifestations of Löfgren's Syndrome. *American Journal of Respiratory and Critical Care Medicine*, 175(1), 40–4.

Grutters, J. C., Sato, H., Pantelidis, P., Lagan, A. L., McGrath, D. S., Lammers, J.-W. J., et al. (2002). Increased Frequency of the Uncommon Tumor Necrosis Factor -857T Allele in British

and Dutch Patients with Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 165(8), 1119–24.

Grutters, J. C., Sato, H., Pantelidis, P., Ruven, H. J. T., McGrath, D. S., Wells, A. U., et al. (2003). Analysis of IL6 and IL1A Gene Polymorphisms in UK and Dutch Patients with Sarcoidosis. *Sarcoidosis, Vasculitis and Diffuse Lung Diseases*, 20(1), 20–7.

Guleva, I. & Seitzer, U. (2000). Vitamin D Receptor Gene Polymorphism in Patients with Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 162(2 Pt 1), 760–1.

Haber, J. E. (1992). Exploring the Pathways of Homologous Recombination. *Current Opinion in Cell Biology*, 4(3), 401–12.

Hastings, P. J., Ira, G. & Lupski, J. R. (2009). A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. *PLoS Genetics*, 5(1), e1000327.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. (2009). Mechanisms of Change in Gene Copy Number. *Nature Reviews Genetics*, 10(8), 551–64.

Hattori, T., Konno, S., Takahashi, A., Isada, A., Shimizu, K., Shimizu, K., et al. (2010). Genetic Variants in Mannose Receptor Gene (MRC1) Confer Susceptibility to Increased Risk of Sarcoidosis. *BMC Medical Genetics*, 11(1), 151.

Heron, M., Grutters, J. C., van Moorsel, C. H. M., Ruven, H. J. T., Huizinga, T. W. J., van der Helm-van Mil, a H. M., et al. (2009). Variation in IL7R Predisposes to Sarcoid Inflammation. *Genes and Immunity*, 10(7), 647–53.

Heron, M., Grutters, J. C., Van Moorsel, C. H. M., Ruven, H. J. T., Kazemier, K. M., Claessen, A. M. E., et al. (2009). Effect of Variation in ITGAE on Risk of Sarcoidosis, CD103 Expression, and Chest Radiography. *Clinical Immunology*, 133(1), 117–25.

Heron, M., van Moorsel, C. H. M., Grutters, J. C., Huizinga, T. W. J., van der Helm-van Mil, A. H. M., Nagtegaal, M. M., et al. (2011). Genetic Variation in GREM1 Is a Risk Factor for Fibrosis in Pulmonary Sarcoidosis. *Tissue Antigens*, 77(2), 112–7.

Hill, M. R., Papafili, A., Booth, H., Lawson, P., Hubner, M., Beynon, H., et al. (2006). Functional Prostaglandin-Endoperoxide Synthase 2 Polymorphism Predicts Poor Outcome in Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 174(8), 915–22.

Hizawa, N., Yamaguchi, E., Furuya, K., Jinushi, E., Ito, A. & Kawakami, Y. (1999). The Role of the C-C Chemokine Receptor 2 Gene Polymorphism V64I (CCR2-64I) in Sarcoidosis in a Japanese Population. *American Journal of Respiratory and Critical Care Medicine*, 159(6), 2021–3.

Hofmann, S., Fischer, A., Nothnagel, M., Jacobs, G., Schmid, B., Wittig, M., et al. (2013). Genome-Wide Association Analysis Reveals 12q13.3-q14.1 as New Risk Locus for Sarcoidosis. *The European Respiratory Journal*, 41(4), 888–900.

Hofmann, S., Fischer, A., Till, A., Quernheim, J. M.-, Häsler, R., Franke, A., et al. (2011). A Genome-Wide Association Study Reveals Evidence of Association with Sarcoidosis at 6p12.1. *European Respiratory Journal*, 38(5), 1127–35.

Hofmann, S., Franke, A., Fischer, A., Jacobs, G., Nothnagel, M., Gaede, K. I., et al. (2008). Genome-Wide Association Study Identifies ANXA11 as a New Susceptibility Locus for Sarcoidosis. *Nature Genetics*, 40(9), 1103–6.

Hunt, M. C., Yamada, J., Maltais, L. J., Wright, M. W., Podesta, E. J. & Alexson, S. E. H. (2005). A Revised Nomenclature for Mammalian Acyl-CoA Thioesterases/hydrolases. *Journal of lipid research*, 46(9), 2029–32.

Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F. & Barillot, E. (2004). Analysis of Array CGH Data: From Signal Ratio to Gain and Loss of DNA Regions. *Bioinformatics*, 20(18), 3413–22.

Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. (2008). The Functional Impact of Structural Variation in Humans. *Trends in Genetics*, 24(5), 238–45.

Hutyrová, B., Pantelidis, P., Drábek, J., Zůrková, M., Kolek, V., Lenhart, K., et al. (2002). Interleukin-1 Gene Cluster Polymorphisms in Sarcoidosis and Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 165(2), 148–51.

Iafrate, a J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., et al. (2004). Detection of Large-Scale Variation in the Human Genome. *Nature Genetics*, 36(9), 949–51.

Iannuzzi, M. C. & Rybicki, B. a. (2007). Genetics of Sarcoidosis: Candidate Genes and Genome Scans. *Proceedings of the American Thoracic Society*, 4(1), 108–16.

Iannuzzi, M. C., Rybicki, B. A. & Teirstein, A. S. (2007). Sarcoidosis. *New England Journal of Medicine*, 357(21), 2153–65.

Ishihara, M., Ohno, S., Ishida, T., Mizuki, N., Ando, H., Naruse, T., et al. (1995). Genetic Polymorphisms of the TNFB and HSP70 Genes Located in the Human Major Histocompatibility Complex in Sarcoidosis. *Tissue Antigens*, 46(1), 59–62.

Jacobs, P. a, Browne, C., Gregson, N., Joyce, C. & White, H. (1992). Estimates of the Frequency of Chromosome Abnormalities Detectable in Unselected Newborns Using Moderate Levels of Banding. *Journal of Medical Genetics*, 29(2), 103–8.

Janssen, R., Grutters, J. C., Ruven, H. J. T., Zanen, P., Sato, H., Welsh, K. I., et al. (2004). No Association between Interleukin-18 Gene Polymorphisms and Haplotypes in Dutch Sarcoidosis Patients. *Tissue Antigens*, 63(6), 578–83.

Janssen, R., Sato, H., Grutters, J. C., Ruven, H. J. T., du Bois, R. M., Matsuura, R., et al. (2004). The Clara cell10 adenine38guanine Polymorphism and Sarcoidosis Susceptibility in Dutch and Japanese Subjects. *American Journal of Respiratory and Critical Care Medicine*, 170(11), 1185–7.

Jonth, A. C., Silveira, L., Fingerlin, T. E., Sato, H., Luby, J. C., Welsh, K. I., et al. (2007). TGF- 1 Variants in Chronic Beryllium Disease and Sarcoidosis. *The Journal of Immunology*, 179(6), 4255–4262.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., et al. (1992). Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. *Science*, 258(5083), 818–21.

Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., et al. (2009). Genome-Wide Association of Early-Onset Myocardial Infarction with Single Nucleotide Polymorphisms and Copy Number Variants. *Nature Genetics*, 41(3), 334–41.

Kato, M., Kawaguchi, T., Ishikawa, S., Umeda, T., Nakamichi, R., Shapero, M. H., et al. (2010). Population-Genetic Nature of Copy Number Variations in the Human Genome. *Human Molecular Genetics*, 19(5), 761–73.

Kawamura, Y., Otowa, T., Koike, A., Sugaya, N., Yoshida, E., Yasuda, S., et al. (2011). A Genome-Wide CNV Association Study on Panic Disorder in a Japanese Population. *Journal of Human Genetics*, 56(12), 852–6.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, a. M., et al. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6), 996–1006.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., et al. (2008). Mapping and Sequencing of Structural Variation from Eight Human Genomes. *Nature*, 453(7191), 56–64.

Kim, H. S., Choi, D., Lim, L. L., Allada, G., Smith, J. R., Austin, C. R., et al. (2011). Association of Interleukin 23 Receptor Gene with Sarcoidosis. *Disease Markers*, 31(1), 17–24.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., et al. (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science*, 318(5849), 420–6.

Korbel, J. O., Urban, A. E., Grubert, F., Du, J., Royce, T. E., Starr, P., et al. (2007). Systematic Prediction and Validation of Breakpoints Associated with Copy-Number Variants in the Human Genome. *Proceedings of the National Academy of Sciences of the United States of America*, 104(24), 10110–5.

Krawczak, M., Nikolaus, S., von Eberstein, H., Croucher, P. J. P., El Mokhtari, N. E. & Schreiber, S. (2006). PopGen: Population-Based Recruitment of Patients and Controls for the Analysis of Complex Genotype-Phenotype Relationships. *Community Genetics*, 9(1), 55–61.

Kruit, A., Grutters, J. C., Ruven, H. J. T., van Moorsel, C. H. M., Weiskirchen, R., Mengsteab, S., et al. (2006). Transforming Growth Factor-Beta Gene Polymorphisms in Sarcoidosis Patients with and without Fibrosis. *Chest*, 129(6), 1584–91.

Kuo, M. T., Vyas, R. C., Jiang, L. X. & Hittelman, W. N. (1994). Chromosome Breakage at a Major Fragile Site Associated with P-Glycoprotein Gene Amplification in Multidrug-Resistant CHO Cells. *Molecular and Cellular Biology*, 14(8), 5202–11.

Labunski, S., Posern, G., Ludwig, S., Kundt, G., Bröcker, E. B. & Kunz, M. (2001). Tumour Necrosis Factor-Alpha Promoter Polymorphism in Erythema Nodosum. *Acta dermato-venereologica*, 81(1), 18–21.

Lee, J. A., Carvalho, C. M. B. & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, 131(7), 1235–47.

Levin, a M., Iannuzzi, M. C., Montgomery, C. G., Trudeau, S., Datta, I., McKeigue, P., et al. (2013). Association of ANXA11 Genetic Variation with Sarcoidosis in African Americans and European Americans. *Genes and Immunity*, 14(1), 13–8.

Li, J., Yang, T., Wang, L., Yan, H., Zhang, Y., Guo, Y., et al. (2009). Whole Genome Distribution and Ethnic Differentiation of Copy Number Variation in Caucasian and Asian Populations. *PLoS ONE*, 4(11), e7958.

Li, Y., Pabst, S., Kubisch, C., Grohé, C. & Wollnik, B. (2010). First Independent Replication Study Confirms the Strong Genetic Association of ANXA11 with Sarcoidosis. *Thorax*, 65(10), 939–40.

Li, Y., Wollnik, B., Pabst, S., Lennarz, M., Rohmann, E., Gillissen, A., et al. (2006). BTNL2 Gene Variant and Sarcoidosis. *Thorax*, 61(3), 273–4.

Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. (2003). Mechanism and Regulation of Human Non-Homologous DNA End-Joining. *Nature Reviews Molecular Cell Biology*, 4(9), 712–20.

Lopez-Campos, J. L., Rodriguez-Rodriguez, D., Rodriguez-Becerra, E., Alfageme Michavila, I., Guerra, J. F., Hernandez, F. J. G., et al. (2009). Cyclooxygenase-2 Polymorphisms Confer Susceptibility to Sarcoidosis but Are Not Related to Prognosis. *Respiratory Medicine*, 103(3), 427–33.

Lopez-Campos, J. L., Rodriguez-Rodriguez, D., Rodriguez-Becerra, E., Michavila, I. A., Guerra, J. F., Hernandez, F. J. G., et al. (2008). Association of the 3050G>C Polymorphism in the Cyclooxygenase 2 Gene with Systemic Sarcoidosis. *Archives of Medical Research*, 39(5), 525–30.

Loth, D. W., Artigas, M. S., Gharib, S. a, Wain, L. V, Franceschini, N., Koch, B., et al. (2014). Genome-Wide Association Analysis Identifies Six New Loci Associated with Forced Vital Capacity. *Nature Genetics*, 46(7), 669–77.

Macias, M. J., Hyvönen, M., Baraldi, E., Schultz, J., Sudol, M., Saraste, M., et al. (1996). Structure of the WW Domain of a Kinase-Associated Protein Complexed with a Proline-Rich Peptide. *Nature*, 382(6592), 646–9.

Maher, B. (2008). Personal Genomes: The Case of the Missing Heritability. *Nature*, 456(7218), 18–21.

Makrythanasis, P., Tzetis, M., Rapti, A., Papatheodorou, A., Tsipi, M., Kitsiou, S., et al. (2010). Cystic Fibrosis Conductance Regulator, Tumor Necrosis Factor, Interferon Alpha-10, Interferon Alpha-17, and Interferon Gamma Genotyping as Potential Risk Markers in Pulmonary Sarcoidosis Pathogenesis in Greek Patients. *Genetic Testing and Molecular Biomarkers*, 14(4), 577–84.

Maliarik, M. J., Chen, K. M., Sheffer, R. G., Rybicki, B. a, Major, M. L., Popovich, J., et al. (2000). The Natural Resistance-Associated Macrophage Protein Gene in African Americans with Sarcoidosis. *American Journal of Respiratory Cell and Molecular Biology*, 22(6), 672–5.

Maliarik, M. J., Rybicki, B. a, Malvitz, E., Sheffer, R. G., Major, M., Popovich, J., et al. (1998). Angiotensin-Converting Enzyme Gene Polymorphism and Risk of Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 158(5 Pt 1), 1566–70.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature*, 461(7265), 747–53.

Maver, A., Medica, I., Salobir, B., Sabovic, M., Tercelj, M. & Peterlin, B. (2007). Polymorphisms in Genes Coding for Mediators in the Interleukin Cascade and Their Effect on Susceptibility to Sarcoidosis in the Slovenian Population. *International Journal of Molecular Medicine*, 20(3), 385–90.

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., et al. (2006). Common Deletion Polymorphisms in the Human Genome. *Nature Genetics*, 38(1), 86–92.

McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., et al. (2008). Deletion Polymorphism Upstream of IRGM Associated with Altered IRGM Expression and Crohn's Disease. *Nature Genetics*, 40(9), 1107–12.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., et al. (2008). Integrated Detection and Population-Genetic Analysis of SNPs and Copy Number Variation. *Nature Genetics*, 40(10), 1166–74.

McDougal, K. E., Fallin, M. D., Moller, D. R., Song, Z., Cutler, D. J., Steiner, L. L., et al. (2009a). Variation in the Lymphotoxin-Alpha/tumor Necrosis Factor Locus Modifies Risk of Erythema Nodosum in Sarcoidosis. *The Journal of Investigative Dermatology*, 129(8), 1921–6.

McDougal, K. E., Fallin, M. D., Moller, D. R., Song, Z., Cutler, D. J., Steiner, L. L., et al. (2009b). Variation in the Lymphotoxin-Alpha/tumor Necrosis Factor Locus Modifies Risk of Erythema Nodosum in Sarcoidosis. *The Journal of Investigative Dermatology*, 129(8), 1921–6.

McGrath, D. S., Foley, P. J., Petrek, M., Izakovicova-Holla, L., Kolek, V., Veeraraghavan, S., et al. (2001). Ace Gene I/D Polymorphism and Sarcoidosis Pulmonary Disease Severity. *American Journal of Respiratory and Critical Care Medicine*, 164(2), 197–201.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., et al. (2011). Mapping Copy Number Variation by Population-Scale Genome Sequencing. *Nature*, 470(7332), 59–65.

Milman, N., Nielsen, O. H., Hviid, T. V. F. & Fenger, K. (2007). CARD15 Single Nucleotide Polymorphisms 8, 12 and 13 Are Not Increased in Ethnic Danes with Sarcoidosis. *Respiration*, 74(1), 76–9.

Milman, N., Svendsen, C. B., Nielsen, F. C. & van Overeem Hansen, T. (2011). The BTNL2 A Allele Variant Is Frequent in Danish Patients with Sarcoidosis. *The Clinical Respiratory Journal*, 5(2), 105–11.

Morais, A., Lima, B., Peixoto, M. J., Alves, H., Marques, A. & Delgado, L. (2012). BTNL2 Gene Polymorphism Associations with Susceptibility and Phenotype Expression in Sarcoidosis. *Respiratory Medicine*, 106(12), 1771–7.

Morais, A., Lima, B., Peixoto, M., Melo, N., Alves, H., Marques, J. A., et al. (2013). Annexin A11 Gene Polymorphism (R230C Variant) and Sarcoidosis in a Portuguese Population. *Tissue Antigens*, 82(3), 186–91.

Morganella, S., Cerulo, L., Viglietto, G. & Ceccarelli, M. (2010). VEGA: Variational Segmentation for Copy Number Detection. *Bioinformatics*, 26(24), 3020–7.

Morohashi, K., Takada, T., Omori, K., Suzuki, E. & Gejyo, F. (2003). Vascular Endothelial Growth Factor Gene Polymorphisms in Japanese Patients with Sarcoidosis. *Chest*, 123(5), 1520–6.

Moss, S. E. & Morgan, R. O. (2004). The Annexins. *Genome biology*, 5(4), 219.

Mrazek, F., Stahelova, A., Kriegova, E., Fillerova, R., Zurkova, M., Kolek, V., et al. (2011). Functional Variant ANXA11 R230C: True Marker of Protection and Candidate Disease Modifier in Sarcoidosis. *Genes and Immunity*, 12(6), 490–4.

Müller-Quernheim, J. (1998). Sarcoidosis: Immunopathogenetic Concepts and Their Clinical Application. *European Respiratory Journal*, 12(3), 716–738.

Müller-Quernheim, J., Prasse, A. & Zissel, G. (2012). Pathogenesis of Sarcoidosis. *Presse Médicale*, 41(6 Pt 2), e275–87.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. & Erlich, H. (1986). Specific Enzymatic Amplification of DNA in Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51 Pt 1, 263–73.

Mumford, D. & Shah, J. (1989). Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems. *Communications on Pure and Applied Mathematics*, 42(5), 577–685.

Muraközy, G., Gaede, K. I., Zissel, G., Schlaak, M. & Müller-Quernheim, J. (2001). Analysis of Gene Polymorphisms in Interleukin-10 and Transforming Growth Factor-Beta 1 in Sarcoidosis. *Sarcoidosis, Vasculitis and Diffuse Lung Diseases*, 18(2), 165–9.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., et al. (2005). A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays. *Cancer Research*, 65(14), 6071–9.

Newman, L. S., Rose, C. S., Bresnitz, E. a, Rossman, M. D., Barnard, J., Frederick, M., et al. (2004). A Case Control Etiologic Study of Sarcoidosis: Environmental and Occupational Risk Factors. *American Journal of Respiratory and Critical Care Medicine*, 170(12), 1324–30.

Nguyen, T., Liu, X. K., Zhang, Y. & Dong, C. (2006). BTNL2, a Butyrophilin-like Molecule That Functions to Inhibit T Cell Activation. *Journal of Immunology*, 176(12), 7354–60.

Niimi, T., Sato, S., Sugiura, Y., Yoshinouchi, T., Akita, K., Maeda, H., et al. (2002). Transforming Growth Factor-Beta Gene Polymorphism in Sarcoidosis and Tuberculosis Patients. *The International Journal of Tuberculosis and Lung Disease*, 6(6), 510–5.

Niimi, T., Tomita, H., Sato, S., Kawaguchi, H., Akita, K., Maeda, H., et al. (1999). Vitamin D Receptor Gene Polymorphism in Patients with Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 160(4), 1107–9.

Nunez, M. I., Ludes-Meyers, J. & Aldaz, C. M. (2006). WWOX Protein Expression in Normal Human Tissues. *Journal of Molecular Histology*, 37(3-4), 115–25.

Ohchi, T., Shijubo, N., Kawabata, I., Ichimiya, S., Inomata, S., Yamaguchi, A., et al. (2004). Polymorphism of Clara Cell 10-kD Protein Gene of Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 169(2), 180–6.

Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. (2004). Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. *Biostatistics*, 5(4), 557–72.

Pabst, S., Baumgarten, G., Stremmel, A., Lennarz, M., Knüfermann, P., Gillissen, A., et al. (2006). Toll-like Receptor (TLR) 4 Polymorphisms Are Associated with a Chronic Course of Sarcoidosis. *Clinical and Experimental Immunology*, 143(3), 420–6.

Pabst, S., Bradler, O., Gillissen, A., Nickenig, G., Skowasch, D. & Grohe, C. (2013). Toll-like Receptor-9 Polymorphisms in Sarcoidosis and Chronic Obstructive Pulmonary Disease. *Advances in Experimental Medicine and Biology*, 756, 239–45.

Pabst, S., Fränken, T., Schönau, J., Stier, S., Nickenig, G., Meyer, R., et al. (2011). Transforming Growth Factor-B Gene Polymorphisms in Different Phenotypes of Sarcoidosis. *The European Respiratory Journal*, 38(1), 169–75.

Pabst, S., Golebiewski, M., Herms, S., Karpushova, A., Díaz-Lacava, A., Walier, M., et al. (2011). Caspase Recruitment Domain 15 Gene Haplotypes in Sarcoidosis. *Tissue Antigens*, 77(4), 333–7.

Pabst, S., Karpushova, A., Diaz-Lacava, A., Herms, S., Walier, M., Zimmer, S., et al. (2010). VEGF Gene Haplotypes Are Associated with Sarcoidosis. *Chest*, 137(1), 156–63.

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., et al. (2006). High-Resolution Genomic Profiling of Chromosomal Aberrations Using Infinium Whole-Genome Genotyping. *Genome Research*, 16(9), 1136–48.

Pereira, R. M., Martinez, G. J., Engel, I., Cruz-Guilloty, F., Barboza, B. a., Tsagaratou, A., et al. (2014). Jarid2 Is Induced by TCR Signalling and Controls iNKT Cell Maturation. *Nature Communications*, 5, 1–14.

Petrek, M., Drábek, J., Kolek, V., Zlámal, J., Welsh, K. I., Bunce, M., et al. (2000). CC Chemokine Receptor Gene Polymorphisms in Czech Patients with Pulmonary Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 162(3 Pt 1), 1000–3.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., et al. (2011). Comprehensive Assessment of Array-Based Platforms and Calling Algorithms for Detection of Copy Number Variants. *Nature Biotechnology*, 29(6), 512–20.

Piotrowski, W. J., Górski, P., Pietras, T., Fendler, W. & Szemraj, J. (2011). The Selected Genetic Polymorphisms of Metalloproteinases MMP2, 7, 9 and MMP Inhibitor TIMP2 in Sarcoidosis. *Medical Science Monitor*, 17(10), CR598–607.

Pique-Regi, R., Cáceres, A. & González, J. R. (2010). R-Gada: A Fast and Flexible Pipeline for Copy Number Analysis in Association Studies. *BMC Bioinformatics*, 11, 380.

Pollack, J. R., Perou, C. M., Alizadeh, a a, Eisen, M. B., Pergamenschikov, A., Williams, C. F., et al. (1999). Genome-Wide Analysis of DNA Copy-Number Changes Using cDNA Microarrays. *Nature Genetics*, 23(1), 41–6.

Polzehl, J. & Spokoiny, V. G. (2000). Adaptive Weights Smoothing with Applications to Image Restoration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2), 335–354.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabin, D. L., Thompson, B., Brown, K. M., Judson, M. a, Huang, X., Lackland, D. T., et al. (2004). Sarcoidosis: Social Predictors of Severity at Presentation. *The European Respiratory Journal*, 24(4), 601–8.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global Variation in Copy Number in the Human Genome. *Nature*, 444(7118), 444–54.

Ried, K., Finnis, M., Hobson, L., Mangelsdorf, M., Dayan, S., Nancarrow, J. K., et al. (2000). Common Chromosomal Fragile Site FRA16D Sequence: Identification of the FOR Gene Spanning FRA16D and Homozygous Deletions and Translocation Breakpoints in Cancer Cells. *Human Molecular Genetics*, 9(11), 1651–63.

Rybicki, B. A., Iannuzzi, M. C., Frederick, M. M., Thompson, B. W., Rossman, M. D., Bresnitz, E. A., et al. (2001). Familial Aggregation of Sarcoidosis. A Case-Control Etiologic Study of

Sarcoidosis (ACCESS). *American Journal of Respiratory and Critical Care Medicine*, 164(11), 2085–91.

Rybicki, B. A., Levin, A. M., McKeigue, P., Datta, I., Gray-McGuire, C., Colombo, M., et al. (2011). A Genome-Wide Admixture Scan for Ancestry-Linked Genes Predisposing to Sarcoidosis in African-Americans. *Genes and Immunity*, 12(2), 67–77.

Rybicki, B. A., Maliarik, M. J., Poisson, L. M. & Iannuzzi, M. C. (2004). Sarcoidosis and Granuloma Genes: A Family-Based Study in African-Americans. *European Respiratory Journal*, 24(2), 251–257.

Rybicki, B. A., Walewski, J. L., Maliarik, M. J., Kian, H. & Iannuzzi, M. C. (2005). The BTNL2 Gene and Sarcoidosis Susceptibility in African Americans and Whites. *American Journal of Human Genetics*, 77(3), 491–9.

Sakuyama, K., Meguro, A., Ota, M., Ishihara, M., Uemoto, R., Ito, H., et al. (2012). Lack of Association between IL10 Polymorphisms and Sarcoidosis in Japanese Patients. *Molecular Vision*, 18(February), 512–8.

Sato, H., Williams, H. R. T., Spagnolo, P., Abdallah, A., Ahmad, T., Orchard, T. R., et al. (2010). CARD15/NOD2 Polymorphisms Are Associated with Severe Pulmonary Sarcoidosis. *The European Respiratory Journal*, 35(2), 324–30.

Sato, T., Tanigami, A., Yamakawa, K., Akiyama, F., Kasumi, F., Sakamoto, G., et al. (1990). Allelotype of Breast Cancer: Cumulative Allele Losses Promote Tumor Progression in Primary Breast Cancer. *Cancer research*, 50(22), 7184–9.

Schlattl, A., Anders, S., Waszak, S. M., Huber, W. & Korbel, J. O. (2011). Relating CNVs to Transcriptome Data at Fine Resolution: Assessment of the Effect of Variant Size, Type, and Overlap with Functional Regions. *Genome Research*, 21(12), 2004–13.

Schürmann, M., Albrecht, M., Schwinger, E. & Stuhrmann, M. (2002). CFTR Gene Mutations in Sarcoidosis. *European Journal of Human Genetics*, 10(11), 729–32.

Schürmann, M., Reichel, P., Müller-Myhsok, B., Schlaak, M., Müller-Quernheim, J. & Schwinger, E. (2001). Results from a Genome-Wide Search for Predisposing Genes in Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 164(5), 840–6.

Schürmann, M., Valentonyte, R., Hampe, J., Muller-Quernheim, J., Schwinger, E. & Schreiber, S. (2003). CARD15 Gene Mutations in Sarcoidosis. *European Respiratory Journal*, 22(5), 748–754.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong Association of de Novo Copy Number Mutations with Autism. *Science*, 316(5823), 445–9.

Seitzer, U., Swider, C., Stüber, F., Suchnicki, K., Lange, a, Richter, E., et al. (1997). Tumour Necrosis Factor Alpha Promoter Gene Polymorphism in Sarcoidosis. *Cytokine*, 9(10), 787–90.

Seyhan, E. C., Cetinkaya, E., Altin, S., Gunluoglu, M. Z., Demir, A., Koksal, V., et al. (2008). Vascular Endothelial Growth Factor Gene Polymorphisms in Turkish Patients with Sarcoidosis. *Tissue Antigens*, 72(2), 162–5.

Shaffer, L. G. & Lupski, J. R. (2000). Molecular Mechanisms for Constitutional Chromosomal Rearrangements in Humans. *Annual Review of Genetics*, 34, 297–329.

Sharma, O. P. (2005). Definition and History of Sarcoidosis, in: *European Respiratory Monograph 32: Sarcoidosis*, (pp. 1–12). European Respiratory.

Sharp, A. J., Cheng, Z. & Eichler, E. E. (2006). Structural Variation of the Human Genome. *Annual Review of Genomics and Human Genetics*, 7, 407–42.

Smith, C. E., Llorente, B. & Symington, L. S. (2007). Template Switching during Break-Induced Replication. *Nature*, 447(7140), 102–5.

Somosközi, A., Zissel, G., Seitzer, U., Gerdes, J., Schlaak, M. & Müller-QuernheimJ. (1999). Polymorphisms at Position -308 in the Promoter Region of the TNF-Alpha and in the First Intron of the TNF-Beta Genes and Spontaneous and Lipopolysaccharide-Induced TNF-Alpha Release in Sarcoidosis. *Cytokine*, 11(11), 882–7.

Song, Z., Marzilli, L., Greenlee, B. M., Chen, E. S., Silver, R. F., Askin, F. B., et al. (2005). Mycobacterial Catalase-Peroxidase Is a Tissue Antigen and Target of the Adaptive Immune Response in Systemic Sarcoidosis. *The Journal of Experimental Medicine*, 201(5), 755–67.

Spagnolo, P., Renzoni, E. a, Wells, A. U., Copley, S. J., Desai, S. R., Sato, H., et al. (2005). C-C Chemokine Receptor 5 Gene Variants in Relation to Lung Disease in Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 172(6), 721–8.

Spagnolo, P., Renzoni, E. a, Wells, A. U., Sato, H., Grutters, J. C., Sestini, P., et al. (2003). C-C Chemokine Receptor 2 and Sarcoidosis: Association with Lofgren's Syndrome. *American Journal of Respiratory and Critical Care Medicine*, 168(10), 1162–6.

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., et al. (2008). Large Recurrent Microdeletions Associated with Schizophrenia. *Nature*, 455(7210), 232–6.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., et al. (2007). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science*, 315(5813), 848–53.

Suzuki, H., Ota, M., Meguro, A., Katsuyama, Y., Kawagoe, T., Ishihara, M., et al. (2012). Genetic Characterization and Susceptibility for Sarcoidosis in Japanese Patients: Risk Factors of BTNL2 Gene Polymorphisms and HLA Class II Alleles. *Investigative Ophthalmology & Visual Science*, 53(11), 7109–15.

Sverrild, a, Backer, V., Kyvik, K. O., Kaprio, J., Milman, N., Svendsen, C. B., et al. (2008). Heredity in Sarcoidosis: A Registry-Based Twin Study. *Thorax*, 63(10), 894–6.

Takada, T., Suzuki, E., Morohashi, K. & Gejyo, F. (2002). Association of Single Nucleotide Polymorphisms in the IL-18 Gene with Sarcoidosis in a Japanese Population. *Tissue Antigens*, 60(1), 36–42.

The 1000 Genomes Project Consortium. (2010). A Map of Human Genome Variation from Population-Scale Sequencing. *Nature*, 467(7319), 1061–73.

The 1000 Genomes Project Consortium. (2012). An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature*, 491, 56–65.

The American Thoriacic Society (ATS), The European Respiratory Society (ERS) & The World Association of Sarcoidosis and other Granulomatous Disorders (WASOG). (1999). Statement on Sarcoidosis. *American Journal of Respiratory and Critical Care Medicine*, 160(2), 736–55.

The International HapMap Consortium. (2005). A Haplotype Map of the Human Genome. *Nature*, 437(7063), 1299–320.

The Wellcome Trust Case Control Consortium. (2010). Genome-Wide Association Study of CNVs in 16,000 Cases of Eight Common Diseases and 3,000 Shared Controls. *Nature*, 464(7289), 713–20.

Thomas, K. W. & Hunninghake, G. W. (2003). Sarcoidosis. *Journal of the American Medical Association*, 289(24), 3300–3.

Tsuda, H., Callen, D. F., Fukutomi, T., Nakamura, Y. & Hirohashi, S. (1994). Allele Loss on Chromosome 16q24.2-Qter Occurs Frequently in Breast Cancers Irrespectively of Differences in Phenotype and Extent of Spread. *Cancer research*, 54(2), 513–7.

Tuzun, E., Sharp, A. J., Bailey, J. a, Kaul, R., Morrison, V. A., Pertz, L. M., et al. (2005). Fine-Scale Structural Variation of the Human Genome. *Nature Genetics*, 37(7), 727–32.

Valentonyte, R., Hampe, J., Croucher, P. J. P., Müller-Quernheim, J., Schwinger, E., Schreiber, S., et al. (2005). Study of C-C Chemokine Receptor 2 Alleles in Sarcoidosis, with Emphasis on Family-Based Analysis. *American Journal of Respiratory and Critical Care Medicine*, 171(10), 1136–41.

Valentonyte, R., Hampe, J., Huse, K., Rosenstiel, P., Albrecht, M., Stenzel, A., et al. (2005). Sarcoidosis Is Associated with a Truncating Splice Site Mutation in BTNL2. *Nature Genetics*, 37(4), 357–64.

Valeyre, D., Bernaudin, J.-F., Uzunhan, Y., Kambouchner, M., Brillet, P.-Y., Soussan, M., et al. (2014). Clinical Presentation of Sarcoidosis and Diagnostic Work-Up. *Seminars in Respiratory and Critical Care Medicine*, 35(3), 336–51.

Valeyre, D., Prasse, A., Nunes, H., Uzunhan, Y., Brillet, P. & Müller-Quernheim, J. (2013). Sarcoidosis. *Lancet*, 6736(13), 1–13.

Vasakova, M., Sterclova, M., Kolesar, L., Slavcev, A., Skibova, J. & Striz, I. (2010). Cytokine Gene Polymorphisms in Sarcoidosis. *Sarcoidosis, Vasculitis and Diffuse Lung Diseases*, 27(1), 70–5.

Veltkamp, M., Van Moorsel, C. H. M., Rijkers, G. T., Ruven, H. J. T., Van Den Bosch, J. M. M. & Grutters, J. C. (2010). Toll-like Receptor (TLR)-9 Genetics and Function in Sarcoidosis. *Clinical and Experimental Immunology*, 162(1), 68–74.

Visscher, P. M., Brown, M. a, McCarthy, M. I. & Yang, J. (2012). Five Years of GWAS Discovery. *American Journal of Human Genetics*, 90(1), 7–24.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. a, et al. (2007). PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data. *Genome Research*, 17(11), 1665–74.

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C. & Marth, G. (2002). Human Diallelic Insertion/deletion Polymorphisms. *American journal of human genetics*, 71(4), 854–62.

Wennerström, A., Pietinalho, A., Vauhkonen, H., Lahtela, L., Palikhe, A., Hedman, J., et al. (2012). HLA-DRB1 Allele Frequencies and C4 Copy Number Variation in Finnish Sarcoidosis Patients and Associations with Disease Prognosis. *Human Immunology*, 73(1), 93–100.

Westfall, P. H. & Young, S. S. (1993). *Resampling-Based Multiple Testing*. New York: Wiley.

Wijnen, P. a, Voorter, C. E., Nelemans, P. J., Verschakelen, J. a, Bekers, O. & Drent, M. (2011). Butyrophilin-like 2 in Pulmonary Sarcoidosis: A Factor for Susceptibility and Progression? *Human Immunology*, 72(4), 342–7.

Wilson, A. G., di Giovine, F. S., Blakemore, A. I. & Duff, G. W. (1992). Single Base Polymorphism in the Human Tumour Necrosis Factor Alpha (TNF Alpha) Gene Detectable by NcoI Restriction of PCR Product. *Human Molecular Genetics*, 1(5), 353.

Winchester, L., Yau, C. & Ragoussis, J. (2009). Comparing CNV Detection Methods for SNP Arrays. *Briefings in Functional Genomics & Proteomics*, 8(5), 353–66.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., et al. (2009). BioGPS: An Extensible and Customizable Portal for Querying and Organizing Gene Annotation Resources. *Genome Biology*, 10(11), R130.

Yendamuri, S., Kuroki, T., Trapasso, F., Henry, A. C., Dumon, K. R., Huebner, K., et al. (2003). WW Domain Containing Oxidoreductase Gene Expression Is Altered in Non-Small Cell Lung Cancer. *Cancer research*, 63(4), 878–81.

Yunis, J. J. & Soreng, A. L. (1984). Constitutive Fragile Sites and Cancer. *Science*, 226(4679), 1199–204.

Zhang, D., Qian, Y., Akula, N., Alliey-Rodriguez, N., Tang, J., Gershon, E. S., et al. (2011). Accuracy of CNV Detection from GWAS Data. *PLoS ONE*, 6(1), e14511.

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D. & Lupski, J. R. (2009). The DNA Replication FoSTeS/MMBIR Mechanism Can Generate Genomic, Genic and Exonic Complex Rearrangements in Humans. *Nature Genetics*, 41(7), 849–53.

Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. (2013). Computational Tools for Copy Number Variation (CNV) Detection Using next-Generation Sequencing Data: Features and Perspectives. *BMC Bioinformatics*, 14 Suppl 1(Suppl 11), S1.

Zhao, X., Li, C., Paez, J. G., Chin, K., Jänne, P. A., Chen, T., et al. (2004). An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Research*, 64(9), 3060–71.

Zhou, Y., Yamaguchi, E., Hizawa, N. & Nishimura, M. (2005). Roles of Functional Polymorphisms in the Interleukin-18 Gene Promoter in Sarcoidosis. *Sarcoidosis, Vasculitis and Diffuse Lung Diseases*, 22(2), 105–13.

Zissel, G., Prasse, A. & Müller-Quernheim, J. (2010). Immunologic Response of Sarcoidosis. *Seminars in Respiratory and Critical Care Medicine*, 31(4), 390–403.

Zorzetto, M., Bombieri, C., Ferrarotti, I., Medaglia, S., Agostini, C., Tinelli, C., et al. (2002). Complement Receptor 1 Gene Polymorphisms in Sarcoidosis. *American Journal of Respiratory Cell and Molecular Biology*, 27(1), 17–23.

Zorzetto, M., Ferrarotti, I., Campo, I., Trisolini, R., Poletti, V., Scabini, R., et al. (2005). NOD2/CARD15 Gene Polymorphisms in Idiopathic Pulmonary Fibrosis. *Sarcoidosis, Vasculitis and Diffuse Lung Diseases*, 22(3), 180–5.

Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. (2012). The Mystery of Missing Heritability: Genetic Interactions Create Phantom Heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4), 1193–8.

B Appendix

B.1 CNV Software Benchmark

Appendix Tabelle 1: Eigenschaften der CNV-Vorhersagen in europäischen Proben

Software	Anzahl	Median der Länge [kb]	Median der kumulativen Länge [kb]	Median der Anzahl der Marker	Median der Distanz der Marker [kb]	DDR
APT	87,0 (81,0-94,0)	10,2 (9,3-10,7)	5,0 (4,1-6,4)	13,8 (11,1-15,4)	0,22 (0,19-0,31)	4,9 (3,7-5,8)
GLAD	161,5 (133,8-192,5)	7,5 (6,8-8,6)	6,1 (4,4-8,6)	8,0 (6,0-8,9)	0,20 (0,16-0,24)	1,9 (1,5-2,3)
PennCNV	67,0 (54,2-76,8)	24,4 (18,8-33,6)	5,3 (4,3-6,4)	26,0 (25,0-30,0)	0,20 (0,15-0,22)	4,5 (3,8-5,5)
QuantiSNP	137,0 (122,0-147,5)	10,0 (8,8-11,2)	7,7 (4,4-22,4)	6,2 (6,0-7,0)	0,27 (0,21-0,30)	2,6 (2,2-3,3)
R-gada	200,0 (174,8-214,0)	9,3 (7,6-10,4)	254,7 (45,1-390,1)	9,0 (7,6-12,5)	0,31 (0,25-0,39)	4,1 (3,4-5,4)
VEGA	145,5 (119,0-159,2)	7,4 (6,9-8,1)	6,5 (4,7-8,0)	8,0 (6,2-9,0)	0,27 (0,22-0,30)	3,4 (3,1-4,6)
Algorithmus						
HMM	87,0 (81,0-94,0)	10,6 (9,9-12,0)	5,0 (4,3-6,3)	13,8 (11,1-15,4)	0,22 (0,19-0,27)	4,1 (3,3-5,1)
Segmentation	162,0 (144,0-186,5)	7,8 (7,1-8,7)	7,1 (5,3-10,3)	8,0 (7,2-9,4)	0,26 (0,22-0,32)	3,5 (3,1-4,2)

Angegeben ist der Median und in Klammern der Interquartilsabstand (IQR)für die Proben der Kinder der 30 HapMap CEU Trios. **DDR:** Verhältnis von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*).

Appendix Tabelle 2: Eigenschaften der CNV-Vorhersagen in afrikanischen Proben

Software	Anzahl	Median der Länge [kb]	Median der kumulativen Länge [kb]	Median der Anzahl der Marker	Median der Distanz der Marker [kb]	DDR
APT	111,0 (103,2-120,0)	8,1 (7,6-8,9)	4,3 (3,5-5,4)	8,5 (7,5-9,9)	0,22 (0,20-0,24)	3,8 (2,9-4,4)
GLAD	197,0 (148,2-230,2)	6,9 (6,3-7,6)	6,2 (4,4-7,8)	6,0 (5,0-7,0)	0,21 (0,18-0,27)	2,8 (2,4-3,1)
PennCNV	73,0 (63,5-79,8)	19,8 (15,9-21,9)	4,1 (3,4-5,3)	25,5 (23,1-27,8)	0,17 (0,14-0,20)	6,6 (5,8-8,3)
QuantiSNP	165,5 (134,2-177,8)	8,5 (7,5-9,9)	15,8 (7,1-26,3)	6,0 (5,0-7,0)	0,23 (0,21-0,24)	3,4 (2,9-3,7)
R-gada	229,5 (189,0-251,5)	7,4 (6,7-8,4)	73,7 (13,9-139,2)	7,0 (6,0-7,0)	0,26 (0,23-0,29)	4,6 (3,8-5,4)
VEGA	185,0 (157,2-195,2)	6,5 (6,1-7,4)	5,5 (4,6-7,6)	5,0 (5,0-7,0)	0,28 (0,26-0,32)	3,8 (3,1-4,6)
Algorithmus						
HMM	108,0 (100,2-119,5)	8,9 (8,5-10,0)	4,6 (3,8-5,4)	8,8 (7,5-10,0)	0,21 (0,19-0,23)	3,8 (3,1-4,4)
Segmentation	204,0 (178,2-225,0)	7,0 (6,5-7,6)	7,0 (5,0-8,2)	6,0 (6,0-7,0)	0,26 (0,22-0,30)	3,8 (3,0-4,5)

Angegeben ist der Median und in Klammern der Interquartilsabstand (IQR)für die Proben der Kinder der 30 HapMap YRI Trios. **DDR:** Verhältnis von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*).

Appendix Tabelle 3: Validierungsraten in europäischen Proben

Software	Validierung CNVs [%]	Validierung Deletionen [%]	Validierung Duplicationen [%]	DDR, validierte CNVs	Validierte kumulative Sequenz [%]
APT	58.8 (52.2-62.9)	58.9 (54.8-65.3)	57.1 (47.1-66.7)	1.0 (0.9-1.2)	54.3 (39.5-63.9)
GLAD	50.8 (41.1-56.2)	49.9 (40.9-57.3)	56.0 (43.3-61.3)	1.1 (1.0-1.3)	51.9 (34.0-58.0)
PennCNV	61.1 (51.7-65.8)	64.5 (53.0-72.0)	56.9 (48.6-65.0)	1.1 (0.9-1.3)	48.1 (34.8-61.1)
QuantiSNP	48.1 (45.3-53.4)	49.0 (43.8-55.0)	44.1 (38.9-53.4)	1.1 (1.0-1.4)	41.2 (28.3-79.7)
R-gada	29.9 (19.5-40.2)	30.7 (20.1-43.2)	27.6 (18.9-36.1)	0.9 (0.8-1.2)	4.1 (1.1-16.0)
VEGA	46.0 (42.3-50.6)	45.6 (41.3-51.1)	49.2 (42.0-55.5)	0.9 (0.7-1.2)	46.2 (30.2-57.7)
Algorithmus					
HMM	57.1 (50.5-61.7)	57.0 (52.2-62.3)	55.6 (47.9-58.2)	1.1 (0.9-1.2)	50.1 (38.7-61.1)
Segmentation	44.9 (38.9-50.4)	45.2 (37.6-52.0)	45.5 (39.3-53.5)	1.0 (0.9-1.1)	44.7 (24.4-51.0)

Angegeben ist der Median und in Klammern der Interquartilsabstand (IQR)für die Proben der Kinder der 30 HapMap CEU Trios. **DDR:** Verhältnis von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*).

Appendix Tabelle 4: Validierungsraten in afrikanischen Proben

Software	Validated CNVs [%]	Validated deletions [%]	Validated duplications [%]	DDR, confined to validated CNVs	Validated cumulative sequence [%]
APT	53.1 (48.8-59.2)	53.7 (50.6-57.3)	62.6 (48.7-68.2)	0.9 (0.8-1.0)	56.7 (47.1-68.8)
GLAD	43.1 (36.5-49.3)	39.9 (34.5-49.5)	53.6 (40.8-58.5)	1.1 (1.0-1.5)	42.5 (28.2-57.1)
PennCNV	59.1 (53.0-64.5)	62.5 (54.2-68.9)	52.6 (45.2-60.4)	1.2 (1.0-1.5)	51.9 (40.4-63.6)
QuantiSNP	48.9 (45.1-53.5)	50.2 (46.7-53.5)	39.0 (33.3-51.6)	0.9 (0.8-1.0)	66.9 (38.2-86.0)
R-gada	37.0 (25.0-44.0)	35.6 (24.8-41.9)	36.7 (24.8-54.4)	0.7 (0.5-0.9)	5.7 (2.1-16.6)
VEGA	34.3 (24.6-39.5)	32.5 (22.6-37.1)	42.7 (29.9-51.8)	0.8 (0.6-1.0)	27.0 (15.6-49.6)
Algorithmus					
HMM	53.6 (48.8-56.4)	53.6 (48.8-57.7)	50.0 (43.4-58.5)	1.1 (1.0-1.4)	54.5 (48.7-68.7)
Segmentation	38.7 (33.5-43.9)	36.0 (31.4-40.7)	42.7 (34.4-53.4)	0.8 (0.7-0.9)	26.9 (15.8-44.6)

Angegeben ist der Median und in Klammern der Interquartilsabstand (IQR)für die Proben der Kinder der 30 HapMap YRI Trios. DDR: Verhältnis von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*).

Appendix Tabelle 5: Probenspezifische Eigenschaften der nicht validierten CNV-Vorhersagen

Software	Anzahl	Median der Länge [kb]	Median der kumulativen Länge [kb]	Median der Anzahl der Marker	Median der Distanz der Marker [kb]	DDR
APT	43.0 (34.8 - 57.2)	8.1 (6.4 - 10.0)	1.9 (1.4 - 3.3)	6.2 (4.0 - 10.0)	0.30 (0.22 - 0.37)	4.7 (3.3 - 6.5)
GLAD	92.0 (63.5 - 127.8)	6.6 (5.7 - 7.5)	3.3 (1.9 - 4.9)	4.0 (4.0 - 5.5)	0.30 (0.23 - 0.40)	2.0 (1.2 - 2.9)
PennCNV	29.0 (18.0 - 35.2)	26.2 (17.3 - 44.8)	2.0 (1.5 - 3.4)	23.0 (19.4 - 29.0)	0.25 (0.17 - 0.35)	4.8 (3.7 - 6.6)
QuantiSNP	75.0 (60.0 - 92.2)	7.9 (6.9 - 10.2)	3.0 (2.1 - 5.8)	4.0 (4.0 - 5.5)	0.31 (0.26 - 0.40)	2.9 (2.4 - 4.1)
R-gada	130.0 (107.5 - 169.2)	7.6 (6.2 - 8.9)	100.6 (16.4 - 266.8)	6.0 (5.0 - 8.1)	0.35 (0.31 - 0.44)	5.5 (3.9 - 7.4)
VEGA	89.5 (72.8 - 120.2)	6.0 (5.2 - 7.4)	3.6 (2.5 - 5.9)	5.0 (4.0 - 6.0)	0.37 (0.29 - 0.43)	4.5 (3.1 - 6.2)
Algorithmus						
HMM	43.0 (34.0 - 52.0)	9.3 (7.8 - 11.4)	2.2 (1.7 - 3.6)	7.0 (5.0 - 10.0)	0.29 (0.23 - 0.36)	4.3 (3.2 - 5.2)
Segmentation	103.0 (86.2 - 141.5)	6.7 (5.7 - 7.6)	4.2 (3.0 - 6.0)	5.0 (4.0 - 6.0)	0.35 (0.28 - 0.41)	4.3 (3.1 - 5.2)

Angegeben ist der Median und in Klammern der Interquartilsabstand. **DDR:** Verhältnis von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*).

Appendix Tabelle 6: Probenspezifische erweiterte Validierungsraten (Validierung durch alle Softwares)

Software	Validated CNVs [%]	Validated deletions [%]	Validated duplications [%]	DDR, confined to validated CNVs	Validated cumulative sequence [%]
APT	71.0 (65.3 - 75.4)	71.3 (65.0 - 75.5)	71.4 (61.4 - 83.8)	1.0 (0.8 - 1.2)	73.2 (59.7 - 83.1)
GLAD	63.1 (55.8 - 70.2)	61.8 (52.8 - 69.1)	71.4 (65.7 - 78.4)	1.2 (1.0 - 1.4)	67.3 (57.2 - 77.5)
PennCNV	76.7 (73.0 - 81.8)	82.0 (76.6 - 85.8)	69.8 (58.9 - 76.6)	1.1 (0.9 - 1.2)	72.5 (57.8 - 79.8)
QuantiSNP	63.6 (59.3 - 69.0)	64.9 (58.1 - 69.6)	59.5 (50.0 - 66.7)	1.1 (0.9 - 1.2)	69.6 (42.7 - 83.7)
R-gada	45.0 (31.9 - 52.3)	45.5 (31.8 - 52.8)	41.4 (28.4 - 56.7)	0.9 (0.7 - 1.0)	6.5 (1.9 - 23.2)
VEGA	54.1 (50.5 - 60.2)	52.5 (46.8 - 59.6)	62.2 (53.7 - 72.1)	0.9 (0.7 - 1.0)	59.9 (42.0 - 71.3)
Algorithmus					
HMM	70.5 (67.0 - 75.4)	71.4 (66.1 - 75.5)	68.0 (58.3 - 72.3)	1.1 (0.9 - 1.2)	72.0 (59.7 - 79.8)
Segmentation	53.9 (50.9 - 60.4)	52.7 (47.9 - 59.6)	61.8 (55.1 - 70.4)	0.9 (0.8 - 1.0)	57.8 (35.9 - 66.5)

Angegeben ist der Median und in Klammern der Interquartilsabstand. **DDR:** Verhältnis von Deletionen zu Duplikationen (engl. *deletions-to-duplications ratio*).

Appendix Tabelle 7: Anteil paarweise konkordanter CNV-Vorhersagen

	Vorhersage						
Verifikation	**APT**	**GLAD**	**PennCNV**	**QuantiSNP**	**R-gada**	**VEGA**	**3+**
Grenzwert zu Verifikation: 95%							
APT	-	45.8	42.9	55.6	44.5	64.8	74.0
GLAD	60.8	-	48.9	56.2	55.8	57.4	80.0
PennCNV	52.8	41.9	-	66.7	40.3	44.0	71.4
QuantiSNP	34.9	30.4	38.1	-	35.5	35.2	52.8
R-gada	36.6	38.6	30.1	44.8	-	51.0	56.5
VEGA	56.5	42.7	33.3	48.4	61.5	-	67.4
3+	74.3	71.9	59.9	75.9	85.7	82.8	-
Grenzwert zu Verifikation: 90%							
APT	-	49.6	44.3	58.1	45.8	66.4	74.2
GLAD	61.3	-	50.3	58.1	56.2	58.7	80.0
PennCNV	55.4	47.9	-	68.3	42.5	46.1	72.3
QuantiSNP	36.2	33.0	38.9	-	36.2	36.5	52.8
R-gada	38.8	41.3	32.5	46.0	-	51.4	56.5
VEGA	59.2	45.9	34.1	50.0	62.4	-	67.5
3+	77.3	79.3	63.1	78.8	90.4	87.2	-
Grenzwert zu Verifikation: 80%							
APT	-	54.0	45.9	60.9	47.5	69.9	76.7
GLAD	62.6	-	51.4	59.2	57.3	61.3	80.7
PennCNV	58.0	54.6	-	70.7	46.7	51.0	75.4
QuantiSNP	38.5	36.4	39.0	-	38.9	39.6	55.0
R-gada	39.9	43.4	33.7	48.8	-	52.0	57.3
VEGA	61.1	49.6	36.4	50.6	62.4	-	69.4
3+	78.5	83.2	64.2	81.4	91.8	90	-
Grenzwert zu Verifikation: 50%							
APT	-	59.9	49.7	62.9	50.7	72.7	78.3
GLAD	63.8	-	53.3	61.0	59.2	63.0	82.8
PennCNV	62.2	61.1	-	74.7	50.5	55.4	80.0
QuantiSNP	41.1	41.0	40.0	-	41.6	42.6	57.4
R-gada	40.6	47.3	36.8	50.8	-	52.1	58.3
VEGA	62.7	54.4	37.9	53.1	63.0	-	71.0
3+	80.3	88.2	66.7	84.3	94.1	92.8	-

Median über den Anteil an verifizierten CNVs je Probe. **3+:** Konsensus-Regionen die durch mindestens drei Softwares als CNV vorhergesagt wurden.

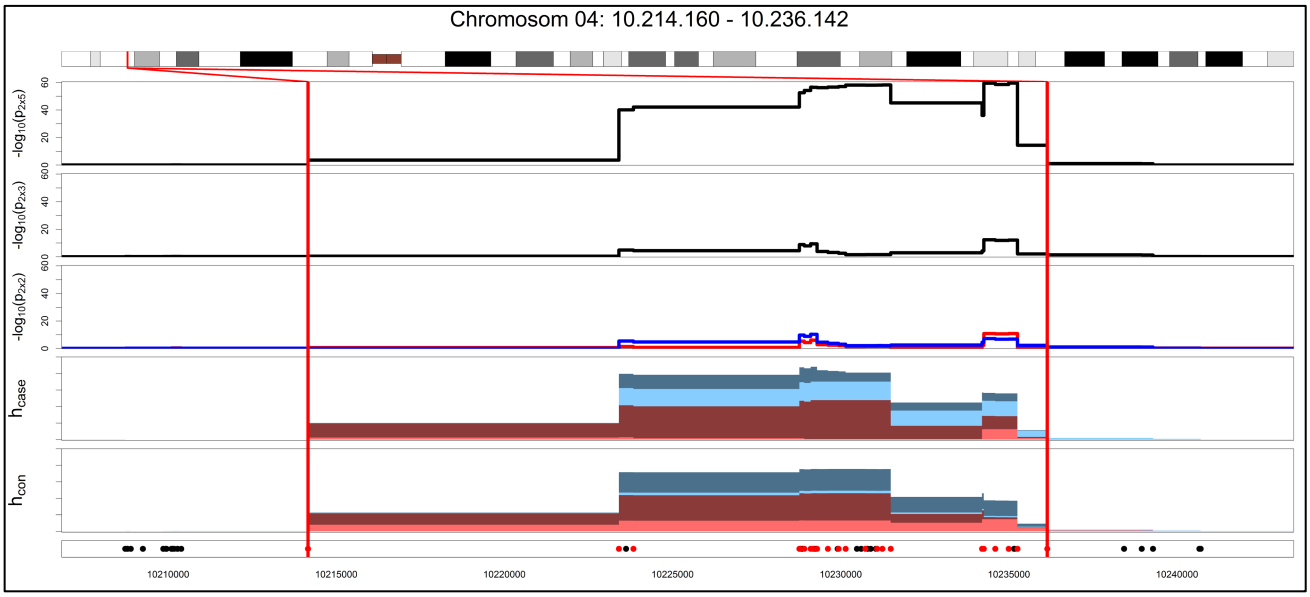
B.2 Assoziationsanalyse

Appendix Tabelle 8: Kandidaten-Regionen

#	Chr.	Start	Ende	Länge	n_M	$\tilde{d}_M$	p_{2x5}^{min}	p_{2x5}^{sum}	p_{2x3}^{min}	p_{2x3}^{sum}	p_{del}^{min}	p_{del}^{min}	$\tilde{h}_n$
1	4	10.214.160	10.236.142	21.983	36	124	3,45E-60	36	4,41E-13	35	1.27E-11	3.98E-11	0,773
2	1	112.695.674	112.697.378	1.705	25	35	3,94E-43	25	2,47E-45	23	8.69E-24	9.14E-28	0,229
3	16	78.372.428	78.384.753	12.326	17	302	8,32E-43	17	1,20E-44	17	1.36E-26	1.04E-28	0,343
4	2	34.699.315	34.738.020	38.706	49	83	6,90E-39	44	4,34E-40	1	2.75E-05	1.78E-38	0,515
5	1	152.555.176	152.586.582	31.407	36	204	4,28E-23	1	4,06E-25	2	8.51E-25	2.15E-03	0,551
6	8	24.974.431	24.984.333	9.903	26	55	3,69E-18	26	3,34E-20	6	4.87E-11	2.63E-12	0,369
7	11	4.968.117	4.978.382	10.266	47	37	6,13E-15	39	2,72E-02	6	7.73E-02	1.17E-02	0,277
8	6	103.737.964	103.762.049	24.086	46	65	3,03E-11	46	2,72E-01	0	3.22E-01	2.97E-01	0,499
9	1	196.725.233	196.816.718	91.486	103	128	8,52E-10	1	2,87E-11	4	1.52E-06	2.17E-07	0,146
10	3	162.512.645	162.626.591	113.947	74	455	1,65E-09	74	3,79E-06	72	4.96E-05	9.40E-05	0,440
11	12	9.633.858	9.731.637	97.780	26	2.177	1,23E-08	2	6,15E-10	2	4.15E-04	6.07E-09	0,516
12	9	44.201.245	44.855.725	654.481	75	2.897	1,69E-08	54	1,31E-03	39	6.02E-04	1.51E-02	0,256
13	6	32.452.789	32.570.353	117.565	33	2.171	2,52E-08	27	8,38E-08	26	1.50E-04	1.33E-06	0,283
14	11	55.374.020	55.453.564	79.545	64	484	4,35E-08	50	1,08E-07	59	1.71E-05	2.85E-05	0,283
15	5	57.326.015	57.340.073	14.059	33	91	1,02E-07	2	5,67E-09	2	1.90E-04	4.60E-07	0,438
16	4	34.778.848	34.824.713	45.866	45	200	1,19E-07	38	4,31E-04	36	8.45E-05	3.35E-02	0,473
17	6	78.969.053	79.035.173	66.121	114	79	3,31E-07	114	1,44E-06	114	3.01E-02	3.00E-07	0,610
18	1	152.761.911	152.768.688	6.778	36	44	4,65E-07	36	7,17E-03	36	2.75E-02	6.14E-03	0,417
19	6	77.439.868	77.452.804	12.937	22	100	6,22E-07	10	1,10E-06	13	4.80E-07	1.21E-01	0,145
20	13	38.072.024	38.084.745	12.722	15	203	1,91E-06	15	1,42E-07	15	6.66E-01	1.96E-08	0,160
21	7	133.785.165	133.806.244	21.080	39	41	2,46E-06	26	2,03E-07	30	9.88E-03	4.88E-08	0,279
22	6	67.785.652	67.787.996	2.345	12	190	5,12E-06	12	2,31E-07	12	2.14E-08	7.60E-01	0,390
23	7	126.045.470	126.046.869	1.400	17	56	9,78E-06	17	2,36E-03	11	1.15E-03	1.19E-01	0,132
24	14	74.001.111	74.020.967	19.857	9	1.665	1,00E-05	9	1,70E-06	9	4.60E-04	3.13E-05	0,143
25	20	1.557.178	1.593.881	36.704	67	104	1,13E-05	7	4,99E-07	16	1.05E-06	4.71E-03	0,524
26	15	76.891.229	76.895.763	4.535	21	68	3,02E-05	10	7,92E-02	0	1.16E-01	4.06E-02	0,482
27	7	142.476.707	142.486.548	9.842	32	72	3,68E-05	32	1,29E-05	32	7.14E-04	1.93E-05	0,409
28	3	129.762.847	129.808.799	45.953	64	166	7,52E-05	56	8,26E-06	56	5.04E-04	4.00E-05	0,399
29	2	146.864.392	146.866.922	2.531	23	85	5,34E-04	23	2,65E-01	0	2.93E-01	2.41E-01	0,414
30	8	39.235.591	39.388.287	152.697	60	680	7,89E-04	59	8,96E-02	0	4.99E-02	6.47E-02	0,478
31	18	38.260.408	38.265.377	4.970	18	14	8,33E-04	18	4,69E-02	7	4.02E-02	2.18E-01	0,177
32	4	64.697.457	64.708.246	10.790	29	71	8,48E-04	7	9,65E-02	0	1.12E-01	4.53E-02	0,122
33	7	97.395.448	97.402.463	7.016	15	23	8,55E-04	7	8,12E-05	7	9.02E-02	5.02E-05	0,650
34	19	35.855.341	35.861.836	6.496	21	76	9,76E-04	6	5,53E-03	6	3.88E-03	1.32E-01	0,128
35	14	19.422.521	20.423.360	1.000.840	197	1.400	9,95E-04	119	3,18E-04	143	6.08E-05	3.29E-02	0,225
36	7	109.433.735	109.454.259	20.525	41	14	1,01E-03	12	1,13E-03	1	1.68E-03	4.05E-02	0,170
37	5	180.378.698	180.418.091	39.394	18	1.102	2,08E-03	18	3,54E-03	18	6.56E-01	7.82E-04	0,184
38	6	31.286.264	31.296.439	10.176	11	253	2,23E-03	11	1,01E-03	11	1.46E-02	3.85E-03	0,570
39	7	64.594.053	64.595.686	1.634	5	226	3,44E-03	5	3,39E-01	0	7.73E-01	1.56E-01	0,133
40	17	39.418.360	39.430.519	12.160	26	47	3,64E-03	3	3,80E-04	14	4.11E-03	3.33E-03	0,488
41	4	115.175.311	115.182.163	6.853	22	40	5,85E-03	22	6,21E-02	0	1.77E-01	2.99E-02	0,332
42	14	106.530.352	106.569.451	39.100	18	1.614	5,94E-03	6	4,11E-02	1	1.34E-02	1.77E-01	0,344

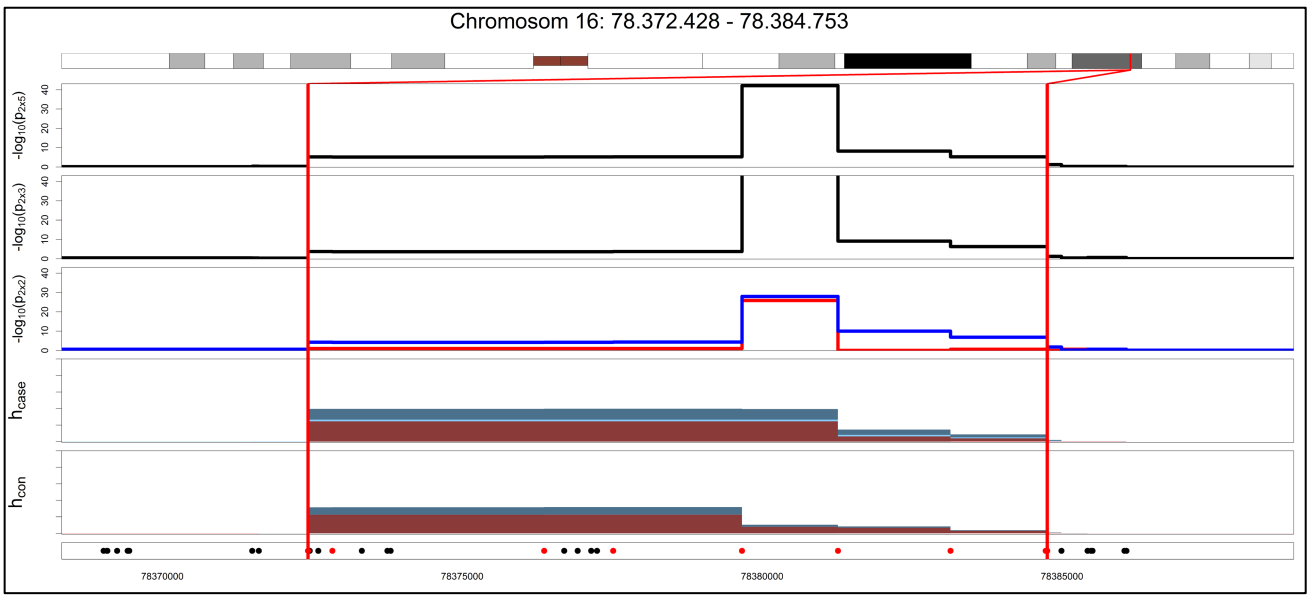
43	10	67.307.911	67.314.434	6.524	27	50	6,05E-03	27	1,29E-01	0	1.74E-01	9.51E-02	0,129
44	22	16.102.330	16.248.694	146.365	32	128	6,08E-03	16	4,10E-04	32	1.69E-02	9.09E-04	0,480
45	4	161.880.006	161.885.021	5.016	26	84	7,28E-03	14	4,01E-02	13	1.47E-02	3.70E-01	0,123
46	19	53.519.526	53.547.349	27.824	20	843	7,76E-03	20	2,86E-02	14	7.27E-01	8.65E-03	0,410
47	2	98.140.735	98.162.177	21.443	8	2.731	8,60E-03	8	3,93E-03	8	9.44E-04	5.91E-01	0,800
48	11	4.250.001	4.328.160	78.160	26	475	1,21E-02	12	3,43E-01	0	1.53E-01	6.38E-01	0,610
49	16	55.796.377	55.822.432	26.056	21	416	1,72E-02	20	3,10E-02	11	1.02E-02	2.06E-01	0,143
50	3	192.877.890	192.882.891	5.002	19	110	1,80E-02	17	2,69E-02	19	1.46E-01	3.07E-02	0,285
51	10	46.917.303	47.704.625	787.323	404	263	1,90E-02	21	1,84E-02	3	4.32E-02	2.61E-02	0,990
52	14	106.874.349	106.918.637	44.289	53	214	2,04E-02	21	5,28E-03	53	2.83E-02	4.28E-03	0,271
53	14	106.922.519	106.931.353	8.835	4	550	2,17E-02	3	1,10E-01	0	4.33E-02	6.01E-01	0,440
54	11	18.949.060	18.962.351	13.292	68	101	2,50E-02	3	8,62E-03	7	2.19E-03	2.21E-01	0,182
55	8	11.987.714	12.443.043	455.330	129	395	2,57E-02	11	1,02E-02	22	2.84E-02	7.97E-03	0,109
56	17	18.355.380	18.474.634	119.255	31	2.870	2,62E-02	2	7,01E-03	1	2.80E-03	2.43E-01	0,430
57	18	63.907.422	63.911.577	4.156	11	354	2,76E-02	11	2,06E-01	0	8.30E-02	7.94E-01	0,520
58	1	25.593.116	25.663.332	70.217	34	504	3,23E-02	3	1,40E-02	5	9.68E-02	4.09E-03	0,371
59	6	29.871.636	29.904.865	33.230	20	448	3,57E-02	7	1,65E-01	0	6.23E-02	4.87E-01	0,130
60	6	67.008.811	67.048.629	39.819	57	114	3,89E-02	1	5,91E-03	1	1.24E-03	7.21E-01	0,149
61	14	106.783.032	106.821.724	38.693	11	1.328	4,07E-02	1	4,98E-01	0	2.68E-01	5.51E-01	0,123
62	2	35.977.778	35.987.935	10.158	7	504	4,21E-02	7	7,90E-03	7	4.02E-03	2.66E-01	0,420
63	2	52.754.456	52.781.530	27.075	39	55	4,50E-02	8	8,52E-02	0	6.01E-01	2.69E-02	0,502
64	12	869.097	874.017	4.921	17	77	4,62E-02	1	1,08E-02	17	2.87E-03	7.09E-01	0,174
65	1	72.768.406	72.811.136	42.731	47	215	4,94E-02	2	2,82E-01	0	4.58E-01	1.78E-01	0,524

$\boldsymbol{n_M}$**:** Anzahl der Marker, $\boldsymbol{\tilde{d}_M}$**:** Median des Abstands der Marker, $\boldsymbol{p_{2x5}^{min}}$**:** Minimum aller p-Werte der in der Region durchgeführten 2×5-χ^2-Tests, $\boldsymbol{p_{2x5}^{sum}}$**:** Anzahl der Marker mit lokal signifikantem p-Wert ($p_{2x5} \leq 0{,}05$), $\boldsymbol{p_{2x3}^{min}}$**:** Minimum aller p-Werte der in der Region durchgeführten 2×3-χ^2-Tests, $\boldsymbol{p_{2x3}^{sum}}$**:** Anzahl der Marker mit lokal signifikantem p-Wert ($p_{2x3} \leq 0{,}05$), $\boldsymbol{p_{del}^{min}}$**:** Minimum aller p-Werte der in der Region durchgeführten 2×2-χ^2-Tests in Bezug auf Deletionen, $\boldsymbol{p_{del}^{min}}$**:** Minimum aller p-Werte der in der Region durchgeführten 2×2-χ^2-Tests in Bezug auf Duplikationen, $\boldsymbol{\tilde{h}_n}$**:** Median der relativen Häufigkeit von CNV-Vorhersagen ($cn \neq 2$)



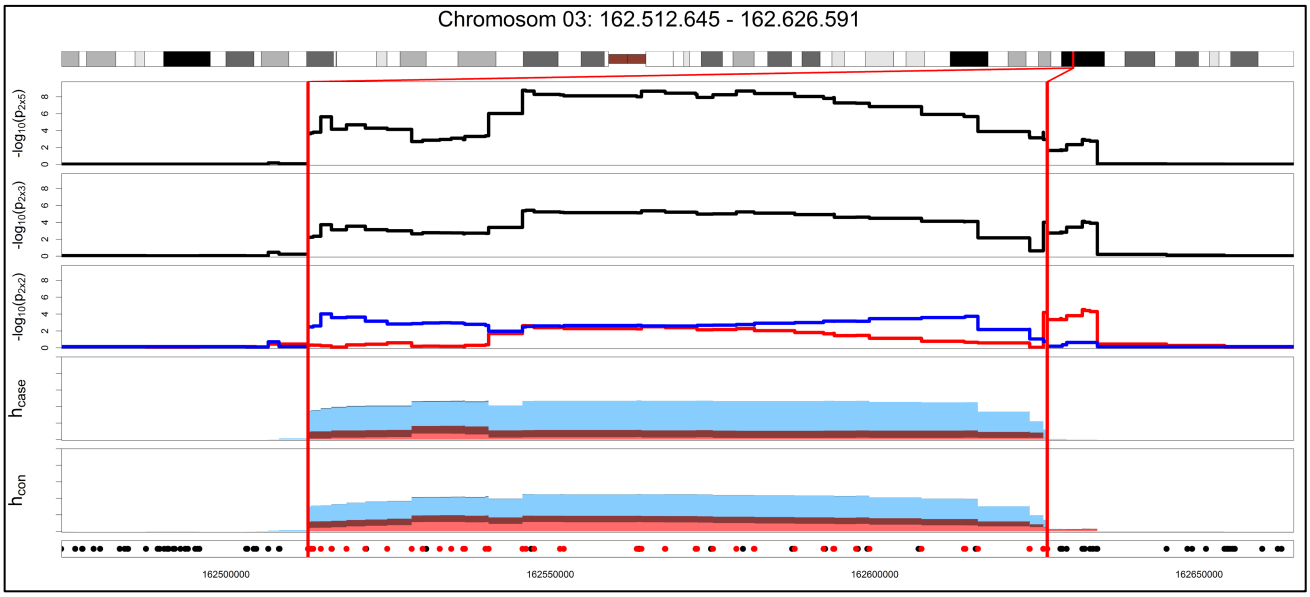
Appendix Abbildung 1: Kandidaten-Region #1

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



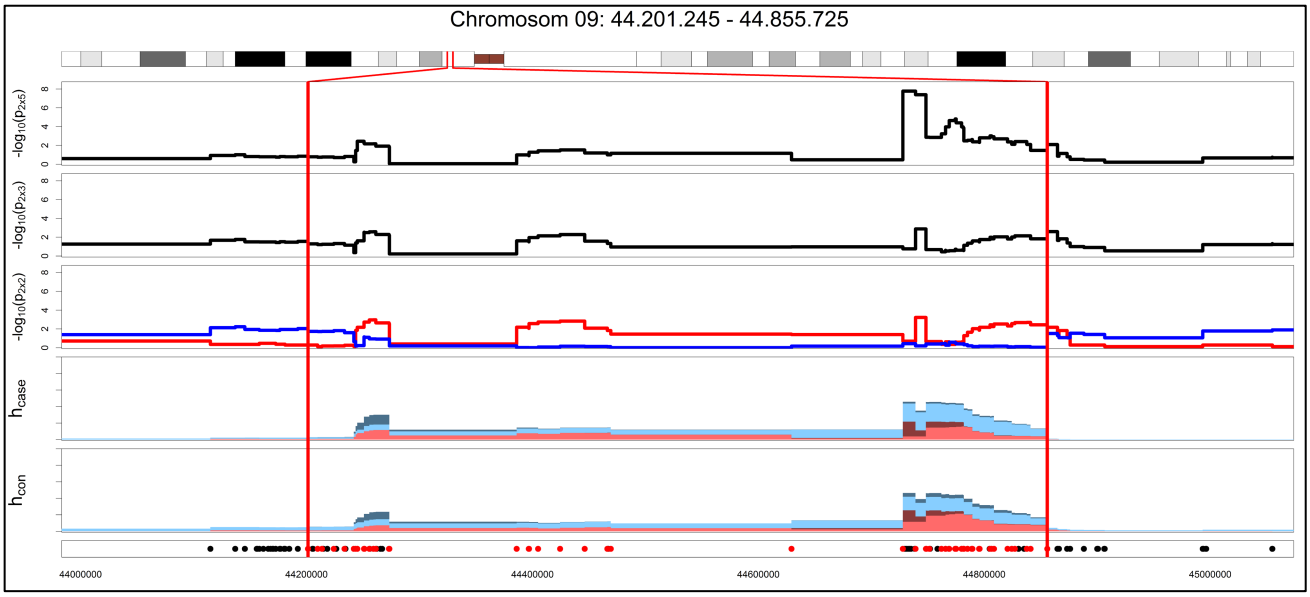
Appendix Abbildung 2: Kandidaten-Region #3

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



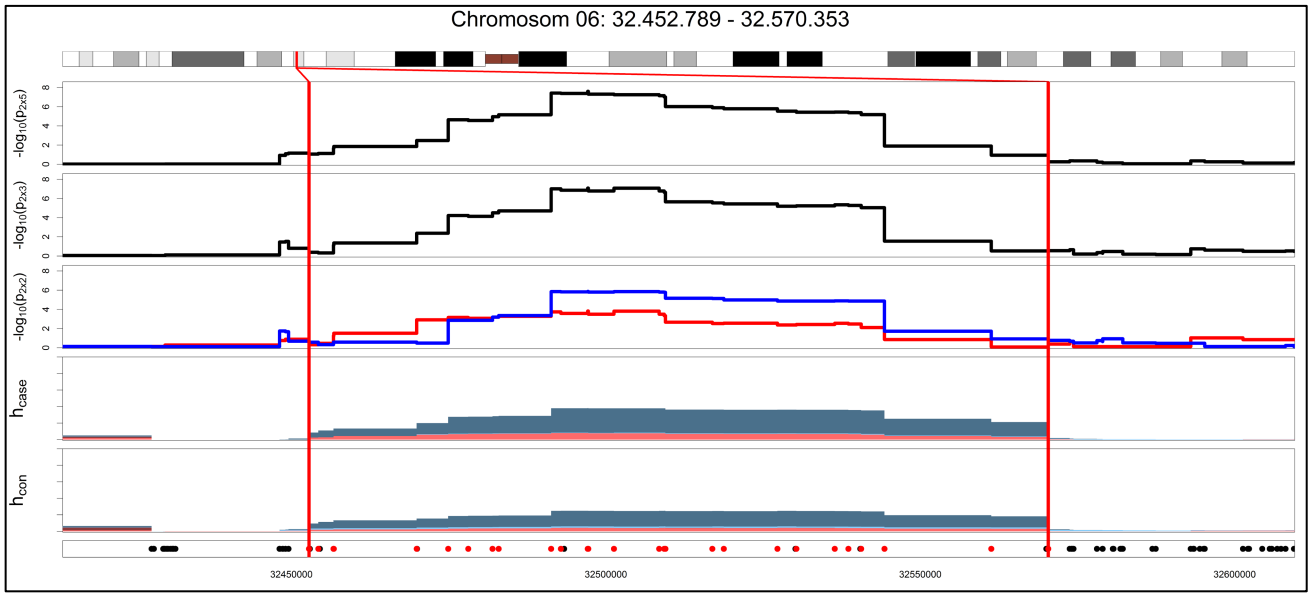
Appendix Abbildung 3: Kandidaten-Region #10

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



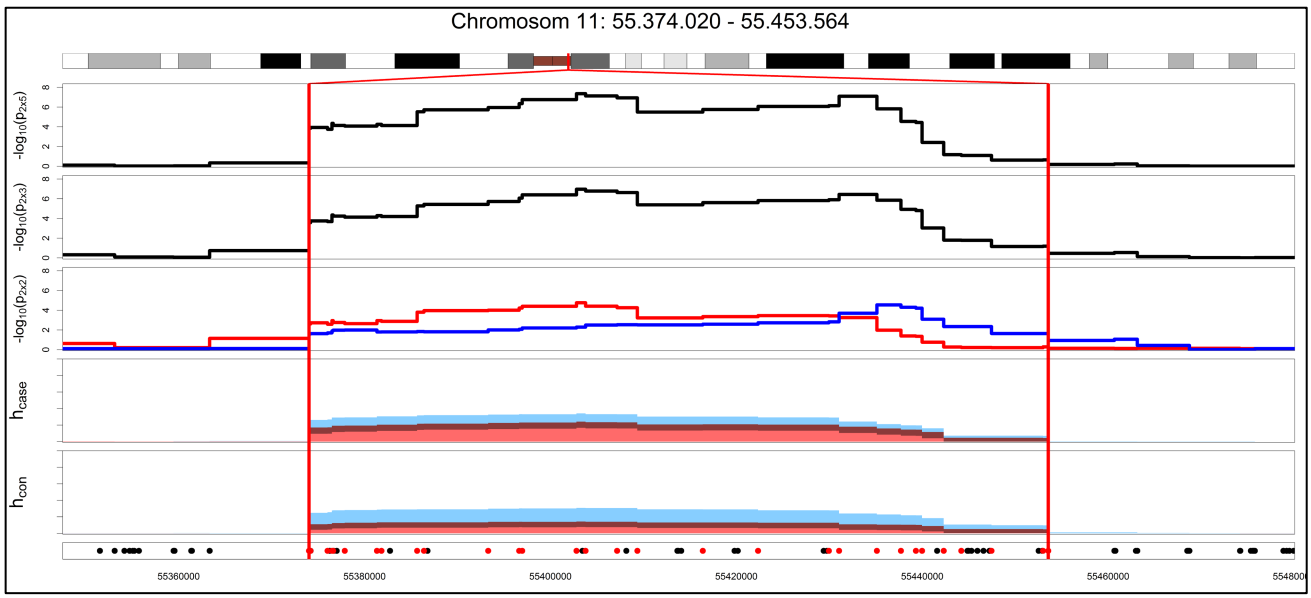
Appendix Abbildung 4: Kandidaten-Region #12

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



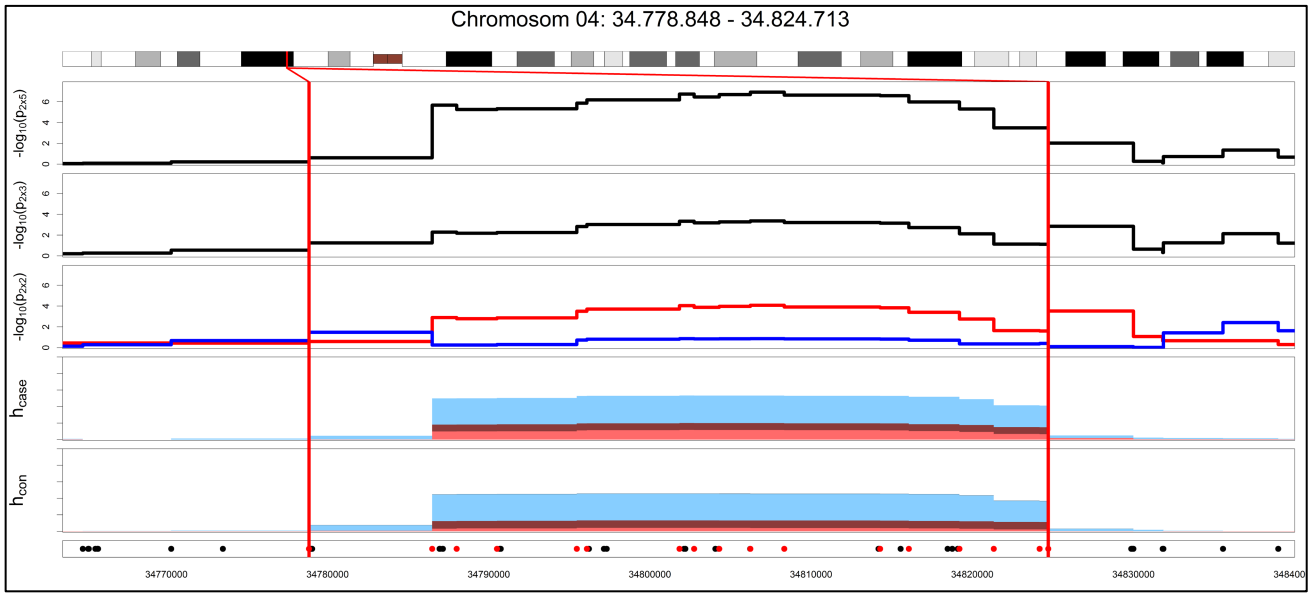
Appendix Abbildung 5: Kandidaten-Region #13

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



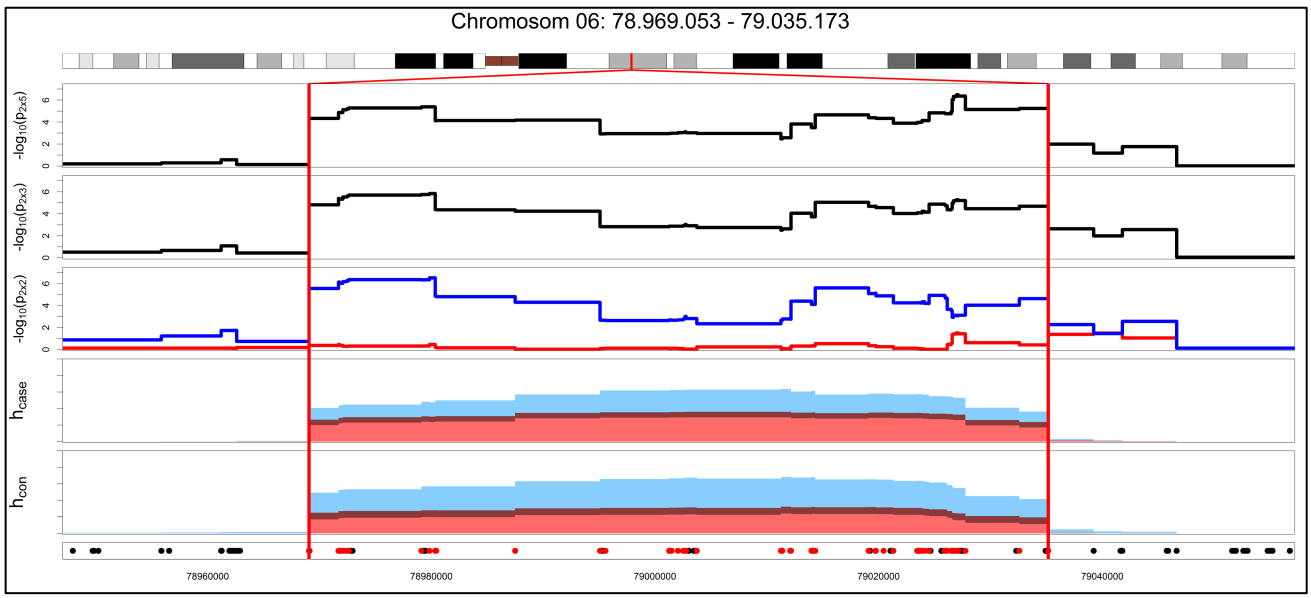
Appendix Abbildung 6: Kandidaten-Region #14

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



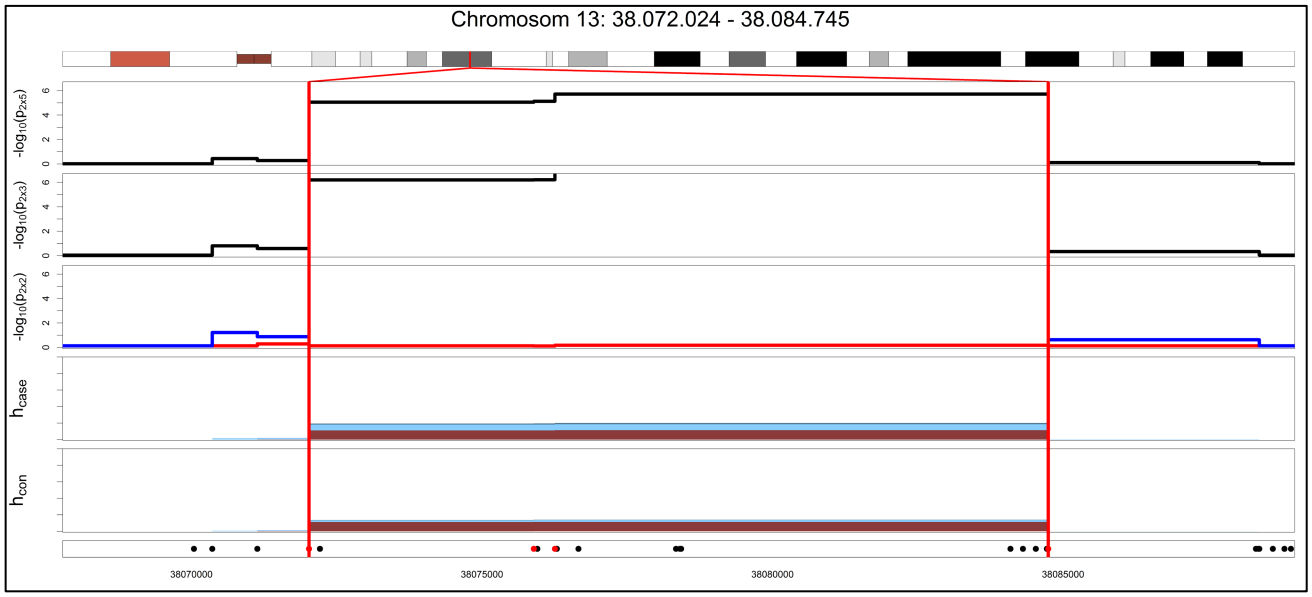
Appendix Abbildung 7: Kandidaten-Region #16

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



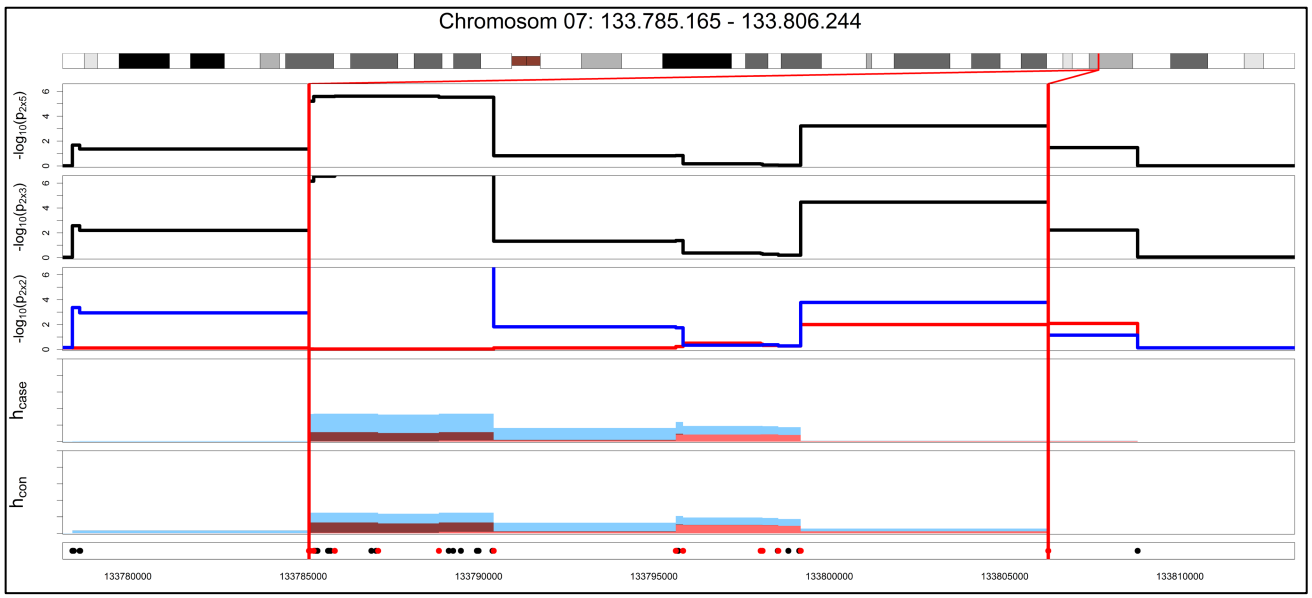
Appendix Abbildung 8: Kandidaten-Region #17

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



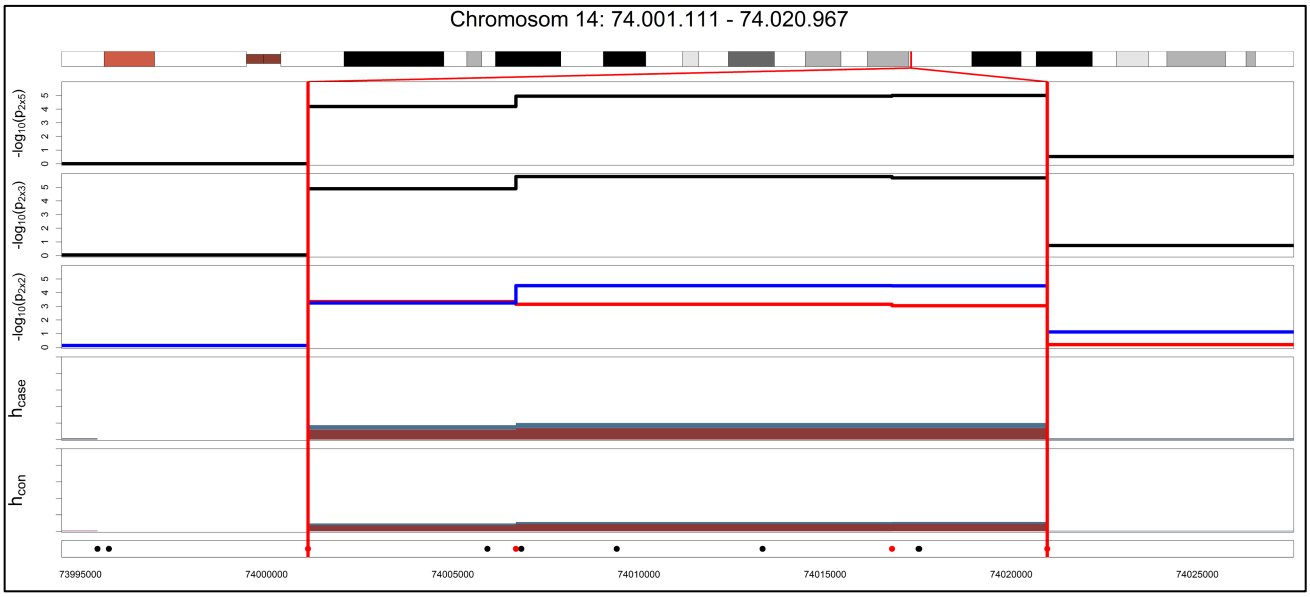
Appendix Abbildung 9: Kandidaten-Region #20

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



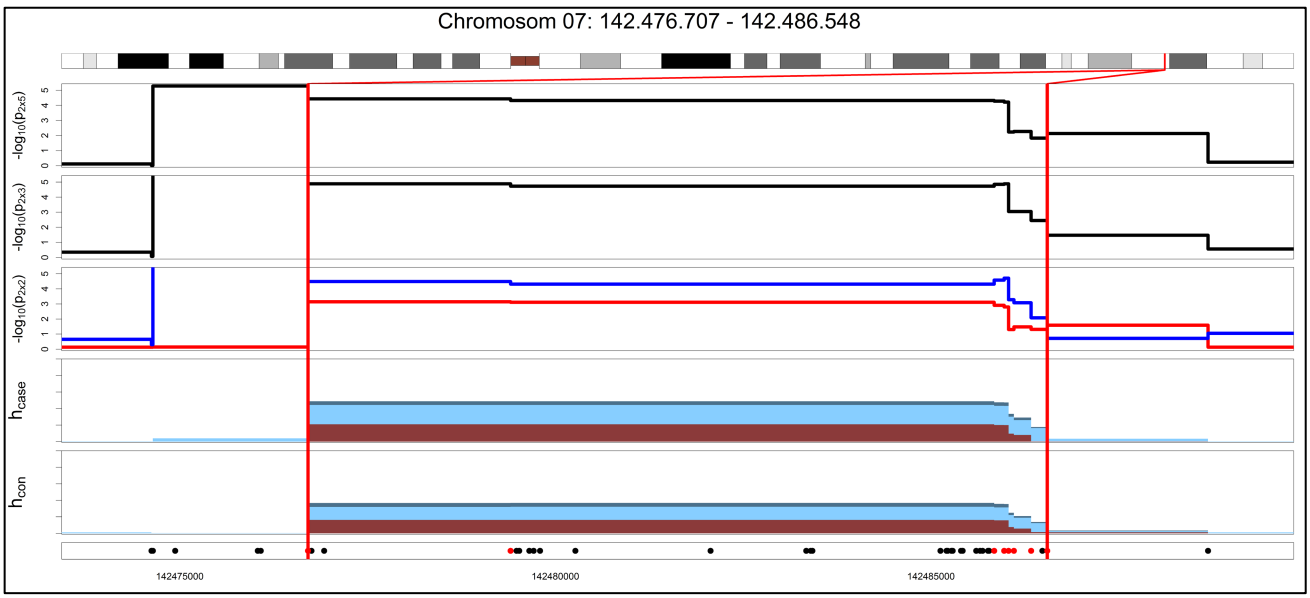
Appendix Abbildung 10: Kandidaten-Region #21

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



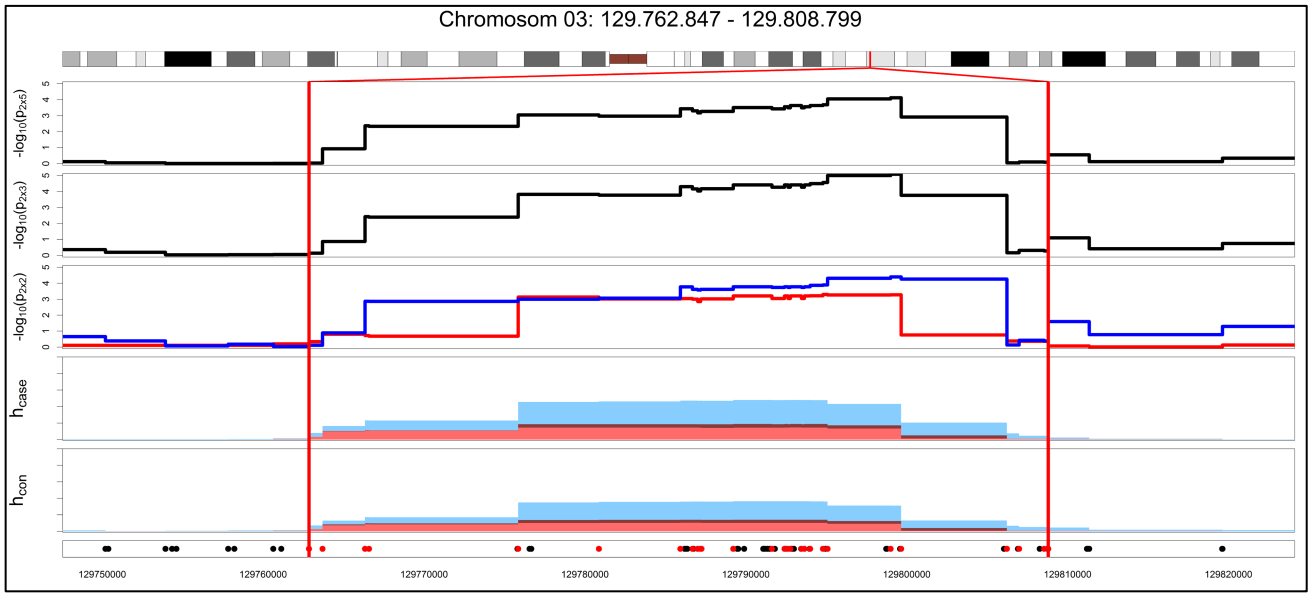
Appendix Abbildung 11: Kandidaten-Region #24

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



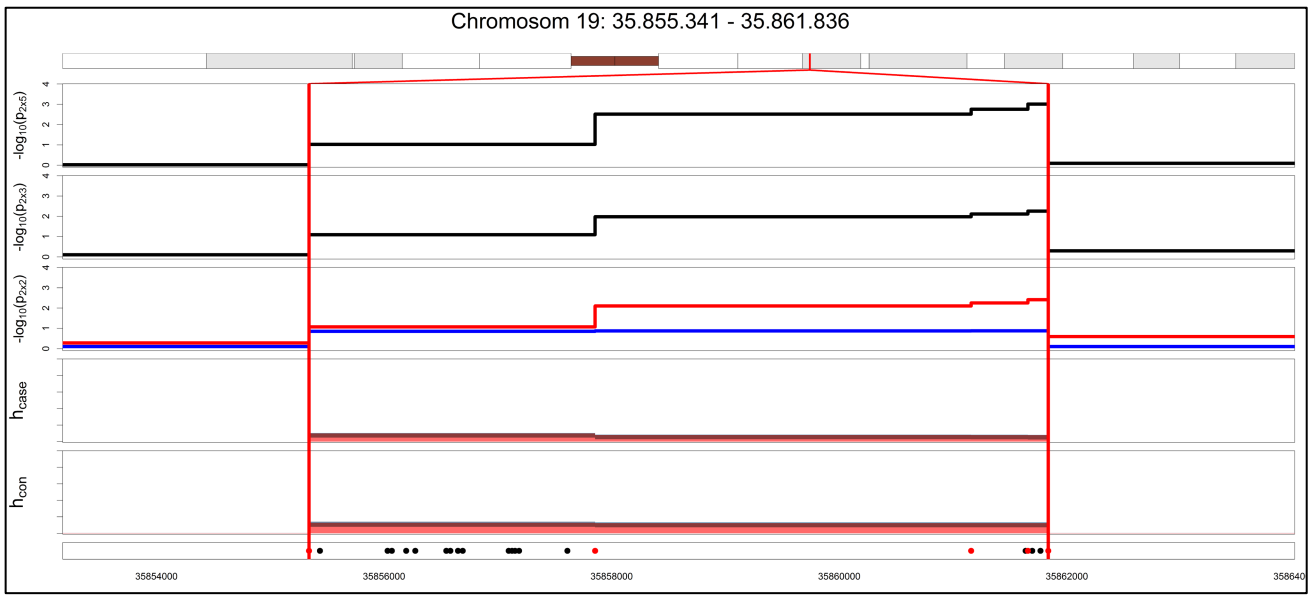
Appendix Abbildung 12: Kandidaten-Region #27

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



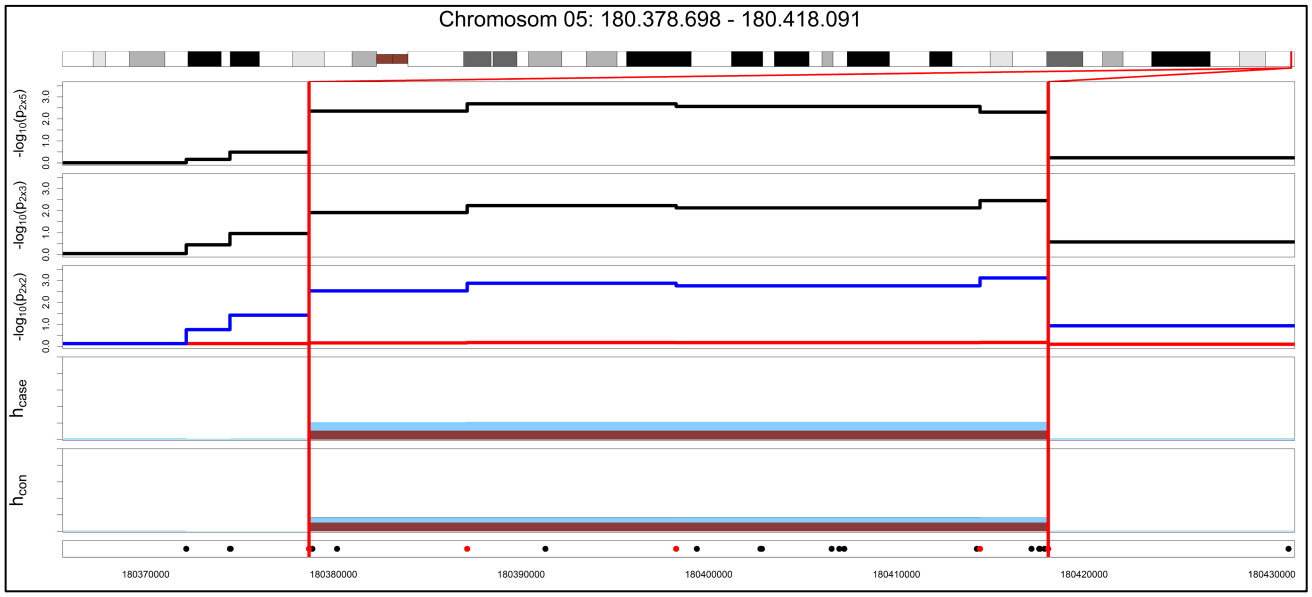
Appendix Abbildung 13: Kandidaten-Region #28

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



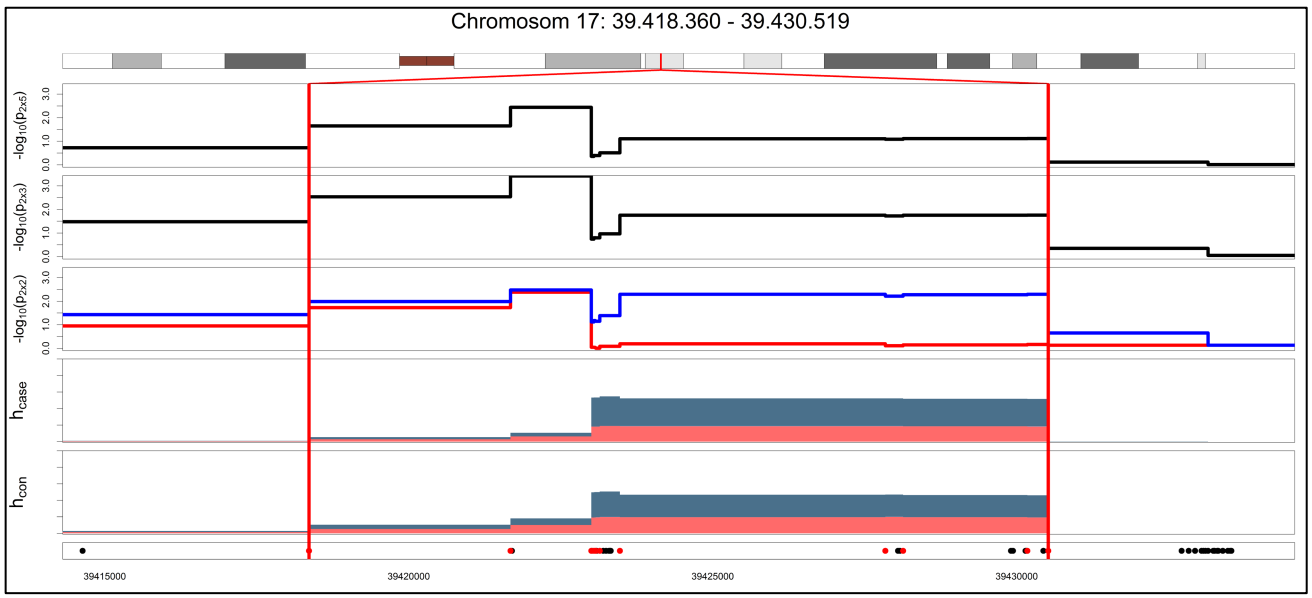
Appendix Abbildung 14: Kandidaten-Region #34

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



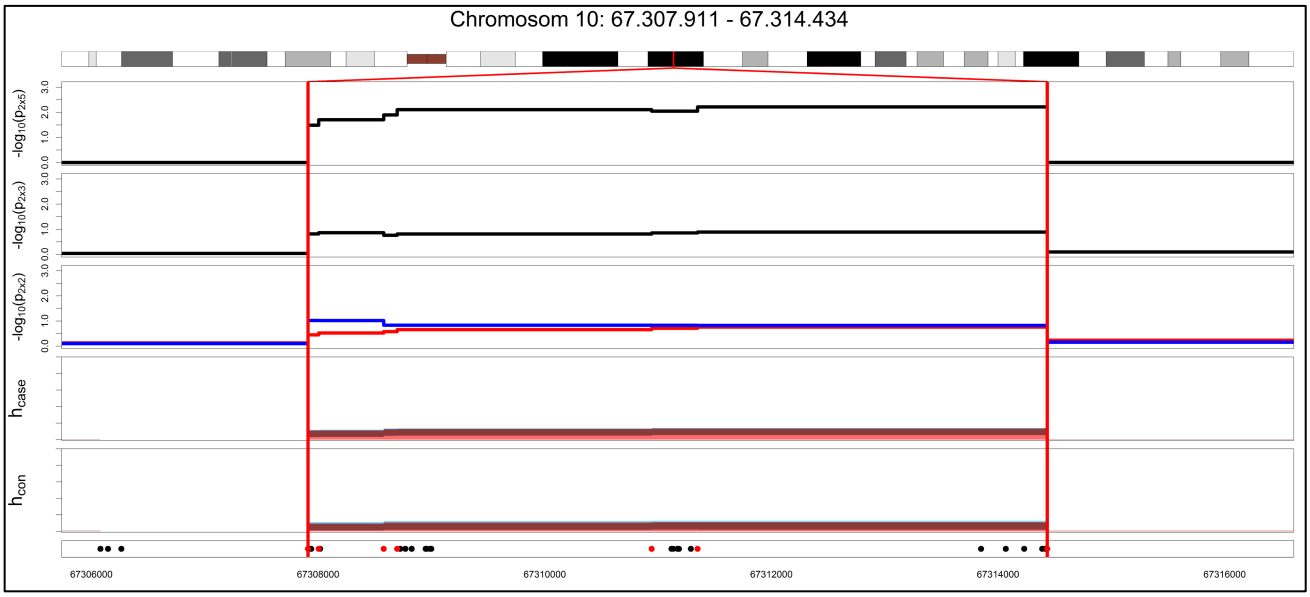
Appendix Abbildung 15: Kandidaten-Region #37

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



Appendix Abbildung 16: Kandidaten-Region #40

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).



Appendix Abbildung 17: Kandidaten-Region #43

$-\log_{10}(p)$**:** negativer dekadischer Logarithmus der p-Werte. p_{2x5}**:** Test der CN-Genotypen. p_2x3**:** Test der CN-Klassen. p_2x2**:** Test der Deletionen (rot) und Duplikationen (blau) h**:** relative Häufigkeiten. $case$**:** Fälle. con**:** Kontrollen. **Farbgebung:** einfache Deletion (hellrot), doppelte Deletion (dunkelrot), einfache Duplikation (hellblau), mehrfache Duplikation (dunkelblau).

C Lebenslauf

Persönliche Daten

Name:	**Marcel Elie Kokou Nutsua**
Geburtsdatum:	17. Oktober 1984
Geburtsort:	Hamburg

Ausbildung & Studium

10/2010 - heute	**Promotion am Institut für Klinische Molekularbiologie, Christian-Albrechts Universität zu Kiel** Titel der Dissertation: "Identification of copy-number variations associated with the complex inflammatory lung disease sarcoidosis." Stipendium über den *Exzellenzcluster Entzündungsforschung*
10/2005 - 12/2010	**Studium der Mathematik an der Universität Bremen** Nebenfach: Philosophie Abschluss: Diplom (Gut) Titel der Diplomarbeit: "Analyse varianzanalytischer Verfahren zur Feststellung familiärer Aggregation" Betreuung: Prof. Dr. Pigeot
09/2005 - 03/2006	**Grundwehrdienst**, Wachsoldat, 1./LogBtl 161 in Delemenhorst und als Teil der *German Force Protection* in Mannheim
06/2005 - 09/2005	**Grundwehrdienst,** Grundausbildung, 8./LwAusbRgt 1 in Goslar
08/2001 - 07/2004	**Allgemeine Hochschulreife** Leistungskurse: Mathematik & Physik Abschlussnote: Gut Schulzentrum Walliser Straße, Bremen

Akademische Tätigkeiten & Forschungserfahrung

01/2014 – 08/2014	**Wissenschaftliche Hilfskraft**, Institut für Klinische Molekularbiologie, Christian-Albrechts Universität zu Kiel Arbeitsgruppe *Ancient DNA*
10/2009 - 09/2010	**Studentische wissenschaftliche Hilfskraft**, Bremer Institut für Präventionsforschung und Sozialmedizin (BIPS) Abteilung für Biometrie und EDV Forschungsgruppe *Genetische Epidemiologie und Bioinformatik*

01/2008 - 10/2008	**Studentische wissenschaftliche Hilfskraft**, STAr Group Bremen (Kooperationsprojekt der Universität Bremen und der Europäischen Weltraumorganisation)
05/2006 - 10/2009	**Technische Hilfskraft für Software-Entwicklung,** Rheinmetall Defence Electronics

Veröffentlichungen

Nutsua M, Fischer A, Nebel A, Hofmann S, Schreiber S, Krawczak M, Nothnagel M (2014) Family-based benchmarking of copy number variation detection software. (Manuskript zur Veröffentlichung eingereicht)

Teilnahme an Kursen und Konferenzen

08/2014	**International Symposium on Biomolecular Archeology (ISBA),** Basel, Schweitz
04/2014	**European Mathematical Genetics Meeting (EMGM),** Köln
04/2013	**European Mathematical Genetics Meeting (EMGM),** Leiden, Niederlande Poster-Präsentation: “Family-based benchmarking of copy-number variation detection software”
04/2012	**Training in Genetic Epidemiology,** Universität zu Lübeck
04/2012	**European Mathematical Genetics Meeting (EMGM),** Göttingen Poster-Präsentation: “A comparison of CNV calling algorithms and analysis software”
09/2011	**International Genetic Epidemiology Society (IGES) Conference,** Heidelberg Poster-Präsentation: “A comparison of CNV calling algorithms and analysis software”

D Eidesstattliche Erklärung

Hiermit versichere ich, Marcel Nutsua, an Eides statt, dass ich die vorliegende Arbeit 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, Prof. Dr. Michael Nothnagel und Prof. Dr. Hinrich Schulenburg, in Anspruch genommen. Die Arbeit wurde bis jetzt weder vollständig noch in Teilen an anderer Stelle im Rahmen eines Prüfungsverfahren vorgelegt. Zudem erkläre ich, keine früheren Promotionsversuche unternommen zu haben. Für Teile dieser wurde Arbeit das Manuskript „Family-based benchmarking of copy number variation detection software“ angefertigt und zur Veröffentlichung eingereicht.

Kiel,

Marcel Nutsua

E Danksagung

Mein Dank gilt allen, die mich bei der Entstehung dieser Arbeit begleitet und unterstützt haben. Zuerst möchte ich daher Prof. Dr. Schreiber dafür danken, dass ich die Möglichkeit hatte die dieser Arbeit zugrunde liegende Studie am Institut für Klinische Molekularbiologie (IKMB) an der Christian-Albrechts-Universität zu Kiel durchzuführen. Auch bei Prof. Dr. Michael Nothnagel, Prof. Dr. Schulenburg und Prof. Dr. Almut Nebel möchte ich mich für die Betreuung, das Interesse an meinem Thema sowie der konstruktiven Kritik und reichlich Unterstützung bei meiner Arbeit bedanken.

Bei Dr. Annegret Fischer, Dr. Ben Krause-Kyora und Dr. Friderike Flachsbart bedanke ich mich ausdrücklich für den fachlichen Rat und ihre Hilfe bei der Lösung von Problemen jeglicher Art. Ihr großes Engagement sowie ihre wertvollen Anregungen und Hinweise haben maßgeblich zum erfolgreichen Verlauf dieser Arbeit beigetragen. Ein ganz herzliches Dankeschön auch für die sehr intensive und kritische Durchsicht des Manuskripts!

Des Weiteren möchte ich mich bei Prof. Dr. Michael Krawczak für die wertvolle Unterstützung bei der Projektplanung bedanken sowie bei Sanaz Sedghpour-Sabet, Tanja Wesse, Anja Eggert und allen Mitarbeitern in den Laboren für die ausgezeichnete Zusammenarbeit. Für eine angenehme Atmosphäre und viel Spaß bei der Arbeit möchte ich mich bei Alex, Benni, Lilli, Gregor, Caro, Jojo und all meinen Kollegen am IKMB bedanken.

Abschließend möchte ich den wichtigsten Menschen in meinem Leben für Beistand, Aufmunterung und Motivation in allen Lebenslagen danken. Für die beste Freundschaft die ich mir vorstellen kann danke ich Tom, Kubi, Mario, Sören, Schnubbi und Timmi, die immer bei mir sind, egal welche Entfernungen zwischen uns liegen. Mein ganz besonderer Dank gilt Jasmin für den gemeinsamen Aufbruch ins Unbekannte und all die Träume die ich mit ihr teilen durfte, und meiner Mutter, die mir schon früh die Augen für all die Wunder dieser Welt geöffnet hat.