# "Next Generation Sequencing" basierte Prozessentwicklung zur systematischen Patientengenomanalyse am Beispiel der chronisch entzündlichen Darmerkrankungen

**Dissertation zur Erlangung des Doktorgrades**

der Mathematisch-Naturwissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel

vorgelegt von

**Dipl. Biol. Sandra May**

Lübeck, Mai 2013

| | |
|---|---|
| Referent/in | Prof. Dr. Andre Franke |
| Korefferent/in | Prof. Dr. Manuela Dittmar |
| Tag der mündlichen Prüfung | 4.07.2013 |
| Zum Druck genehmigt | 4.07.2013 |

gez. Prof. Dr. Wolfgang J. Duschl

Der Dekan

*Du holde Kunst, in wieviel grauen Stunden,*
*Wo mich des Lebens wilder Kreis umstrickt,*
*Hast du mein Herz zu warmer Lieb' entzunden,*
*Hast mich in eine beßre Welt entrückt!*

*Oft hat ein Seufzer, deiner Harf' entflossen ,*
*Ein süßer, heiliger Akkord von dir*
*Den Himmel beßrer Zeiten mir erschlossen,*
*Du holde Kunst, ich danke dir dafür!*

*Franz von Schober*

*Für meine Familie*

## Inhaltsverzeichnis

**VERZEICHNIS DER ABBILDUNGEN UND TABELLEN ..... IV**

**ABKÜRZUNGSVERZEICHNIS ..... VI**

**VERZEICHNIS ENGLISCHER FACHTERMINI ..... VII**

**1 EINLEITUNG ..... 1**

1.1 CHRONISCH ENTZÜNDLICHE DARMERKRANKUNGEN ..... 1

*1.1.1 Phänotyp der chronisch entzündlichen Darmerkrankungen (CED) ..... 1*

*1.1.2 Epidemiologie ..... 2*

1.2 URSACHEN ..... 3

*1.2.1 Genetik ..... 3*

*1.2.2 Umweltfaktoren ..... 4*

*1.2.3 Lebensstil ..... 4*

*1.2.4 Nicht-pathogene intestinale Darmflora ..... 5*

*1.2.5 Pathogene intestinale Darmflora ..... 6*

1.3 GENETISCHER HINTERGRUND DER CHRONISCH ENTZÜNDLICHEN DARMERKRANKUNGEN ..... 6

*1.3.1 Genetische Variationen ..... 6*

*1.3.2 Bekannte genetische Risikoloci für chronisch entzündliche Darmerkrankungen 8*

1.4 GENETISCHE VARIATIONEN – DETEKTION UND BEDEUTUNG ZUR AUFKLÄRUNG DES GENETISCHEN HINTERGRUNDES KOMPLEXER ERKRANKUNGEN ..... 10

*1.4.1 Gezieltes Resequenzieren ..... 13*

*1.4.2 Exom-Sequenzierung ..... 13*

*1.4.3 Genomsequenzierung ..... 14*

*1.4.4 Geschichte der Sequenzierung ..... 15*

1.5 ZIELE DER ARBEIT ..... 16

**2 MATERIAL UND METHODEN FÜR DIE RESEQUENZIERUNG DER GWAS-LOCI ..... 18**

2.1 DNA SELEKTION ..... 18

2.2 KOPPLUNGSUNGLEICHGEWICHT UND HAPLOTYPANALYSE ..... 18

2.3 DNA EXTRAKTION AUS BLUT ..... 19

2.4 NEXT-GENERATION-SEQUENZIERUNG MIT SOLID™ TECHNOLOGIE ..... 20

*2.4.1 Selektion der Ziel-Region und Anreicherung durch Long-Range-PCR (LR-PCR) 20*

*2.4.2 Aufreinigung der PCR Produkte ..... 23*

*2.4.3 Multiplex-Ansatz und Sequenzierstrategie ..... 23*

2.5 HERSTELLUNG UND SEQUENZIERUNG VON SOLID 50 BP FRAGMENT-LIBRARIES ..... 24

*2.5.1 Librarypräperation ..... 25*

*2.5.2 Emulsions-PCR, Enrichment und 3´ End-Modifikation ..... 25*

*2.5.3 SOLiD- Sequenzierung ..... 26*

2.6 ALIGNMENT GEGEN DIE REFERENZSEQUENZ ..... 27

2.7 SNP-CALLING MIT DIBAYSE UND PILEUP (SAM TOOLS) ..... 28

2.8 Annotation der SNPs ..... 28
2.9 Identifikation von falsch-positiven SNPs unter Verwendung von pibase und IGV ..... 29
*2.9.1 Analyse mit dem Validierungstool pibase ..... 29*
*2.9.2 Inspektion mit IGV ..... 30*
2.10 Sanger Sequenzierung ..... 30
2.11 Assoziationsexperimente ..... 32
*2.11.1 Selektion von SNPs für das sich anschließende Assoziationsexperiment I ..... 32*
*2.11.2 Selektion der SNPs für das Assoziationsexperiment II ..... 34*
*2.11.3 Genotypisierung mittels Sequenom ..... 35*
*2.11.4 Statistische Methoden und Qualitätskontrolle der Genotyp-Daten ..... 37*
**3 MATERIAL UND METHODEN FÜR DIE GESAMT-GENOM-SEQUENZIERUNG ..... 38**
3.1 Patientin ..... 38
3.2 DNA-Isolierung ..... 39
3.3 Library-Herstellung ..... 39
*3.3.1 Konstruktion von 50 bp Fragment-Libraries ..... 39*
*3.3.2 Konstruktion von 2x50 bp Mate-Paired-Libraries ..... 39*
*3.3.3 Konstruktion von 2x25 bp Mate-Paired-Libraries ..... 41*
3.4 Emulsions-PCR, Enrichment und 3`End-Modifikation ..... 42
3.5 Sequenzierung ..... 42
3.6 Sequenz-Alignment ..... 42
3.7 SNP-Detektion und Annotation ..... 43
3.8 Small InDel-Detektion und Annotation ..... 44
3.9 CNV-Analyse ..... 44
3.10 Validierung der ausgewählten SNPs mit Hilfe der Sanger- Technologie ..... 45
**4 ERGEBNISSE- RESEQUENZIERUNG DER GWAS LOCI ..... 46**
4.1 Mapping Ergebnisse ..... 46
4.2 Coverages ..... 48
4.3 Konkordanzen für HapMap-Proben ..... 52
*4.3.1 Konkordanz zwischen Datensätzen des Internationalen HapMap-Projektes und SOLiD-Sequenzierung ..... 53*
*4.3.2 Konkordanz zwischen Datensätzen des Internationalen HapMap-Projektes und SOLiD-Sequenzierung nach Filterprozess und Reduktion auf CED-Loci ..... 54*
4.4 Ergebnisse der SNP-Detektion ..... 55
4.5 Ergebnisse der Analyse für Transkriptionsfaktorbindestellen ..... 64
4.6 Ergebnisse des Assoziationsexperimentes I ..... 64
4.7 Ergebnisse Assoziationsexperiment II in einer größeren Analysepopulation ..... 68
**5 ERGEBNISSE DER GESAMTGENOM-SEQUENZIERUNG ..... 70**
5.1 Datenproduktion und Mappingergebnisse ..... 70
5.2 Zusammenfassende Statistik der SNP-Detektion und Annotation ..... 74
5.3 SNP-calling –Vergleich mit publizierten Genomen ..... 75
5.4 Zusammenfassende Statistik der Small InDel Detektion und Annotation ..... 76
5.5 Zusammenfassende Statistik der CNV-Analyse ..... 79

5.6 VALIDIERUNG AUSGEWÄHLTER VARIANTEN ..... 83
5.7 AUSGEWÄHLTE KANDIDATEN ALS SUSZEPTIBILITÄTSVARIANTEN ..... 83
5.8 RISIKOVARIANTEN FÜR MORBUS CROHN ..... 91
5.9 *IN SILICO* VORHERSAGE-PROGRAMME ..... 95

**6 DISKUSSION ZUR RESEQUENZIERUNG DER GWAS-LOCI ..... 97**
6.1 KONKORDANZ MIT HAPMAP TRIOS ..... 98
6.2 HERAUSFORDERUNG TARGETED ENRICHMENT ..... 99
6.3 MAPPING GEGEN DIE ZIEL-REGION UND SNP-DETEKTION ..... 101
6.4 COVERAGE ..... 102
6.5 SEQUENZIERSTRATEGIE UND MULTIPLEXANSATZ ..... 102
6.6 SNP-VALIDIERUNG UND ASSOZIATIONSEXPERIMENT ..... 103

**7 DISKUSSION GESAMT-GENOM-SEQUENZIERUNG ..... 106**
7.1 DATENERZEUGUNG UND DETEKTION VON GENETISCHEN VARIATIONEN ..... 106
7.2 CHARAKTERISIERUNG DES CROHN-GENOMS ..... 107
7.3 RISIKOVARIANTEN, DIE MIT MORBUS CROHN ASSOZIIERT SIND UND POTENTIELLE RISIKOVARIANTEN ..... 109
7.4 STRUKTURELLE VARIANTEN ..... 110
7.5 NACHWEIS VON *MYCOBACTERIUM AVIUM* UND SUSZEPTIBILITÄT ..... 111

**8 SCHLUSSFOLGERUNG UND AUSBLICK ..... 113**

**9 ZUSAMMENFASSUNG ..... 115**

**10 LITERATURVERZEICHNIS ..... 118**

**11 ERKLÄRUNG ..... 130**

**12 LEBENLAUF ..... 131**

**13 DANKSAGUNG ..... 133**

**14 ANHANG ..... 134**

# Verzeichnis der Abbildungen und Tabellen


| | | |
|---|---|---|
| **Abbildung 1-1.** | Vergleich zwischen Morbus Crohn und Colitis Ulcerosa | 2 |
| **Abbildung 1-2.** | Sinkende Kosten für die Sequenzierung | 16 |
| **Abbildung 2-1.** | Ablaufschema des Enrichmentprozesses | 18 |
| **Abbildung 2-2.** | Bestimmung von htSNPs zur Charakterisierung von Haplotypblöcken | 19 |
| **Abbildung 2-3.** | Long-Range-PCR basiertes Enrichment | 22 |
| **Abbildung 2-4.** | Darstellung der Multiplex- und Sequenzierstrategie | 24 |
| **Abbildung 2-5.** | Herstellung einer Fragmentlibrary | 25 |
| **Abbildung 2-6.** | Schematische Darstellung der Sequenzierung | 27 |
| **Abbildung 2-7.** | Ablauf der Sequenzierung | 27 |
| **Abbildung 2-8.** | Genotypisierung mittels Sequenom-Technologie | 36 |
| **Abbildung 3-1.** | Herstellung von SOLiD 2x50bp Mate-Paired-Libraries | 40 |
| **Abbildung 3-2.** | Herstellung der 2x25 bp SOLiD Mate-Paired-Libraries | 42 |
| **Abbildung 4-1.** | Mappingergebnisse für die GWAS-Loci | 47 |
| **Abbildung 4-2.** | Coverage-Diagramm für alle CED-Loci | 52 |
| **Abbildung 4-3.** | Sequenzierergebnisse der Sequenzierung des Locus *ATG16L1* | 57 |
| **Abbildung 4-4.** | Sequenzierergebnisse der Sequenzierung des Locus *IL10* | 58 |
| **Abbildung 4-5.** | Sequenzierergebnisse für die Sequenzierung des Locus *IL23R* | 59 |
| **Abbildung 4-6.** | Sequenzierergebnisse der Sequenzierung des Locus *IRGM* | 60 |
| **Abbildung 4-7.** | Sequenzierergebnisse für den Locus *NOD2* | 61 |
| **Abbildung 4-8.** | die Sequenzierergebnisse für den Locus *NKX2-3* | 62 |
| **Abbildung 4-9.** | Sequenzierergebnisse für den Locus *STAT3* | 63 |
| **Abbildung 5-1.** | Gesamtheit aller Rohdaten für die Gesamt-Genom-Sequenzierung | 73 |
| **Abbildung 5-2.** | Coverage-Diagramm für die Gesamt-Genom-Sequenzierung | 74 |
| **Abbildung 5-3.** | Überlappungen mit verschiedenen Referenzgenomen | 75 |
| **Abbildung 5-4.** | Längenverteilung der detektierten InDels | 78 |

| | | |
|---|---|---|
| **Abbildung 5-5.** | Ergebnisse für *in silico* Vorhersageprogrammen | 96 |
| **Abbildung 14-1.** | Coverage-Diagramm für den CED-Locus *IL23R* | 139 |
| **Abbildung 14-2.** | Coverage-Diagramm für den CED-Locus *ATG16L1* | 140 |
| **Abbildung 14-3.** | Coverage-Diagramm für den CED-Locus *NKX2-3* | 141 |
| **Abbildung 14-4.** | Coverage-Diagramm für den CED-Locus *NOD2* | 142 |
| **Abbildung 14-5.** | Coverage-Diagramm für den CED-Locus *STAT3* | 143 |
| **Abbildung 14-6.** | Coverage-Diagramm für den CED-Locus *IL10* | 144 |
| **Abbildung 14-7.** | Coverage-Diagramm für den CED-Locus *IRGM* | 145 |

| | | |
|---|---|---|
| **Tabelle 2-1.** | Ausgewählte GWAS-Loci für das Enrichment | 21 |
| **Tabelle 2-2.** | Liste der ausgewählten SNPs für das Assoziationsexperiment I | 32 |
| **Tabelle 4-1.** | Coverage-Tabelle für GWAS-Loci | 49 |
| **Tabelle 4-2.** | Übersicht über Datensätze für die Konkordanzberechnungen | 53 |
| **Tabelle 4-3.** | Tabelle zur Konkordanzberechnung | 54 |
| **Tabelle 4-4.** | Tabelle zur Konkordanzberechnung | 55 |
| **Tabelle 4-5.** | SNP-Tabelle für CED-Loci vor und nach dem Filterprozess | 55 |
| **Tabelle 4-6.** | SNP-Tabelle für CED-Loci nach der Annotation mit SnpActs | 55 |
| **Tabelle 4-7.** | Ergebnisse des Assoziationsexperimentes I | 66 |
| **Tabelle 4-8.** | Ergebnisse des Assoziationsexperimentes II | 68 |
| **Tabelle 5-1.** | Überblick über die Datenproduktion | 71 |
| **Tabelle 5-2.** | Zusammenfassende Statistik der SNP-Detektion und Annotation | 75 |
| **Tabelle 5-3.** | Zusammenfassung der Small InDel-Detektion und Annotation | 77 |
| **Tabelle 5-4.** | Übersicht der CNVs nach visueller Inspektion | 79 |
| **Tabelle 5-5.** | Deletionen nach dem Filterprozess | 81 |
| **Tabelle 5-6.** | Übersicht der CNVs | 81 |
| **Tabelle 5-7.** | Ergebnisse der Sanger-Validierung | 83 |
| **Tabelle 5-8.** | Validierungsergebnisse für ausgewählte SNPs (Exom und Sanger) | 85 |
| **Tabelle 5-9.** | Risikoallele für das Crohn-Genom | 92 |

**Tabelle 14-1.** GWAS-Katalog 134

**Tabelle 14-2.** Primersequenzen für das Enrichment der CED-Loci 146

## Abkürzungsverzeichnis

| | |
|---|---|
| Abb. | Abbildung |
| BAM | binary sequence alignment map (binäres Sequenzalignment Datei) |
| CED | chronisch entzündliche Darmerkrankung |
| CNV | engl. copy number variation |
| DNA | engl. Abbkürzung Desoxyribonukleinsäure |
| ePCR | Emulsions-PCR |
| GWAS | engl. genome-wide-association-study (Genomweite Assoziationsstudie) |
| HWE | Hardy-Weinberg-Equilibrium |
| InDel | Insertion oder Deletion |
| LRR | engl. leucine-rich-repeat |
| LR-PCR | Long-Range-PCR |
| MAP | *Mycobacterium avium* subspecies *paratuberculosis* |
| NGS | engl. Next Generation Sequencing |
| OMIM | online mendelian inheritancein man |
| PCR | Polymerasekettenreaktion (engl. poymerase chain reaction) |
| SNP | engl. single nucleotide polymorphism |
| SNV | engl. single nucleotide variation |
| Tab. | Tabelle |
| UTR | untranslatierter Bereich (engl. untranslated region) |

## Verzeichnis englischer Fachtermini

Folgende Fachtermini aus dem englischen Sprachgebrauch wurden verwendet, deren Übersetzung ins Deutsche schwierig und in der molekularbiologischen Literatur auch nicht üblich ist.

| | |
|---|---|
| Annealing | Anlagerung der Primer an einzelsträngige DNA aufgrund komplementärer Basenpaarung |
| Beads | Kleine Kügelchen, die in der Molekularbiologie vielfach und unterschiedlich zum Einsatz kommen, z.B. in der Emulsions-PCR als Trägersubstanz von komplementären Sequenzen zur Adaptorsequenz der Librarymoleküle, die während der Emulsions-PCR an diese binden und an der die Vermehrung über PCR stattfindet. Diese Librarymoleküle-tragenden Beads werden später in der Sequenzierung auf den dafür eigens modifizierten Objektträger gebracht und sequenziert |
| Haplotype tagging SNP | SNPs, die die häufigsten Haplotypen eines bestimmten DNA-Abschnittes charakterisieren |
| (Targetet) Enrichment | (gezielte) Anreicherung bestimmter Regionen aus dem Genom zur anschließenden Sequenzierung |
| Kit | Kommerziell erhältliche Zusammenstellung aller Komponenten, die für eine Versuchsdurchführung benötigt werden |
| (DNA-) Library | beherbergt die Gesamtheit der zu sequenzierenden genetischen Information in definierten Teilstücken. |
| Primer | Kurzes Oligonukleotid, komplementär zu DNA- oder RNA-Bereichen, dient der gerichteten enzymatischen Amplifikation spezifischer DNA- oder RNA-Abschnitte |
| Slide | speziell beschichteter Objektträger, auf den die Beads für die Sequenzierung kovalent gebunden werden |

| | |
|---|---|
| Splice site | Beim Spleißvorgang stellt sie die Grenze zwischen Exon und Intron dar |

# 1 Einleitung

## 1.1 Chronisch entzündliche Darmerkrankungen

### *1.1.1 Phänotyp der chronisch entzündlichen Darmerkrankungen (CED)*

Morbus Crohn und Colitis Ulcerosa repräsentieren die Haupterscheinungsformen der chronisch entzündlichen Darmerkrankungen (MIM266600)[1]. Die Krankheiten teilen einige Gemeinsamkeiten, sind aber auch durch verschiedene Charakteristika unterscheidbar.

Morbus Crohn und Colitis Ulcerosa sind chronische Erkrankungen, bei denen zwischen akuten Krankheitsphasen auch Phasen der Remission auftreten können. Die Patienten leiden in akuten Phasen unter ständigen, zum Teil blutigen Durchfällen, Bauchschmerzen und Gewichtsverlust. Komplikationen wie Fisteln, Abszesse und Stenosen sind häufig mit Morbus Crohn verbunden[2,3]. Bei bis zu 25% der Patienten treten zudem extraintestinale Manifestationen in Form von Gelenkentzündungen, Hautentzündungen oder auch Augenentzündungen auf[2,4]. Die gefährlichste Komplikation für Patienten, die unter Colitis Ulcerosa leiden, ist das sogenannte toxische Megacolon, das durch eine fulminante Dilatation des Dickdarms charakterisiert ist und zu einem septisch-toxischen Zustand führt[5]. Für Patienten beider Erkrankungen, Morbus Crohn und Colitis Ulcerosa, besteht ein erhöhtes Risiko an Darmkrebs zu erkranken[6]. Colitis Ulcerosa Patienten haben eine normale Lebenserwartung, während die Lebenserwartung für Morbus Crohn Patienten leicht verringert ist[2].

Bei Morbus Crohn können die wiederkehrenden Entzündungen typischerweise im gesamten Magen-Darm-Trakt von der Mundhöhle bis zum After auftreten[2]. Der untere Dünndarm (terminales Ileum) und der Dickdarm (Colon) sind dabei am häufigsten befallen. Auch Mund und Speiseröhre können in selteneren Fällen betroffen sein. Charakteristisch für Morbus Crohn ist der diskontinuierliche Befall der Darmschleimhaut, d.h. es sind gleichzeitig mehrere Darmabschnitte entzündet, die durch gesunde Abschnitte unterbrochen sind. Außerdem gehen mit Morbus Crohn häufig Komplikationen wie Fisteln, Abszesse und Stenosen einher. Colitis Ulcerosa hingegen weist Entzündungen auf, die auf das Colon beschränkt sind[7]. Die Entzündungen beginnen typischerweise im Mastdarm (Rektum) und breiten sich dann kontinuierlich in Richtung Colon aus. Häufig treten Kryptenabzesse auf. Histologisch zeigen sich bei Colitis Ulcerosa oberflächlich entzündliche Veränderungen, die

auf die Mucosa und Submucosa beschränkt sind, wohingegen die histologischen Veränderungen durch die Entzündungen bei Morbus Crohn transmural sind, d.h. sie betreffen alle Schichten der Darmwand (siehe Abbildung 1-1). Granulome sind hier häufig anzutreffen[2]. Obwohl therapeutische Fortschritte die Mortalität für chronisch entzündliche Darmerkrankungen deutlich herabgesenkt haben, werden die Gesundheitssysteme weiterhin durch diese Erkrankungen belastet, da vor allem junge Erwachsene betroffen sind, die sich im erwerbsfähigem Alter befinden[8]. Außerdem bedeutet diese Erkrankung zeitweilig, je nach Verlaufsform, für den Betroffenen einen hohen Verlust von Lebensqualität.

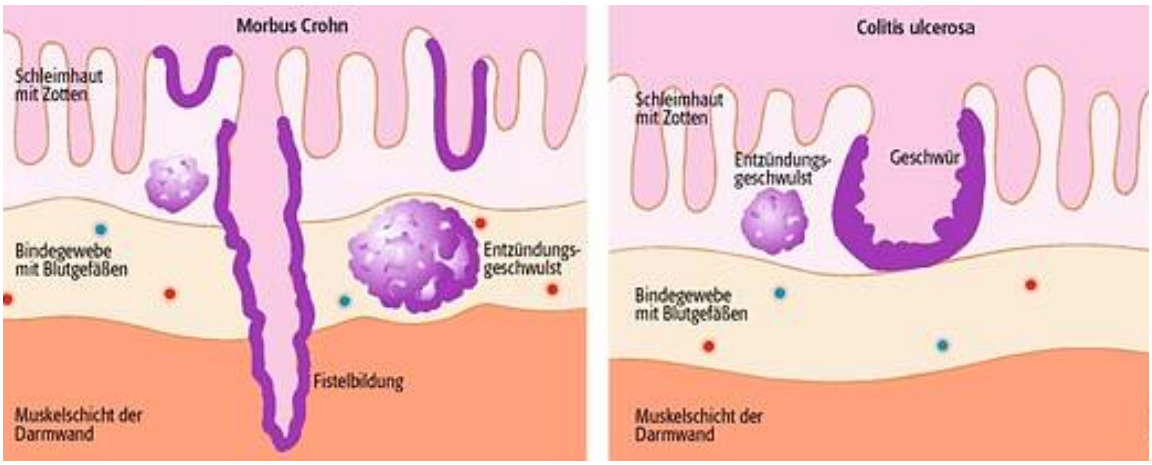



**Abbildung 1-1.** Vergleich zwischen Morbus Crohn und Colitis Ulcerosa. Bei Morbus Crohn betrifft die Entzündung alle Schichten des Darmes (transmural), während bei Colitis Ulcerosa meist nur die Mucosa und Submucosa betroffen sind (die Abbildung wurde aus der Online-Ausgabe der Pharmazeutischen Zeitung entnommen).

### *1.1.2 Epidemiologie*

Die höchsten Inzidenzraten und Prävalenzen von Colitis Ulcerosa und Morbus Crohn werden in Europa, Großbritannien und Nord Amerika beobachtet. In diesen Industrieländern erkranken jährlich etwa 7-8 Personen pro 100.000 Einwohner neu an Morbus Crohn[9]. Die Prävalenz liegt bei ca. 150 Erkrankten pro 100.000 Einwohner. Betroffen sind in erster Linie junge Erwachsene zwischen 20 und 30 Jahren oder ältere Menschen ab 60 Jahre. Von Colitis Ulcerosa sind ca. 200 Personen pro 100.000 Einwohner betroffen, die Inzidenzrate liegt bei 3-7 Neuerkrankungen im Jahr pro 100.000 Einwohner. Frauen und Männer sind hier gleichermaßen betroffen, die Krankheit tritt am häufigsten bei Menschen zwischen dem 20. und 40. Lebensjahr auf[3,10]. Innerhalb Europas wurde ein Nord-Süd Gradient in den Inzidenzraten beobachtet. Die Inzidenz in nordeuropäischen Ländern wie Island ist für Colitis Ulcerosa um 40% höher als in den südlicheren gelegenen europäischen Staaten. Für Morbus Crohn ist die Inzidenzrate sogar um 80% höher[10]. Interessanterweise treten auch innerhalb

einer geographischen Region Unterschiede in der Prävalenz unter verschiedenen ethnischen Gruppen auf, wie eine Studie an einer nicht-jüdischen und einer Ashkenazi-Juden Population zeigte[11]. In Nord Amerika ist die Prävalenz für Morbus Crohn unter Asiaten und Lateinamerikanern geringer als unter Individuen europäischer Abstammung[12].

## 1.2 Ursachen

Morbus Crohn gilt als Erkrankung ungeklärter Äthiologie[2], wenn auch verschiedene Erklärungsansätze als Ursache für Morbus Crohn diskutiert werden. Burrill Bernhard Crohn, ein US-amerikanischer Gastroenterologe, der die Krankheit 1932 erstmalig beschrieb und nach dem sie später auch benannt wurde, hielt den Morbus Crohn für eine durch intrazelluläre Bakterien oder Viren ausgelöste Erkrankung. Eine These, die bis heute nicht bewiesen werden konnte. Der fehlende Nachweis spezifischer Erreger und das gute Ansprechen der Krankheit auf das immun-suppressive Medikamente Cortison und Azathioprin, die das Immunsystem hemmen, sprechen für eine Autoimmunkrankheit der Darmschleimhaut. Die dominierende Hypothese basierend auf epidemiologischen Daten geht von einer Dysregulation der Immunantwort in der Darmschleimhaut in Menschen mit genetischer Disposition aus[13]. Neueste Erkenntnisse weisen allerdings auf einen großen Einfluss von Mycobakterien in der Ätiologie hin[14].

### *1.2.1 Genetik*

Von einer familiären Häufung von chronisch entzündlichen Darmerkrankungen wurde das erste Mal in den 1930er Jahren berichtet[15]. Die deutlichsten Hinweis auf eine genetische Komponente zeigten Studien an monozygoten Zwillingen[16,17]. In monozygoten Zwillingen beträgt die Konkordanzrate zwischen 30-35%[18]. Das relative Geschwisterrisiko, also das Risiko eines Geschwisterteils eines Erkrankten selbst zu erkranken ist im Vergleich zur allgemeinen Bevölkerung für Morbus Crohn ($\lambda S$) 15-35-fach[19,20] und für Colitis Ulcerosa 6-9-fach erhöht[21,22]. Zusätzlich tragen Verwandte eines an Colitis Ulcerosa Erkrankten ein immerhin noch bis zu 4-fach erhöhtes Risiko an Morbus Crohn zu erkranken[21]. Diese Zahlen sprechen für eine Überlappung der krankheitsassoziierten Suszeptibilitätsloci. Klinische Befunde, epidemiologische Daten und genetische Hinweise sprechen dafür, dass Colitis Ulcerosa und Morbus Crohn verwandte polygenetische Erkrankungen sind[23], das heißt, es sind im Gegensatz zu monogenetischen Erkrankungen, mehrere bis viele Risikoloci für das Auftreten der Erkrankung mitverantwortlich.

### *1.2.2 Umweltfaktoren*

Die höchsten Inzidenzraten und Prävalenzen für Colitis Ulcerosa und Morbus Crohn werden in Nordamerika und Nordeuropa verzeichnet, die niedrigsten hingegen in Südamerika, Südostasien und Afrika (mit Ausnahme Südafrika)[9]. Diese Daten zeigen nicht nur einen existierenden Nord-Süd Gradienten, sie könnten außerdem auf einen Unterschied in Zugang und Qualität zu Gesundheitsvorsorge oder auf eine verschieden stark ausgeprägte Industrialisierung und Hygiene hindeuten[1]. Die Ursachen für die verschiedenen Inzidenzraten in unterschiedlichen Gebieten der Erde könnten auch in den ungleichen genetischen Hintergründen der Einwohner zu suchen sein. Bedeutendere Faktoren scheinen jedoch eher Umweltfaktoren zu sein. Diese Hypothese wird durch steigende Inzidenzraten unter Immigranten unterstützt, die aus Regionen niedriger Inzidenz in Gebiete hoher Inzidenz eingewandert sind, sowie auch einer Korrelation der Inzidenzraten mit der zunehmenden Industrialisierung Hong Kongs und Chinas[1,24]. Einige große epidemiologische Studien zeigten, dass in Nordamerika und Europa eine Häufung von chronisch entzündlichen Darmerkrankungen vor allem in urbanen Lebensräumen verglichen mit ländlichen Regionen auftritt[25,26].

### *1.2.3 Lebensstil*

Die Assoziation zwischen exzessiver Kohlenhydrataufnahme und der Entwicklung einer chronisch entzündlichen Darmerkrankung deutet auf Zucker als einen starken zur Erkrankung beitragenden Faktor[27]. So zeigen z.B. die Asiaten mit deutlich niedriger Inzidenzrate auch einen deutlich niedrigeren Kohlenhydratkonsum als Nordamerika und Westeuropa. Ein erhöhtes Risiko zu erkranken konnte auch mit der vermehrten Aufnahme von mehrfach ungesättigten Fettsäuren assoziiert werden[28]. Die traditionell geringe Inzidenz für chronisch entzündliche Darmerkrankung und andere chronisch entzündliche Erkrankungen in Entwicklungsländern, die nun beginnt zu steigen, könnte in Verbindung mit veränderten Hygienebedingungen stehen[29]. Ein geringeres Erkrankungsrisiko, speziell für Morbus Crohn, wurde beispielsweise für ein vermindertes Geburtsgewicht, fehlenden Zugang zu sauberem Trinkwasser und Heißwasser, große oder arme Familien mit mehreren Kindern, beengte Wohnverhältnisse sowie die Konsumierung verunreinigter Nahrungsmittel beobachtet[30,31]. Ein extrem hoher Hygienestandard könnte somit zu einer limitierten Exposition von Antigenen aus der Umgebung führen, was den Reifeprozess des Immunsystems negativ

beeinflussen könnte, was letztlich zu einer überschießenden Immunreaktion führen könnte, sollte der Körper diesen Antigenen zu einem späteren Zeitpunkt erneut ausgesetzt sein[1].

Ein weiterer Faktor, der immer wieder als möglicher Risikofaktor betrachtet wird, ist das Rauchen. Dieser scheint jedoch zu einer weniger häufigen Verschlechterung der Symptomatik für Colitis Ulcerosa Patienten zu führen, wohingegen bei Morbus Crohn das Rauchen eine Verschlimmerung der Symptomatik zu bewirken scheint[32]. Für das passive Rauchen ist die Datenlage widersprüchlich. Es gibt Studien, die ein verringertes Risiko an Colitis Ulcerosa zu erkranken mit dem Ausgesetztsein des Rauchens in der Kindheit assoziieren konnten. Zusätzlich zeigen andere Daten, dass eine solche Exposition in der frühen Kindheit mit einem erhöhten Krankheitsrisiko sowohl für Colitis Ulcerosa als auch für Morbus Crohn verbunden ist[33]. Eine Metaanalyse von Jones et al. (2008) konnte allerdings keine Korrelation von passivem Rauchen in der Kindheit und einem erhöhten Erkrankungsrisiko für Colitis Ulcerosa oder Morbus Crohn finden[34].

### *1.2.4 Nicht-pathogene intestinale Darmflora*

Der wahrscheinlich einflussreichste Umgebungsfaktor zumindest für die Entstehung von Morbus Crohn scheint die Zusammensetzung der intestinalen Bakterienflora zu sein. Die Epithelzellen sind, nur getrennt durch eine dünne Mucusschicht, in ständigem engen Kontakt zu potentiell pathogenen Substanzen aus dem Darmlumen. Unter normalen Umständen besteht zwischen dem gesunden Wirt und der Darmflora eine symbiontische Beziehung. Änderungen in der Zusammensetzung der Mikroflora können diese Symbiose empfindlich stören und eine Krankheit verursachen. Eindrucksvoll wurde der Einfluss von Mikroben in Experimenten mit verschiedenen Tiermodellen demonstriert, die unter keimfreien Bedingungen aufgezogen wurden. In den meisten Tiermodellen für CED schützte die keimfreie Aufzucht die Tiere vor der Entwicklung einer intestinalen Entzündung. Die Konfrontation dieser Mutanten mit Mikroben resultierte schnell in der Entstehung einer entzündlichen Darmerkrankung[35,36]. Die Abhängigkeit zwischen Wirt und Mikroflora ist durch verschiedene Experimente bewiesen[37,38]. Das Vorhandensein oder Fehlen von bestimmten Mikroben triggert die Wirtsantwort, dass wiederum kann die Mikroflora ändern. So scheinen die Zusammensetzung des Mikrobioms und die Diversität der Mikroflora einen großen Einfluss zu haben[39,40]. Es ist noch nicht geklärt, ob Änderungen der Mikroflora eine chronisch-entzündliche Darmerkrankung verursachen oder ob die genetische Veranlagung eines Wirtes Ursache für die CED ist und die Veränderungen in der Mikroflora begünstigt.

Sehr wahrscheinlich ist, dass beides eine Rolle spielt, deshalb ist es auch nicht überraschend, dass viele Mechanismen in Morbus Crohn und Colitis Ulcerosa beeinträchtigt sind, die mit mikrobieller Abwehr assoziiert sind.

### *1.2.5 Pathogene intestinale Darmflora*

Die normale intestinale Mikroflora beherbergt auch verschiedene pathogene Bakterien. *Enterobacteriaceae*, *Enterococci* oder *Clostridium perfringens*, die in der Lage sind, Krankheiten wie Abszesse, Endokarditis oder Sepsis zu verursachen, liegen in hohen Konzentrationen vor. Diese Bakterien sind durch eine Mucusschicht von der Darmwand getrennt und werden deshalb vom Wirtsorganismus toleriert. In Patienten mit Morbus Crohn oder Colitis Ulcerosa ist diese Trennschicht zerstört und die Bakterien können sich an die Darmwand heften. Die Anheftung und das darauf folgende Eindringen in die Epithelzellen kann in einer Entzündungsantwort resultieren[41]. Gemutmaßt werden verschiedene pathogene Mikroorganismen als Mitwirkende in der Entstehung von CED. Studien geben Hinweise auf eine mögliche Beteiligung von *Mycobacterium tuberculosis*, *Listeria monocytogenes*, *Escherichia coli* und einige andere[25]. Eine kontrovers diskutierte Frage ist die Rolle von *Mycobacterium paratuberculosis* in der Ätiologie von Morbus Crohn. Das Bakterium wurde vor mehr als 20 Jahren aus Morbus Crohn befallenem Gewebe isoliert[42,43]. Seitdem sind in vielen Untersuchungen die Rolle von *M. paratuberculosis* untersucht worden, die Frage bleibt bis heute kontrovers[44–47].

## 1.3 Genetischer Hintergrund der chronisch entzündlichen Darmerkrankungen

### *1.3.1 Genetische Variationen*

Im humanen Genom kommen genetische Variationen in vielerlei Gestalt vor; einschließlich Einzelbasenpolymorphismen (engl. single nucleotide polymorphisms; SNPs), Tandem Repeats, Insertionen und Deletionen. Insertionen oder Deletionen, die die Kopienzahl eines größeren Segments einer DNA-Sequenz verändern, werden als copy number variations (CNVs) bezeichnet. Andere chromosomale Umordnungen z. B. Translokationen oder Inversionen, bezeichnet man als sogenannte copy neutral variations[48]. Diese genetischen Variationen umfassen ein Größenspektrum von einer Einzelbase bis hin zu Megabasen. Werden bei Einzelbasenpolymorphismen, wie im Namen impliziert (SNPs), einzelne Basen ausgetauscht, sind bei CNVs, Inversionen und Translokationen größere Bereiche von der

Veränderungen betroffen. Die Größen rangieren hier zwischen 1 kb und mehreren Megabasen[49].

Genetische Variationen des humanen Genoms eignen sich nicht nur als Marker in der forensischen Medizin oder für medizinische Routinetests (z.B. vor Organtransplantationen)[50], sondern sie werden auch vielfach und erfolgreich als genetische Marker zur Identifizierung von Krankheitsgenen bei monogenetischen und komplexen Erkrankungen eingesetzt, z.B. in Kopplungsstudien oder Genomweiten Assoziationsstudien (GWAS)[51]. Für Kopplungsstudien und GWAS spielen vor allem die SNPs eine bedeutende Rolle. SNPs sind biallelische Marker mit einer hohen Prävalenz im humanen Genom. 2001 identifizierte die „International SNP Map working group" 1.42 Millionen SNPs im Humanen Genom[52]. Hochdurchsatz-Genotypisierungstechnologien machten es im folgenden möglich, eine große Anzahl von SNPs in großen Studienpopulationen zu typisieren. Mehr als 3 Millionen SNPs sind im Zuge des internationalen HapMap-Projektes (http://hapmap.ncbi.nlm.nih.gov/) in Individuen verschiedener Abstammung genotypisiert worden. Mittlerweile sind mehr als 20 Millionen SNPs in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) dokumentiert. Bei einer so großen Anzahl sind Fehler in den Datenbanken nicht zu vermeiden. Es wird geschätzt, dass die falsch-positiv Rate für dbSNP zwischen 15 und 17% liegt[53]. Um SNPs von Punktmutationen abzugrenzen, wird für SNPs eine Frequenz des selteneren Allels (engl. minor allele frequency (MAF)) von mindestens 1% angenommen, für häufige SNPs (engl. common SNPs) wird eine MAF von >5% festgelegt.

SNPs klassifiziert man außerdem nach ihrer Lokalisation im Gen. So können sie in der kodierenden Region eines Gens (kodierende SNPs; cSNP für engl. „coding SNP"), in den nicht-kodierenden Regionen (ncSNPs für engl. „non-coding SNP") eines Gens (intronische oder 5` oder 3´untranslatierte Bereiche) oder aber auch ganz außerhalb von Genen (intergenetische Bereiche) vorkommen. Diese beeinflussen zwar nicht die Aminosäuresequenz, können aber einen regulatorischen Einfluss auf die Transkription haben. Kodierende SNPs können weiter in nonsense (resultierendes Triplett ist ein Stoppcodon), missense (resultiert in Aminosäureänderung), in synonyme SNPs (keine Änderung auf Aminosäurelevel) und nicht-synonyme SNPs (nsSNPs; Änderung auf Aminosäurelevel) klassifiziert werden.

### *1.3.2 Bekannte genetische Risikoloci für chronisch entzündliche Darmerkrankungen*

Eine starke familiäre Häufung von chronisch entzündlichen Darmerkrankungen wurde, wie schon erwähnt (Kapitel 1.2.1), früh beobachtet und seit den 1980er Jahren systematisch erforscht[19]. Verglichen mit monogenetischen Erkrankungen ist es allerdings schwieriger die genetischen Ursachen einer komplexen polygenetischen Erkrankung aufzuklären[5]. Dieses Unterfangen galt in der Erforschung der chronisch entzündlichen Darmerkrankungen lange als nahezu unmöglich. Allerdings konnten mit der Entwicklung neuer molekularbiologischer Methoden in den letzten Jahren viele Risikoloci mit Morbus Crohn und Colitis Ulcerosa assoziiert werden.

Durch Kopplungsstudien betroffener Geschwisterpaare[54] konnten mehrere Loci ausgemacht werden, die die Identifizierung einzelner Krankheitsgene für Morbus Crohn ermöglichten[55]. Das erste beschriebene und damit am besten untersuchte und replizierte assoziierte Gen für Morbus Crohn ist *NOD2*[55–57]. Das *NOD2/CARD15* Gen ist ein Mitglied einer intrazellulärer Rezeptorfamilie, der sogenannten „pattern-recognition receptors“ (PRRs), die mikrobielle Komponenten erkennen. Die LRR-Domäne (leucine rich repeat) des *NOD2* Gens erkennt die Zellwandkomponente gram-negativer und gram-positiver Bakterien Myramyl-Dipeptid (MDP)[58]. Drei codierende Varianten, R702W, G908R und 1007fsInsC, die das Risiko an Morbus Crohn zu erkranken um bis zu 40-fach erhöhen, konnten in *NOD2* durch diese Studien bestimmt werden. Ogura et al. berichteten über eine Assoziation der 1007fsInsC in *NOD2* mit Morbus Crohn und vermuteten, dass das daraus resultierende verkürzte *NOD2* Protein zu einer Funktionsbeeinträchtigung in der Abwehr von Bakterien führt[56]. Mutationsanalyen von *NOD2* in britischen und deutschen Patienten konnten die Mutation 1007fsInsC als krankheitsassoziierte Variante bestätigen[57]. Diese Befunde führten zu einem völlig neuen Verständnis der Krankheit. *NOD2* spielt eine große Rolle im angeborenen Immunsystem und sorgt für eine intakte Barrierefunktion des Darmes. Das Protein fungiert als ein intrazellulärer Sensor für das in der Zellwand von Bakterien vorkommende, hochkonservierte Myramyl-Dipeptid (MDP) und aktiviert einen wichtigen Transkriptionsfaktor, NF-κB, für die Entzündungsantwort[55]. Die Crohn-assoziierte Variante im *NOD2* Gen und das daraus resultierende verkürzte Protein führt zu einer Beeinträchtigung der Proteinfunktion[59], d.h. einer Abschwächung der NF-κB Aktivierung[60]. Diese Funktionsbeeinträchtigung in der Funktion von *NOD2* scheint spezifisch für Morbus Crohn

Patienten zu sein und könnte zu einem komplexen Defekt der Barrierefunktion des Darmes beitragen[61].

Die polygenetische Ätiologie komplexer Erkrankungen wie Morbus Crohn und Colitis Ulcerosa systematisch aufzuklären wurde auch durch Hochdurchsatz-Genotypisierungsmethoden möglich. So wurde im Oktober 2006 das Suszeptibilitätsgen *IL23R* identifiziert[62]. Während *NOD2* ein für Morbus Crohn spezifisches Risikogen ist, konnte *IL23R*, welches ursprünglich als Suszeptibilitätslocus für Morbus Crohn identifiziert wurde, auch mit Colitis Ulcerosa assoziiert werden[63]. Im Sommer 2007 konnte, kurz nachdem das Autophagiegen *ATG16L1* als mit Morbus Crohn assoziiert beschrieben wurde, ein weiteres mit dem Prozess der Autophagie verbundenes Gen, *IRGM*, als Risikogen für Morbus Crohn identifiziert werden[64]. So brachten GWAS den Durchbruch in der Identifizierung vieler Risikoloci für Morbus Crohn einschließlich der auch in dieser Arbeit sequenzierten Gene/Loci *IL23R, ATG16L1, IRGM, STAT3,* und *NKX2-3*[62,65–68].

In Replikationsstudien mit Morbus Crohn assoziierten Varianten in Colitis Ulcerosa Patienten und anschließenden Metaanalysen konnten sowohl gemeinsame Varianten für Morbus Crohn und Colitis Ulcerosa als auch neue, nur für eine Erkrankung spezifische Varianten, identifiziert werden[69]. Die Zahl der bekannten Risikoloci für Morbus Crohn ist mittlerweile auf 71 gestiegen[70]. Jüngst publizierte Metaanalysen konnten 99 nicht überlappende Risikoloci für Chronisch entzündliche Darmerkrankungen identifizieren, 28 sind sowohl mit Morbus Crohn als auch mit Colitis Ulcerosa assoziiert[71]. 23% des erblichen Risikos für Morbus Crohn und 16% für Colitis Ulcerosa lassen sich durch die Befunde der GWAS erklären. Die identifizierten Gene betreffen verschiedene Funktionswege, die der intestinalen Homeostase, einschließlich Barrierefunktion, mikrobielle Abwehr, Regulation der angeborenen und erworbenen Immunität, Autophagie sowie metabolische Stoffwechselprozesse, die mit zellulärer Homeostase assoziiert sind [13]. So sind *IL23R* und *STAT3* Gene, die die Ausdifferenzierung von CD4+ über T- Helferzellen (Th1 und Th2 Zellen) hin zu TH17-Lymphozyten mitkoordinieren[72]. In anderen durch GWAS identifizierten Suszeptibilitätloci[73] finden sich die Autophagiegene *ATG16L1*[65] und *IRGM* [66].

## 1.4 Genetische Variationen – Detektion und Bedeutung zur Aufklärung des genetischen Hintergrundes komplexer Erkrankungen

Für monogene vererbbare Krankheiten, deren Ursachen mit konventionellen genetischen Studien aufgeklärt werden konnten, wie z.B. die Cystische Fibrose, fand man selten mutierte Gene, die eine vollständige Penetranz haben, bei welcher es immer zur Ausprägung des Merkmals kommt[74]. Schnell erkannte man aber, dass diese Varianten nicht die Ursache komplexer Erkrankungen sind. Genetischer Hintergrund komplexer Erkrankungen sind Varianten mit unvollständiger Penetranz, d.h. trotz vorhandenen Genotyps muss sich das Merkmal phänotypisch nicht ausprägen und eher das Zusammenwirken verschiedener Varianten und Umweltfaktoren bedingen gemeinsam den Phänotyp.

In den letzten Jahren haben sich die Herangehensweisen an die Untersuchung genetischer Ursachen komplexer Erkrankungen beachtlich weiterentwickelt. Der Ansatz der Kandidatengenstudien, der beschränkt ist auf die Analyse einiger SNPs in einer moderaten Anzahl von Probanden, wurde von dem Ansatz der hypothesenfreien GWAS verdrängt, in denen eine riesige Anzahl von SNPs in vielen Individuen typisiert werden können[75]. Durch Kenntnisse der Haploblockstruktur des menschlichen Genoms lässt sich der Aufwand zum Auffinden krankheits- bzw. phänotypassoziierter SNPs reduzieren. Realisiert wird dies im sogenannten HapMap-Projekt (www.hapmap.org), dass sich zur Aufgabe gemacht hat, die Blockstruktur des humanen Genoms aufzuklären und für Wissenschaftler aus aller Welt zur Verfügung zu stellen. Als Haplotyp wird die Kombination von SNP-Allelen in einem kleinen chromosomalen Segment bezeichnet, das wenig Hinweis auf meiotische Rekombination gibt. Die verschiedenen Haplotypen lassen sich meist durch eine kleine Zahl von Polymorphismen, sogenannte *Haplotype-Tagging-SNPs* (htSNPs), sicher voneinander unterscheiden. Somit braucht, um einen Haplotypblock vollständig zu charakterisieren, nur ein geringer Anteil der vorhandenen Polymorphismen typisiert zu werden (siehe auch Kapitel 2.2).

Mit Hilfe von GWAS konnten in den letzten Jahren hunderte genetischer Varianten, die mit komplexen Erkrankungen assoziiert sind, identifiziert werden. Die GWAS, in denen mehrere Hunderttausend bis mehr als eine Million SNPs in mehreren Tausend Individuen genotypisiert wurden, wurden somit lange Zeit zu einem mächtigen Werkzeug in der Erforschung des genetischen Hintergrundes komplexer Erkrankungen[76]. Allerdings haben GWAS auch ihre Limitationen. Die für die GWAS verwendeten SNP-Chips repräsentieren zwar sehr viele, aber längst nicht alle im humanen Genom vorkommenden SNPs und

genetischen Variationen. Das Design der GWAS und die Auswahl der SNPs für die kommerziell erhältlichen Genotypisierungsassays wurden von der sogenannten „common disease-common variation“ Hypothese angetrieben (häufige Erkrankung-häufige Variation)[76]. Diese stellte für lange Zeit die populärste Hypothese dar. Hier wird postuliert, dass genetische Variationen mit hoher Frequenz (MAF>5%) in einer Population die Hauptbeitragenden für die genetische Veranlagung häufig zu findender Erkrankungen, zwar mit geringem bis moderaten Effekt, sind[77]. Ein bekanntes Beispiel, das diese Hypothese stützt, ist das Gen *APOE*, in dem eine einzelne häufige Variante zu einem hohen Risiko von Alzheimer und Herzinfarkt führt[78,79]. Die meisten durch GWAS identifizierten Varianten leisten allerdings einen relativ kleinen Beitrag zum erblichen Risiko an einer bestimmten Krankheit zu erkranken und erklären so nur zu einem kleinen Teil das familiäre Clustern komplexer Erkrankungen[80]. So sind durch die 99 identifizierten Krankheitsloci für chronisch-entzündliche Darmerkrankungen nur ungefähr 23% des familiären erblichen Risikos erklärbar[71]. Das lässt vermuten, dass die häufig in einer Population vorkommenden Varianten wahrscheinlich nicht gänzlich die Heritabilität häufiger chronischer Erkrankungen begründen. Dies führt unweigerlich zu der Frage, wo die fehlende Heritabilität zu finden ist. Viele Erklärungsansätze für diese Frage sind diskutiert worden, einschließlich der Hypothese einer noch größeren Anzahl von Varianten mit kleinem Effekt.

Ein weiterer Erklärungsansatz ist die der „Common disease-common variant“ Hypothese gegenüberstehende oder auch ergänzende „Common disease-rare variant“ Hypothese (Häufige Erkrankung-seltene Variante). Ihr zufolge leisten mehrere seltene Varianten, jede mit relativ hoher Penetranz, den Hauptbeitrag für eine genetische Suszeptibilität[81]. Beide Hypothesen haben in der aktuellen Forschung ihre Berechtigung. Denkbar sind auch häufige Varianten mit so geringer Penetranz, dass sie mit Hilfe von GWAS statistisch nicht mit der Erkrankung in Verbindung gebracht werden können. Dies würde bedeuten, dass theoretisch jede Variante eines jeden Gens einen minimalen Beitrag zur Erkrankung liefern könnte.

Eine gute Möglichkeit fehlende Heritabilität für verschiedene Krankheiten aufzuspüren scheint die Methode der Sequenzierung zu sein. So konnten Romeo et al. im Jahr 2007 durch Sequenzierung des Gens *ANGPTLA4* in 3500 Individuen (ein Gen, das durch vorhergehende Studien mit der Konzentration von Cholesterol und Triglyceriden in Verbindung stand) mehrere unbekannte Varianten identifizieren, die einen tiefgreifenden Effekt auf die Konzentration dieser Lipide zu haben scheinen[82]. Nejentsev et al.[83] bewiesen mit ihrer

Resequenzierungsstudie an dem Gen *IFIH1*, in der sie eine seltene protektive Variante gegen Diabetes Typ I identifizieren konnten, dass die Resequenzierung durch GWAS lokalisierte Regionen kausative Gene und Varianten genau bestimmen kann. Seltene Varianten, für die es durch GWAS keine regionalen Vorbefunde gibt, werden schwerer zu finden zu sein.

Die Entdeckung von Mutationen, die den Phänotyp determinieren ist eine fundamentale Voraussetzung von genetischer Grundlagenforschung und wird gewaltig durch „Next Generation Sequencing", sowohl für gezielte als auch für genomweite Mutationsdetektion, vorangetrieben[84]. Mit dem technischen Fortschritt und Erweiterung der Sequenzierungsmöglichkeiten und der damit verbundenen sinkenden Sequenzierungskosten sind größere Sequenzierungsprojekte die logische Konsequenz der GWAS. Einen Anfang machte „The 1000-Genomes-Project", dessen Ziel es ist, DNA-Variationen zu entdecken und korrekte Haplotypinformation für sämtliche DNA-Variationen für verschiedene humane Populationen zu liefern[85].

Viele Wissenschaftler zielen es nun auch auf die sog. copy number variations (CNVs) ab, einige Basen bis zu mehreren Hunderten Basenpaaren lange Deletionen oder Insertionen. Variationen mit diesen Eigenschaften konnten einen Teil der Vererbbarkeit für Erkrankungen wie Autismus und Schizophrenie erklären[86]. Diese strukturellen Varianten könnten einen großen Teil der genetischen Variabilität von Person zu Person begründen.

Epigenetische Effekte, welche die Genexpression beeinflussen, wie z. B. Methylierung verkomplizieren die Aufklärung der Vererbbarkeit komplexer Erkrankungen. Denkbar ist auch, dass sogenannte häufige Erkrankungen gar nicht häufig sind. Die medizinische Wissenschaft versucht komplexe Symptome zusammenzufassen und sie unter einem Krankheitsnamen zu führen. Sollten aber tatsächlich tausende verschiedene von seltenen Varianten zu einer Krankheit beitragen, und die genetische Untermauerung von Person zu Person stark variieren, wie häufig ist die Erkrankung dann wirklich? Oder sind dies sogar eigentlich verschiedene Erkrankungen?[87].

Wie erwähnt ist eine Möglichkeit Varianten aufzuspüren, die die fehlende Heritabilität komplexer Erkrankungen erklären, die Sequenzierung. Dabei kommen verschiedene Sequenzierungsstrategien zur Anwendung, die in den letzten Jahren entwickelt worden sind. Man kann dabei folgende Ansätze unterscheiden:

- gezieltes Resequenzieren (engl. targeted resequencing)
- Exomsequenzierung (engl. exome-sequencing)
- Gesamt-Genom-Sequenzierung (engl. whole genome sequencing)

### *1.4.1 Gezieltes Resequenzieren*

SNPs resultierend aus GWAS sind eng korreliert mit anderen SNPs in einer Region und möglicherweise sogar in LD (siehe Kapitel 2.2) mit der kausalen Variante, so dass teilweise ein komplettes Katalogisieren aller Varianten in dieser Region für das Auffinden der kausalen Variante nötig ist. Sogar in den dichten Referenzdaten des „1000-Genomes-Project“ sind die meisten GWAS-Hits nicht mit einer offensichtlichen funktionellen Variante korreliert[88]. Durch die Anreicherung, sog. Enrichment, spezifischer genomischer Regionen mittels unterschiedlicher Technologien und anschließender Resequenzierung dieser krankheitsassoziierten Loci, die zuvor durch GWAS identifiziert worden sind, ist es so möglich, große Regionen sehr genau aufzulösen und so mögliche funktionelle Varianten zu finden. Der Vorteil des „Targeted Resequencing“ besteht darin, dass man die Kapazität des Sequenziergerätes komplett für die gezielten Bereiche ausnutzen kann, somit können mehr Individuen und schneller für diese genomischen Bereiche sequenziert werden als beim Sequenzieren ganzer Genome einzelner Individuen. Diese Methode wird aufgrund immer weiter fallender Kosten für die Herstellung der Libraries und Sequenzierung und einer beständig steigenden Generierung von Daten durch die Sequenzierung ganzer Exome bzw. Genome zunehmend verdrängt[75].

### *1.4.2 Exom-Sequenzierung*

Exomsequenzierung – das gezielte Sequenzieren des Teils des Genoms (ca. 1% des humanen Genoms), das für Proteine kodiert, ist eine kosteneffektive und vielversprechende, relativ neue Methode, um genetische Ursachen von Krankheiten und Charakteristika, die schwer oder gar nicht durch konventionelle gentechnische Methoden aufzuklären sind, zu identifizieren[89]. Mit der Exomsequenzierung lassen sich besonders gut Ursachen monogenetischer Krankheiten erforschen, sie wird aber auch immer mehr verwendet, um seltene Allele, die die Heritabilität komplexer Erkrankungen mitverursachen, aufzufinden. Dieser Ansatz ist sehr vielversprechend, da SNPs oder Mutationen in kodierenden Bereichen oft informativer sind als SNPs in nicht-kodierenden Bereichen[90]. Mit der Sequenzierung ganzer Exome konnten schon einige Erfolge, d.h. die Identifizierung von Krankheitsgenen

bzw. krankheitsassoziierter Varianten, erzielt werden. So konnte beispielsweise durch Exomsequenzierung das mit der Krankheit „Kombinierte Hypolipidämie“ (Gesamtcholesterin normal bis leicht erhöht; LDL-Cholesterin ist meist erhöht, HDL-Cholesterin oft erniedrigt) assoziierte Gen *ANGPTL3* identifiziert werden[91]. Jüngste Untersuchungen konnten mit Hilfe von Exomsequenzierungen in an Morbus Crohn erkrankten Kindern sehr seltene Varianten in Genen, die für chronisch entzündliche Darmerkrankungen bekannt sind, finden. *In-silico* Analysen zeigten große Wahrscheinlichkeiten für schädigende Wirkungen dieser Varianten[92].

Dem eigentlichen Sequenzieren des Exoms geht noch ein sog. Enrichmentschritt voraus, d.h. die Anreicherung möglichst aller Exone aus dem gesamten Genom. Für diesen Enrichmentschritt ist in den letzten Jahren viel Aufwand betrieben wurden. Mittlerweile stehen viele sog. Kits basierend auf verschiedenen Strategien (PCR-basiert/ Chip-basiert) auf dem Markt kommerziell zur Verfügung, so dass Exomsequenzierungsdaten von großen Populationen schnell und kosteneffektiv generiert werden können.

### *1.4.3 Genomsequenzierung*

Mit Hilfe von „Next Generation Sequencing“-Methoden kann effektiv die Gesamtheit genetischer Variation ganzer Genome detektiert werden. Sie dienen damit als ein probates Mittel für die systematische Aufklärung des gesamten Spektrums an Varianten (häufige, seltene bis sehr seltene) im Genom und verbessern somit die Möglichkeiten, die verbleibende genetische Varianz, die mittels GWAS (und der gezielten Sequenzierung genomischer Bereiche/Exomsequenzierung) nicht identifiziert wurde, aufzuklären[93]. Die Sequenzierung ganzer Genome erfordert kein Anreichern bestimmter genomischer Bereiche. Trotzdem ist es schwierig, eine über das gesamte Genom ausgeglichene und einheitliche Abdeckung durch die Sequenzierung (engl. coverage) zu erhalten. Es sind sogar einige, v.a. in sogenannten repeat-reichen Regionen und GC-reichen Abschnitten des Genoms, Lücken zu erwarten. Um wirklich das gesamte Genom mit einer angemessenen Coverage abzudecken, ist u.U. ein ergänzendes Sequenzieren mit konventionellen Kapillarsequenzierern erforderlich, die auf Grund ihrer Leselängen und zielgerichteten Ansatzes gewisse Lücken zu schließen in der Lage sind. In „proof-of-principle“ Studien konnten mit der Sequenzierung ganzer Genome genetische Ursachen mendelnder Erkrankungen identifiziert werden[94,95]. Das Sequenzieren ganzer Genome wurde ebenfalls dazu verwendet, um die Ergebnisse von GWAS zu vervollständigen und um eine seltene Variante im Gen *MYH6*, welches für die schwere Aminosäurekette des α-Myosin kodiert, als ein Risikoallel für das Sick-Sinus-Syndrom

(Synonym: Sinusknotensyndrom) zu identifizieren[96]. Trotz der technischen Machbarkeit ist und bleibt das Sequenzieren ganzer Genome eine enorme Herausforderung auf Grund der gewaltigen Datenmengen, Datensicherung und nicht zuletzt durch die Komplexität der Daten. Mit dem vorhergesagten Sinken der Sequenzierungskosten und Fortschritte in der Bioinformatik wird das Sequenzieren ganzer Genome wohl der vielversprechendste Ansatz werden, die genetische Variabilität aufzuklären. Allerdings wird, um mit Hilfe der Gesamt-Genom-Sequenzierung die genetischen Ursachen komplexer Krankheiten und Phänotypen zu skizzieren, die Sequenzierung einer großen Anzahl an Genomen erforderlich sein[75].

### *1.4.4 Geschichte der Sequenzierung*

Maxam und Gilbert entwickelten im Jahre 1977 eine Methode zur Sequenzierung von DNA, die auf basenspezifischer chemischer Spaltung der DNA beruht. Da diese Methode aufwendig und schlecht automatisierbar war, und außerdem aufgrund der verwendeten Chemikalien auch Gefahren in sich birgt, hat sie sich nie ganz durchgesetzt. Fast parallel zu Maxam und Gilbert etablierten Sanger et al. die Didesoxynukleotidsequenzierung, die auf der sogenannten Kettenabbruchreaktion beruht. 1977 wurde das erste virale Genom, das vollständig sequenziert wurde, publiziert[97]. Drei Jahre später sequenzierte die gleiche Gruppe das komplette humane mitochondriale Genom[98]. Seitdem durchlebte die DNA Sequenzierung eine regelrechte Metamorphose von einer kleinen Labormethode hin zu einer im großen Stil produzierenden Industrie, die eine spezialisierte und funktionierende Infrastruktur, d. h. Roboter, Bioinformatiker und Computerdatenbanken, erforderte[99]. Im Zuge dieser Metamorphose sanken neben dem zeitlichen Aufwand auch die finanziellen Kosten für die Sequenzierung rapide (Abbildung 1-2). Durch die zunehmende Automatisierung der Prozesse und fallende Kosten konnten immer größere Sequenzierungsprojekte angestoßen werden. Moderne Kapillarsequenzierer lassen es zu, 96 Proben in weniger als 24 Stunden parallel zu sequenzieren. Dadurch wurde es 1991 möglich, das Humane Genom Projekt zu realisieren, das zehn Jahre später in der „Entschlüsselung“ der humanen DNA seinen ersten Höhepunkt hatte[100] und 2003 fertig gestellt wurde. Die Sequenzierung nach der Methode Sanger gilt auch heute noch als Goldstandard in der Sequenzierung und vor allem bei „de novo“-Sequenzierung als Methode der Wahl[101]. Da die DNA-Sequenzierung im Laufe der Jahre immer mehr an Bedeutung gewann und das Interesse an Hochdurchsatzmethoden stärker wurde, ist in den letzten Jahren viel Aufwand aus Richtung der Industrie betrieben worden, neue molekularbiologische Methoden zur Sequenzierung zu entwickeln. Diese sind unter

dem Namen „Next Generation Sequencing“ Methoden zusammengefasst. Mehrere unterschiedliche Ansätze kommen dabei zur Anwendung. Die in dieser Arbeit verwendete Methode des „Next Generation Sequencing“, ABI SOLiD, wird im Methodenteil dieser Arbeit detailliert erläutert.

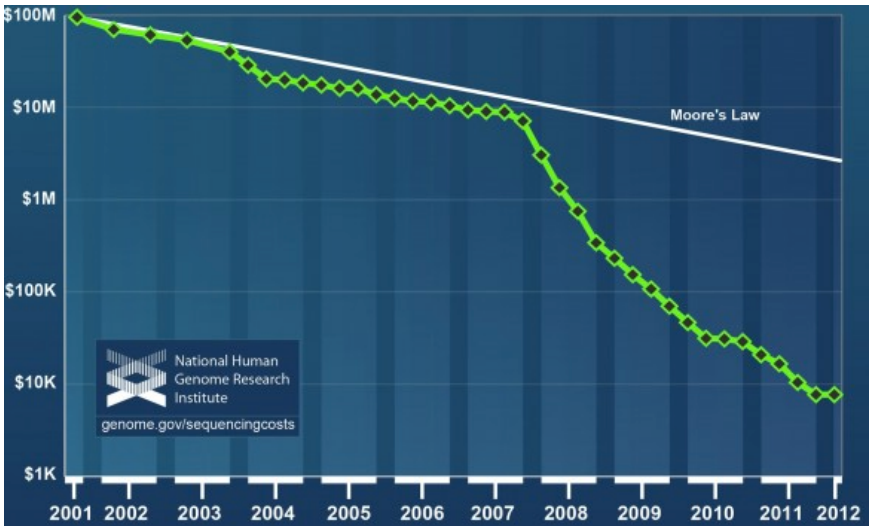



**Abbildung 1-2.** Die Abbildung zeigt die sinkenden Kosten für die Sequenzierung. Die Sequenzierung eines Genoms kostete zur Jahrtausendwende 100 Mio Dollar, 12 Jahre später sind diese Kosten um den Faktor 1000 gefallen (Abbildung entnommen aus http://scienceblogs.de/weitergen/2012/10/es-wird-immer-billiger-kommerzielle-dna-sequenzierung-zur-vorhersage-von-krankheiten).

## 1.5 Ziele der Arbeit

Die Arbeit unterteilt sich in zwei Teilprojekte, die jedes für sich ein Pilotprojekt zur Etablierung der „Next Generation Sequencing“-Technologie beinhalteten:

Im ersten Teil der Arbeit war es Ziel, durch GWAS identifizierte Risikoloci für CED durch Hochdurchsatz-Sequenzierungstechnologie aufzulösen und neue seltene Varianten, die in Verbindung mit den untersuchten Phänotypen stehen, zu identifizieren. Dafür sollten für jeden Locus 50 Individuen (20 Kontrollen, 30 Fälle), die mittels Haplotypanalysen selektiert wurden, in dem jeweiligen Locus nach erfolgtem Enrichment der genomischen Bereiche, sequenziert werden. Dabei sollten nicht nur die Exone Beachtung finden, sondern es sollten die gesamten Regionen, über die sich die GWAS-Assoziationssignale erstrecken, sequenziert werden. Um finanzielle und zeitliche Ressourcen zu schonen, wurde als weiteres Ziel der Arbeit eine Strategie für die Sequenzierung entwickelt, die erlaubte, zehn Individuen parallel auf einem Spot zu sequenzieren dabei aber hinterher die Sequenzierergebisse den einzelnen Individuen zuzuordnen. Mit Hilfe von gängigen Genotypisierungsverfahren sollten

ausgewählte neue identifizierte Varianten in einer größeren Fall/Kontrollstudie verifiziert und auf evtl. Assoziation getestet werden.

Im zweiten Teil der Arbeit wurde ein vollständiges Genom einer an Morbus Crohn erkrankten Patientin sequenziert. Hier gab es einerseits methodische Ziele. So sollte möglichst das gesamte Genom mit einer für SNP-Detektion erforderlichen Abdeckung sequenziert werden.

Für die Gesamt-Genom-Sequenzierung wurde eine Patientin ausgesucht, deren Krankheitsverlauf als schwer zu bezeichnen ist. Das hauptsächliche Anliegen war es, das Genom dieser Patientin zu charakterisieren und in einer rein deskriptiven Statistik aller im Genom detektierten Variationen. Kann dieses Genom in seiner Gesamtheit als ein „krankes“ Genom von anderen „gesunden“ Genomen unterschieden werden?

Ein weiteres Ziel war es, neue bisher unbekannte Varianten zu identifizieren, die diesen Phänotyp erklären könnten. Hierzu kommen sowohl Varianten in schon mit Morbus Crohn assoziierten Genen in Frage als auch Varianten in Genen, die bisher in noch keinem Zusammenhang mit Morbus Crohn stehen. Dabei standen nicht nur die Detektion von Einzelbasenpolymorphismen zur Debatte sondern es wurde auch nach strukturellen Varianten gesucht.

# 2 Material und Methoden für die Resequenzierung der GWAS-Loci

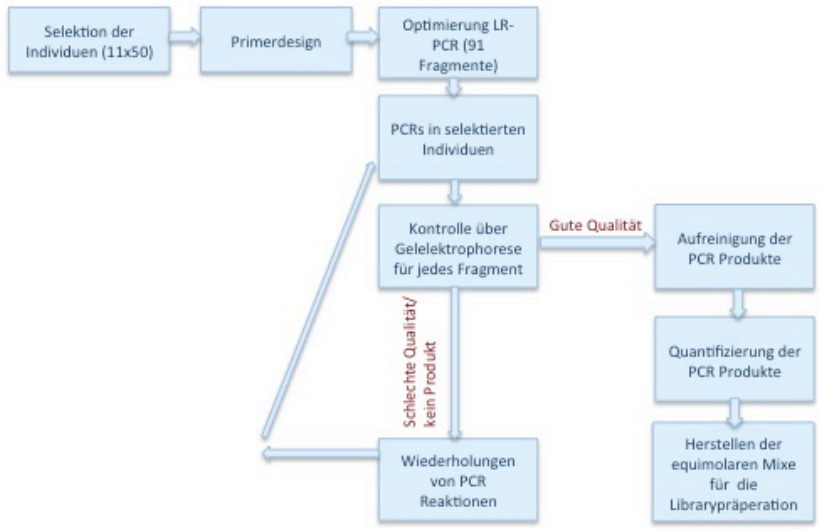



**Abbildung 2-1.** Das Schema zeigt den Ablauf des Enrichmentprozesses. Viele Schritte mussten wiederholt werden, was den Aufwand weiter steigen ließ.

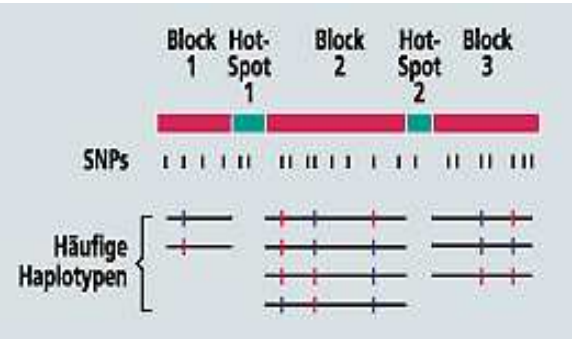
## 2.1 DNA Selektion

Für jede der zehn Zielregionen wurden 30 Fälle, 20 Kontrollen sowie 2 HapMap-Trios als technische Kontrollen ausgewählt. Um eine möglichst große Variation an Haplotypen zu sequenzieren und um seltene Varianten anzureichern, wurden die Kontrollen und Patienten basiert auf einer Haplotypanalyse (siehe Kapitel 2.2), die mit Hilfe der frei zugänglichen Software famhap (online unter: http://famhap.meb.uni-bonn.de/) und vorangegangenen GWAS durchgeführt wurden, selektiert.

## 2.2 Kopplungsungleichgewicht und Haplotypanalyse

Treten bestimmte Allele an zwei Genloci in einer Population überzufällig häufig gemeinsam auf, spricht man vom sogenannten Kopplungsungleichgewicht (oder engl. Linkage Disequilibrium=LD) . Somit kann aus dem Wissen über das Vorliegen des ersten Allels mit einer großen Wahrscheinlichkeit auf das Vorliegen des zweiten Allels geschlossen werden. Ist eines der Allele ein Suszeptibilitätsallel für eine bestimmte Krankheit, und tritt dieses häufiger in der Gruppe der Betroffenen auf als in zufällig ausgewählten gesunden Individuen, werden auch die gekoppelten Allele häufiger in der Gruppe der betroffenen Personen zu

finden sein. Diese Allele können ebenfalls als Markerallele zur Detektierung von genetischen Unterschieden zwischen zwei Gruppen (krank/gesund) benutzt werden. Viele Untersuchungen lassen Inseln mit Kopplungsungleichgewicht erkennen, die durch sog. Rekombinations-Hotspots getrennt sind[102]. In diesen Hotspots ereigneten sich nahezu alle Rekombinationsereignisse der Populationsgeschichte, während in den sich im Kopplungsungleichgewicht befindenden Blöcken nur wenige Rekombinationsereignisse stattgefunden haben. Die Anzahl möglicher SNP-Allelkombinationen in diesen Bereichen ist durch die gegenseitige Abhängigkeit von Allelen stark reduziert. Die beobachteten Kombinationen werden als Haplotyp bezeichnet. Die Haplotypen lassen sich meist durch eine geringe Anzahl von Polymorphismen, sog. *Haplotype-Tagging SNPs (htSNPs)*, voneinander unterscheiden. Somit ist es möglich, durch die Typisierung von nur wenigen Polymorphismen die jeweiligen Haplotypen zu charakterisieren. Die Kenntnisse über die Blockstruktur lassen sich auch in der Identifizierung von krankheitsassoziierten Haplotypen durch eine deutliche Aufwandreduktion zu Nutze machen. Die Haplotypanalyse und die Berechnung des LD erfolgte für die biallelischen SNP-Marker paarweise mit dem Programm Haploview[103] unter Verwendung des $r^2$ [104]. Als perfektes LD wird ein $r^2$ von 1 bezeichnet.




Ausgedehnte Blöcke (rot), in denen nahezu keine Rekombinationsereignisse stattfanden, wechseln sich mit so genannten Hotspots der Rekombination (grün) ab. Die häufigen Haplotypen der resultierenden Blöcke können mittels der Allele weniger ausgewählter Polymorphismen (Kombination blauer und roter Balken) charakterisiert werden. Die Allele der übrigen Polymorphismen (schwarze Balken) stehen mit diesen Varianten im Kopplungsungleichgewicht und lassen sich aus den Allelen der ausgewählten Polymorphismen ableiten.

**Abbildung 2-2.** Bestimmung von htSNPs zur Charakterisierung von Haplotypblöcken[105].

Für das in diesem Rahmen durchgeführte Experiment wurden Patienten und Kontrollen anhand von Haplotypanalysen ausgewählt.

## 2.3 DNA-Extraktion aus Blut

Die genomische DNA wurde aus Blut mit dem Invitek DNA-Extraktionskit gewonnen. 14 mL Vollblut wurden für die Erythrozytenlyse mit Lyse-Puffer versetzt. Nach 10 minütiger Inkubation erfolgt ein Zentrifugationsschritt, der Überstand wird vorsichtig abgenommen und

verworfen. Die Probe wird mit dem gleichen Puffer erneut gewaschen und bei 1000 g für 5 Minuten zentrifugiert. Auch diesmal wird der Überstand vorsichtig abgenommen. Es sollten nun die weißen, Nukleus-beinhaltenden Leukozyten als Pellet am Boden des Eppendorfgefäßes zu sehen sein. Mit einem weiteren Lysepuffer und Proteinase- K werden bei 60 °C Inkubationstemperatur die Leukozyten und der Nukleus enzymatisch aufgeschlossen und die darin enthaltende DNA freigesetzt. Zu dem Lysat wird nun die DNA Präzipitationslösung zugegeben, vorsichtig gemischt und fünf Minuten bei 2000 g zentrifugiert. Der Überstand über dem DNA-Pellet wird wiederum sehr vorsichtig abpipettiert. Das Pellet wird nun mit 70% Ethanol gewaschen und getrocknet, bevor es in Niedrigsalzpuffer TE zur Langzeitlagerung gelöst wird.

## 2.4 „Next Generation Sequencing“-Sequenzierung mit SOLiD™ Technologie

### *2.4.1 Selektion der Ziel-Region und Anreicherung durch Long-Range-PCR (LR-PCR)*

Für diese Studie wurden zehn durch GWAS identifizierte und validierte Risikoloci für verschiedene Phänotypen, v.a. CED ausgewählt. Insgesamt umfasst die Summe aller ausgewählten Ziel-Loci eine Größe von ca. 700 kb. Sämtliche LR-PCR-Primer sind mit der Software Primer3 (http://frodo.wi.mit.edu/primer3/) erstellt worden[106]. Für jedes erwünschte Amplikon wurden mehrere Alternativprimerpaare generiert, um eine möglichst hohe Erfolgsquote zu erzielen. Das Primerpaar, das für das erwartete Amplikon die besten Ergebnisse lieferte, wurde für die PCR auf den eigentlichen Proben DNAs benutzt. Die HPLC-aufgereingten Primer wurden bei der Firma Metabion (Martinsried, Deutschland) bestellt. Nicht für alle Regionen konnten funktionierende Primer erstellt werden, so dass ca 600 kb der ursprünglich anvisierten Zielregion sequenziert werden konnten (siehe Tabelle 2-1). Kodierende Regionen, die nicht durch LR-PCR abgedeckt werden konnten, wurden mittels Sanger-Sequenzierung sequenziert.

**Tabelle 2-1.** Die Tabelle zeigt die Risikoloci, die für das Resequenzierungsprojekt ausgewählt worden sind. Die ursprünglich anvisierte Größe der Ziel-Region konnte nicht komplett realisiert werden, so dass sich für die Loci Lücken ergeben.

| Locus | Phänotyp | Genomische Position der ursprünglichen Ziel-Region | Größe [bp] | Genomische Position der angereicherten Ziel-Region | Größe der Ziel-Region [bp] | Anzahl der PCR Produkte |
|---|---|---|---|---|---|---|
| *IL10* | CED | chr1:204983095-205015648 | 32553 | IL10 chr1:204983095-205015648 | 32553 | 5 |
| *NKX2-3* | CED | chr10:101258700-101288268 | 29568 | NKX2-3 chr10:101259344-101284361 | 25017 | 4 |
| *STAT3* | CED | chr17: 37718869-37800000 | 81131 | STAT3_1 chr17:37719018-37773124 | 54106 | 8 |
| | | | | STAT3_2 chr17: 37780266-37799837 | 19571 | 3 |
| *IL23R* | CED | chr1:67376513-67497961 | 121448 | IL23R_1 chr1:67376513-67393978 | 17465 | 2 |
| | | | | IL23R_2 chr1:67404546-67432961 | 28415 | 2 |
| | | | | IL23R_3 chr1:67437222-67440720 | 3498 | 1 |
| | | | | IL23R_4 chr1:67440915-67448416 | 7501 | 1 |
| | | | | IL23R_5 chr1:67457673-67458069 | 396 | 1 |
| | | | | IL23R_6 chr1:67464201-67471806 | 7605 | 1 |
| | | | | IL23R_7 chr1:67478210-67481458 | 3248 | 1 |
| | | | | IL23R_8 chr1:67486379-67497220 | 10841 | 1 |
| *IRGM* | CED | chr5:150164616-150254191 | 89575 | IRGM_1 chr5:150164616-150181646 | 17030 | 2 |
| | | | | IRGM_2 chr5:150204143-150213581 | 9438 | 1 |
| | | | | IRGM_3 chr5: 150221168-150254191 | 33023 | 5 |
| *ATG16L1* | CED | chr2:233820329-233869785 | 49456 | ATG16L1 chr2:233820329-233869785 | 49456 | 8 |
| *NOD2* | CED | chr16:49285038-49325336 | 40298 | NOD2 chr16:49285038-49325336 | 40298 | 6 |
| *ANXA11* | Sarcoidose | chr10:81902860-81957308 | 54448 | ANXA11 chr10:81902956-81949577 | 46621 | 8 |
| *MDR3* | PCS | chr7:86863322-86947270 | 83948 | MDR3 chr7:86863322-86947270 | 83948 | 12 |
| *IL12B* | Psoriasis | chr5:158673369-158800000 | 126631 | IL12B chr5:158673417-158673417 | 126500 | 19 |
| ***Total*** | | | **709056** | **616530** | **616530** | **91** |

Die Länge der erwarteten LR-PCR-Amplikons variiert zwischen fünf und elf kb. Nebeneinanderliegende PCR-Amplikons überlappen mit mindestens 100 bp um möglichst keine Lücken zu generieren.

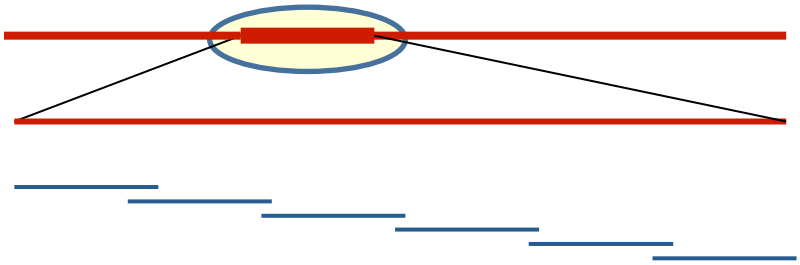


**Abbildung 2-3.** Schema eines LR-PCR basierten Enrichments einer definierten Region aus dem Genom. Die PCR-Produkte (hier blau) überlappen sich idealerweise in einem kleinen Bereich, so dass möglichst keine Lücken entstehen.

Jedes der verwendeten LR-PCR-Primerpaare wurde auf einer Kontroll-DNA (eingestellt auf 50 ng/µl) optimiert, d.h. es wurde versucht, für jedes Primerpaar die optimalen Bedingungen für die PCR-Reaktionen zu finden, unter denen ein verwertbares Amplikon generiert wurde. Für sämtliche Primer wurde im ersten Schritt mit einem Standardansatz eine PCR mit einem Temperaturgradient durchgeführt. Für jedes Primerpaar wurde der Standardansatz in 12 Parallellreaktionen aufgeteilt und für diese gleichzeitig bei unterschiedlichen Primer-Annealingtemperaturen eine PCR durchgeführt. Die Reaktionsbedingen sind also für alle Parallelansätze identisch, ausgenommen der Annealingtemperatur. Die PCR-Primeroptimierung wurde mit einer Biometra-Gradient-PCR-Maschine (Biometra, Göttingen, Deutschland) durchgeführt. Im Idealfall findet man allein mit dieser Methode optimale Bedingungen für die PCR. Da die Primer sehr unterschiedlich zu optimieren waren, mussten für einen Großteil der Primerpaare mehrere Optimierungsläufe durchgeführt werden und mehrere die PCR beeinflussenden Parameter und Reaktionskomponenten variiert werden. Solche Parameter sind Konzentration von $Mg^{2+}$-Ionen, Konzentration der Ausgangs-DNA, Konzentration der Primer, Zusatz von PCR-Additiven wie DMSO, sowie alternative Polymerasen.

Ein Standardansatz für die durchgeführten LR-PCRs enthielt 50 ng genomische DNA, 1,5 mL (10 mM) des Vorwärts und Rückwärtsprimers, 5 mL „GeneAmp High Fidelity 10x PCR Puffer“ (Applied Biosystems) sowie 2,5 Units des „GeneAmp High Fidelity Enzyme Mix“ (Applied Biosystems). Die PCR wurde mit folgenden Parametern durchgeführt: 2 Minuten bei 94 °C, 35 Zyklen bei 94 °C für 15 Sekunden (initiale Denaturierung), 30 Sekunden bei einem Temperaturgradienten von 54–63 °C (Primerannealing), 9 Minuten bei 68 °C (Elongation), gefolgt von einer finalen Elongation bi 72 °C für 7 Minuten.

### *2.4.2 Aufreinigung der PCR Produkte*

Alle PCR-Produkte wurden anschließend an die PCR mittels einer Agarosegel-Elektrophorese überprüft und mit Hilfe eines Qiagen Roboters 8000 (Qiagen, Hilden, Deutschland) und Qiaquick Aufreinigungs-Kits (Qiagen, Hilden, Deutschland) aufgereinigt. Diese Aufreinigung bewirkt eine effiziente Entfernung der Primer und nicht verwendeter Nukleotide. Die Proben wandern unter einem Vakuum durch eine Silica-Membran, die DNA adsorbiert unter Hochsalzbedingungen an der Membran, während die Kontaminanten ungehindert durch die Membran gehen. Verunreinigungen werden effizient weggewaschen und die reinen DNA-Fragmente werden mit 100 µl Tris Puffer eluiert. Die aufgereinigten PCR Produkte wurden mit Picogreen (Invitrogen, Carlsbad, CA, USA) quantifiziert und equimolar nach folgender Formel miteinander gemischt:

$$\text{Menge des PCR-Produkt [ng]} = \frac{\text{Länge des Fragments [bp]} \times \text{Menge der input-DNA für library [ng]}}{\text{Länge des target [bp]}}$$

### *2.4.3 Multiplex-Ansatz und Sequenzierstrategie*

Um die Kapazität der Achtelsequenzierspots eines Sequenzier-Slides zu nutzen, mussten die über Long-Range-PCR generierten Amplikons auf eine spezielle Art und Weise gemischt werden, damit die Sequenzen hinterher den einzelnen Individuen zugeordnet werden können. Insgesamt mussten über 90 Amplikons in 56 Individuen generiert werden. Insgesamt waren das weit über 5000 Amplikons. Tabelle 2-1 zeigt die für die jeweiligen Loci benötigten Fragmente, die pro Individuum für diese Region 56 mal amplifiziert werden mussten, um die Region abzudecken. Alle PCR-Produkte wurden einer Gelkontrolle unterzogen und die Konzentration der PCR-Produkte fluorometrisch bestimmt. Nach der Konzentrations-bestimmung mussten die PCR-Produkte in einem bestimmten Verhältnis zueinander gemischt werden. Die PCR-Produkte eines Individuums wurden als erstes zu „individuellen Pools“ gemischt. 56 solcher „individuellen Pools“ gibt es pro Locus. Diese „individuellen Pools“ wurden zu sogenannten „Gen Pools“ gemischt. Dazu wurde aus jedem Locus ein „individueller Pool“ genommen und mit je einem „individuellen Pool“ aus den anderen Loci gemischt. Es wurden insgesamt 56 solcher sogenannten „Gen Pools“ hergestellt. Jeder „Gen Pool“ bestand somit aus 10 „individuellen Pools“, pro Locus ein Individuum. (siehe Abbildung 2-4). Aus diesen „Gen-Pools“ wurden anschließend die 56 SOLiD-Sequenzierlibraries generiert. Mit Hilfe dieses Multiplexing-Ansatzes war es möglich ohne

Barkodierung der Sequenzierlibraries mit nur insgesamt sieben Sequenzierungsläufe zehn krankheitsassoziierte Loci in 56 Individuen zu sequenzieren.

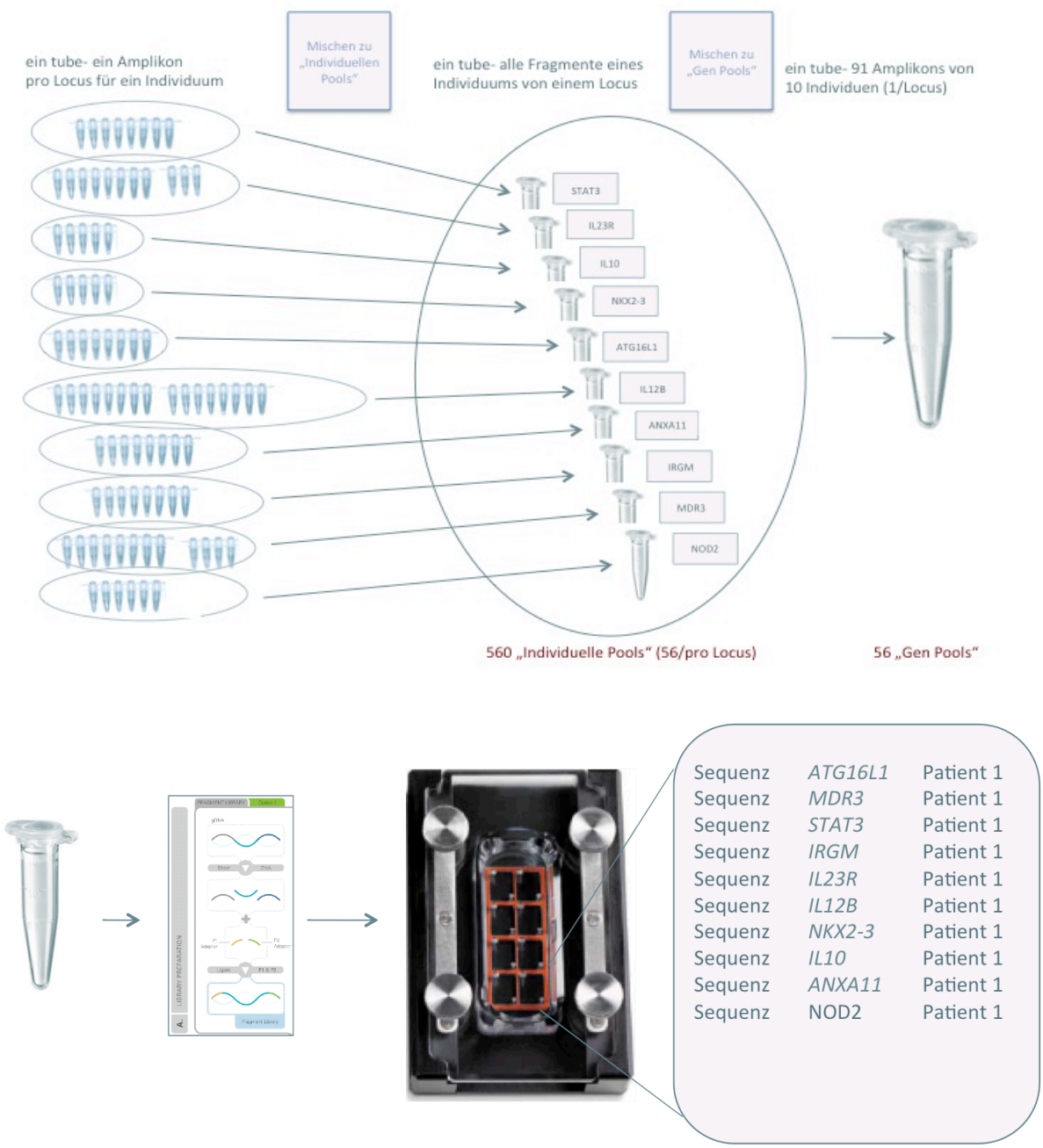



**Abbildung 2-4.** Darstellung der Multiplex- und Sequenzierstrategie. Beschreibung im Text.

## 2.5 Herstellung und Sequenzierung von SOLiD 50 bp Fragment-libraries

Die SOLiD-Sequenzierung (Sequencing by Oligonucleotide Ligation and Detection) hat seinen Ursprung in einem System, was durch Shendure et al[107] beschrieben wurde und durch die Arbeit von McKernan und Kollegen von Agencourt Personal Genomics (Beverly, MA,

USA) (2006 gekauft durch Applied Biosystems (Foster City, CA, USA) zur Reife gebracht wurde.

### *2.5.1 Librarypräperation*

Für die Herstellung der sog. SOLiD 50 bp Fragment-library wurden 2 µg pro Pool als Ausgangsmaterial eingesetzt. Die PCR-Amplikons wurden mit Hilfe des Covaris™S2 Systems in kleinere Fragmente mit einer Zielgröße von 125 bp geschert. Die Enden der resultierenden Fragmente wurden mit *End Polishing enzymes 1 und 2* „repariert" und mit *Pure Link™columns* (Invitrogen™ Corporation) aufgereinigt. Danach wurde die verbleibende Menge der DNA mit Hilfe eines Nanodrop®ND-1000 Spectrophotometer (Thermo Scientific) vermessen. Danach erfolgte die Ligation der doppelsträngigen Adaptoren P1 und P2. Nach der Ligation wurden die Fragmente mit einer korrekten Größe aus einem 2%-igen Agarosegel ausgeschnitten und aufgereinigt. Im folgenden Schritt wurde eine Nicktranslation der Libraries durchgeführt. Die Libraries wurden anschließend unter Benutzung der Library Primer 1 und 2 und mit Hilfe des *Platinum® PCR Amplification Mix* (Invitrogen™ Corporation) PCR-basiert vermehrt und mittels quantitativer PCR quantifiziert.

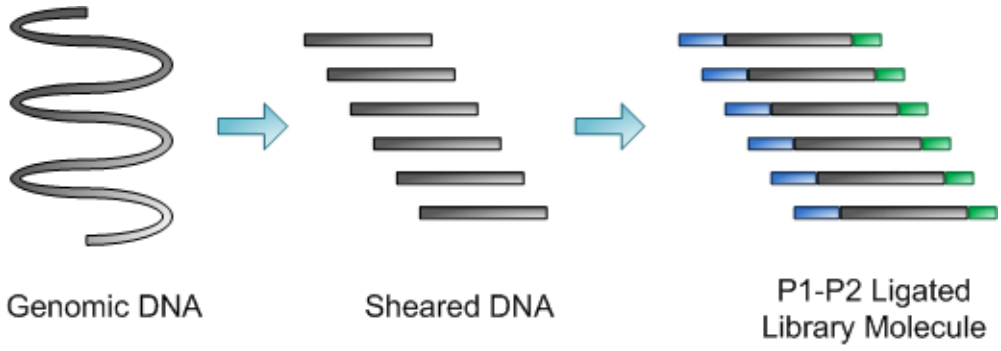



**Abbildung 2-5.** Schematische Darstellung für die Herstellung einer Fragmentlibrary (Abbildung entnommen aus Applied Biosystems SOLiD ™ 4 System Library Preparation Guide[108])

### *2.5.2 Emulsions-PCR, Enrichment und 3´ End-Modifikation*

Jedes Librarymolekül wurde klonal an einem sog. „Bead" (kleines Kügelchen) in einer Emulsions-PCR amplifiziert. Um aus der wässrigen Phase und Ölphase eine Emulsion zu generieren, wurde das ULTRA-TURRAX™ Tube Drive der Firma IKA® genutzt. Für die Herstellung der Ölphase für die Emulsion wurden 1,8 ml Emulsionsstabilisator 1, 0,4 ml Emulsionsstabilisator 2 in 37,8 ml Emulsionsöl gegeben. 9 ml von dieser Ölphase und 5,6 ml des Emulsions-PCR-Mix (Wasser, PCR Puffer, dNTPs Mix, $MgCl^{2+}$, Emulsions-PCR-Primer 1 und 2, AmpliTaq Gold UP, P1-Beads und Library) wurden für die Herstellung der Emulsion

benutzt. Nach der Emulsions-PCR wurden die Emulsionen mit der Zugabe von 2-Butanol gebrochen. Die Beads wurden gewaschen und es wurden mit sog. Enrichmentbeads die Beads angereichert, die ein Template-Molekül tragen. Durch Ultraschallbehandlung mit Hilfe des Covaris™S2 System wurden die Beads voneinander getrennt. Die magnetischen DNA-tragenden Beads wurden magnetisch von den Enrichmentbeads getrennt. Die sich anschließende 3´ End-Modifikation mit einem Beadlinker durch eine terminale Transferase erlaubt eine kovalente Bindung des Beads an die Oberfläche des Slides, ein speziell beschichteter Objektträger.

### *2.5.3 SOLiD- Sequenzierung*

Die SOLiD- Sequenzierung ist eine auf Ligation beruhende Reaktion. Ein Universalprimer hybridisiert an die komplementäre P1-Adaptor-Sequenz. Ein Pool von fluoreszenzmarkierten sog. *dibase Probes*, Oktamere mit vier unterschiedlich fluoreszierenden Farbstoffen, wetteifern um die Ligation an die Sequenzierprimer. Jeder fluoreszierende Farbstoff repräsentiert vier der 16 möglichen Dinukleotid-Sequenzen. Komplementäre *probes* werden während des Sequenzierungsprozesses an die Sequenzierprimer ligiert, anschließend wird die Fluoreszenz gemessen. Multiple Zyklen von Ligation, Detektion und Abspaltung werden durchgeführt. Die Leselänge ist determiniert durch die Anzahl der Zyklen. Nach einer Reihe von Ligationszyklen, wird der synthetisierte DNA-Strang entfernt und ein neuer Primer, der komplementär zur Position n-1 ist, hybridisiert für eine zweite Runde der Ligationszyklen. Dieses Zurücksetzten der Primer wird fünfmal wiederholt. (siehe Abbildung 2-6). Für einen 50 bp-Fragment-Sequenzierlauf schließt jeder Primerzyklus zehn Ligationen der floureszensmarkierten Oktamere ein. Dieses Prinzip der Sequenzierung mit den Dibase Probes erlaubt das zweimalige Abfragen einer jeden sequenzierten Base (siehe Abbildung 2-7).

benutzt. Nach der Emulsions-PCR wurden die Emulsionen mit der Zugabe von 2-Butanol gebrochen. Die Beads wurden gewaschen und es wurden mit sog. Enrichmentbeads die Beads angereichert, die ein Template-Molekül tragen. Durch Ultraschallbehandlung mit Hilfe des Covaris™S2 System wurden die Beads voneinander getrennt. Die magnetischen DNA-tragenden Beads wurden magnetisch von den Enrichmentbeads getrennt. Die sich anschließende 3´ End-Modifikation mit einem Beadlinker durch eine terminale Transferase erlaubt eine kovalente Bindung des Beads an die Oberfläche des Slides, ein speziell beschichteter Objektträger.

### *2.5.3 SOLiD- Sequenzierung*

Die SOLiD- Sequenzierung ist eine auf Ligation beruhende Reaktion. Ein Universalprimer hybridisiert an die komplementäre P1-Adaptor-Sequenz. Ein Pool von fluoreszenzmarkierten sog. *dibase Probes*, Oktamere mit vier unterschiedlich fluoreszierenden Farbstoffen, wetteifern um die Ligation an die Sequenzierprimer. Jeder fluoreszierende Farbstoff repräsentiert vier der 16 möglichen Dinukleotid-Sequenzen. Komplementäre *probes* werden während des Sequenzierungsprozesses an die Sequenzierprimer ligiert, anschließend wird die Fluoreszenz gemessen. Multiple Zyklen von Ligation, Detektion und Abspaltung werden durchgeführt. Die Leselänge ist determiniert durch die Anzahl der Zyklen. Nach einer Reihe von Ligationszyklen, wird der synthetisierte DNA-Strang entfernt und ein neuer Primer, der komplementär zur Position n-1 ist, hybridisiert für eine zweite Runde der Ligationszyklen. Dieses Zurücksetzten der Primer wird fünfmal wiederholt. (siehe Abbildung 2-6). Für einen 50 bp-Fragment-Sequenzierlauf schließt jeder Primerzyklus zehn Ligationen der floureszensmarkierten Oktamere ein. Dieses Prinzip der Sequenzierung mit den Dibase Probes erlaubt das zweimalige Abfragen einer jeden sequenzierten Base (siehe Abbildung 2-7).

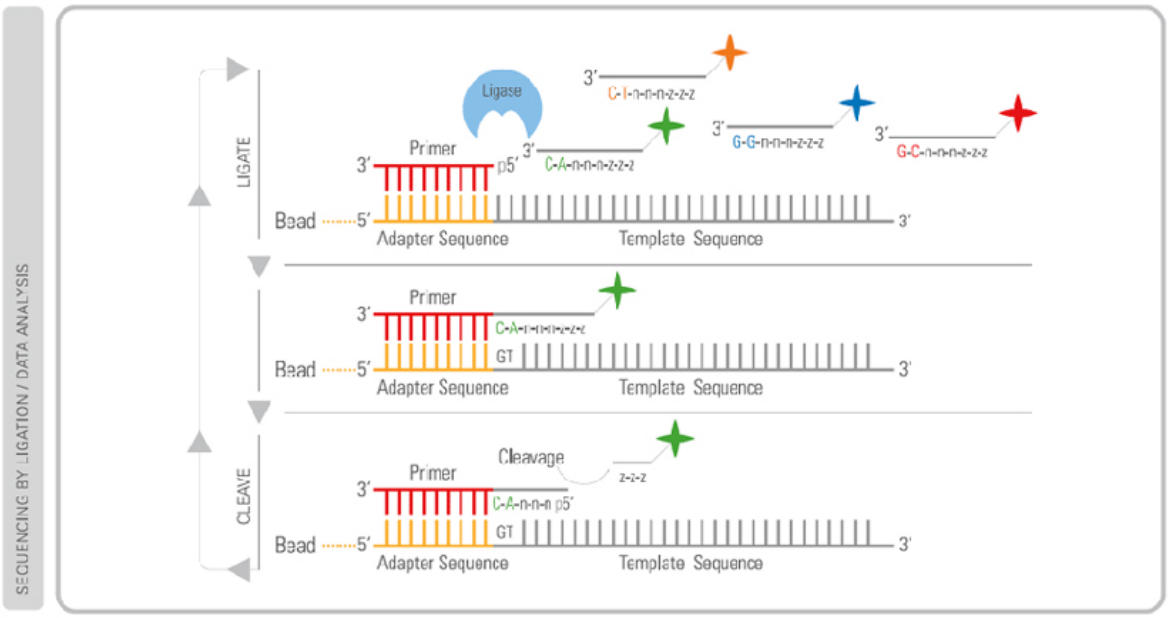



**Abbildung 2-6.** Schematische Darstellung des Ablaufs des Sequenzierungsvorgangs **(**Abbildung entnommen aus Applied Biosystems SOLiD ™ 4 System Library Preparation Guide[108]).

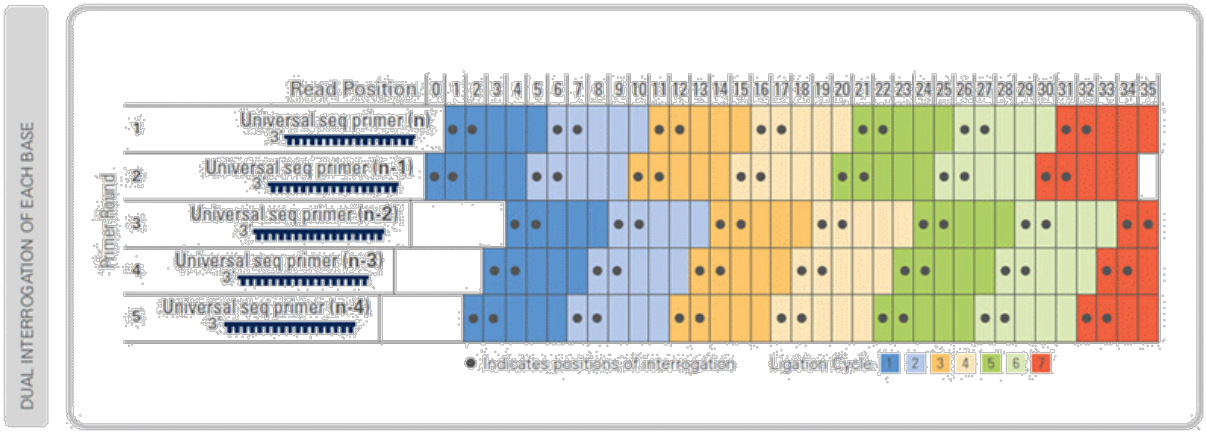



**Abbildung 2-7.** Die Abbildung zeigt, dass bei der SOLiD-Sequenzierung jede Basenposition zweimal sequenziert wird (Abbildung entnommen aus Applied Biosystems SOLiD ™ 4 System Library Preparation Guide[108]).

## 2.6 Alignment gegen die Referenzsequenz

Die 50mer Sequenzier-Reads, wurden gegen NCBI36/hg18 Referenzsequenz aligniert, was dem gesamten humanen Genom entspricht. Auf Grund einer schlechten Abdeckung und großer Lücken in der Zielregion, wurden die Reads gegen die Ziel-Region gemappt. So wurde erreicht, dass mehr Reads in der Ziel-Region landen, auch z. B. jene, die nicht einmalig im Genom mappen würden, sondern mehrfach, auch außerhalb der Ziel-Region. Für dieses sog. Target-Region-Mapping wurde eigens für dieses Projekt eine FASTA-Referenzsequenz erstellt, die aus größeren Blöcken der Target-Region, jeweils getrennt durch einen Block von

50Ns, besteht. Die Blöcke des Targets bestanden aus den zusammengesetzten Referenzsequenzen (entnommen aus UCSC Genome Browser) der PCR-Produkte. Dabei wurde die Sequenz des Forward-Primers des ersten PCR-Produktes eines Blockes und die Sequenz des letzten Reverse-Primers eines zusammenhängenden Blockes als Start-und Endpunkt eines Referenzsequenzblockes genutzt.

Das Mappen der Reads wurde für die Daten von jeden Sequenzierungslauf und jede Library separat mit Standardeinstellungen der BioScope Software[109] ausgeführt. Alle Mappingschritte und weiteren Analysen, die an die Software Bioscope geknüpft waren, wurden auf dem Linux Hochleistungsrechencluster des Rechenzentrums der Universität Kiel (http://www.rz.uni-kiel.de/hpc/rzcluster) durchgeführt.

## 2.7 SNP-calling mit diBayse und pileup (SAM tools)

Die SNP- bzw. SNV-Identifikation erfolgte in zwei Schritten und mit Hilfe zweier Programme. Zuerst wurde das SNP-calling mit dem diBayes SNP-caller, der in BioScope implementiert ist (User Guide[109], S. 173-179) für alle 56 Libraries unter der Verwendung der BAM-files aus dem Mappingschritt mit den Standardparametern durchgeführt.

Das SNP-calling ist auf Grund fehlender Erfahrungen mit dem mitgelieferten SNP-caller diBayes zusätzlich mit dem SNP-caller „pileup“ basierend auf der Software SAMtools[110] durchgeführt worden. Das SNP-calling wurde mit den Einstellungen für die Konsensus Quality 40 (basierend auf Phred-Scala), für die SNP Qualität 40 (basierend auf Phred Scala) und die minimale Coverage 8x gestartet. Die Ergebnisse beider SNP-callings sind zusammengefasst worden, da beide Programme in der SNP-Detektion potentiell falsch negative Ergebnisse lieferten. Die Ergebnisse der zusammengeführten SNP-callings waren die Grundlage weiterer Analysen.

## 2.8 Annotation der SNPs

Das Annotationsprogramm SnpActs ist ein Datenbank-basiertes Softwarepacket um SNPs zu analysieren und kategorisieren. Die SNPs, die im vorangegangenem diBayes-Analyseschritt detektiert worden sind, werden in SnpActs unter dem Gesichtspunkt ihrer Funktionalität eingeteilt. Für exonische SNPs werden „Scores“ mit Hilfe von Vorhersageprogrammen errechnet, um die potentielle Auswirkung auf die Änderung der Aminosäuresequenz abzuschätzen. Außerdem sammelt SnpActs Informationen über SNPs, z.B. von UCSC, dem

„the 1000-Genomes-Project“, HGMD oder diverser Vorhersageprogrammen. SnpActs wurde am IKMB in Kiel entwickelt und wird ständig erweitert.

Die deskriptive Statistik für die detektierten SNPs wurde ebenfalls mit SnpActs durchgeführt (siehe Ergebnisse).

## 2.9 Identifikation von falsch-positiven SNPs unter Verwendung von pibase und IGV

Von diesem Analyseschritt an wurde sich mit weiteren Analyseschritten auf die CED-Loci *ATG16L1, IL10, NOD2, NKX2-3, STAT3, IL23R* und *IRGM* beschränkt. Die Ergebnisse der verbleibenden Loci wurden an die entsprechenden Arbeitsgruppen zur Weiterverarbeitung übergeben.

### *2.9.1 Analyse mit dem Validierungstool pibase*

Für die pibase-Analyse wurde eine Liste mit potentiellen SNV-Positionen generiert. Es wurden alle Positionen für SNVs, die mindestens einmal in den 56 Libraries detektiert worden sind, mit Hilfe dieses Programmes für jede Library abgefragt.

Das Programm pibase[111] analysiert für alle in dieser Liste vorhandenen Koordinaten die BAM-files der Libraries und die dem SNP-calling zugrunde liegende Referenzsequenz und generiert für alle abgefragten Koordinaten und Libraries eine Zusammenfassungstabelle. Die Tabelle beinhaltet viele Informationen zur Bewertung des SNP-callings: Kontext in Referenzsequenz, Allel-Coverage, unique Startpoints für jeweils A, C, G, T, N auf jedem Strang und für fünf Filterstufen. In der ersten Filterstufe werden Reads mit Indels (Insertionen/Deletionen, gewöhnlich Artefakte im Alignment) entfernt, in der zweiten Filterstufe ist ein sogenannter „Base quality filter“ hinzugefügt (Standardeinstellung: PHRED-like score ≥ 20), durch den viele potentielle Sequenzierungsfehler entfernt werden. Die dritte Filterstufe filtert für die Länge der Reads, die ins Alignment eingeschlossen wurden (Standardeinstellung: ≥ 49 Nukleotide). Dieser Filter eliminiert viele falsch gemappte Reads in weniger komplexen Regionen. Das vierte Filterlevel integriert ein „Mismatch-per-Read“ Filter (Standardeinstellung ist ein Mismatch) und entfernt weitere fehlerhaft gemappte Reads, das fünfte Filterlevel hat zusätzlich einen „Mapping quality Filter“ (Standardeinstellung: MapQV 20) und es werden dadurch zufällig gemappte Reads entfernt. Mit Hilfe der pibase Ausgabedatei ist es möglich zehn verschiedene Ausführungen für die Anzahl der Nukleotidesignale je nach Filterstufe miteinander zu vergleichen und das SNP-calling zu evaluieren. Die Parameter für die Filtereinstellungen können verändert werden.

Das pibase-consensus-Tool errechnet außerdem auf Basis der BAM-files und Filter einen sogenannten „best-genotype“ für jede genomische Koordinate. Die Genotypen „best genotype“ werden nach folgenden Standardregeln (variierbar) detektiert: Ein Allel wird detektiert, wenn mindestens 2,2 % der Reads dieses Allel zeigen und wenn mindestens 4% der unique Startpoints dieses Allel unterstützen. Der „best genotype“ wird in einer nebenstehenden Spalte qualitativ mit ?1 bis ?8 bewertet, wobei die Qualität von ?1 zu ?8 ansteigt. Die „best genotypes“ mit ?1 wurden nochmal genauer inspiziert, die „best genotypes“ mit Qualitätswert ?2 und höher wurden als als Grundlage für den Abgleich mit dem SNP-calling genutzt. Ist kein Qualitätslabel neben der Spalte „best genotype“ zu finden, ist dieser als „best genotype“ errechnete Genotyp für alle Filtermethoden bei einer minimalen Coverage von acht Reads und minimal vier unique Startpoints. Die durch pibase errechneten „best genotypes“ dienten als Grundlage zur Eliminierung von falsch-positiven SNP-calls, die mit relativ geringen Aufwand gegen die SNP-Detektionsergebnisse verglichen werden konnten. Eliminiert wurden SNP-calls, bei denen der „best genotype“ der pibase Ausgabedatei homozygot fürs Referenzallels war oder gar keine Genotypen aufgrund mangelnder qualitativ hochwertiger Reads oder fehlender Coverage errechnet wurde. Abweichungen in der Homo- bzw Heterozygotie von SNPs wurden nicht berücksichtigt, da die Sequenzierung in erster Linie für die Detektion von (potentiellen) SNP Positionen benutzt wurde.

### *2.9.2 Inspektion mit IGV*

Mit Hilfe des Integrative Genomics Viewer (frei zugänglich unter http://www.broadinstitute.org/igv/) wurden sich gezielt potentielle SNP-Positionen angeschaut, um die Qualität einzelner SNP-calls zu evaluieren. Mit der Software ist es möglich, das auf BAM-Files basierende Alignment visuell darzustellen und sich theoretisch jede einzelne Base anzusehen (z. B. Coverage, Allelecounts, Coverage der SNV-Umgebung).

## 2.10 Sanger-Sequenzierung

Für die Abdeckung von exonischen Bereichen, die durch das Enrichment mit Long-Range-PCR und der nachfolgenden SOLiD-Sequenzierung nicht abgedeckt wurden, wurden diese konventionell nach der Sanger-Methode sequenziert. Sanger Sequenzierung ist seit der Originalpublikation von Sanger et al. von 1977[99] eine Standardmethode in der Sequenzanalyse.

Für die Sanger-Sequenzierung werden die Zielbereiche im Genom, also diejenigen Bereiche der DNA, an denen man interessiert ist und die sequenziert werden sollen, über PCR amplifiziert. Für diese Bereiche wurden alle Primer mit dem frei zugänglichen Programm Primer3 (http://frodo.wi.mit.edu/) designt. Die HPLC aufgereinigten Primer sind bei der Firma Metabion (Martinsried) bestellt worden. Die Länge der Amplikons variierte zwischen 150 und 600 bp. Es wurde für alle Primerpaare ein Standardprotokoll mit dem AmpliTaq Gold Kit durchgeführt. Die 25 µl PCR Reaktionen wurden mit 5 ng genomischer DNA, 0,4 µl des Forward- und Reverseprimers (10 µM), 0,5 µl dNTPs (10 mM), 2,5 µl 10x PCR-Puffer (Applied Biosystems), 3 µl MgCl2 (25 mM) und 0,15 µl AmpliTaq Gold (5 u/µl) durchgeführt. Die PCR Reaktionen wurden in 96-well GeneAmp®PCR System 9700 PCR Maschinen mit folgendem touch-down PCR-Programm durchgeführt: 95 °C für 5 min gefolgt von 16 Zyklen 95 °C für 30 sec (Denaturierung), 66 °C (touch-down -0,5 °C jeder Zyklus) für 30 sec (Annealing), 72 °C für 30 sec (Extension), gefolgt von 19 Zyklen 95 °C für 30 sec (Denaturation), 58 °C für 30 sec (Annealing), 72 °C für 30 sec (Extension) und beendet durch 72 °C für 10 min (finale Extension). Die Amplikons wurden nach der PCR-Reaktion gelelektrophoretisch evaluiert. (Gelelektrophoresebedingung 1,5% Gel, 110 V, 400 mA, 30 min). Die Amplikons wurden mit SAP und Exo enzymatisch aufgereinigt. Für die sich anschließende Sequenzierreaktion wurde das BigDye Terminator v1.1 Cycle Sequencing Kit benutzt. Dieser Ansatz enthält neben den für eine gewöhnliche PCR üblichen Desoxynukleosidtriphosphaten (dNTPs) auch sogenannte Didesoxynukleosidtriphosphate (ddNTP), die jeweils unterschiedlich fluoreszenzmarkiert sind. Diese ddNTPs besitzen keine 3`-Hydroxygruppe. Werden diese in den neusynthetisierten DNA-Strang eingebaut, ist eine weitere Strangverlängerung durch die Polymerase nicht mehr möglich, die Reaktion bricht ab. Da dieser Einbau der ddNTPs gegenüber der dNTPs rein zufällig ist, entstehen in der Folge der Reaktion DNA-Fragmente unterschiedlichster Länge. Diese Sequenzierreaktion wurde mit folgendem Protokoll durchgeführt: BigDye 0,7 µl, 5x 1,5 µl Sequenzierpuffer, 0,32 µl Forward- oder Reverseprimer (10 µM), 5,48 µl H2O, 2 µl des aufgereinigten PCR-Produkts. Für die Sequenzierreaktion wurde folgendes PCR Programm benutzt: 96 °C 1 min, gefolgt von 25 Zyklen 96 °C für 10 sec (Denaturierung), 5 sec bei 50 °C (Annealing) and 4 min bei 60 °C (Extension). Die Proben wurden über Sephadex-Säulen aufgereinigt und mit 3730xl Kapillarsequenzierern der Firma Applied Biosystems kapillarelektrophoretisch nach ihrer Größe aufgetrennt, durch einen Laser zur Fluoreszenz angeregt und die Signale über einen Detektor abgenommen. Das resultierende Chromatogramm gibt direkt die Basenabfolge des sequenzierten DNA-Strangs.

Das SNP-calling wurde durchgeführt mit Hilfe des halbautomatisierten Programm novoSNP. Frei zugänglich online unter www.molgen.ua.ac.be/bioinfo/novosnp/).

## 2.11 Assoziationsexperimente

### *2.11.1 Selektion von SNPs für das sich anschließende Assoziationsexperiment I*

Insgesamt sind 88 SNPs in 364 in Morbus Crohn bzw. an Colitis Ulcerosa erkrankten Individuen und in 368 deutschen Kontrollindividuen genotypisiert worden.

Alle SNP-Positionen, die durch die SOLiD-Sequenzierung und der Sanger-Sequenzierung detektiert worden sind, wurden in das institutsinterne Tool *SnpActs* gegeben. Für alle detektierten kodierenden Varianten (bekannt/unbekannt) sind Genotypisierungsassays für das anschließende Assoziationsexperiment mit der Sequenom™ Technologie erstellt worden.

Außerdem wurden die genomischen Koordinaten aller detektierten SNVs an die Arbeitsgruppe von Thomas Manke (Max Planck Institut für Molekulare Genetik, Berlin) übergeben. Das bioinformatorisches Softwaretool sTRAP (strap.molgen.mpg.de) berechnet mögliche Konsequenzen von Sequenzvariationen in regulativen Netzwerken. sTRAP analysiert die Variationen in der DNA-Sequenz und kann quantitative Änderungen der Bindungsaffinität von allen Transkriptionsfaktoren vorhersagen, für das Affinitätsmodelle existieren[113]. In das erste Genotypisierungsverfahren sind zusätzlich all jene SNPs eingeflossen, die eine Änderung in der Affinität für Transkriptionsfaktoren, die eine Rolle in Stoffwechselwegen von CED spielen könnten, zur Folge hatten.

In die Follow-Up Genotypisierung flossen außerdem noch alle SNPs aus der Publikation Daly et al[88] [Nature 2011] mit ein, die in dieser untersuchten Regionen liegen. Insgesamt 14 SNPs sind dieser Publikation entnommen, von denen drei SNPs aber auch in den eigenen Sequenzierdaten gefunden worden sind.

Eine Übersicht der Variationen, die für das Follow-Up Genotyping ausgewählt wurden, findet sich in Tabelle 2-2.

**Tabelle 2-2.** Liste der ausgewählten SNPs, die im Assoziationsexperiment I mit Hilfe der Sequenom-Technologie genotypisiert worden sind.

| | Locus | Chr | Pos | Herkunft | Grund | Name |
|---|---|---|---|---|---|---|
| 1 | IL23R | 1 | 67384201 | SOLiD-Seq | TF-Bindestelle | rs4655683 |
| 2 | IL23R | 1 | 67384867 | SOLiD-Seq | TF-Bindestelle | SNP_46 |
| 3 | IL23R | 1 | 67385090 | SOLiD-Seq | TF-Bindestelle | SNP_47 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | IL23R | 1 | 67386595 | SOLiD-Seq | TF-Bindestelle | SNP_1 |
| 5 | IL23R | 1 | 67386703 | SOLiD-Seq | TF-Bindestelle | SNP_2 |
| 6 | IL23R | 1 | 67388595 | SOLiD-Seq | TF-Bindestelle | SNP_48 |
| 7 | IL23R | 1 | 67390806 | SOLiD-Seq | TF-Bindestelle | SNP_51 |
| 8 | IL23R | 1 | 67391598 | SOLiD-Seq | TF-Bindestelle | SNP_52 |
| 9 | IL23R | 1 | 67391921 | SOLiD-Seq | TF-Bindestelle | SNP_4 |
| 10 | IL23R | 1 | 67392833 | SOLiD-Seq | TF-Bindestelle | SNP_53 |
| 11 | IL23R | 1 | 67392837 | SOLiD-Seq | TF-Bindestelle | SNP_54 |
| 12 | IL23R | 1 | 67406400 | SOLiD-Seq | exonisch | rs1884444 |
| 13 | IL23R | 1 | 67406551 | SOLiD-Seq | TF-Bindestelle | rs11465770 |
| 14 | IL23R | 1 | 67408538 | SOLiD-Seq | TF-Bindestelle | rs2295359 |
| 15 | IL23R | 1 | 67444019 | SOLiD-Seq | TF-Bindestelle | SNP_7 |
| 16 | IL23R | 1 | 67445873 | SOLiD-Seq | TF-Bindestelle | SNP_9 |
| 17 | IL23R | 1 | 67446087 | SOLiD-Seq | TF-Bindestelle | SNP_58 |
| 18 | IL23R | 1 | 67448316 | SOLiD-Seq | TF-Bindestelle | rs10889669 |
| 19 | IL23R | 1 | 67457857 | Daly et al | Validierung | SNP_114 |
| 20 | IL23R | 1 | 67457975 | SOLiD-Seq | exonisch | rs7530511 |
| 21 | IL23R | 1 | 67486458 | SOLiD-Seq | TF-Bindestelle | SNP_60 |
| 22 | IL23R | 1 | 67486654 | SOLiD-Seq | TF-Bindestelle | SNP_61 |
| 23 | IL23R | 1 | 67488758 | SOLiD-Seq | TF-Bindestelle | SNP_62 |
| 24 | IL23R | 1 | 67497416 | SOLiD-Seq | exonisch | rs11465827 |
| 25 | IL23R | 1 | 67497521 | SOLiD-Seq | exonisch | SNP_124 |
| 26 | IL23R | 1 | 67497708 | SOLiD-Seq | exonisch | rs10889677 |
| 27 | IL23R | 1 | 67497795 | SOLiD-Seq | exonisch | SNP_126 |
| 28 | IL10 | 1 | 204984177 | SOLiD-Seq | TF-Bindestelle | SNP_63 |
| 29 | IL10 | 1 | 204993863 | SOLiD-Seq | TF-Bindestelle | rs61815630 |
| 30 | IL10 | 1 | 204994237 | SOLiD-Seq | TF-Bindestelle | SNP_17 |
| 31 | IL10 | 1 | 204997334 | SOLiD-Seq | TF-Bindestelle | rs4579758 |
| 32 | IL10 | 1 | 205008810 | SOLiD-Seq | TF-Bindestelle | SNP_18 |
| 33 | ATG16L1 | 2 | 233823325 | SOLiD-Seq | TF-Bindestelle | SNP_34 |
| 34 | ATG16L1 | 2 | 233829561 | Daly et al | Validierung | SNP_119 |
| 35 | ATG16L1 | 2 | 233830146 | SOLiD-Seq | TF-Bindestelle | SNP_81 |
| 36 | ATG16L1 | 2 | 233830746 | SOLiD-Seq | TF-Bindestelle | SNP_35 |
| 37 | ATG16L1 | 2 | 233836551 | SOLiD-Seq | exonisch | rs13011156 |
| 38 | ATG16L1 | 2 | 233838333 | Daly et al | Validierung | SNP_120 |
| 39 | ATG16L1 | 2 | 233840262 | SOLiD-Seq | TF-Bindestelle | SNP_36 |
| 40 | ATG16L1 | 2 | 233846431 | SOLiD-Seq | exonisch | SNP_122 |
| 41 | ATG16L1 | 2 | 233848107 | SOLiD-Seq | exonisch | rs2241880 |
| 42 | ATG16L1 | 2 | 233851452 | SOLiD-Seq | TF-Bindestelle | SNP_37 |
| 43 | ATG16L1 | 2 | 233852754 | SOLiD-Seq | TF-Bindestelle | SNP_85 |
| 44 | ATG16L1 | 2 | 233856269 | SOLiD-Seq | TF-Bindestelle | rs2278610 |
| 45 | IRGM | 5 | 150181543 | SOLiD-Seq | TF-Bindestelle | SNP_86 |
| 46 | IRGM | 5 | 150206334 | SOLiD-Seq | TF-Bindestelle | SNP_87 |
| 47 | IRGM | 5 | 150206548 | SOLiD-Seq | TF-Bindestelle | SNP_43 |
| 48 | IRGM | 5 | 150207929 | Daly et al | Validierung | SNP_116 |
| 49 | IRGM | 5 | 150208020 | Daly et al | Validierung | SNP_117 |
| 50 | IRGM | 5 | 150208159 | Daly et al | Validierung | SNP_118 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 51 | IRGM | 5 | 150208191 | SOLiD-Seq | exonisch | SNP_131 |
| 52 | IRGM | 5 | 150228330 | SOLiD-Seq | TF-Bindestelle | SNP_89 |
| 53 | IRGM | 5 | 150230561 | SOLiD-Seq | TF-Bindestelle | rs1277463 |
| 54 | IRGM | 5 | 150242995 | SOLiD-Seq | TF-Bindestelle | SNP_91 |
| 55 | IRGM | 5 | 150244285 | SOLiD-Seq | TF-Bindestelle | SNP_92 |
| 56 | IRGM | 5 | 150244718 | SOLiD-Seq | TF-Bindestelle | SNP_93 |
| 57 | IRGM | 5 | 150252921 | SOLiD-Seq | TF-Bindestelle | SNP_94 |
| 58 | NKX2-3 | 10 | 101260253 | SOLiD-Seq | TF-Bindestelle | SNP_66 |
| 59 | NKX2-3 | 10 | 101266896 | SOLiD-Seq | TF-Bindestelle | SNP_19 |
| 60 | NKX2-3 | 10 | 101278795 | SOLiD-Seq | TF-Bindestelle | SNP_20 |
| 61 | NKX2-3 | 10 | 101283846 | SOLiD-Seq | TF-Bindestelle | rs884144 |
| 62 | NOD2 | 16 | 49291360 | SOLiD-Seq | exonisch | rs2067085 |
| 63 | NOD2 | 16 | 49299292 | Daly et al | Validierung | SNP_130 |
| 64 | NOD2 | 16 | 49302066 | SOLiD-Seq | exonisch | SNP_129 |
| 65 | NOD2 | 16 | 49302125 | SOLiD-Seq | exonisch | rs2066842 |
| 66 | NOD2 | 16 | 49302189 | Daly et al;SOLiD-Seq | exonisch | rs5743271 |
| 67 | NOD2 | 16 | 49302393 | Daly et al | Validierung | SNP_98 |
| 68 | NOD2 | 16 | 49302615 | Daly et al | Validierung | SNP_99 |
| 69 | NOD2 | 16 | 49302618 | SOLiD-Seq | exonisch | rs2076754 |
| 70 | NOD2 | 16 | 49302700 | SOLiD-Seq | exonisch | rs2066843 |
| 71 | NOD2 | 16 | 49303084 | SOLiD-Seq | exonisch | rs1861759 |
| 72 | NOD2 | 16 | 49303156 | SOLiD-Seq | exonisch | rs61736932 |
| 73 | NOD2 | 16 | 49303302 | SOLiD-Seq | exonisch | SNP_104 |
| 74 | NOD2 | 16 | 49303373 | Daly et al | Validierung | rs5743276 |
| 75 | NOD2 | 16 | 49303427 | SOLiD-Seq | exonisch | rs2066844 |
| 76 | NOD2 | 16 | 49308311 | Daly et al | Validierung | SNP_109 |
| 77 | NOD2 | 16 | 49308343 | Daly et al;SOLiD-Seq | exonisch | SNP_110 |
| 78 | NOD2 | 16 | 49314041 | Daly et al;SOLiD-Seq | exonisch | rs2066845 |
| 79 | NOD2 | 16 | 49314777 | SOLiD-Seq | exonisch | rs5743291 |
| 80 | STAT3 | 17 | 37724627 | SOLiD-Seq | TF-Bindestelle | rs8066464 |
| 81 | STAT3 | 17 | 37733350 | SOLiD-Seq | TF-Bindestelle | SNP_70 |
| 82 | STAT3 | 17 | 37749602 | SOLiD-Seq | TF-Bindestelle | SNP_73 |
| 83 | STAT3 | 17 | 37753998 | SOLiD-Seq | TF-Bindestelle | SNP_26 |
| 84 | STAT3 | 17 | 37761955 | SOLiD-Seq | TF-Bindestelle | SNP_76 |
| 85 | STAT3 | 17 | 37772557 | SOLiD-Seq | TF-Bindestelle | SNP_29 |
| 86 | STAT3 | 17 | 37782849 | SOLiD-Seq | TF-Bindestelle | SNP_77 |
| 87 | STAT3 | 17 | 37793614 | SOLiD-Seq | TF-Bindestelle | SNP_78 |
| 88 | STAT3 | 17 | 37799123 | SOLiD-Seq | TF-Bindestelle | SNP_33 |

### *2.11.2 Selektion der SNPs für das Assoziationsexperiment II*

Für das zweite Assoziationsexperiment wurden 27 Varianten aus dem Assoziationsexperiment I ausgewählt und in 2800 deutschen Kontrollen und 2500 Crohnpatienten genotypisiert. Ausgewählt wurden SNPs, die im ersten Experiment einen $p < 0.05$ erreicht hatten und von guter Qualität sind, dabei hatten neue unbekannte Varianten eine höhere Priorität als bekannte

SNPs. SNPs, von denen Genotypisierungsdaten durch das Immunochip-Projekt existieren, wurden im zweiten Validierungsschritt nicht erneut genotypisiert. Unter den 27 ausgewählten Varianten wurden 25 unbekannte Varianten (11 davon waren im Assoziationsexperiment I monomorph) genotypisiert. Die Tabelle 4-7 zeigt neben den Ergebnissen des Assoziationsexperimentes I, welche SNPs aus dieser Studie für das zweite, größer angelegte Experiment ausgesucht wurden.

### *2.11.3 Genotypisierung mittels Sequenom*

Die SNP-Genotypisierung für die Assoziationsexperimente wurde innerhalb dieses Projektes mittels Sequenom Technologie durchgeführt, die auf MALDI-TOF Massenspektrometrie beruht. Der genomische Bereich um die zu untersuchenden SNPs wird mit Hilfe einer PCR amplifiziert. Die für diese Amplifikation verwendeten „Sense“- und „Antisense“- Primer weisen neben dem sequenz-spezifischen Bereich jeweils ein Motiv von 10 Basen auf. Dieses zusätzliche Sequenzmotiv, der sogenannte „tag“, hat mehrere Funktionen. Einerseits verläuft in einer Multiplex-PCR (eine Reaktion, wenn mehrere PCR Reaktionen in einem Gefäß parallel ablaufen) die Amplifikationsreaktion gleichmäßiger, andererseits ist es nötig, die Masse dieser Primer mit Hilfe des „tags“ soweit zu erhöhen, dass sie bei der anschließenden MALDI-TOF Analyse nicht im Massenbereich der Primer-Extensionsprodukte liegen und somit außerhalb des eingestellten Massenfensters sind, da die überschüssigen, in der PCR nicht verbrauchten Primer nicht entfernt werden. Nach erfolgter PCR werden freie, nicht in das Amplifikat eingebaute Desoxynukleotide durch das Enzym SAP (shrimp alkaline phosphatase) enzymatisch abgebaut. Bei der Primer-Extensionsreaktion wird der sogenannte Extensionsprimer so gewählt, dass das 3`Ende unmittelbar vor dem zu detektierenden SNP liegt. Entsprechend der Basenabfolge auf dem Template-Strang wird das 3`Ende mit einer speziellen DNA-Polymerase verlängert. Durch die Zugabe einer bestimmten Mischung aus Di-Desoxynukleotiden und Desoxynukleotiden entstehen je nach Zustand des zu detektierenden Allels unterschiedliche Primer-Extensionsprodukte.

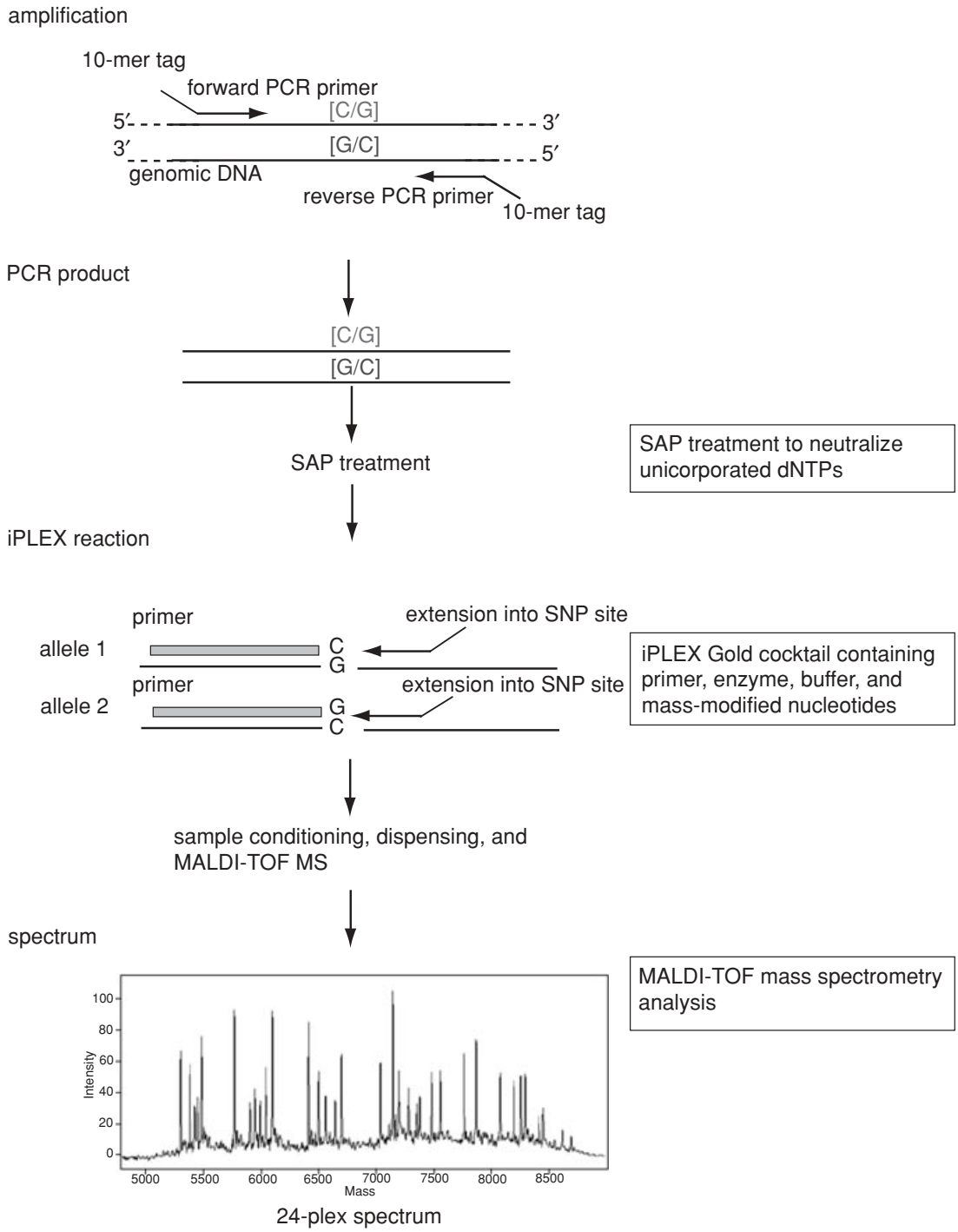



**Abbildung 2-8.** Schematische Übersicht über die iPlex-Reaktion und die Genotypisierung eines C-G SNPs. (Abbildung entnommen aus Gabriel et al. 2009[114])

Mit Hilfe des Spectro-Point-Nanoliter-Pipetting Systems™ der Firma Sequenom (Hamburg, Deutschland) wurden nach erfolgter Probenaufbereitung 1-2 nl des Analytengemisches auf einen Siliziumchip gespottet. Nach Beladung des Chips wurde dieser auf einen metallischen Probenträger transferiert und in die Vakuumschleuse des MassARRAY™

Massenspektrometers eingeführt. Die Vermessung der Proben und Analyse sowie die Kalibrierung des Systems erfolgten automatisch nach Standardeinstellungen.

### *2.11.4 Statistische Methoden und Qualitätskontrolle der Genotyp-Daten*

Eine indirekte Kontrolle der Qualität der Genotypisierungsdaten ist das Bestimmen der Abweichung vom Hardy-Weinberg-Gleichgewicht. Dem zugrunde liegt das Hardy-Weinberg-Modell, was auch für die meisten menschlichen Populationen seine Gültigkeit hat[115]. Wenn die Annahme dieses Modells für eine Population zutrifft und die Allel- und Genotypfrequenzen über viele Generationen stabil bleiben, befindet sich diese Population im Hardy-Weinberg-Equilibrium (HWE); Signifikante Abweichungen von diesem Maß sind entweder ein Hinweis auf Genotypisierungsfehler oder dass auf dieses Allel ein hoher evolutiver Druck ausgeübt wird. Marker, welche in der Kontrollgruppe signifikant vom Hardy-Weinberg-Gleichgewicht abwichen ($p < 0{,}05$), wurden deshalb von der weiteren Analyse ausgeschlossen. Darüber hinaus sind auch solche Marker von der weiteren Analyse ausgeschlossen worden, deren Genotypisierungserfolgsrate unter 95% lag.

Die Analyse der Allel- und Genotypdaten wurde als Fall-Kontroll-Assoziationsanalyse (case *control*, CC) unter Berechnung eines $\chi 2$ und einer Schätzung der *Odds Ratio* (OR) durchgeführt. Die Allel- und Genotypfrequenzen für die einzelnen Assoziationsmarker wurden jeweils getrennt für die Kontroll- und Patientengruppe berechnet und miteinander verglichen.

# 3 Material und Methoden für die Gesamt-Genom-Sequenzierung

## 3.1 Patientin

Die Morbus Crohn-Patientin, die ihre DNA für die Gesamt-Genom-Sequenzierung zur Verfügung stellte, ist Geburtsjahrgang 1963 mit einer Krankengeschichte für Morbus Crohn seit 1985. In der Ambulanz für chronisch entzündliche Darmerkrankungen am Universitätsklinikum Schleswig-Holstein, Campus Kiel, wurde sie das erste mal 2002 wegen rezidivierenden aktiven Morbus Crohn vorstellig. Die Probe dieser Patientin ist von hieran mit der Proben-ID "Amb132" benannt wurden. In der Vergangenheit wurde sie mit Mesalamin und/oder Corticosteroiden therapiert. Im Februar 2000 wurde eine ileozökale Resektion durchgeführt, bei der vorrübergehend ein Ileostoma angelegt wurde, welches im Oktober des selben Jahres zurückverlegt wurde. Die Patientin ist verheiratet, Mutter zweier Kinder und war Raucherin (100 Packungen/Jahr). Beide Kinder zeigen bisher keinerlei Symptome, die für Morbus Crohn sprechen.

Bei der Erstvorstellung hatte die Patientin keinerlei Medikation, 12 Stuhlgänge pro Tag, starke abdominale Schmerzen und ein reduziertes Allgemeinbefinden. Eine Koloskopie zeigte den für Morbus Crohn typischen segmentalen Befall des gesamten Kolons und des terminalen Ileums. Sie bekam TNF-Binde-Protein-1 (TBP-1) im Rahmen einer randomisierten Placebo-kontrollierten klinischen Studie. Innerhalb von zwei Monaten war die Patientin in stabiler Remission mit einer Reduktion der täglichen Stuhlfrequenz von mehr als zehn flüssiger Stühle auf einen geformten Stuhl pro Tag, was unter der TBP-1 Therapie bis Juli 2003 erhalten blieb, als sie erneut auf Grund massiver abdominaler Krämpfe ins Krankenhaus aufgenommen werden musste und ein Ileus diagnostiziert wurde. Als Folge dieser Diagnose erfolgte eine Sigmoidektomie. Der weitere Krankheitsverlauf war gekennzeichnet durch Stenosen, sowohl in Kolon als auch in der Anastomoseregion. Intensivierte Anti-inflammatorische Therapie mit Azathioprin und Infliximab konnte die Entzündung nicht stoppen. Nach einem anaphylaktischen Schock während einer Infusion im Juli 2005 wurde die Inflaximab-Therapie unterbrochen und der Dickdarm wurde im November desselben Jahres chirurgisch entfernt. Im August 2008 entwickelte sie einen Ileus, 20 cm des Dünndarms wurden daraufhin entfernt. Weitere abdominale Operationen wurden auf Grund von Abzessbildungen nötig und im Verlauf der nächsten Monate verlor die Patientin 20 kg Körpergewicht. Von November 2008 bis Juli wechselte die Patientin zu parenteraler Ernährung, mit der Hoffnung die Krankheit in Remission zu bringen. Die letzte

endoskopische Untersuchung erfolgte im Januar 2010, bei der keinerlei Entzündungen im Dünndarm gesehen werden konnten (weder makroskopisch noch histologisch).

## 3.2 DNA-Isolierung

Genomische DNA wurde aus EDTA-Blut mit Hilfe eines Präperationsroboters (Autopure; Qiagen, Hilden, Deutschland) isoliert und aufgereinigt. Die DNA-Probe wurde anschließend qualitativ über eine Gelelektrophorese überprüft und die Konzentration wurde mit Hilfe von Picogreen (Molecular Probes; Life Technologies, Carlsbad, CA, USA) bestimmt.

## 3.3 Library-Herstellung

### *3.3.1 Konstruktion von 50 bp Fragment-Libraries*

Die Herstellung der 50 bp erfolgte entsprechend dem Kapitel 2.5.1.

### *3.3.2 Konstruktion von 2x50 bp Mate-Paired-Libraries*

10 µg genomische DNA wurden in Fragmente einer spezifischen Größe geschert. Für die Mate-Paired-Library mit einer Insertgröße <500 bp wurde die DNA mit einem Nebulizer eine Minute bei 2,5 bar fragmentiert. Um die DNA für die Mate-Paired-Libraries mit einer Insertgröße zwischen 3-4 kb zu fragmentieren, wurde der Hydroshear der Firma Genomic Solutions® benutzt, ein System, dass die hydrodynamischen Scherkräfte für die Fragmentierung nutzt (10 Zyklen; Speedcode 15). Die erhaltenen Fragmente wurden aufgereinigt und ein sog. *end-repair* durchgeführt. Dies ist nötig, um eine effiziente Ligation der doppelsträngigen LMP-CAP-Adaptoren an die DNA-Fragmente zu gewährleisten. Auf die Ligation der Adaptoren folgte eine Selektion der Fragmente nach der richtigen Größe. Die DNA-Fragmente werden dafür einer Gel-Elektrophorese unterzogen (0,8 oder 1,5% Agarose Gel, abhängig von Fragmentgröße). Die Fragmente mit der korrekten Größe wurden aus dem Gel ausgeschnitten und mit dem PureLink™ QuickGelExtraction-Kit der Firma Life Technologies (CA, USA) aufgereinigt und die eluierte DNA mit einem Nanodrop® ND-1000 Spectrophotometer (Thermo Scientific; MA, USA) quantifiziert. Im nächsten Schritt wurden die erhaltenen Fragmente mit einem doppelsträngigen, biotinylierten Adaptor zirkularisiert. Eine „Epicentre® Plasmid-Safe™ATP-Dependent DNase“ (Biozym, Hessen, Deutschland) wurde eingesetzt, um unzirkularisierte Fragmente zu entfernen. Die zirkularisierten Moleküle wurden vor einer Quantifizierung mit dem Nanodrop® ND-1000 Spectrophotometer unter Verwendung des PureLink PCR Micro Kit (Life Technologies) aufgereinigt. Es erfolgte eine

Nick-Translation mittels *E.coli* DNA-Polymerase I. Ein Verdau mit T7-Exonuklease und S1-Nuklease (Life Technologies) wurde durchgeführt. Der Nick in der zirkularisierten DNA wird von der T7-Exonuklease erkannt. Durch den Verdau des unligierten Stranges mit ihrer 5`-zu 3`-Exonukleaseaktivität wird eine Lücke in die Sequenz erzeugt. Die S1-Nuklease erkennt die Einzelstrangregion und spaltet das Librarymolekül von dem zirkularisierten Template. Um die Librarymoleküle von Nebenprodukten frei zu bekommen, wurden die Librarymoleküle an Streptavidin-Beads gebunden. Diese Dynabeads® MyOne™Streptavidin-C1 (Life Technologies) binden spezifisch an den Biotin-gelabelten internen Adaptor im Librarymolekül. P1- und P2-Adaptoren wurden an die Library ligiert und die an die Streptavidin-Beads gebundenen Librarymoleküle wurden in Puffer gewaschen und von Nebenprodukten entfernt. Die Library wurde unter Verwendung des Platinum® PCR-Amplification-Mix amplifiziert. Die Library wurde über ein Lonza Flash Gel 4% gelelektrophoretisch aufgetrennt und aus dem Gel aufgereinigt. Die Library-Bande (250-350bp) wurde ausgeschnitten, aufgereingt und für die sich anschließende Emulsions-PCR über eine quantitative PCR quantifiziert.

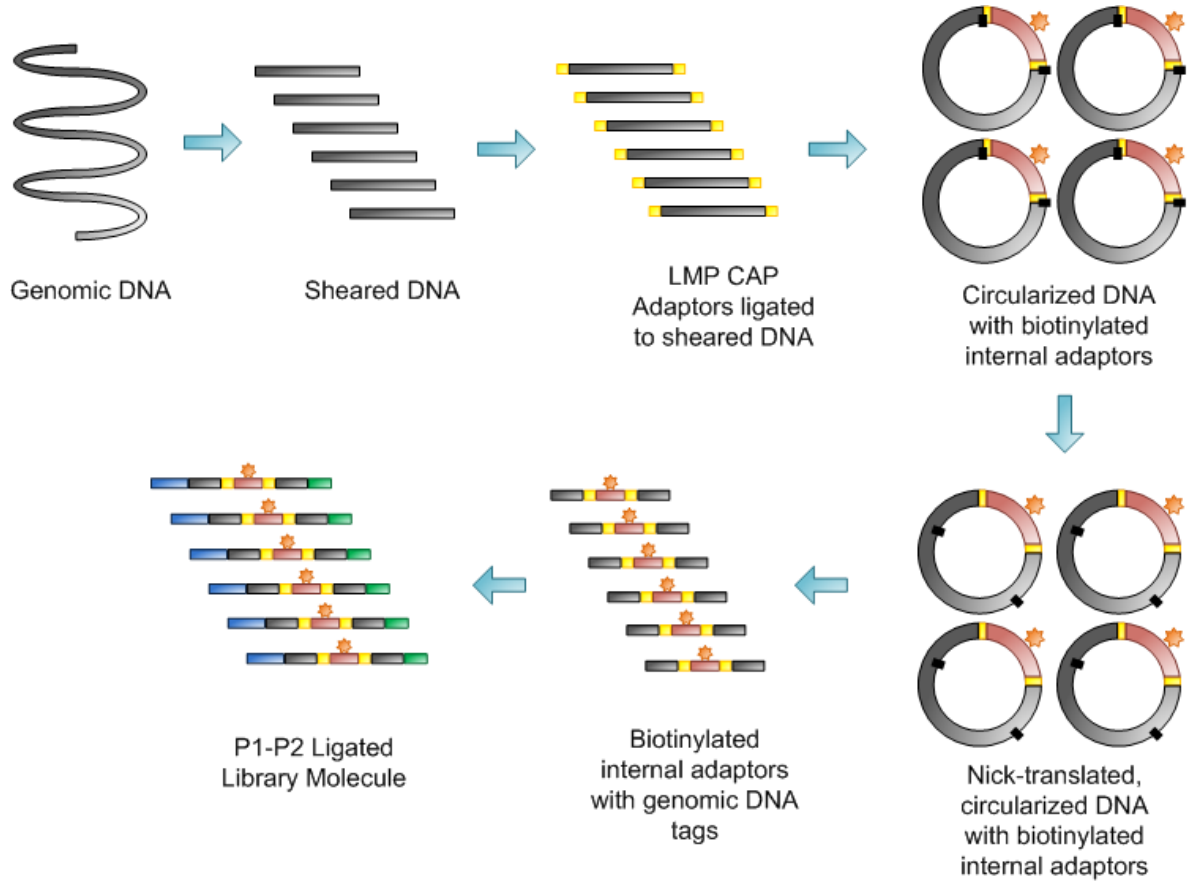



**Abbildung 3-1.** Die schematische Darstellung zeigt die Herstellung von SOLiD 2x50bp Mate-Paired-Libraries (Abbildung entnommen aus Applied Biosystems SOLiD ™ 4 System Library Preparation Guide[108] ).

### *3.3.3 Konstruktion von 2x25 bp Mate-Paired-Libraries*

15 µg genomische DNA wurden für die Herstellung von 2x25 bp mate-paired-Libraries mit einer Insertgröße zwischen 4-5 kb und 51 µg für die Konstruktion einer mate-paired-Library 0,5 und 1 kb fragmentiert. Für das Fragmentieren wurde der Hydroshear® benutzt. Nach der Aufreinigung wurde eine Agarosegel-Elektrophorese-Größenselektion durchgeführt (1,5% Agarosegel für Größenselektion bis 1 kb). Für Fragmente zwischen 4-5 kb wurde ein 0,8% Agarosegel für die Größenselektion genutzt). *EcoP15I*-CAP-Adaptoren (Life Technologies) wurden an die fragmentierte, aufgereinigte und methylierte DNA ligiert. Nach erfolgter Ligation der Adaptoren wurden die Fragmente mit einem biotinylierten internen Adaptor zirkularisiert. Die Methylierung der *EcoP15I*-Seiten vor der Adaptorligation in der Ziel-DNA stellt sicher, dass *EcoP15I* nur unmethylierte Restriktionstellen im CAP-Adaptor während des Verdaus erkennt. Bei dem Verdau mit *EcoP15I* schneidet das Enzym 25-27bp von der unmethylierten Restriktionserkennungssequenz im CAP-Adaptor. Nach einem Aufreinigungsschritt mit Dynabeads® MyOne™ Streptavidin-C1 (Life Technologies) erfolgte die Ligation der P1- und P2-Adaptoren (siehe Abbildung 3-2).

Die Library wurde anschließend amplifiziert und für eine zweite Größenselektion über ein 4% Lonza Flash Gel laufen gelassen. Die Library-Bande (154-156 bp) wurde aus dem Gel ausgeschnitten, unter Verwendung von Qiaquick Purification Kit (Qiagen; Hilden, Germany) aufgereinigt und die Konzentration wurde über eine quantitative PCR bestimmt.

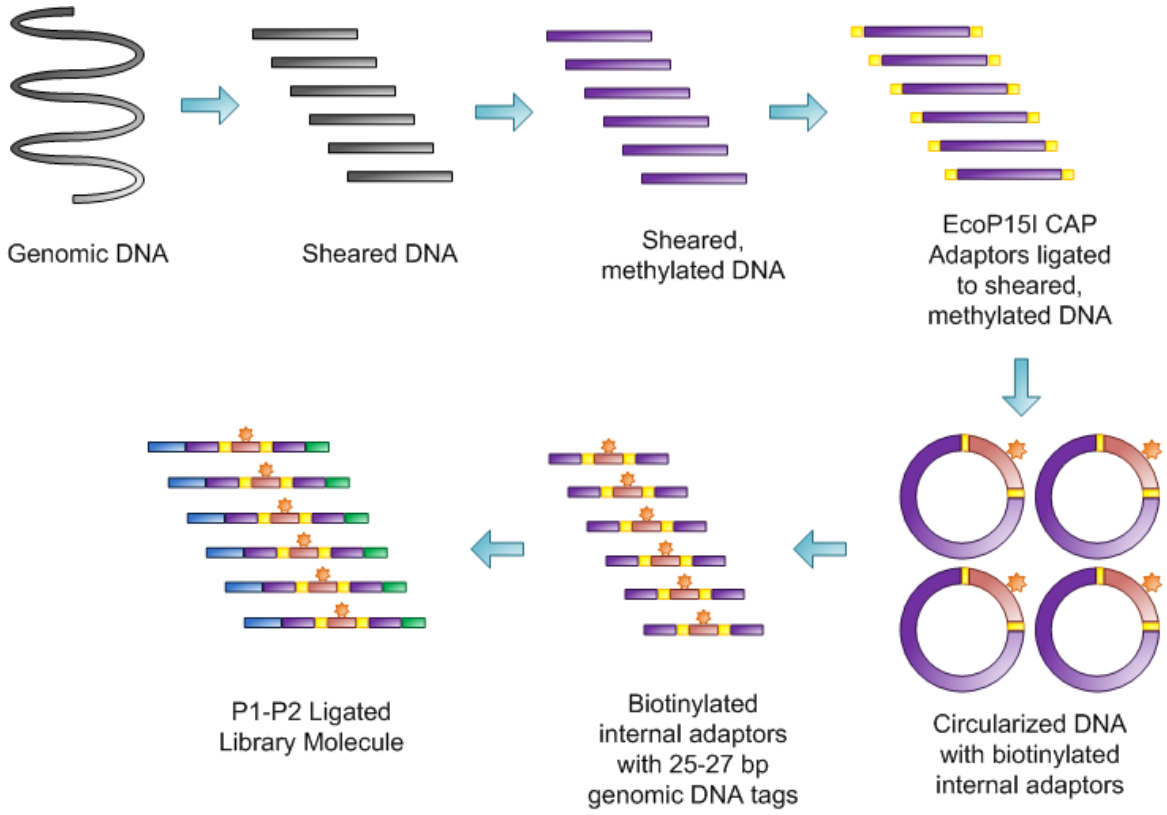



**Abbildung 3-2.** Schematische Darstellung der 2x25 bp SOLiD Mate-Paired-Libraries (Abbildung entnommen aus Applied Biosystems SOLiD ™ 4 System Library Preparation Guide[108]).

## 3.4 Emulsions-PCR, Enrichment und 3`End-Modifikation

Die Themen Emulsions-PCR, Enrichment und 3`End-Modifikation werden ausführlich im Kapitel 2.5.2 beschrieben und wurde bei der Gesamt-Genom-Sequenzierung entsprechend durchgeführt.

## 3.5 Sequenzierung

Im Kapitel 2.5.3 wird die SOLiD-Sequenzierung beschrieben.

## 3.6 Sequenz-Alignment

Die genomische Referenzsequenz, die für das Alignment benutzt wurde, basiert auf der humanen Genomassemblierung Version 36.1 (hg18, März 2006) des internationalen humanen Genomprojekt[116]. Aus der Referenzsequenz wurde die Sequenzen für die Chromosomen 1-22, X und die mitochondriale DNA herausgelöst, die Sequenz für das Y-Chromosom wurde ausgelassen, da die Patientin weiblich ist.

Alle Mappingschritte und weiteren Analysen, die an die Software BioScope geknüpft waren, wurden auf dem Linux Hochleistungsrechencluster des Rechenzentrums der Universität Kiel (http://www.rz.uni-kiel.de/hpc/rzcluster) durchgeführt.

Die SOLiD-Sequenzierung generiert pro Lauf verschiedene Plattform-spezifische Textdateien, die für den Mapping-Analyseschritt benötigt werden. Für jeden Primer wird eine Datei, das die Reads im sog. SOLiD *color space format* beinhaltet und eine Datei mit den damit korrespondieren Qualitätswerten generiert. Für die 16 Mate-pair-Sequenzierläufe und zwei Fragment-Sequenzierläufe wurden 68 Dateien generiert.

Das Read-Mapping wurde für die Daten für jeden Sequenzierlauf separat durchgeführt. Es wurden die Standardeinstellungen des Programms genutzt (User Guide[109], S.121-125). Die Ergebnisse des Mapping sind BioScope spezifische Dateien (`.ma` and `.stat`). Für die beiden Fragment-Läufe 17 und 18 wurden diese *Output-files* ins `BAM`-format[110] unter Verwendung des `maToBam` und `bamReport` Analyseschritts konvertiert. Für die 16 Mate-pair-Läufe folgte der sog. `pairing` Analyseschritt (User Guide[109], S149-154), wieder unter Verwendung der Standardparameter (User Guide[109] S. 157-160). Dieser Schritt generiert ebenso `BAM`-files wie für die Fragment-Läufe beschrieben. Für die folgenden Analysen mit BioScope wurden die `BAM` -files mit den gemappten Sequenzen als Inputfiles benutzt.

### 3.7 SNP-Detektion und Annotation

Die SNP-Detektion wurde in zwei Schritten ausgeführt, beide unter der Verwendung des diBayes SNP-caller, der in der BioScope Software implementiert ist (User guide[109], S173-179). Als erstes wurden das SNP-calling für alle 18 Läufe separat unter Verwendung der BAM-files vom Mappingschritt durchgeführt. Es wurden die Standardparameter angewendet (User Guide[109], S. 184-186) mit folgenden Änderungen: Mit `het.skip.high.coverage=1` wird das SNP-calling für die genomischen Proben 1-17 deaktiviert, wenn die Coverage einer Position zu hoch verglichen mit allen anderen SNP-Positionen ist. Für die Exom-Library (SureSelect) Nummer 18 war es erforderlich, dass ein Allel durch beide Stränge repräsentiert wird (`call.stringency=high`), während für die anderen Libraries diese Strangbedingungen nicht aktiviert waren (`call.stringency=medium`). Für alle Läufe/Libraries galt die Bedingung, dass die Reads, die einen SNP „unterstützen“, uniquely gemappt sein mussten (`reads.only.unique=1`). Im zweiten Schritt des SNP-callings, das zusammengefasste

SNP-calling, wurden die gleichen Bedingungen und Parameter wie für die genomischen Libraries eingestellt, wie oben beschrieben.

Die Annotation erfolgte mit dem SNP-Kategorisierungsprogramm *SnpActs* (siehe 2.8).

Die deskriptive Statistik für die detektierten SNPs wurde unter Verwendung von *SnpActs* durchgeführt. Zusätzlich wurden sie mit anderen SNP-Sets verglichen und mit anderen Referenzen in Beziehung gesetzt: mit den SNPs der öffentlichen und frei zugänglichen Datenbank dbSNP[117] (build130), denen des „1000-Genomes-Project“ und „The Human Gene Mutation Database (HGMD)“[118], (Professional Version2011.1). Für diesen Vergleich wurden die SNPs der Patientin bezüglich ihrer Neuheit und ihres schädigenden Effekts gruppiert.

### 3.8 Small InDel-Detektion und Annotation

Für das Mapping für die InDel-Detektion wurden die Mapping-Daten aller 18 Läufe genutzt. Entsprechend den Anweisungen des „Small InDelTool“ (User Guide[109], ab S. 263 ) wurden die BAM-files der 16 mate-pair Läufe direkt für das zusammengefasste InDel-calling benutzt. Dafür wurde das `smallIndelFrag` gestartet, um Deletionen und Insertionen mit einer Länge bis zu 11 resp. 3 Basen zu detektieren. Die resultierenden Dateien wurden wiederum zu BAM-files konvertiert. Beide Schritte wurden mit Standardparametern durchgeführt. Das Gleiche wurde angewendet für das folgende zusammengefasste InDel-calling, das auf allen 18 BAM-files beruht. Dieses resultierte in 3,611,003 InDel-Kandidaten.

Annovar[119] ist eine auf Kommandozeilen basierte Software zur Annotation genetischer Varianten. Die Annotation kann auf UCSC, ENSEMBLE; RefSeq, GENCODE und anderen Referenzen basieren. Annovar kann mit verschiedenen Formaten für genetische Variationen arbeiten, wie z.Bsp. samtools, Gff3, Soap, VCF. Letzteres wurde für die Annotation für die InDels der Crohn-Patientin benutzt.

### 3.9 CNV-Analyse

Die Daten aus der Sequenzierung der vier verschiedenen Mate-Paired-Libraries aus insgesamt 14 Läufen wurden zur Detektion von Copy Number Variation (CNVs) mit Hilfe des sog. Paired-End-Mapping analysiert. Dieses wurde in Zusammenarbeit mit der Arbeitsgruppe um Jan Korbel (EMBL Heidelberg) durchgeführt.

Als Ergebnis gab es eine Liste mit möglichen CNVs mit der Empfehlung auf Evaluation.

Aus der Liste wurden als erstes diejenigen entfernt, die die minimale Coverage von 5x, die dieses Ereignis unterstützen, unterschritten. Die verbleibenden möglichen CNVs wurden mit Hilfe des Integrative Genomics Viewer (http://www.broadinstitute.org/igv/) genauer analysiert. Dabei wurden folgende Kriterien aufgestellt: Als valide galten diejenigen, die mindestens in zwei von den vier Libraries detektiert wurden. In einem weiteren Schritt wurden die aus dieser visuellen Inspektion übriggebliebenen CNVs mit Hilfe des institutsinternen Genome-Browser auf Heterozygotie/Homozygotie, bekannt/unbekannt und betrifft Gen/intergenetische Bereiche analysiert. Eine Tabelle mit den Ergebnissen aus dieser Analyse findet sich im Ergebnisteil.

### 3.10 Validierung der ausgewählten SNPs mit Hilfe der Sanger- Technologie

Für die Validierung diverser SNPs sind alle detektierten Varianten als erstes mit Hilfe des SNP-Annotationstools *SnpActs* nach verschiedenen Kriterien gefiltert worden. Es wurden SNPs selektiert, die a) zu diesem Zeitpunkt unbekannt und b) in den durch die im Programm implementierten Vorhersageprogrammen zur Auswirkung auf den Organismus eine potentiell schädigende Wirkung hatten. Die in diesem Projekt zur SNP-Validierung angewendete Sanger-Sequenzierung ist im Kapitel 2.10 ausführlich beschrieben.

# 4 Ergebnisse- Resequenzierung der GWAS Loci

## 4.1 Mapping Ergebnisse

Die Abbildung 4-1 zeigt die Ergebnisse des Mappings gegen die Ziel-Region für jede der 56 sequenzierten Libraries. Pro Library sind die sogenannten uniquely mapped Reads sowie die Gesamtheit der assemblierten *Reads* dargestellt. Uniquely mapped Reads sind diejenigen, die nur einen Treffer in der Ziel-Region hatten. Diejenigen Reads, die mehrfach in der Ziel-Region assembliert werden konnten, finden in der weiteren Analyse keine Berücksichtigung mehr, da nicht mit Sicherheit gesagt werden kann, ob sie dort auch wirklich hingehören. Das Detektieren von Varianten baut auf den uniquely mapped Reads auf, da man bei diesen sicher sagen kann, dass sie wirklich von dieser Stelle, auf die sie aligniert sind, kommen. Die Abbildung zeigt, dass innerhalb des Experiments die Anzahl der Reads, die ins Mapping eingeschlossen werden konnten, sehr unterschiedlich von Library zu Library sind. Allerdings sind die Libraries, die zusammen auf einem Slide in einem Lauf sequenziert worden sind relativ ähnlich zueinander. Zusammen sequenziert wurden Library1-8, 9-16, 17-24, 25-32, 33-40, 41-48 und 49-56. Ein Teil der großen Unterschiede lässt sich also auch durch die Qualität der Sequenzierung auf dem Gerät erklären.

Allein die Mappingergebnisse versprechen aber schon eine mehr als ausreichende theoretische Coverage für ein erfolgreiches SNV-Calling.

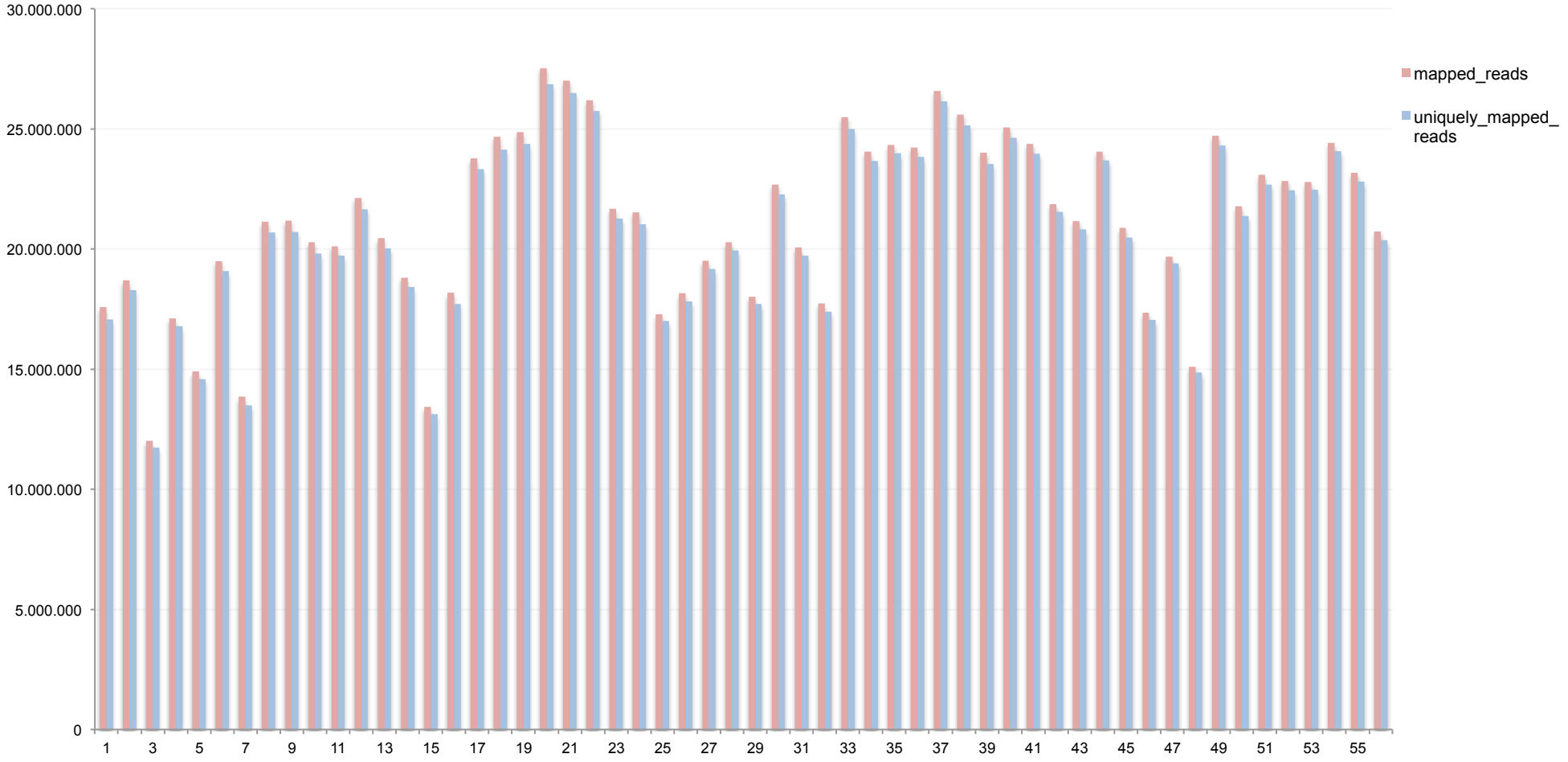



**Abbildung 4-1.** Die Tabelle zeigt die Mapping-Ergebnisse für das Resequenzierungsprojekt. Die Mappingergebnisse sind libraryweise dargestellt.

## 4.2 Coverages

Als Coverage bezeichnet man die Anzahl der Reads, mit der ein sequenzierter Bereich abgedeckt ist. Die Coverage kann also als ein Maß für die Qualität in der Sequenzierung gebraucht werden. Gerade in Experimenten mit „Next Generation Sequencing“ Technologie (Gesamt-Genom-Sequenzierung oder Gezieltes Resequenzierung) ist die Coverage ein Maß für das Gelingen des Experiments. Berechnet wird die durchschnittliche Coverage für einen definierten Bereich über den Quotienten der durchschnittlichen Anzahl der Reads multipliziert mit der durchschnittlichen Länge der Reads und der Größe des zu sequenzierenden Bereiches.

Die Tabelle 4-1 zeigt die durchschnittlichen Coverages und Mediane für die Coverages pro Locus und sequenzierter Library. Zusätzlich sind die prozentualen Werte der zu sequenzierenden Region angegeben, die eine Coverage von mehr als 50x aufweisen. Für ein erfolgreiches Detektieren von SNPs und Mutationen ist eine minimale Coverage erforderlich. Die minimal erforderlichen Coverages für SNP- und Mutationsdetektion sind auf Grund von verbesserter Sequenzierqualität und besserer Assemblierungs- sowie Mutations/SNP Detektionsalgorithmen deutlich unter 50x gesunken. Allerdings konnte unser Pilotprojekt zeigen, dass bei einer Coverage <20x die Wahrscheinlichkeit für falsch-positive/falsch-negative Ergebnisse in der SNP-Detektion steigt[120].

In diesem Experiment konnten für alle Loci befriedigende bis gute Coverages für die meisten Sequenzier-Libraries erzeugt werden. Dass weniger als 90% der Regionen mit einer minimalen Coverage von 50x sequenziert wurden, findet sich in einzelnen Libraries für alle der sequenzierten Regionen.

**Tabelle 4-1.** Die Tabelle zeigt die Coverages (Durchschnittswerte/Median) für jede Library und jeden Locus, sowie die prozentualen Werte der Ziel-Region, die mindestens mit einer 50-fachen Coverage abgedeckt sind.

| | Locus | IL10 | | ATG16L1 | | IL23R | | IRGM | | NKX2-3 | | NOD2 | | STAT3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Library** | | Durchschnitt/Median | % cov>50x | Durchschnitt/Median | % cov>50x | Durchschnitt/Median | % cov>50x | Durchschnitt/Median | % cov>50x | Durchschnitt/Median | % cov>50x | Durchschnitt/Median | % cov>50x | Durchschnitt/Median | % cov>50x |
| **Lib1** | | 1713/160 | 99,69 | 2283/1106 | 82,47 | 926/779 | 86,29 | 845/547 | 98,53 | 2796/1641 | 99,86 | 1575/1322 | 100 | 1675/961 | 90,52 |
| **Lib2** | | 2111/1757 | 99,9 | 2874/1219 | 97,84 | 650/594 | 86,48 | 701/465 | 84,72 | 3082/1886 | 99,89 | 933/827 | 99,98 | 1326/705 | 90,19 |
| **Lib3** | | 1078/866 | 99,66 | 1547/689 | 89,59 | 450/320 | 85,77 | 385/217 | 96,41 | 4073/3282 | 99,91 | 928/559 | 99,91 | 756/433 | 89,59 |
| **Lib4** | | 2708/2061 | 99,79 | 2043/1179 | 98,22 | 663/631 | 86,09 | 541/308 | 97,59 | 2695/1966 | 99,88 | 1433/1066 | 99,96 | 1153/784 | 90,11 |
| **Lib5** | | 1649/1365 | 98,98 | 1269/673 | 98,27 | 589/451 | 83,42 | 754/464 | 96,8 | 3523/3063 | 99,84 | 1115/1085 | 88,68 | 1039/652 | 89,4 |
| **Lib6** | | 2168/1715 | 99,21 | 1062/579 | 97,85 | 588/553 | 82,26 | 908/575 | 83,32 | 3393/2191 | 99,9 | 2201/1646 | 100 | 2122/1061 | 90,43 |
| **Lib7** | | 868/692 | 99,32 | 1157/659 | 99,59 | 454/431 | 72,65 | 691/566 | 99,11 | 3239/2978 | 99,89 | 907/811 | 99,98 | 1434/793 | 89,98 |
| **Lib8** | | 1965/1150 | 99,74 | 2531/1186 | 88,8 | 814/455 | 73,8 | 1077/826 | 99,25 | 3283/1289 | 99,86 | 1617/1460 | 100 | 2125/929 | 90,4 |
| **Lib9** | | 2655/2266 | 99,6 | 2252/1513 | 97,47 | 607/482 | 75,09 | 639/503 | 97,52 | 4019/3022 | 99,86 | 1932/1789 | 99,83 | 1383/632 | 89,44 |
| **Lib10** | | 3219/2822 | 99,89 | 2053/1235 | 98,32 | 1153/976 | 89,39 | 1003/655 | 98,52 | 2545/2579 | 99,75 | 1554/1401 | 99,98 | 753/529 | 89,47 |
| **Lib11** | | 1700/1353 | 99,03 | 2901/1693 | 97,93 | 1177/803 | 97,4 | 839/664 | 97,57 | 2890/2511 | 99,86 | 1824/1500 | 99,73 | 804/593 | 88,85 |
| **Lib12** | | 2185/1749 | 99,82 | 3497/1960 | 99,57 | 834/642 | 83,72 | 913/705 | 93,75 | 2858/2551 | 99,9 | 1812/1597 | 100 | 1285/630 | 88,87 |
| **Lib13** | | 3071/2601 | 99,84 | 1648/1387 | 87,57 | 1037/802 | 77,71 | 1347/653 | 98,86 | 2004/2118 | 99,66 | 2109/1826 | 99,96 | 1342/748 | 90,12 |
| **Lib14** | | 2505/1780 | 99,4 | 2046/1015 | 96,47 | 714/627 | 76,39 | 808/624 | 97,75 | 2449/2453 | 99,41 | 1251/1096 | 99,67 | 1029/554 | 89,32 |
| **Lib15** | | 1205/952 | 99,33 | 1186/453 | 77,9 | 602/447 | 86,46 | 508/297 | 96,22 | 3051/1438 | 99,24 | 1044/921 | 99,75 | 626/248 | 86,04 |
| **Lib16** | | 744/545 | 98,34 | 3146/1417 | 90,39 | 806/775 | 89,27 | 709/517 | 84,24 | 3035/2592 | 74,34 | 2305/1387 | 99,71 | 1137/602 | 89,7 |
| **Lib17** | | 2466/1926 | 99,9 | 5233/1438 | 98,5 | 1550/1412 | 96,45 | 1274/1170 | 99,22 | 2392/2010 | 99,9 | 1749/1651 | 99,88 | 1346/609 | 84,59 |
| **Lib18** | | 2049/1763 | 99,79 | 5062/1952 | 97,13 | 988/848 | 89,07 | 639/325 | 71,54 | 4179/2992 | 99,92 | 1548/1298 | 99,98 | 1607/692 | 83,64 |
| **Lib19** | | 2830/1683 | 99,45 | 5433/1401 | 99,1 | 948/828 | 94,73 | 986/594 | 94,84 | 2915/2646 | 99,83 | 1935/1671 | 99,74 | 943/494 | 89,36 |
| **Lib20** | | 3232/2399 | 99,87 | 6705/2706 | 99,82 | 1231/1120 | 90,19 | 1282/1154 | 99,17 | 4484/3447 | 99,93 | 1950/1977 | 100 | 627/354 | 84,72 |
| **Lib21** | | 2510/1714 | 99,86 | 5200/1485 | 87,57 | 1477/1279 | 99,66 | 1894/1705 | 99,71 | 2667/1909 | 99,89 | 3170/2147 | 100 | 1468/719 | 90,03 |
| **Lib22** | | 2690/1729 | 99,56 | 3744/3030 | 90,41 | 1100/831 | 90,59 | 1927/1027 | 99,38 | 65/6 | 3,72 | 3049/2543 | 99,97 | 1450/1029 | 91 |
| **Lib23** | | 2406/1244 | 99,22 | 2259/1550 | 72,84 | 1472/1174 | 99,57 | 1392/1203 | 94,14 | 1594/838 | 99,71 | 2831/2275 | 100 | 1268/650 | 83,85 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lib24** | | 1997/957 | 99,59 | 1635/423 | 62,49 | 1348/1088 | 99,4 | 437/172 | 89,29 | 1676/1124 | 99,83 | 2174/1783 | 100 | 1460/767 | 88,77 |
| **Lib25** | | 1747/770 | 99,47 | 2503/1788 | 88,18 | 995/794 | 98,21 | 498/232 | 95,88 | 2073/1163 | 99,72 | 1906/1445 | 99,95 | 601/258 | 78,76 |
| **Lib26** | | 1761/1093 | 82,2 | 2552/1733 | 86,85 | 988/831 | 97,89 | 991/291 | 94,44 | 1515/143 | 78,61 | 2308/1666 | 99,96 | 844/508 | 85,7 |
| **Lib27** | | 1182/629 | 98,15 | 1244/14 | 37,13 | 565/245 | 70,14 | 1261/717 | 98,98 | 2404/2310 | 99,57 | 1374/937 | 99,92 | 1118/559 | 86,95 |
| **Lib28** | | 1260/767 | 92,75 | 3188/1698 | 82,88 | 383/119 | 58,73 | 1601/911 | 98,86 | 2470/2101 | 99,85 | 2247/1925 | 94,05 | 1202/593 | 89,12 |
| **Lib29** | | 2033/1189 | 99,67 | 2555/1958 | 88,67 | 684/524 | 89,44 | 1457/1292 | 99,54 | 2598/2237 | 99,86 | 1182/1068 | 99,84 | 895/528 | 84,95 |
| **Lib30** | | 2070/1469 | 99,62 | 2270/1558 | 99,39 | 1205/782 | 99,1 | 1615/1067 | 99,68 | 4542/3520 | 99,86 | 2116/1351 | 99,99 | 1762/796 | 90,04 |
| **Lib31** | | 1513/885 | 99,61 | 3324/2485 | 99,39 | 613/523 | 77,27 | 2815/2253 | 99,85 | 1549/1243 | 99,7 | 1309/974 | 99,98 | 1173/516 | 84,2 |
| **Lib32** | | 1018/696 | 91,83 | 4200/3194 | 99,74 | 401/282 | 63,12 | 896/608 | 83,29 | 1035/628 | 99,12 | 1925/1571 | 100 | 1557/588 | 88,28 |
| **Lib33** | | 1055/658 | 77,13 | 3897/3017 | 99,76 | 908/528 | 64,35 | 2741/1652 | 99,69 | 1880/1064 | 99,44 | 1610/951 | 72,47 | 2524/1188 | 89,75 |
| **Lib34** | | 1306/1112 | 98,62 | 4366/3259 | 99,62 | 700/659 | 69,78 | 1970/1107 | 99,59 | 3319/1714 | 99,87 | 1175/880 | 99,98 | 1960/484 | 83,42 |
| **Lib35** | | 1844/1406 | 99,19 | 2693/2047 | 99,65 | 1067/906 | 83,91 | 877/671 | 98,03 | 2756/1818 | 99,91 | 2921/1441 | 100 | 1403/433 | 90,1 |
| **Lib36** | | 2612/1906 | 99,81 | 3160/1443 | 84,08 | 853/297 | 68,25 | 1342/928 | 99,63 | 3014/3036 | 99,92 | 1522/1352 | 100 | 1385/460 | 79,3 |
| **Lib37** | | 3310/1750 | 99,37 | 442/19 | 30,75 | 1946/1668 | 99,74 | 1335/695 | 99,47 | 2748/2236 | 99,92 | 1466/1378 | 100 | 1955/1045 | 90,47 |
| **Lib38** | | 2342/2025 | 99,68 | 3113/1201 | 84,88 | 1260/1215 | 92,76 | 1966/1356 | 99,77 | 1175/828 | 99,88 | 3845/3514 | 99,96 | 1901/1063 | 91,15 |
| **Lib39** | | 1980/1492 | 99,32 | 2406/1018 | 84,6 | 815/219 | 64,99 | 1986/1619 | 99,88 | 3071/474 | 99,46 | 269/165 | 97,52 | 2584/1164 | 91,37 |
| **Lib40** | | 2234/1426 | 99,67 | 4500/1451 | 84,72 | 1177/1167 | 90,87 | 1609/1070 | 99,79 | 2720/2056 | 99,63 | 1319/1019 | 99,98 | 2020/906 | 91,42 |
| **Lib41** | | 2352/2009 | 99,3 | 1494/530 | 80,78 | 1006/955 | 74,42 | 689/488 | 96,66 | 3243/2942 | 99,88 | 2562/1812 | 99,98 | 1517/850 | 90,58 |
| **Lib42** | | 1928/1459 | 99,38 | 2947/1121 | 84,5 | 695/374 | 72,98 | 1582/996 | 99,21 | 3451/2659 | 99,76 | 2168/1459 | 99,97 | 1355/700 | 90,03 |
| **Lib43** | | 3214/2366 | 99,54 | 342/83 | 74,54 | 1159/1054 | 86,31 | 2128/1115 | 99,5 | 3121/2504 | 99,68 | 1128/836 | 99,88 | 1389/880 | 89,15 |
| **Lib44** | | 3799/2660 | 99,8 | 4456/3616 | 99,74 | 989/683 | 88,7 | 1019/779 | 90,61 | 2293/2000 | 97,53 | 1031/746 | 99,89 | 1522/899 | 89,96 |
| **Lib45** | | 1566/1101 | 99,67 | 3018/2379 | 98,88 | 780/540 | 86 | 1484/1196 | 99,53 | 2577/1868 | 99,84 | 1016/862 | 99,92 | 1287/844 | 91 |
| **Lib46** | | 1311/797 | 99,02 | 1470/919 | 72,46 | 854/698 | 94,5 | 1162/713 | 99,28 | 967/743 | 99,08 | 682/528 | 99,72 | 1266/696 | 90,22 |
| **Lib47** | | 1024/707 | 67,5 | 1713/591 | 63,33 | 933/831 | 73,52 | 1273/1036 | 99,41 | 542/377 | 94,81 | 1045/974 | 99,12 | 1496/877 | 90,07 |
| **Lib48** | | 2109/1382 | 99,64 | 2700/1840 | 99,83 | 815/766 | 84,73 | 778/636 | 98,14 | 859/690 | 93,8 | 648/404 | 99,65 | 1030/492 | 88,12 |
| **Lib49** | | 3253/2400 | 99,75 | 2463/2083 | 99,27 | 1164/917 | 86,49 | 1035/785 | 98,88 | 4732/3117 | 99,82 | 952/651 | 99,77 | 1881/1203 | 91,34 |
| **Lib50** | | 459/11 | 45,7 | 3656/2364 | 82,2 | 1591/1296 | 86,98 | 243/2 | 19,64 | 3374/2787 | 99,9 | 1996/1571 | 100 | 1793/983 | 91,29 |
| **Lib51** | | 2135/1606 | 99,78 | 3033/1293 | 92,09 | 1125/786 | 98,3 | 1266/485 | 80,9 | 2479/1512 | 99,83 | 635/381 | 99,3 | 1632/978 | 90,38 |
| **Lib52** | | 1292/847 | 80,14 | 2452/1366 | 99,05 | 812/278 | 77,81 | 1131/557 | 84,35 | 2094/1339 | 99,83 | 404/379 | 93,66 | 1974/1457 | 90,79 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lib53** | | 2563/2385 | 80,03 | 3148/1520 | 92,53 | 1228/1101 | 99,55 | 1461/677 | 84,11 | 3760/3240 | 99,8 | 679/431 | 99,52 | 1180/816 | 89,35 |
| **Lib54** | | 3234/2799 | 99,81 | 2893/1731 | 92,46 | 1322/1152 | 98,35 | 1224/841 | 83,84 | 1936/1622 | 99,77 | 445/305 | 99,28 | 1628/960 | 91,07 |
| **Lib55** | | 2468/1635 | 99,91 | 1841/1120 | 92,11 | 1315/1074 | 99,93 | 1059/939 | 84,45 | 2269/1864 | 99,63 | 589/502 | 93,92 | 2036/1177 | 91,19 |
| **Lib56** | | 2256/1634 | 99,88 | 2171/1138 | 92,39 | 1130/943 | 90,66 | 818/608 | 84,7 | 435/347 | 84,16 | 928/689 | 99,98 | 2183/1432 | 91,41 |

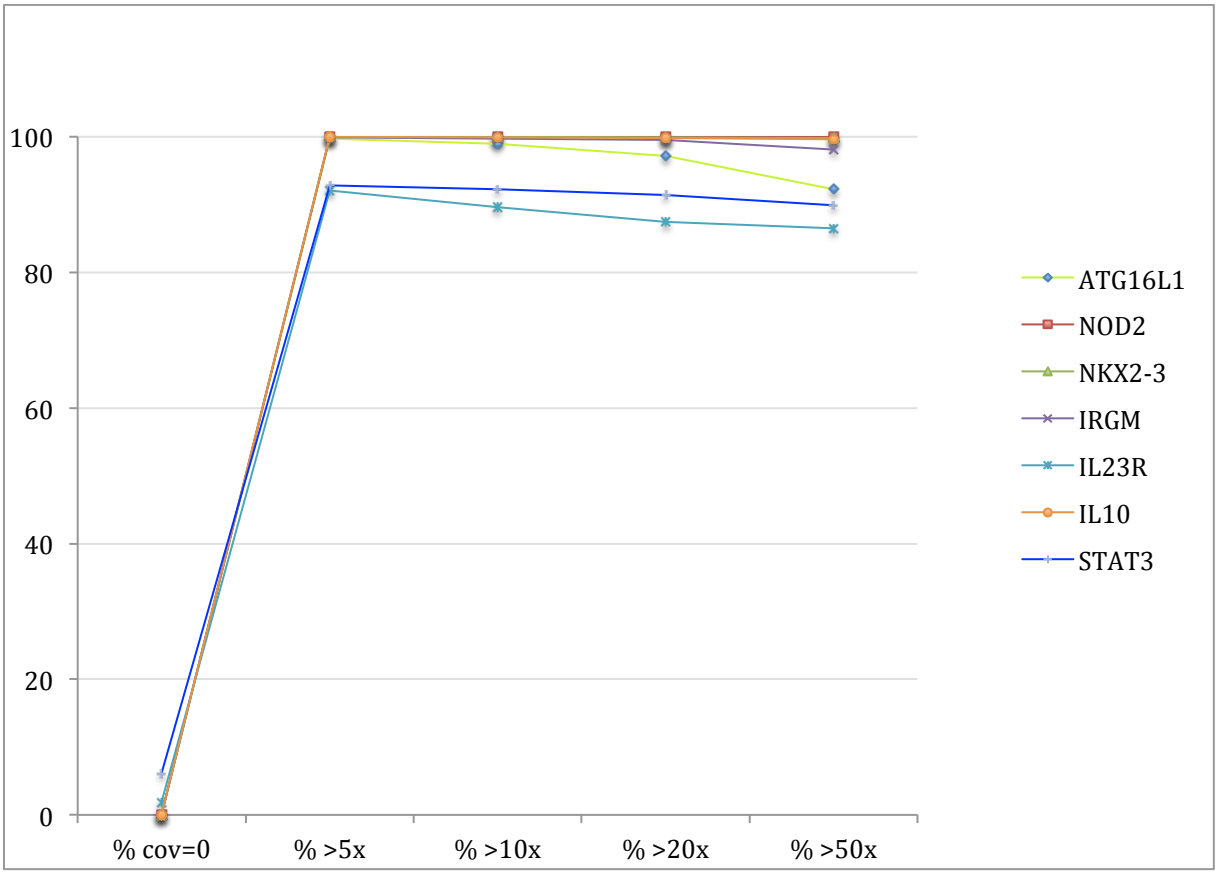



**Abbildung 4-2.** Coverage-Diagramm für alle CED-Loci; aufgetragen sind die Coverages (Y-Achse) gegen die Prozente der Ziel-Region (X-Achse), die mit dieser Coverage minimal sequenziert sind (siehe Text).

Das Diagramm (Abbildung 4-2) zeigt, wieviel Prozent der Ziel-Regionen, die sequenziert wurden, mit einer minimalen Coverage von 50x, 20x, 10x, 5x und Coverage=0 abgedeckt sind. Abgebildet sind pro Locus die Mediane über alle 56 sequenzierten Libraries. An diesem Diagramm sieht man deutlich, welche Loci sehr gut abgedeckt sind und welche einige Lücken in einigen Libraries aufweisen (s. Diagramme im Anhang). Von den Loci *STAT3* und *IL23R* sind 92% (*STAT3*) bzw. 89% (*IL23R*) der Ziel-Region mit einer minimalen Coverage von 10x abgedeckt, alle anderen Loci sind mit jeweils weit über 95% der Ziel-Region mit einer größer als 10-fachen Coverage sequenziert. Eine 5-10 fache Coverage reicht aus, um SNP- und Mutationsdetektion durchzuführen. Nach der Detektion wurden in diesem Experiment alle detektierten Positionen durch eine Art „Kontrollprogramm“ pibase (Kapitel 2.9.1) gegeben, um vermeintliche SNP Detektionsfehler aufzudecken und zu eliminieren. Im Programm pibase ist die Coverage eine wichtige Größe.

## 4.3 Konkordanzen für HapMap-Proben

Die zwei verwendeten HapMap-Trios sind zur Berechnung von Konkordanzen als Kontrollindividuen sequenziert worden. Für diese HapMap-Proben liegen

Vergleichsdatensätze aus dem HapMap-Projekt vor (siehe Tabelle 4-2). Für die HapMap-Proben konnten verschiedene Konkordanzen berechnet werden, um Kontaminationen und Verwechslungen zu erkennen bzw. auszuschließen. Außerdem geben sie Aufschluss über die Qualität des SNP-callings. Allerdings stehen für nur vier der sechs verwendeten HapMap-Proben Datensätze des Internationalen HapMap-Projektes (http://hapmap.ncbi.nlm.nih.gov/) für die Konkordanzberechnung zur Verfügung.

**Tabelle 4-2.** Die Tabelle zeigt, für welche der HapMap-Proben Datensätze aus dem internationalen HapMap-Projekt für die Berechnung von Konkordanzen zur Verfügung stehen.

| HapMap Individuen | Trio | HapMap Datensatz | SOLiD-Sequenzierung | SOLiD-Sequenzierung nach Filterprozess |
|---|---|---|---|---|
| **1** | 1 | + | + | + |
| **2** | 1 | + | + | + |
| **3** | 1 | | + | + |
| **4** | 2 | | + | + |
| **5** | 2 | + | + | + |
| **6** | 2 | + | + | + |

Konkordanzen wurden berechnet für:

a) den HapMap-Datensätzen des Internationalen HapMap-Projektes und Ergebnissen der SOLiD-Sequenzierung
b) den HapMap-Datensätzen des Internationalen HapMap-Projektes und Ergebnissen der SOLiD-Sequenzierung nach Filtern mit pibase und IGV sowie Reduktion der Ziel-Region auf CED-Loci.

### *4.3.1 Konkordanz zwischen Datensätzen des Internationalen HapMap-Projektes und SOLiD-Sequenzierung*

Die Tabelle 4-3 zeigt die Anzahl der detektierten SNPs basierend auf der SOLiD-Sequenzierung für jedes der verwendeten HapMap-Individuen sowie die Anzahl der SNPs für die Proben, für die aus dem Internationalen HapMap-Projekt Daten zur Verfügung stehen. Außerdem zeigt die Tabelle die Überlappung der vorhandenen SNP-Datensätze. Die Konkordanzberechnung geschah basierend auf den überlappenden SNPs. Die Überlappung zwischen den Datensätzen, d. h. die Anzahl der SNPs, für die die Konkordanz berechnet

werden kann, ist, gemessen an den durch die Sequenzierung detektierten SNPs, relativ klein (zwischen 37,72% und 41,79%). Diese geringen Raten an Überlappung hat u.a. den Grund, dass die SNP-Daten des Internationalen HapMap-Projektes auf Genotypisierungstechnologien basieren, bei denen ganz gezielt bestimmte Positionen abgefragt werden, um den Genotyp zu bestimmen. Dabei kommt es vor, dass das Individuum für viele Positionen den Referenzgenotypen hat. Diese Positionen werden bei der SNP-Detektion bei der Auswertung der Sequenzierdaten nicht detektiert, da hierbei nur die Unterschiede zur Referenz angegeben werden. Die Konkordanzraten liegen zwischen 85 und 93%.

**Tabelle 4-3.** Die Tabelle zeigt die absolute Anzahl der SNPs, die für die HapMap-Proben detektiert worden sind, die Anzahl der SNPs, die im internationalen HapMap-Projekt für die Proben in der Ziel-Region genotypisiert worden sind und die entsprechend errechneten Konkordanzen**.**

| Indi-viduum | Anzahl der detektierten SNPs nach SOLiD-Sequenzierung | Anzahl der SNPs aus Int. HapMap Projekt für die sequenzierte Ziel-Region | Überlappung (Anzahl) | Überlappung (%) | Konkor-danz (Anzahl) | Konkordanz (%) |
|---|---|---|---|---|---|---|
| **1** | 1054 | 421 | 184 | 39,94 | 171 | 92,93 |
| **2** | 1158 | 423 | 141 | 36,53 | 126 | 89,36 |
| **3** | 907 | NA | NA | NA | NA | NA |
| **4** | 935 | NA | NA | NA | NA | NA |
| **5** | 1116 | 421 | 175 | 37,72 | 154 | 88 |
| **6** | 1005 | 420 | 163 | 41,79 | 140 | 85,88 |

### *4.3.2 Konkordanz zwischen Datensätzen des Internationalen HapMap-Projektes und SOLiD-Sequenzierung nach Filterprozess und Reduktion auf CED-Loci*

Nach dem Filterprozess mit Hilfe von pibase und IGV für das Auffinden falsch-positiver SNPs sowie durch die Reduktion auf CED-Loci hat sich die Anzahl der übriggebliebenen SNPs aus der SOLiD-Sequenzierung deutlich reduziert. Zum einen liegt das an der Fokussierung auf die CED-Loci zum anderen daran, dass ein Großteil der potentiellen SNPs als falsch-positive Ergebnisse eliminiert werden konnten. Die Tabelle 4-4 zeigt die Anzahl der als „echte“ SNPs identifizierten SNPs innerhalb der CED-Loci sowie die ebenfalls für die CED-Loci reduzierten HapMap-SNPs. Die Konkordanzen zwischen den sich überlappenden SNPs zwischen Genotypisierung und Sequenzierung liegen bei allen HapMap-Proben über 90%.

**Tabelle 4-4.** Die Tabelle zeigt die Konkordanz zwischen Datensätzen des Internationalen HapMap-Projektes (reduziert auf Ziel-Region der CED-Loci) und SOLiD-Sequenzierung nach Filterprozess und Reduktion auf CED-Loci.

| Indi-viduum | SOLiD-SNPs nach Filterprozess und Reduktion auf CED-Loci | HapMap-SNPs | Überlappung (Anzahl) | Konkor-danz (Anzahl) | Konkordanz (%) |
|---|---|---|---|---|---|
| **1** | 395 | 260 | 86 | 81 | 94,18 |
| **2** | 360 | 260 | 61 | 59 | 96,72 |
| **3** | xxx | NA | NA | NA | NA |
| **4** | xxx | NA | NA | NA | NA |
| **5** | 360 | 260 | 78 | 74 | 94,87 |
| **6** | 361 | 259 | 84 | 79 | 94,04 |

## 4.4 Ergebnisse der SNP-Detektion

Die SNP-Detektion konnte aus bioinformatorischen Gründen nicht getrennt für die einzelnen Loci durchgeführt werden. Nach dem SNP-calling wurde sich für diese Arbeit nur auf die SNPs konzentriert, die innerhalb der CED-Loci fielen (siehe Tabelle 4-5). Die SNPs-Detektionsergebnisse für die anderen Loci wurden an die entsprechenden Arbeitsgruppen weitergegeben, die sich näher mit den dazu gehörigen Phänotypen befassen.

**Tabelle 4-5.** Die Tabelle zeigt alle detektierten SNPs für die CED-Loci vor und nach dem Filterpozesss

| | alle | ATG16L1 | IL10 | IL23R | NOD2 | STAT3 | NKX2-3 | IRGM |
|---|---|---|---|---|---|---|---|---|
| Anzahl SNPs total | 6037 | 845 | 317 | 2065 | 102 | 1284 | 220 | 1204 |
| Anzahl SNPs nach Filterprozess | 2075 | 217 | 124 | 752 | 78 | 336 | 156 | 412 |

Die SNPs für die CED-Loci wurden anschließend nach potentiell falsch-positiven SNPs gefiltert. Die SNPs wurden durch das hausinterne Annotationstool SnpActs analysiert. Die Ergebnisse des SNP-callings und des Filterprozesses sind in Tabelle 4-6 dargestellt.

**Tabelle 4-6.** Die Tabelle zeigt die Ergebnisse der SNP- Detektion für die CED-Loci nach dem Filterprozess und nach der Annotation mit SnpActs**.**

| | ATG16L1 | IL10 | IL23R | NOD2 | STAT3 | NKX2-3 | IRGM |
|---|---|---|---|---|---|---|---|
| uniqueSNPs_total | **217** | **124** | **752** | **78** | **336** | **156** | **412** |
| bekannte SNPs | **120** | **81** | **171** | **54** | **103** | **89** | **188** |
| SNPs in Exon_bekannt | 2 | 0 | 4 | 10 | 0 | 1 | 1 |
| 5`UTR_bekannte SNPs | 1 | 0 | 0 | 1 | 0 | 0 | 7 |
| 3`UTR_bekannte SNPs | 6 | 3 | 0 | 3 | 4 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| UTR_splicesite_bekannt | 0 | 0 | 2 | 0 | 44 | 0 | 0 |
| nonsyn_SNPS_bekannt | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| missense_SNPs_bekannt | 1 | 0 | 4 | 7 | 0 | 0 | 0 |
| syn_SNPs_bekannt | 1 | 0 | 0 | 3 | 0 | 1 | 1 |
| SNPs in Intron_bekannt | 88 | 17 | 125 | 33 | 54 | 4 | 0 |
| unbekannt | **97** | **43** | **581** | **24** | **233** | **67** | **224** |
| SNPs in Exon_unbekannt* | 2 | 0 | 1 | 3 | 1 | 0 | 0 |
| 5´UTR_unbekannt | 0 | 0 | 0 | 0 | 0 | 0 | 26 |
| 3´UTR_unbekannt | 3 | 0 | 0 | 0 | 3 | 0 | 0 |
| UTR_splicesite_unbekannt | 0 | 0 | 1 | 0 | 127 | 0 | 8 |
| nonsyn_SNPS_unbekannt | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| missense_SNPs_unbekannt | 1 | 0 | 1 | 3 | 11 | 0 | 8 |
| syn_SNPS_unbekannt | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| SNPs in Intron_unbekannt | 89 | 0 | 289 | 14 | 62 | 2 | 0 |

*unbekannt bedeutet hier zum Zeitpunkt der Datenbankabfrage dbSNP kein Eintrag

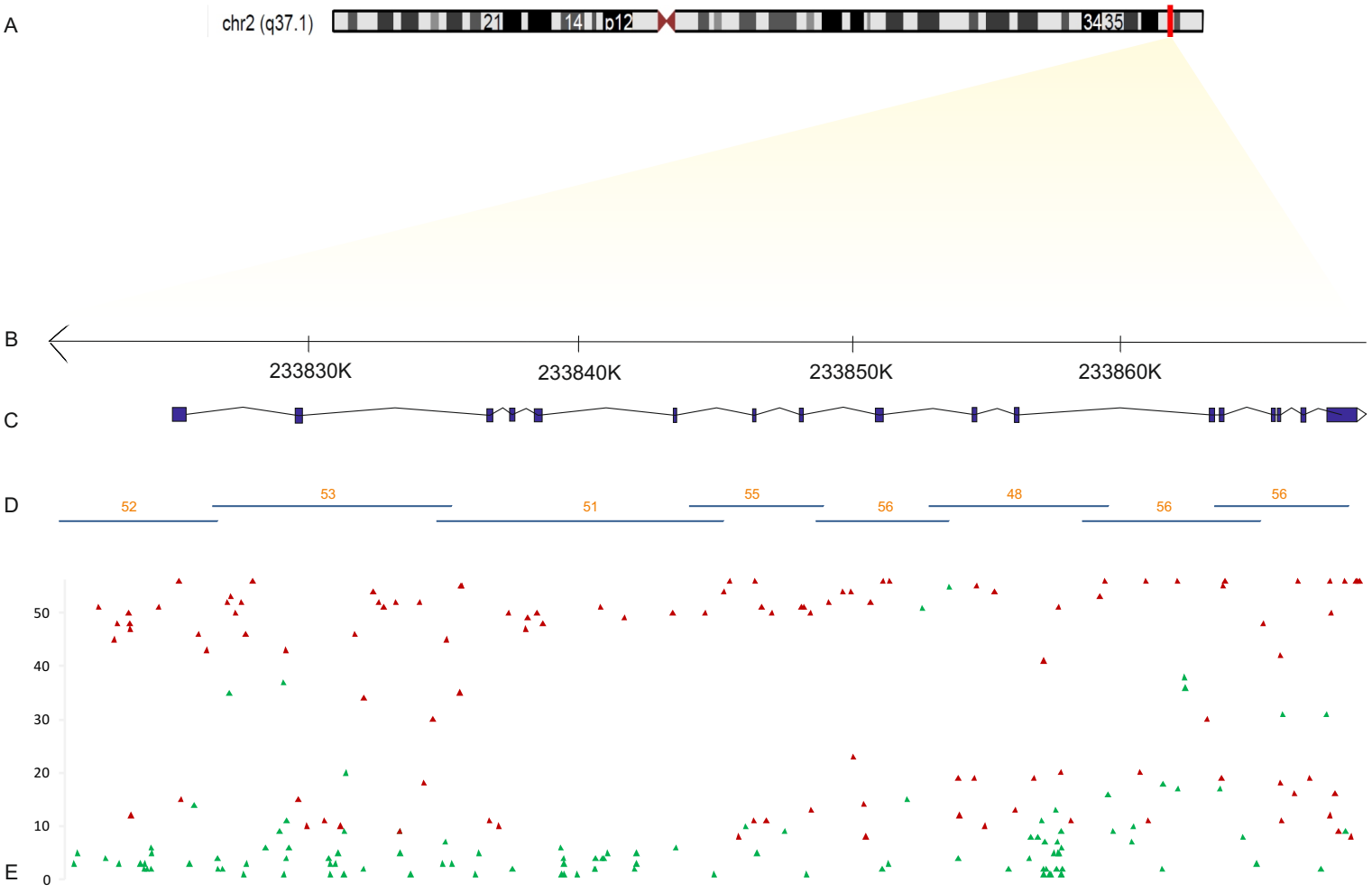



**Abbildung 4-3.** Die Abbildung zeigt die Sequenzierergebnisse der Sequenzierung des Locus *ATG16L1*.

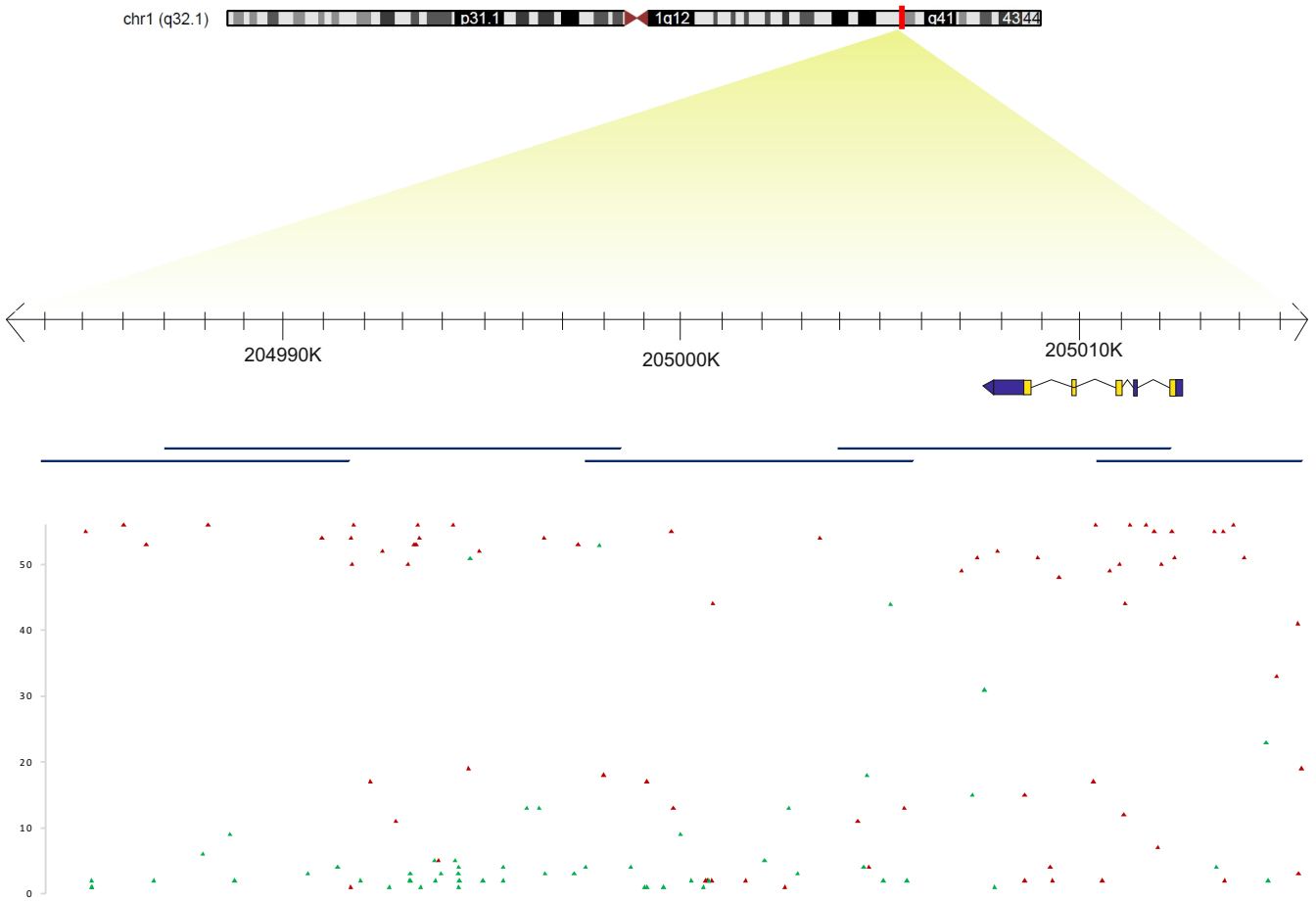



**Abbildung 4-4.** Die Abbildung zeigt die Sequenzierergebnisse der Sequenzierung des Locus *IL10*.

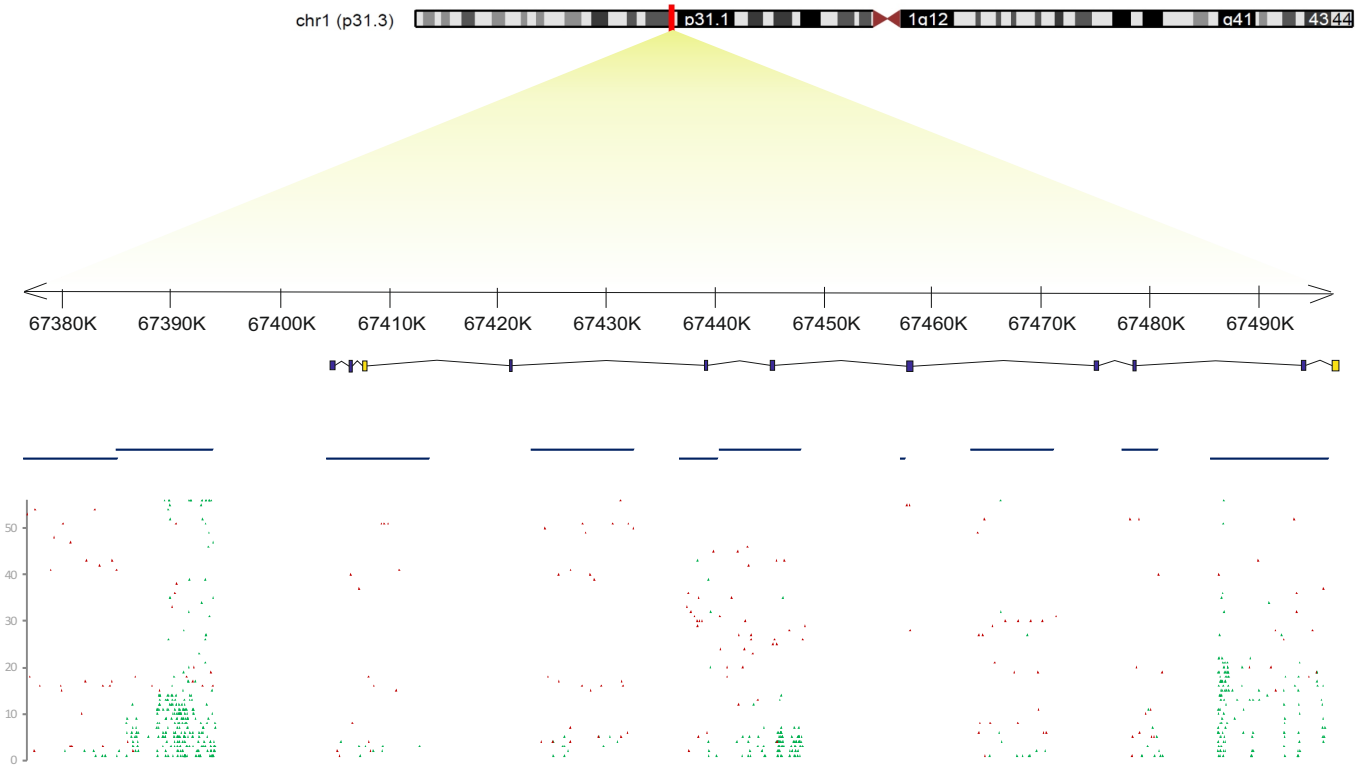



**Abbildung 4-5.** Die Abbildung zeigt die Sequenzierergebnisse für die Sequenzierung des Locus *IL23R*.

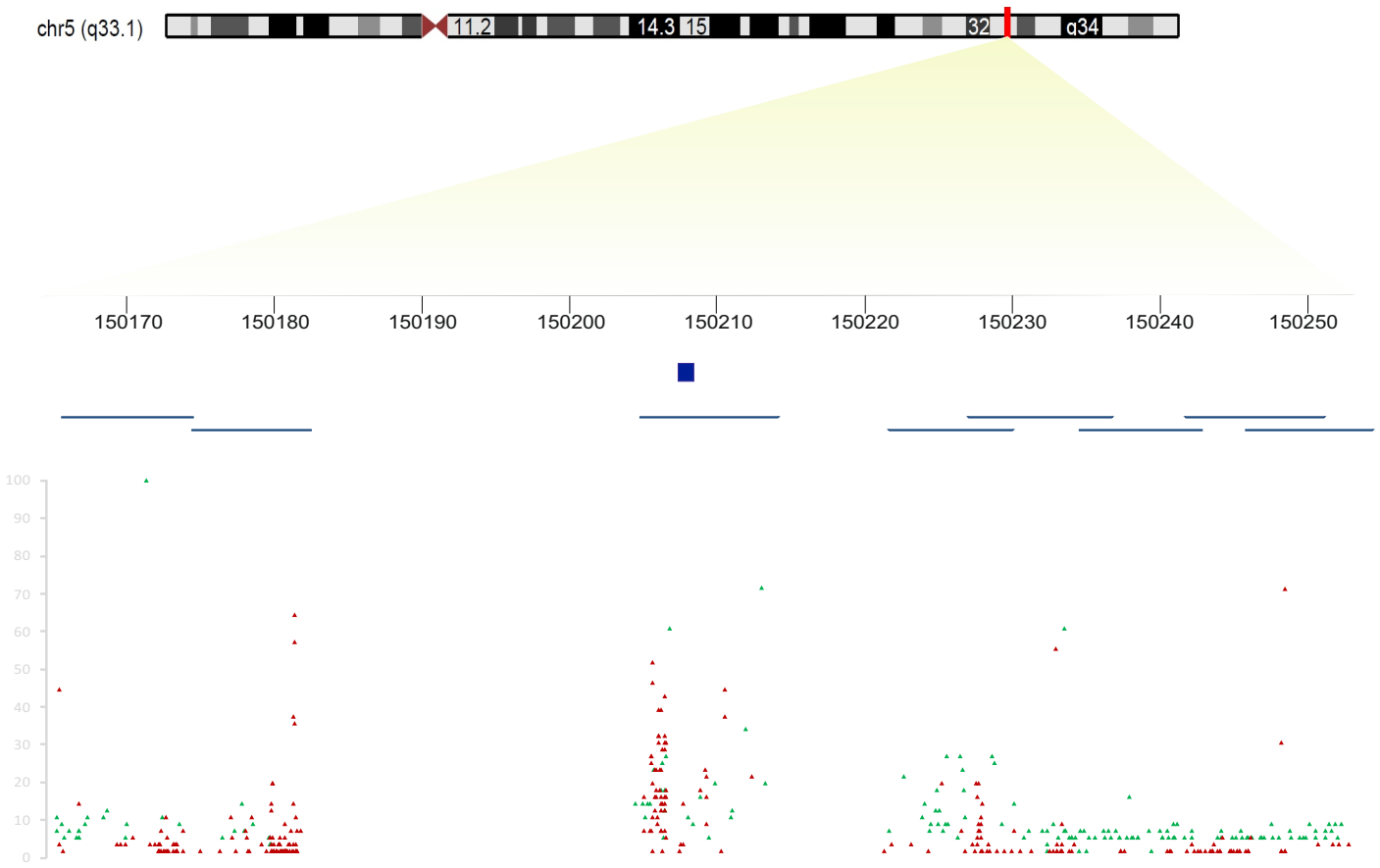



**Abbildung 4-6.** Die Abbildung zeigt die Sequenzierergebnisse der Sequenzierung des Locus *IRGM*.

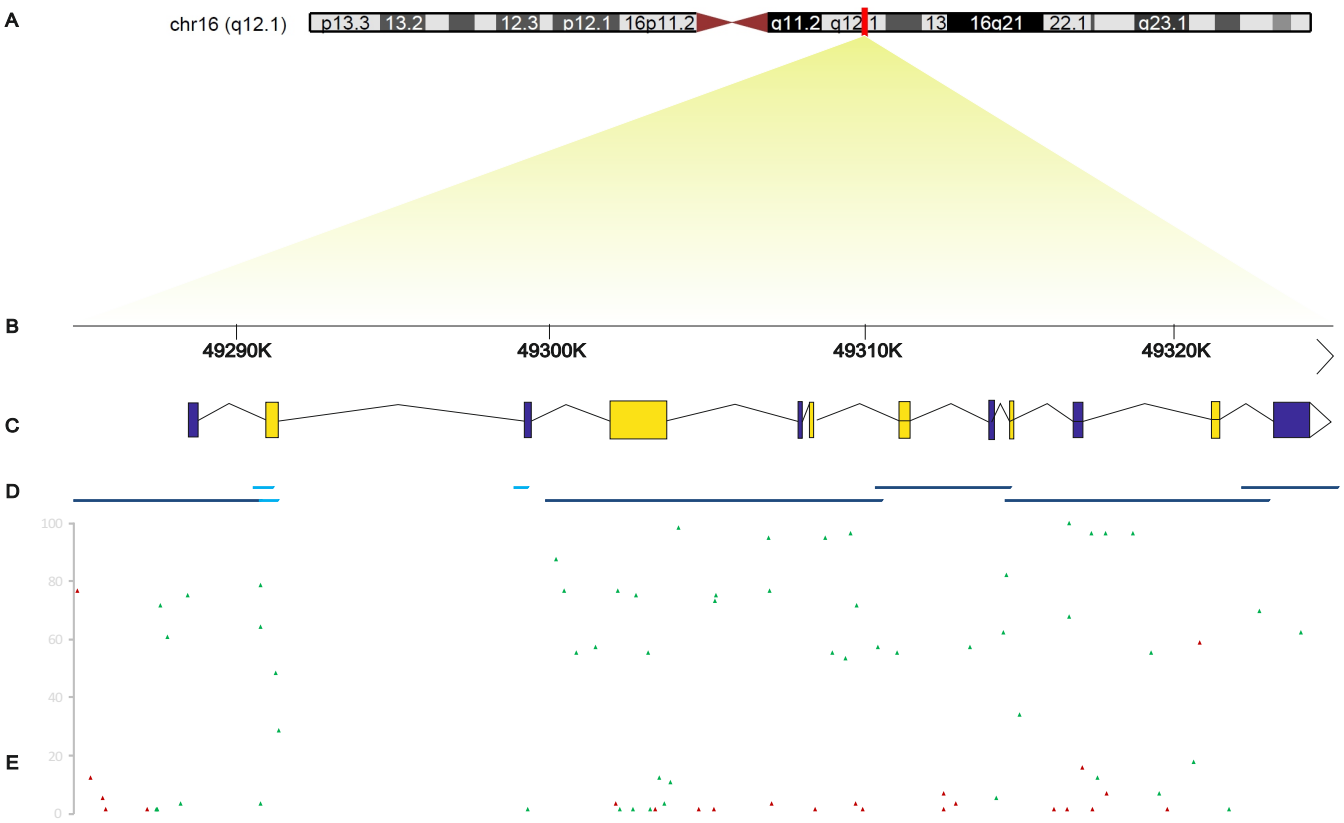



**Abbildung 4-7.** Die Abbildung zeigt die Sequenzierergebnisse für den Locus *NOD2*.

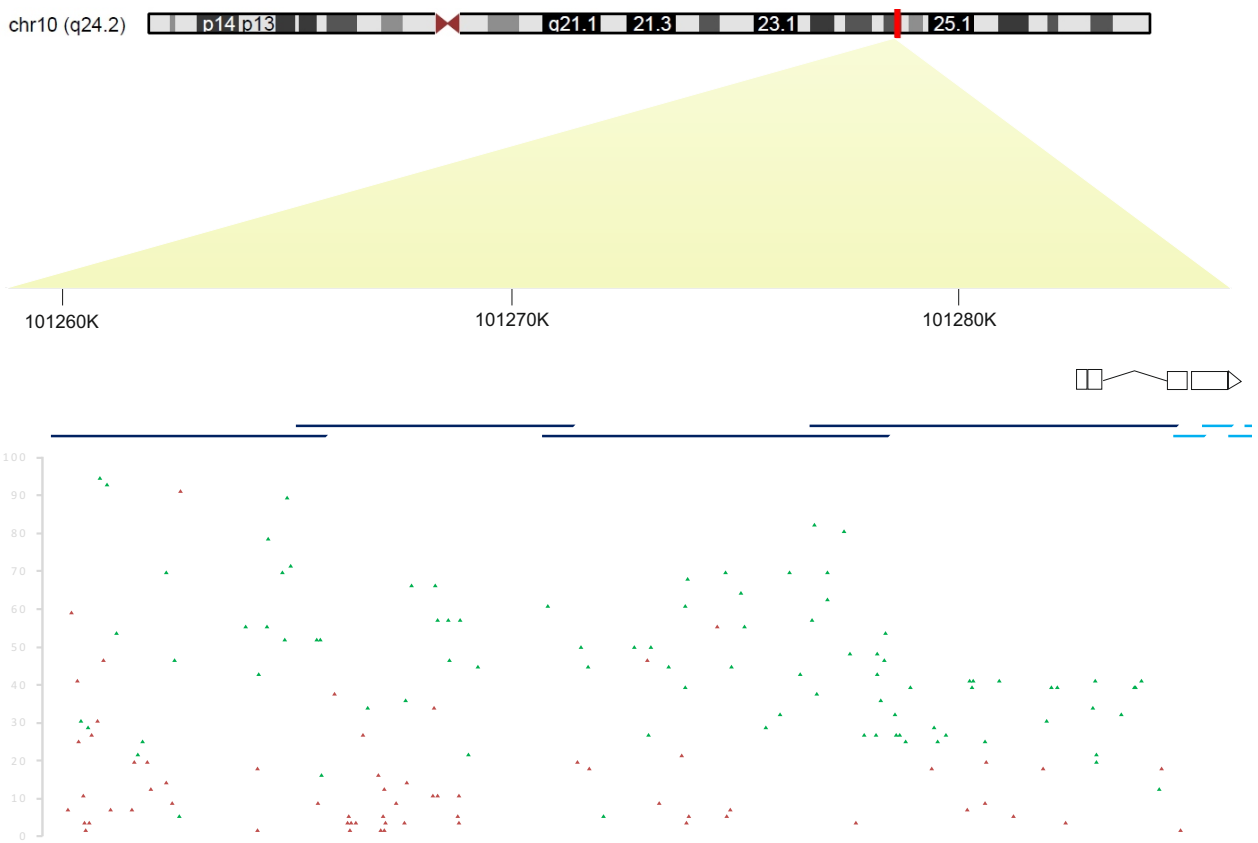



**Abbildung 4-8.** Die Abbildung zeigt die Sequenzeirergebnisse für den Locus *NKX2-3.* Die hellblau dagestellten PCR-Produkte zeigen an, dass diese mit Hilfe der Sanger-Technologie sequenziert wurden sind.

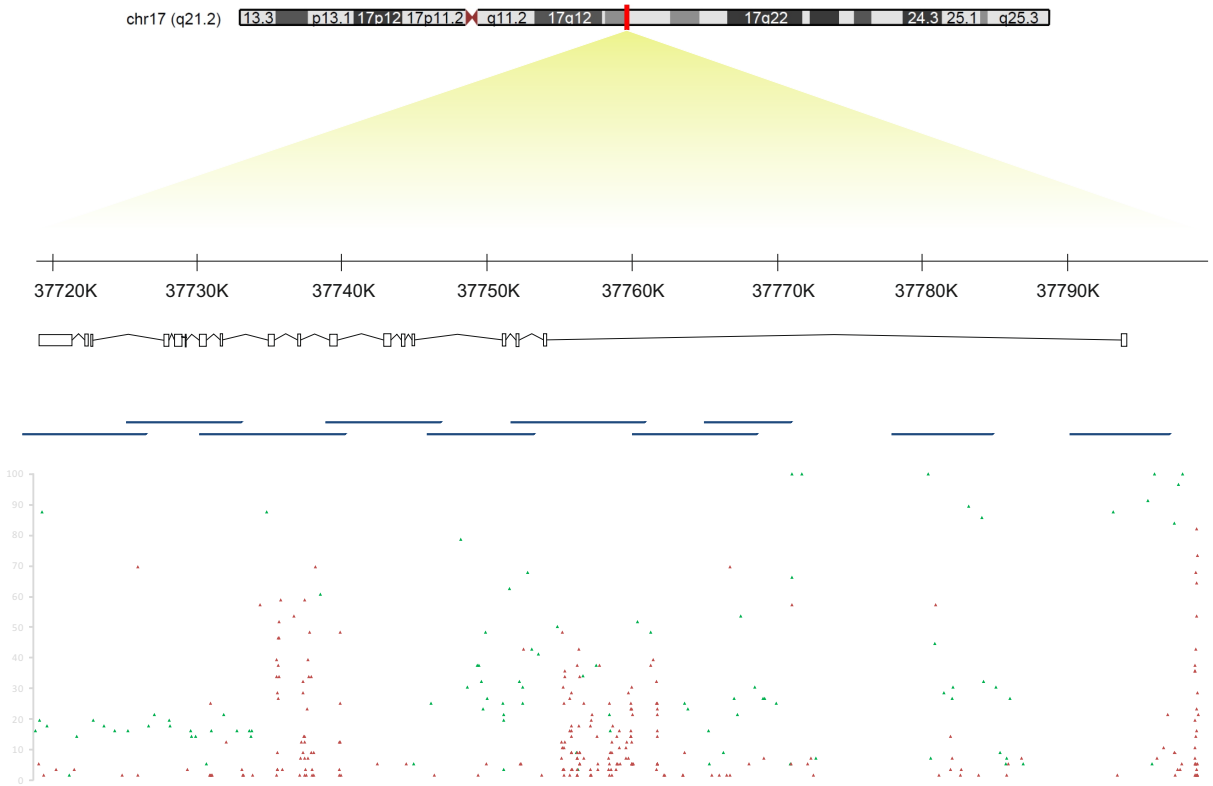



**Abbildung 4-9** Die Abbildung zeigt die Sequenzierergebnisse für den Locus *STAT3*.

Die Abbildungen 4-3 bis 4-9 zeigen die Ergebnisse der Sequenzierung für die verschiedenen CED-Loci. Abbildung 4-3 zeigt beispielsweise die für den *ATG16L1*-Locus erhaltenen Sequenzierergebnisse. Die Abbildungen sind vom Aufbau her gleich, sie repräsentieren lediglich die verschiedenen Loci. Die Abschnitte A-C geben jeweils einen Überblick über Lokalisation des sequenzierten Bereichs im Chromosom und die Genstruktur der betreffenden Gene. Abschnitt D der Abbildungen zeigt die Größe und Lage der PCR-Produkte, die für das Anreichern der gewünschten Bereiche generiert wurden. Die Zahlen, die sich oberhalb der PCR-Produkte zu finden sind, geben an, in wieviel der 56 Proben-DNAs die PCR und die nachfolgenden Laborschritte erfolgreich waren, für die somit Sequenzierdaten vorhanden sind. Abschnitt E zeigt die Lokalisation und Häufigkeit der detektierten SNPs, dabei repräsentieren die roten die SNPs, für die zum Zeitpunkt der Datenbankabfrage bei dbSNP kein Eintrag vorhanden war, die grünen sind SNPs, die bekannt sind.

## 4.5 Ergebnisse der Analyse für Transkriptionsfaktorbindestellen

Die Analyse zur Quantifizierung der Änderung der Bindeaffinität für Transkriptionsfaktoren mit dem Analyseprogramm sTRAP zeigte für einige detektierte Varianten Änderungen im Bindeverhalten verschiedener Transkriptionsfaktoren. Allerdings können solche Berechnungen nur angestellt werden für Transkriptionsfaktoren, für die Affinitätsmodelle existieren. Diese Analyse diente in erster Linie dazu, Varianten zu priorisieren, die in die weiteren Analysen (Weiterverfolgung durch Genotypisierung) einfließen sollten. Dazu wurden die Varianten herausgefiltert, die eine Änderung im Affinitätsverhalten zu denjenigen Transkriptionsfaktoren aufwiesen, die in einem Zusammenhang mit chronisch entzündlichen Darmerkrankungen stehen könnten. 2075 genomische Positionen wurden mit dem Softwaretool sTRAP getestet. Für 461 Positionen konnten Änderungen in der Bindungsaffinität für getestete Transkriptionsfaktoren berechnet werden. Für 94 Varianten konnten Affinitätsänderungen bei Transkriptionsfaktoren festgestellt werden, die nach dem jetzigen Kenntnisstand als gute Kandidaten in den Stoffwechselwegen von CED gewertet wurden und somit in die weiterführende Genotypisierung miteinfließen sollten.

## 4.6 Ergebnisse des Assoziationsexperimentes I

88 Variationen aus der Sequenzierung sind in die weiterführende Genotypisierungsstudie eingeschlossen worden. Für diese 88 Variationen sind insgesamt drei Sequenom-Pools erstellt worden. 368 Kontrollindividuen und 364 an Morbus Crohn erkrankte Individuen sind für dieses 88 Varianten genotypisiert worden. Nach erster Analyse wurden diejenigen Individuen

aus der Analyse ausgeschlossen, bei denen mehr als 30 Assays nicht funktioniert haben. Das betraf 39 Kontrollen und 29 Fälle. Nachdem diese Individuen entfernt worden waren, wurde mit Hilfe der Software plink (http://pngu.mgh.harvard.edu/~purcell/plink) eine Case/Control Assoziationsberechnung durchgeführt. Die Ergebnisse sind in der Tabelle 4-7 dargestellt. Viele der Varianten waren allerdings monomorph und fielen aus der Analyse heraus. Da es sein kann, dass die getesteten Variationen seltene bzw. sehr seltene Varianten sind, heißt das nicht, dass sie in einer größeren/anderen Analysepopulation nicht zu finden wären. Einige, v.a. schon als mit Morbus Crohn assoziierte SNPs lieferten sehr gute p-Werte. So zeigen rs2066843 und rs2066842, sowie rs2066845 aus dem *NOD2*-Gen sehr gute p-Werte (rs2066845;p= 1,19E$^{-10}$, rs2066842 p=2,72E-14; rs2066843 p=1,50E$^{-13}$) wie auch in der Literatur beschrieben[58]. Auch der intronische *IL23R* SNP hat einen p-Wert von 1,85E$^{-09}$, dass die Signifikanzgrenze von 0,05 deutlich übersteigt. Dieser SNP wurde allerdings auch auf dem Immuno-Chip mit genotypisiert, und ging daher nicht mit in das zweite Assoziationsexperiment ein. Insgesamt wurden aus dem ersten Assoziationsexperiment 27 für das sich anschließende zweite Assoziationsexperiment in einer größeren Analysepopulation ausgewählt. Die ausgewählten SNPs sind in der unten angegebenen Tabelle mit * markiert.

**Tabelle 4-7.** Ergebnisse des ersten Assoziationsexperiments. Die SNPs sind nach genomischen Koordinaten geordnet. Die mit * versehenen Varianten, sind in das zweite Assoziationsexperiment miteingegangen.

| | Locus | CHR | Pos | SNP | A1 | Freq_Fälle | Freq_Kontr | A2 | CHISQ | P | OR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IL23R | 1 | 67384201 | rs4655683 | A | 0,2544 | 0,3144 | G | 12,18 | 0,0004824 | 0,7439 |
| 2 | IL23R | 1 | 67384867 | SNP_46* | T | 0,002171 | 0,003639 | C | 0,5132 | 0,4738 | 0,5956 |
| 3 | IL23R | 1 | 67385090 | SNP_47 | T | 0,03515 | 0,03813 | C | 0,175 | 0,6757 | 0,919 |
| 4 | IL23R | 1 | 67386595 | SNP_1* | 0 | 0 | 0 | A | NA | NA | NA |
| 5 | IL23R | 1 | 67386703 | SNP_2 | 0 | NA | NA | | NA | NA | NA |
| 6 | IL23R | 1 | 67388595 | SNP_48* | 0 | 0 | 0 | T | NA | NA | NA |
| 7 | IL23R | 1 | 67390806 | SNP_51* | T | 0 | 0,00289 | C | 3,994 | 0,04566 | 0 |
| 8 | IL23R | 1 | 67391598 | SNP_52 | 0 | NA | NA | | NA | NA | NA |
| 9 | IL23R | 1 | 67391921 | SNP_4 | 0 | NA | NA | | NA | NA | NA |
| 10 | IL23R | 1 | 67392833 | SNP_53* | 0 | 0 | 0 | T | NA | NA | NA |
| 11 | IL23R | 1 | 67392837 | SNP_54 | G | 0,001437 | 0 | A | 1,993 | 0,158 | NA |
| 12 | IL23R | 1 | 67406400 | rs1884444 | G | 0,4161 | 0,4561 | T | 4,521 | 0,03347 | 0,8499 |
| 13 | IL23R | 1 | 67406551 | rs11465770 | T | 0,1013 | 0,1387 | C | 9,213 | 0,002403 | 0,6997 |
| 14 | IL23R | 1 | 67408538 | rs2295359 | A | 0,2594 | 0,3249 | G | 14,43 | 0,0001458 | 0,7276 |
| 15 | IL23R | 1 | 67444019 | SNP_7 | A | 0,02726 | 0,02945 | G | 0,1217 | 0,7272 | 0,9234 |
| 16 | IL23R | 1 | 67445873 | SNP_9* | 0 | 0 | 0 | G | NA | NA | NA |
| 17 | IL23R | 1 | 67446087 | SNP_58* | 0 | 0 | 0 | C | NA | NA | NA |
| 18 | IL23R | 1 | 67448316 | rs10889669 | G | 0,3737 | 0,2668 | T | 36,13 | 1,85E-09 | 1,64 |
| 19 | IL23R | 1 | 67457857 | SNP_114* | 0 | 0 | 0 | G | NA | NA | NA |
| 20 | IL23R | 1 | 67457975 | rs7530511 | T | 0,1381 | 0,1352 | C | 0,04929 | 0,8243 | 1,025 |
| 21 | IL23R | 1 | 67486458 | SNP_60* | 0 | 0 | 0 | G | NA | NA | NA |
| 22 | IL23R | 1 | 67486654 | SNP_61* | 0 | 0 | 0 | T | NA | NA | NA |
| 23 | IL23R | 1 | 67488758 | SNP_62* | 0 | 0 | 0 | G | NA | NA | NA |
| 24 | IL23R | 1 | 67497416 | rs11465827 | G | 0,004342 | 0,007267 | T | 1,023 | 0,3117 | 0,5956 |
| 25 | IL23R | 1 | 67497521 | SNP_124* | 0 | 0 | 0 | A | NA | NA | NA |
| 26 | IL23R | 1 | 67497708 | rs10889677 | A | 0,3665 | 0,2693 | C | 29,97 | 4,40E-08 | 1,57 |
| 27 | IL23R | 1 | 67497795 | SNP_126 | T | 0,007891 | 0,01006 | C | 0,3676 | 0,5443 | 0,7829 |
| 28 | IL10 | 1 | 204984177 | SNP_63 | 0 | 0 | 0 | C | NA | NA | NA |
| 29 | IL10 | 1 | 204993863 | rs61815630 | G | 0,2748 | 0,3048 | A | 3,024 | 0,08205 | 0,8645 |
| 30 | IL10 | 1 | 204994237 | SNP_17 | T | 0,01522 | 0,01381 | C | 0,09567 | 0,7571 | 1,104 |
| 31 | IL10 | 1 | 204997334 | rs4579758 | G | 0,2772 | 0,312 | A | 2,776 | 0,0957 | 0,8458 |
| 32 | IL10 | 1 | 205008810 | SNP_18 | T | 0,01365 | 0,005036 | C | 5,573 | 0,01824 | 2,734 |
| 33 | ATG16L1 | 2 | 233823325 | SNP_34* | A | 0,001433 | 0,0007205 | T | 0,328 | 0,5669 | 1,99 |
| 34 | ATG16L1 | 2 | 233829561 | SNP_119 | A | 0,001437 | 0,00566 | C | 3,301 | 0,06924 | 0,2528 |
| 35 | ATG16L1 | 2 | 233830146 | SNP_81* | 0 | 0 | 0 | C | NA | NA | NA |
| 36 | ATG16L1 | 2 | 233830746 | SNP_35 | G | 0,04748 | 0,05115 | A | 0,1996 | 0,655 | 0,9247 |
| 37 | ATG16L1 | 2 | 233836551 | rs13011156 | G | 0,07143 | 0,05908 | A | 1,734 | 0,1878 | 1,225 |
| 38 | ATG16L1 | 2 | 233838333 | SNP_120 | 0 | 0 | 0 | T | NA | NA | NA |
| 39 | ATG16L1 | 2 | 233840262 | SNP_36* | 0 | 0 | 0 | C | NA | NA | NA |
| 40 | ATG16L1 | 2 | 233846431 | SNP_122 | 0 | 0 | 0 | C | NA | NA | NA |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 41 | ATG16L1 | **2** | 233848107 | rs2241880 | T | 0,411 | 0,4654 | C | 8,331 | 0,003898 | 0,8018 |
| 42 | ATG16L1 | **2** | 233851452 | SNP_37 | A | 0,01149 | 0,01439 | G | 0,4561 | 0,4994 | 0,7965 |
| 43 | ATG16L1 | **2** | 233852754 | SNP_85* | A | 0,2475 | 0,2791 | G | 3,547 | 0,05966 | 0,8495 |
| 44 | ATG16L1 | **2** | 233856269 | rs2278610 | A | 0,03945 | 0,05755 | T | 4,941 | 0,02622 | 0,6726 |
| 45 | IRGM | **5** | 150181543 | SNP_86* | C | 0,01304 | 0,01089 | G | 0,2716 | 0,6022 | 1,201 |
| 46 | IRGM | **5** | 150206334 | SNP_87 | 0 | NA | NA | | NA | NA | NA |
| 47 | IRGM | **5** | 150206548 | SNP_43* | 0 | 0 | 0 | G | NA | NA | NA |
| 48 | IRGM | **5** | 150207929 | SNP_116 | C | 0,005747 | 0,0007194 | G | 5,452 | 0,01955 | 8,029 |
| 49 | IRGM | **5** | 150208020 | SNP_117 | 0 | 0 | 0 | A | NA | NA | NA |
| 50 | IRGM | **5** | 150208159 | SNP_118 | A | 0,04493 | 0,03353 | C | 2,371 | 0,1236 | 1,356 |
| 51 | IRGM | **5** | 150208191 | SNP_131 | T | 0,1075 | 0,0777 | C | 7,339 | 0,006747 | 1,43 |
| 52 | IRGM | **5** | 150228330 | SNP_89* | 0 | 0 | 0 | T | NA | NA | NA |
| 53 | IRGM | **5** | 150230561 | rs1277463* | A | 0,09393 | 0,1102 | G | 1,367 | 0,2424 | 0,8374 |
| 54 | IRGM | **5** | 150242995 | SNP_91 | C | 0,0436 | 0,02908 | T | 3,887 | 0,04867 | 1,522 |
| 55 | IRGM | **5** | 150244285 | SNP_92* | 0 | 0 | 0 | T | NA | NA | NA |
| 56 | IRGM | **5** | 150244718 | SNP_93* | C | 0,00942 | 0,01234 | T | 0,5451 | 0,4603 | 0,7613 |
| 57 | IRGM | **5** | 150252921 | SNP_94* | A | 0,006686 | 0,006747 | G | 0,0003633 | 0,9848 | 0,991 |
| 58 | NKX2-3 | **10** | 101260253 | SNP_66 | C | 0,03736 | 0,0369 | T | 0,003985 | 0,9497 | 1,013 |
| 59 | NKX2-3 | **10** | 101266896 | SNP_19* | 0 | 0 | 0 | A | NA | NA | NA |
| 60 | NKX2-3 | **10** | 101278795 | SNP_20 | T | 0,08921 | 0,09524 | C | 0,3014 | 0,583 | 0,9305 |
| 61 | NKX2-3 | **10** | 101283846 | rs884144 | T | 0,1295 | 0,1322 | C | 0,04541 | 0,8312 | 0,9763 |
| 62 | NOD2 | **16** | 49291360 | rs2067085 | G | 0,3564 | 0,4125 | C | 9,13 | 0,002514 | 0,7886 |
| 63 | NOD2 | **16** | 49299292 | SNP_130* | T | 0,003587 | 0,005029 | C | 0,3377 | 0,5612 | 0,7122 |
| 64 | NOD2 | **16** | 49302066 | SNP_129* | G | 0,001437 | 0 | T | 1,99 | 0,1583 | NA |
| 65 | NOD2 | **16** | 49302125 | rs2066842 | T | 0,41 | 0,2721 | C | 57,93 | 2,72E-14 | 1,859 |
| 66 | NOD2 | **16** | 49302189 | rs5743271 | G | 0,01437 | 0,01149 | A | 0,4503 | 0,5022 | 1,254 |
| 67 | NOD2 | **16** | 49302393 | SNP_98 | 0 | 0 | 0 | A | NA | NA | NA |
| 68 | NOD2 | **16** | 49302615 | SNP_99 | T | 0,006466 | 0,003597 | C | 1,143 | 0,285 | 1,803 |
| 69 | NOD2 | **16** | 49302618 | rs2076754 | T | 0,000717 | 4 0,0007194 | C | 4,13E-06 | 0,9984 | 0,9971 |
| 70 | NOD2 | **16** | 49302700 | rs2066843 | T | 0,41 | 0,2762 | C | 54,58 | 1,50E-13 | 1,821 |
| 71 | NOD2 | **16** | 49303084 | rs1861759 | C | 0,3553 | 0,4058 | A | 7,427 | 0,006424 | 0,8067 |
| 72 | NOD2 | **16** | 49303156 | rs61736932* | T | 0,01014 | 0,007994 | C | 0,3546 | 0,5515 | 1,272 |
| 73 | NOD2 | **16** | 49303302 | SNP_104 | G | 0,000717 | 4 0 | C | 0,9989 | 0,3176 | NA |
| 74 | NOD2 | **16** | 49303373 | rs5743276 | T | 0 | 0,003597 | C | 5,016 | 0,02511 | 0 |
| 75 | NOD2 | **16** | 49303427 | rs2066844 | C | 0,49 | NA | T | NA | NA | NA |
| 76 | NOD2 | **16** | 49308311 | SNP_109 | G | 0 | 0,00218 | A | 3,012 | 0,08265 | 0 |
| 77 | NOD2 | **16** | 49308343 | SNP_110 | G | 0,007902 | 0,002878 | A | 3,274 | 0,07037 | 2,76 |
| 78 | NOD2 | **16** | 49314041 | rs2066845 | C | 0,05548 | 0,01151 | G | 41,49 | 1,19E-10 | 5,044 |
| 79 | NOD2 | **16** | 49314777 | rs5743291 | A | 0,08285 | 0,08573 | G | 0,07467 | 0,7847 | 0,9633 |
| 80 | STAT3 | **17** | 37724627 | rs8066464 | 0 | 0 | 0 | A | NA | NA | NA |
| 81 | STAT3 | **17** | 37733350 | SNP_70 | 0 | NA | NA | | NA | NA | NA |
| 82 | STAT3 | **17** | 37749602 | SNP_73 | T | 0,000718 | 4 0,001439 | G | 0,3351 | 0,5627 | 0,4989 |
| 83 | STAT3 | **17** | 37753998 | SNP_26 | 0 | 0 | 0 | G | NA | NA | NA |
| 84 | STAT3 | **17** | 37761955 | SNP_76 | 0 | 0 | 0 | C | NA | NA | NA |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 85 | STAT3 | **17** | 37772557 | SNP_29 | T | 0,5 | 0,5 | C | 0 | 1 | 1 |
| 86 | STAT3 | **17** | 37782849 | SNP_77 | 0 | NA | NA | | NA | NA | NA |
| 87 | STAT3 | **17** | 37793614 | SNP_78 | 0 | 0 | 0 | C | NA | NA | NA |
| 88 | STAT3 | **17** | 37799123 | SNP_33 | 0 | 0 | 0 | A | NA | NA | NA |

## 4.7 Ergebnisse Assoziationsexperiment II in einer größeren Analysepopulation

Für dieses Assoziationsexperiment wurden 27 SNVs in einer noch größeren Analysepopulation auf Assoziation mit Morbus Crohn getestet. In diesem Experiment wurden die SNVs in 2800 deutschen Kontrollen und 2500 Crohnpatienten genotypisiert. Die Ergebnisse der Genotypisierung sind in Tabelle 4-8 dargestellt. Auch hier liessen sich für einige der SNVs keine Berechnungen anstellen, da sie in der Genotypisierung monomorph waren. Ein SNV (SNP_129) im schon mit Morbus Crohn assoziierten Gen *NOD2* zeigt eine Frequenz in den Fällen von 0,0019, also eine sehr seltene Variante in erkrankten Individuuen, und überhaupt kein Vorkommen in den Kontrollen. Dies ist ein Ergebnis, das weitere Analysen, z. B. eine Typisierung in einer noch größeren Studienpopulation und evtl. auch funktionelle Studien, nach sich ziehen könnte.

**Tabelle 4-8.** Die Tabelle zeigt die Ergebnisse der Genotypisierung von 27 Varianten in 2800 Kontrollen und 2500 Crohn-Patienten. Hervorgehoben ist die Variation mit dem besten p-Wert.

| Nr. | Locus | Chr | Pos | Name | A1 | A2 | Freq_Fälle | Freq_Kontrollen | CHISQ | P | OR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | NOD2 | 16 | 49299292 | SNP_130 | T | C | 0,003341 | 0,004982 | 1,516 | 0,2182 | 0,6695 |
| **2** | **NOD2** | **16** | **49302066** | **SNP_129** | **G** | **T** | **0,001909** | **0** | **10,73** | **0,001053** | **NA** |
| 3 | NOD2 | 16 | 49303156 | rs61736932 | T | C | 0,008831 | 0,009078 | 0,01652 | 0,8977 | 0,9725 |
| 4 | IL23R | 1 | 67384867 | SNP_46 | T | C | 0,00238 | 0,003561 | 1,101 | 0,294 | 0,6675 |
| 5 | IL23R | 1 | 67386595 | SNP_1 | 0 | A | 0 | 0 | NA | NA | NA |
| 6 | IL23R | 1 | 67388595 | SNP_48 | A | T | 0,0002381 | 0,000711 | 1,057 | 0,3038 | 0,3347 |
| 7 | IL23R | 1 | 67390806 | SNP_51 | T | C | 0,0007146 | 0,0001783 | 1,695 | 0,193 | 4,011 |
| 8 | IL23R | 1 | 67392833 | SNP_53 | 0 | T | 0 | 0 | NA | NA | NA |
| 9 | IL23R | 1 | 67445873 | SNP_9 | 0 | G | 0 | 0 | NA | NA | NA |
| 10 | IL23R | 1 | 67446087 | SNP_58 | 0 | C | 0 | 0 | NA | NA | NA |
| 11 | IL23R | 1 | 67457857 | SNP_114 | 0 | G | 0 | 0 | NA | NA | NA |
| 12 | IL23R | 1 | 67486458 | SNP_60 | 0 | G | 0 | 0 | NA | NA | NA |
| 13 | IL23R | 1 | 67486654 | SNP_61 | 0 | T | 0 | 0 | NA | NA | NA |
| 14 | IL23R | 1 | 67488758 | SNP_62 | 0 | G | 0 | 0 | NA | NA | NA |
| 15 | IL23R | 1 | 67497521 | SNP_124 | G | A | 0 | 0,0001781 | 0,7482 | 0,387 | 0 |
| 16 | NKX2-3 | 10 | 101266896 | SNP_19 | 0 | A | 0 | 0 | NA | NA | NA |
| 17 | IRGM | 5 | 150181543 | SNP_86 | C | T | 0,0181 | 0,01508 | 1,365 | 0,2427 | 1,204 |
| 18 | IRGM | 5 | 150206548 | SNP_43 | A | G | 0,0002381 | 0,001601 | 4,39 | 0,03616 | 0,1485 |
| 19 | IRGM | 5 | 150228330 | SNP_89 | C | G | 0,01238 | 0,01069 | 0,6107 | 0,4345 | 1,16 |
| 20 | IRGM | 5 | 150228330 | SNP_89 | A | T | 0 | 0,0001779 | 0,7478 | 0,3872 | 0 |
| 21 | IRGM | 5 | 150230561 | rs1277463 | A | G | 0,1189 | 0,1255 | 0,9681 | 0,3252 | 0,9402 |
| 22 | IRGM | 5 | 150244285 | SNP_92 | C | T | 0 | 0,0001781 | 0,7478 | 0,3872 | 0 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | IRGM | 5 | 150252921 | SNP_94 | A | G | 0,01303 | 0,008565 | 4,588 | 0,03219 | 1,528 |
| 24 | ATG16L1 | 2 | 233829561 | SNP_119 | 0 | C | 0 | 0 | NA | NA | NA |
| 25 | ATG16L1 | 2 | 233830146 | SNP_81 | 0 | C | 0 | 0 | NA | NA | NA |
| 26 | ATG16L1 | 2 | 233851452 | SNP_37 | A | G | 0,01309 | 0,009259 | 3,27 | 0,07054 | 1,419 |
| 27 | ATG16L1 | 2 | 233852754 | SNP_85 | A | G | 0,2591 | 0,272 | 2,051 | 0,1521 | 0,9358 |

# 5 Ergebnisse der Gesamtgenom-Sequenzierung

## 5.1 Datenproduktion und Mappingergebnisse

Für die Gesamt-Genom-Sequenzierung der an Morbus Crohn erkrankten Patientin sind 18 SOLiD-Sequenzierläufe durchgeführt worden. Dafür sind vier verschiedene Typen von Libraries hergestellt worden. Die Ergebnisse und die Menge der Rohdaten der Sequenzierung zeigt die Tabelle 5-1 sowie Abbildung 5-1. Die sich ergebene und Lauf um Lauf zunehmende Coverage für das sequenzierte Genom ist in Abbildung 5-2 dargestellt. Die Abbildung verdeutlicht, dass einige Sequenzierläufe nötig waren, um eine genügende durchschnittliche Coverage für ein genomweites SNP-calling zu erhalten. Mehr als 99,5% des gesamten Genoms wurde mit mindestens einem Read abgedeckt, über 90% des Genoms konnten mit einer mindestens 20-fachen Coverage sequenziert werden. Die durchschnittliche Coverage der Sequenzierung des Crohn-Genom liegt bei 58x, was vergleichbar bzw. sogar besser ist als in anderen vergleichbaren Studien[95,121–124].

**Tabelle 5-1.** Überblick über die Datenproduktion aller sequenzierten Libraries

| Lauf Nr. | Typ / Read-Länge | Library ID | Insert Größe | Library Version | Anzahl der Reads | Gigabasen | Uniquely Mapped Reads | gemappte Gigabasen | Durchschnittliche Coverage |
|---|---|---|---|---|---|---|---|---|---|
| 1 | mate-pair 2×25 bp | 551 | 0,5-1,0 kb | v2 | 377,473,014 | 9,436,825,350 | 198,604,320 | 4,965,108,000 | 1.64 |
| 2 | | | | | 379,646,168 | 9,491,154,200 | 203,113,392 | 5,077,834,800 | 1.68 |
| 3 | | | | | 438,624,583 | 10,965,614,575 | 228,666,152 | 5,716,653,800 | 1.89 |
| 4 | | | | | 368,183,143 | 9,204,578,575 | 201,832,165 | 5,045,804,125 | 1.67 |
| 5 | | | | | 393,535,846 | 9,838,396,150 | 197,543,475 | 4,938,586,875 | 1.63 |
| 6 | | | | | 419,578,232 | 10,489,455,800 | 200,433,675 | 5,010,841,875 | 1.66 |
| 7 | | | | | 289,646,594 | 7,241,164,850 | 168,172,798 | 4,204,319,950 | 1.39 |
| 8 | | 593 | 4-6 kb | | 287,706,210 | 14,385,310,500 | 130,852,297 | 3,271,307,425 | 1.08 |
| 9 | mate-pair 2×50 bp | 611 | 400-500 bp | v3 Rev. B | 426,964,096 | 21,348,204,800 | 237,351,467 | 10,416,680,926 | 3.45 |
| 10 | | | | | 424,060,311 | 21,203,015,550 | 226,722,530 | 9,880,263,802 | 3.27 |
| 11 | | | | | 350,817,343 | 17,540,867,150 | 200,984,305 | 8,844,754,295 | 2.93 |
| 12 | | | | | 433,751,402 | 21,687,570,100 | 239,805,548 | 10,484,173,975 | 3.47 |
| 13 | | 616 | 3-4 kb | | 507,809,229 | 25,390,461,450 | 319,649,390 | 14,523,033,498 | 4.80 |
| 14 | | | | | 479,746,134 | 23,987,306,700 | 313,815,765 | 14,292,762,977 | 4.73 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **15** | | 859 | 2-3 kb | v3+ Rev. A | 711,777,604 | 35,588,880,200 | 518,176,294 | 23,659,506,725 | 7.83 |
| **16** | | 860 | 5-6 kb | | 1,182,149,875 | 59,107,493,750 | 671,251,210 | 28,667,410,995 | 9.48 |
| **17** | fragment 50 bp | 858 | - | | 421,158,120 | 21,057,906,000 | 273,652,481 | 12,712,239,529 | 4.21 |
| **18** | fragment exome 50 bp | 864 | - | v1.5 | 73,554,239 | 3,677,711,950 | 49,931,360 | 2,298,302,376 | 0.76 |
| **Total** | | | | | **7,966,182,143** | **331,641,917,650** | **4,580,558,624** | **174,009,585,948** | **58** |

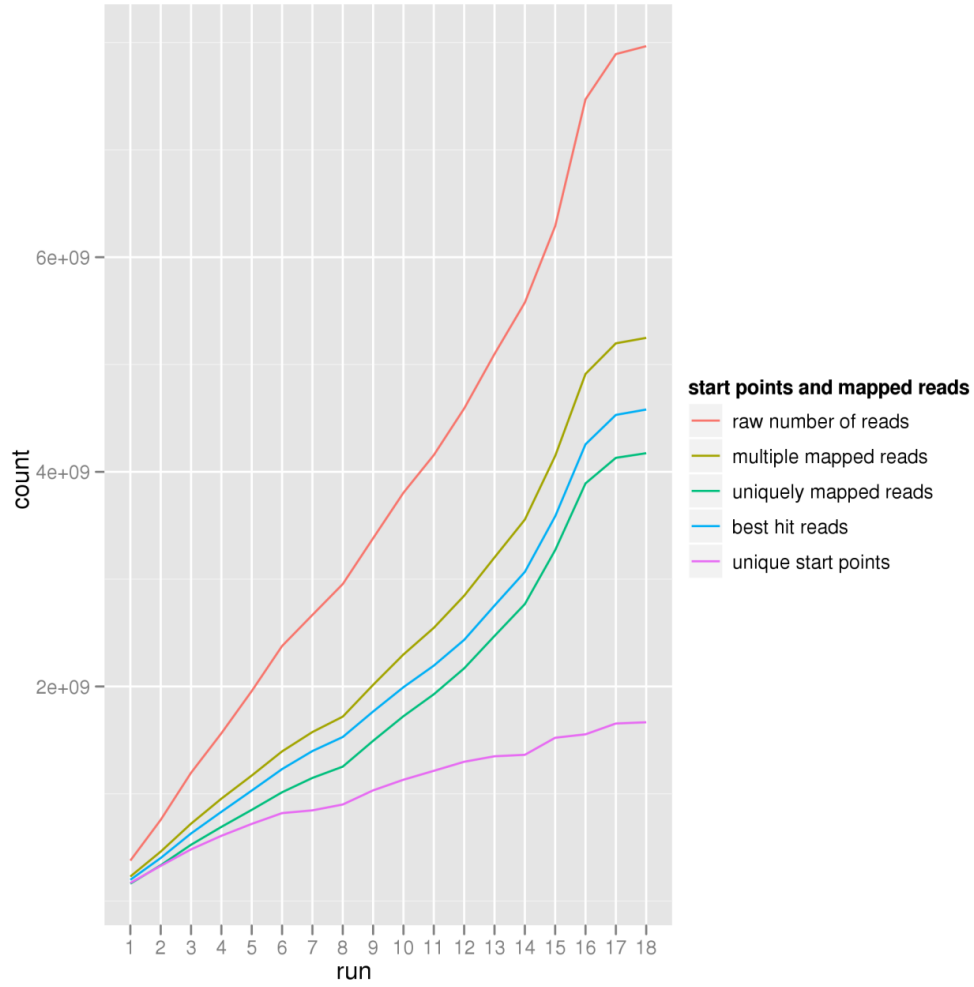



**Abbildung 5-1.** Gemappte Reads. Die Abbildung zeigt die Gesamtheit aller Rohdaten sowie die Anzahl der Reads nach den verschiedenen Filterstufen innerhalb des Mappingprozesses.

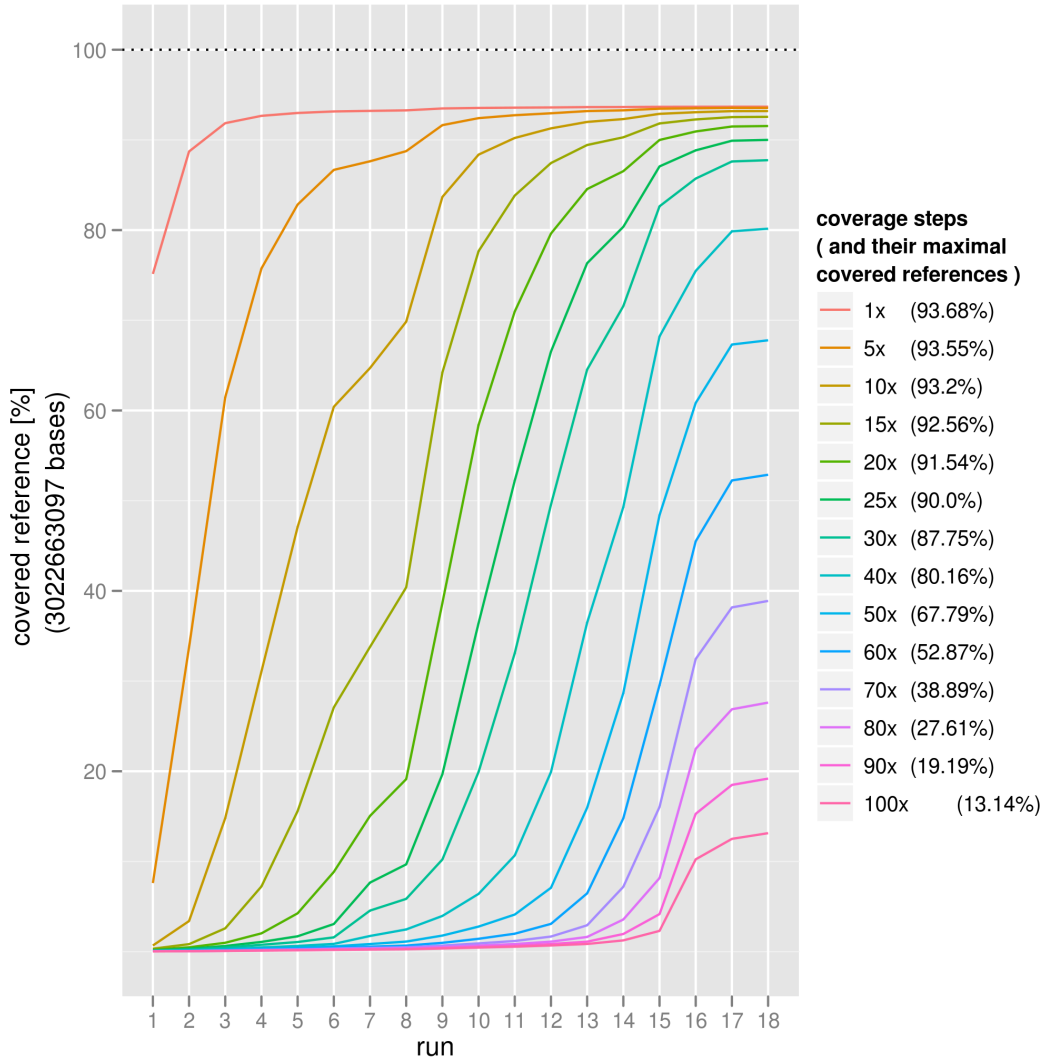



**Abbildung 5-2.** Coverage-Diagramm für die Gesamt-Genom-Sequenzierung.

## 5.2 Zusammenfassende Statistik der SNP-Detektion und Annotation

Eine Gesamtzahl von 3,308,456 SNPs sind für das Crohn-Genom detektiert worden. 9,34% der detektierten SNPs waren zum Zeitpunkt der Datenbankabfrage neu. Es wurden 1,291,234 homozygote SNPs detektiert, 2,017,222 SNPs sind heterozygot. Der Heterozygotiegrad beträgt 0,61%. Die Transition/Transversion-Ratio wurde für den gesamten SNP-Datensatz berechnet und beträgt 2,016. Die quantitative und qualitative Evaluation der identifizierten Varianten wurde durch einen Vergleich mit dbSNP Version 130 und den Daten des „1000-Genomes-Project“ durchgeführt. Die Tabelle 5-2 zeigt eine detaillierte Übersicht der detektierten Varianten.

**Tabelle 5-2.** Zusammenfassende Statistik der SNP-Detektion und Annotation

| **Anzahl, Verhälnisse & Vorhersagen** | | | **SNPs in funktionalen Elementen** | |
|---|---|---|---|---|
| Komplette Anzahl SNPs | 3,308,456 | | synonym-codierend SNPs | 9,119 |
| Komplette Anzahl NOVEL SNPs | 309,216 | | missense SNPs | 8,263 |
| Min. Coverage pro SNP | 2 | | SNPs in Startcodons | 13 |
| Max. Coverage pro SNP | 23,492 | | SNPs in Stopcodons | 25 |
| Durchschnittl. Coverage pro SNP | 20.6933 | | nonsense SNPs | 69 |
| Heterozygosität | 0.609717 | | DAMAGING SNPs | 8370 |
| Ti/Tv ratio | 2.01589 | | SNPs in Splice Acceptor Sites | 13 |
| Komplette Anzahl unbekannte/intergenetische SNPs | 2,091,784 | | SNPs in Splice Donor Sites | 16 |
| Polyphen Prognose | 4,049 | | SNPs in 5'-UTR | 5,592 |
| SNAP Prognose | 2,653 | | SNPS in 3'-UTR | 22,197 |
| SIFT prognose | 0 | | SNPs in CpG-Islands | 19,802 |
| SNPs3d Prognose | 6,772 | | SNPs in Promotor | 326 |
| BAD Polyphen Prognose | 592 | | NOVEL synonymous-coding SNPs | |
| possibly BAD Polyphen Prognose | 426 | | NOVEL missense SNPs | 809 |
| BAD SNAP Prognose | 1,181 | | NOVEL SNPs in Startcodon | 1 |
| BAD SIFT Prognose | 0 | | NOVEL SNPs in Stopcodon | 2 |
| BAD SNPs3d Prognose n | 665 | | NOVEL nonsense SNPs | 22 |
| Any BAD Prognose | 2864 | | NOVEL DAMAGING SNPs | 834 |

## 5.3 SNP-calling –Vergleich mit publizierten Genomen

Die SNPs aus der Sequenzierung des Crohn-Genoms sind mit anderen Genomen, die komplett sequenziert worden sind (und deren Daten öffentlich zugänglich sind) verglichen worden und es sind daraufhin Berechnungen für die Übereinstimmungen bzw. Überlappungen angestellt worden. Dargestellt sind die Ergebnisse im Venn-Diagramm in Abbildung 5-3. Die Abbildung zeigt die Überlappungen mit den verschiedenen Referenzgenomen. Sie zeigt

gewissermaßen, wieviel der detektierten SNPs für jedes sequenzierte Genom mit den SNPs der anderen sequenzierten Genome „geteilt“ werden. Hier zeigt sich, dass die SNPs für das Crohn-Genom (als einziges der vier von einem kranken Individuum), die ausschließlich in diesem Genom detektiert wurden, prozentual nicht mehr ergeben als für die anderen Genome. Für das Crohn-Genom sind das 26%, für Venter 29%, für Quake 22,6% und Watson 18% der SNPs, die „individuell“ waren.

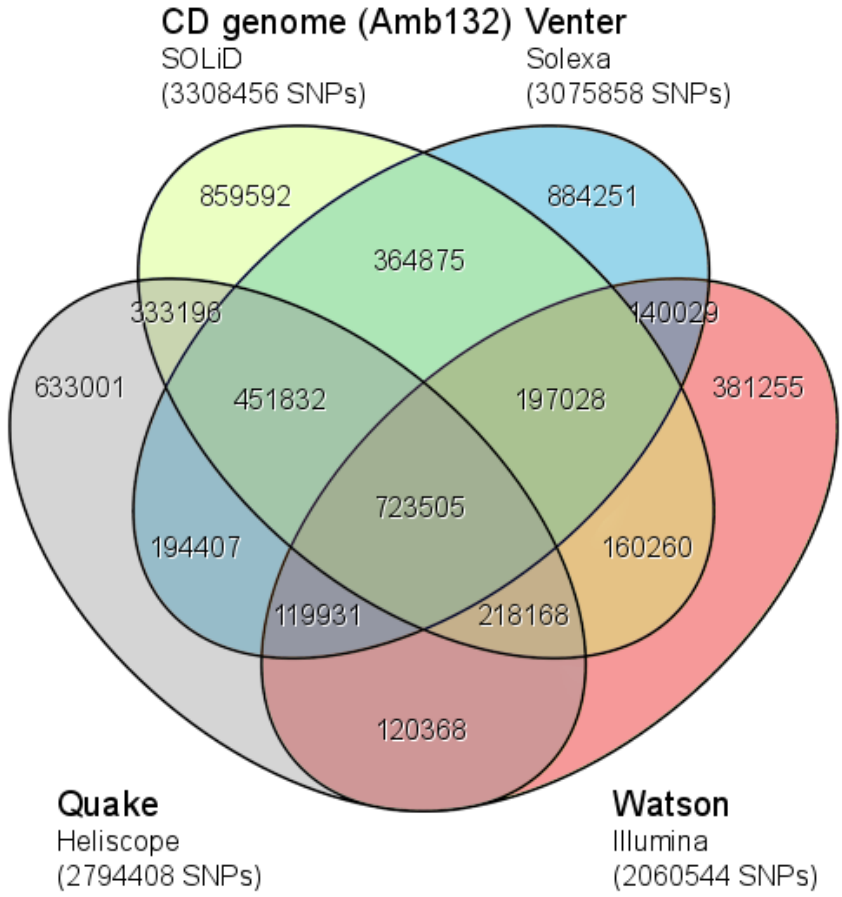



**Abbildung 5-3.** Die Abbildung zeigt die Überlappungen mit den verschiedenen Referenzgenomen. Hier zeigt sich, dass „individuelle“ SNPs im Crohn-Genom Amb132 nicht häufiger auftreten als in den Referenzgenomen.

## 5.4 Zusammenfassende Statistik der Small InDel Detektion und Annotation

Die Tabelle 5-3 zeigt eine zusammenfassende Statistik der Ergebnisse der Small InDel-Detektion. Als Small InDel bezeichnet man i. d. R. kleine Insertionen oder Deletionen, die kleiner als 20 bp sind. Mehr als 300.000 Small InDels sind für das Genom identifiziert worden. Schon auf den ersten Blick ist sichtbar, dass die Small InDels vor allem in nicht-kodierenden Bereichen detektiert worden sind. Nur ein kleiner Anteil der detektierten Small InDels ist in exonischen Bereichen lokalisiert. Die Anzahl exonischer InDels liegt weit unter 1%, was auch erwartet werden konnte, da diese Varianten wahrscheinlich nicht

funktionsfähige Transkripte oder Proteine zur Folge haben. In Abbildung 5-4 sind die detektierten InDels nach funktionellen Kategorien sortiert in ihrer Längenverteilung gezeigt. Der größte Teil der InDels hat eine Länge von 1 bp.

**Tabelle 5-3.** Die Tabelle zeigt eine zusammenfassende Statistik der Small InDel-Detektion und Annotation.

| **InDels** | | | **InDels in Exom** | |
|---|---|---|---|---|
| Komplette Anzahl InDels | 302.547 | | Komplette Anzahl der exonischen InDels | 451 |
| InDels mit der Länge 1 | 202.187 | | exonisch InDels der Länge 1 | 255 |
| InDels mit „Triplett Länge“ | 29.252 | | exonisch InDels mit „Triplett Länge“ | 108 |
| InDels mit „nicht-Triplett Länge“ | 273.295 | | exonisch InDels mit „nicht-Triplett Länge“ | 343 |
| intergenetisch | 175.700 | | | |
| downstream | 2.213 | | **Funktionale Effekte** | |
| upstream | 1.763 | | Nicht-Frameshift Substitution | 17 |
| downstream / upstream | 68 | | Nicht-Frameshift Deletion | 81 |
| intronisch | 110.271 | | Nicht-Frameshift Insertion | 57 |
| UTR3 | 2.783 | | stopgain | 1 |
| UTR5 | 294 | | Frameshift Substitution | 26 |
| UTR3 / UTR3 | 2 | | Frameshift Deletion | 117 |
| ncRNA, intronisch | 8.346 | | Frameshift Insertion | 64 |
| ncRNA, UTR 3 | 32 | | unbekannt | 80 |
| ncRNA, UTR 5 | 4 | | | |
| ncRNA, splicing | 5 | | | |
| exonisch / splicing | 11 | | | |
| splicing | 81 | | | |
| exonisch | 438 | | | |
| | | | | |

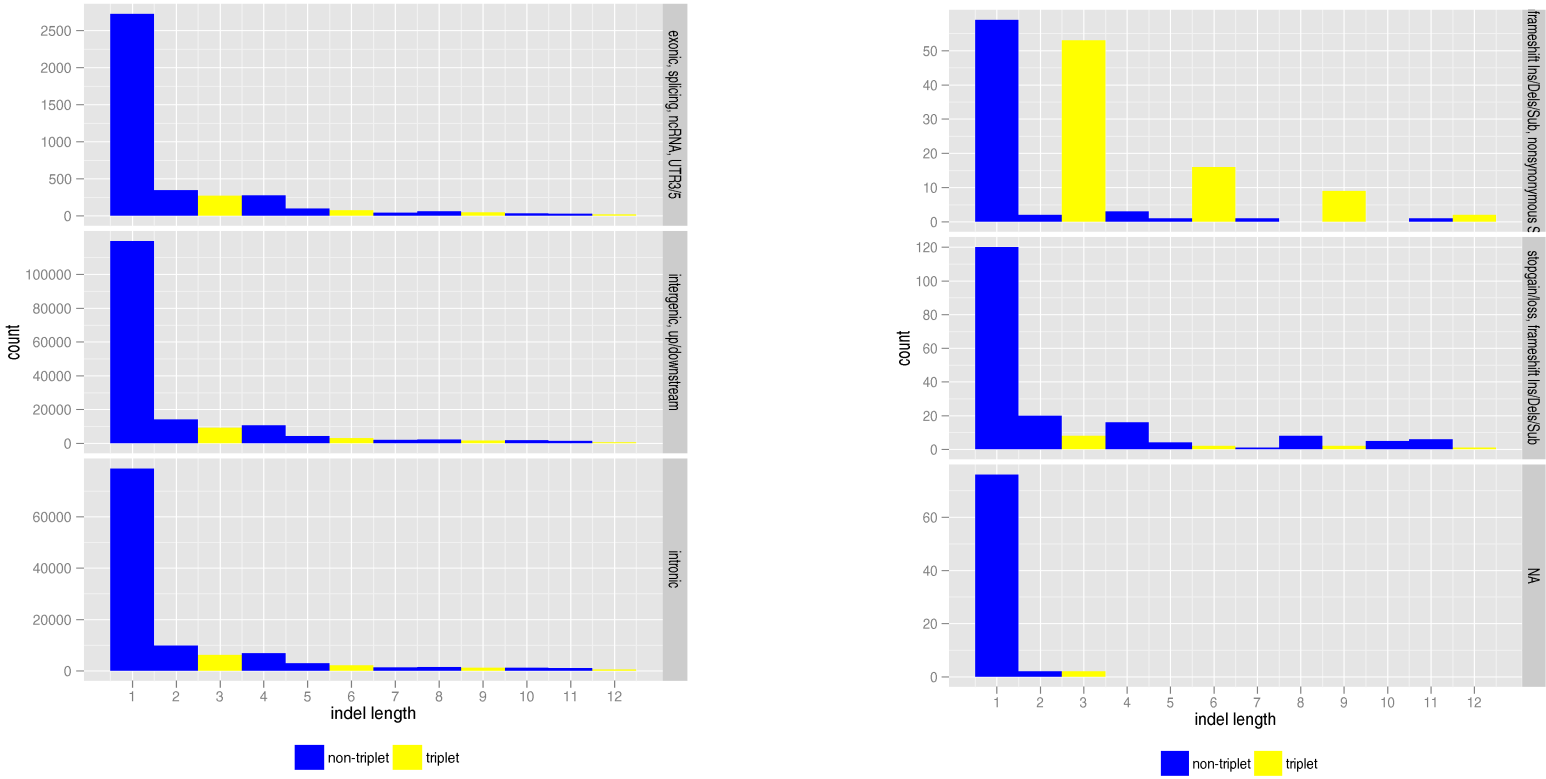



**Abbildung 5-4:** Die Abbildung zeigt die Längenverteilung der detektierten exonischen InDels. In der Abbildung sind die unterschiedlichen funktionellen Kategorien voneinander getrennt dargestellt.

## 5.5 Zusammenfassende Statistik der CNV-Analyse

Die Tabelle 5-4 zeigt eine zusammenfassende Statistik der detektierten copy number variations (CNVs). Spalte A und B zeigen die Anzahl der CNVs vor und nach dem Filtern für eine minimale Coverage von 5x , die CNV-Ereignisse stützen. Spalte C zeigt die Anzahl der CNV-Ereignisse für jede Library, die mit mindestens einer anderen Library für dieses CNV-Ereignis übereinstimmt; Spalte D zeigt alle verbleibenden CNVs zusammenfassend für alle Libraries nach dem Filterprozess.

**Tabelle 5-4.** Die Tabelle zeigt eine Übersicht der CNVs nach erster visueller Inspektion.

| Library ID | Art der SV | A Anzahl der CNVs vor dem Filtern | B Anzahl der CNVs Nach dem Filtern | Überlappung mit anderen libraries | C Anzahl der CNVs pro library Die Überlappung mit CNVs einer anderen Library | verbleibende CNVs | D übrig gebliebene CNVs |
|---|---|---|---|---|---|---|---|
| **551** | **Total** | **2,359** | **1,983** | | **524** | | 100 |
| | Inversionen | 759 | 759 | | 97 | | Inversions |
| | Deletionen | 1,600 | 1,224 | | 425 | | |
| **593** | **Total** | **494** | **148** | | **69** | | |
| | Inversionen | 58 | 39 | | 8 | | 437 |
| | Deletionen | 145 | 87 | | 60 | | Deletions |
| | Insertionen | 291 | 22 | | 1 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **611** | **Total** | **2,767** | **1,603** | | **425** | | |
| | Inversionen | 481 | 481 | | 81 | | 2 |
| | Deletionen | 2,286 | 1,122 | | 349 | | Insertions |
| **616** | **Insgesamt** | **782** | **782** | | **289** | | |
| | Inversionen | 290 | 290 | | 54 | | |
| | Deletionen | 398 | 398 | | 233 | | |
| | Insertionen | 94 | 94 | | 2 | | |
| **Total** | | **6,402** | **4,516** | | | | **539** |

Die Tabelle 5-5 fasst die Deletionen nach der visuellen Inspektion mit dem hausinternen UCSC–Browser zusammen. Mit Hilfe dieses Tools konnten falsch-positive Ergebnisse eliminiert werden, homozygote/heterozygote unterschieden werden und überprüft werden, welche der CNVs schon in der "Database of Genomic Variants" (DGV) annotiert sind bzw. neu sind. Außerdem konnte überprüft werden, welche der Deletionen mit einer alternativen Methode (NimbleGen) in dieser Patientin ebenfalls detektiert werden konnten.

**Tabelle 5-5.** Die Tablelle zeigt einen Überblick über alle nach dem Filterprozess übriggebliebenen Deletionen.

| Nach visueller Prüfung | Hom | Het | DGV | Überlappung NimbleGen | Betroffene Gene | Codierende Regionen betreffend |
|---|---|---|---|---|---|---|
| 346 | 144 | 202 | 338 | 103 | 109 | 8 |

**Tabelle 5-6.** Die Tabelle zeigt alle Deletionen (nach Filterprozess), die entweder kodierende Regionen betreffen oder nicht in DGV (Database of Genetic Variation) annotiert sind.

| Chr | Start | Ende | Hom/het | Größe | Betroffene Gene | DGV |
|---|---|---|---|---|---|---|
| 1 | 6,360,755 | 6,368,503 | het | 7,748 | ACOT7 | + |
| 1 | 147,578,431 | 147,920,425 | het | 341,994 | FCGR1C, FCGR1B, PPIAL4A | + |
| 1 | 150,822,115 | 150,854,484 | hom | 32,639 | LCE3C, LCE3B | + |
| 3 | 148,150,128 | 148,156,313 | het | 6,185 | - | - |
| 4 | 14,290,739 | 14,292,787 | het | 2,051 | MGC4836 | - |
| 6 | 68,740,468 | 68,745,532 | het | 5,064 | - | - |
| 6 | 72,884,427 | 72,891,122 | het | 6,695 | RIMS1 | - |
| 6 | 76,298,770 | 76,323,188 | het | 24,418 | - | - |
| 6 | 32,562,268 | 32,640,097 | het | 77,829 | HLA-DRB5, HLA-DRB1 | + |
| 9 | 109,073,265 | 109,075,295 | hom | 2,030 | FKSG56 | + |
| 10 | 114,102,166 | 114,106,647 | het | 4,481 | GUCY2GP | + |

| | | | | | | |
|---|---|---|---|---|---|---|
| **11** | 97,771,805 | 97,774,301 | het | 2,496 | - | - |
| **11** | 113,930,107 | 113,939,150 | het | 7,043 | FAM55A | + |
| **13** | 56,343,241 | 56,347,423 | het | 4,182 | - | - |
| **14** | 105,519,999 | 105,521,942 | het | 1,943 | Parts of Antibodies | + |
| **20** | 5,576,990 | 5,578,460 | het | 1,470 | - | - |

## 5.6 Validierung ausgewählter Varianten

Im Rahmen dieses Projektes sind zur Evaluierung der SNP-Detektionsergebnisse 545 SNPs mit Hilfe der Sanger Technologie überprüft worden. Die Ergebnisse sind in Tabelle 5-7 dargestellt. Von 545 Varianten konnten 359 Varianten mit kompletter Übereinstimmung zu den Detektionsergebnissen aus der SOLiD-Sequenzierung validiert werden. Das entspricht 66%. Bei 10% der Varianten war der durch die SOLiD-Sequenzierung ermittelte Genotyp falsch bestimmt worden, d.h. an dieser Position wurde in der Sanger-Sequenzierung ebenfalls eine Abweichung zum Referenzallel gezeigt, Homo- bzw. Heterozygotie waren aber nicht übereinstimmend. Für 18% der für die Validierung ausgewählten SNPs wurde ein falsch-positives Detektionsergebnis festgestellt. Für einen kleinen Teil konnten keine eindeutigen Ergebnisse produziert werden. Der Validierungsprozess verdeutlicht, dass vor allem die bisher unbekannten Varianten gefährdet sind, was sowohl die genaue Bestimmung des Genotyps als auch falsch-positive Detektionsergebnisse betrifft.

**Tabelle 5-7.** Sanger-Validierung. Die Tabelle zeigt die Ergebnisse der Sanger-Validierung von insgesamt 545 in der SOLiD-Sequenzierung detektierten Varianten.

| | Gesamtanzahl | bekannte SNPs | neue SNPs |
|---|---|---|---|
| **Eingang in Validierungsprozess** | 545 | 292 | 253 |
| **validiert** | 359 | 226 | 133 |
| **mut-Typ falsch** | 53 | 40 | 13 |
| **falsch-positiv** | 97 | 10 | 87 |
| **Kein Ergebnis** | 35 | 20 | 15 |

## 5.7 Ausgewählte Kandidaten als Suszeptibilitätsvarianten

Da dieses Projekt eine Pilot-Studie ist, war es wichtig zu erfahren, die Qualität der SNP-Detektion zu evaluieren, bevor man die detektierten Varianten in weitere Analysen, die experimentell aufwendig sind, einfließen lässt. Diese SNPs wurden dafür mit der etablierten Sequenzierungsmethode nach Sanger validiert. Von der DNA der Crohn-Patientin wurde im Rahmen einer Exom-Sequenzierungsstudie zudem verschiedene Exom-Enrichmentmethoden und Sequenzierungen mit verschiedenen NGS-Technologien durchgeführt, das hier nicht näher erläutert wird. Die SNP-Detektionsergebnisse dieser Exom-Studien konnten zusätzlich für die vorliegende Arbeit mitverwendet werden, um die Ergebnisse des SNP-callings zu vergleichen. Die Tabelle 5-8 zeigt eine Liste der krankheitsassoziierten HGMD-SNPs (engl.: human gene mutation database) aus der Gesamt-Genom-Sequenzierung und einen Vergleich mit den verschiedenen Exom-Enrichment basierten Sequenzierungsexperimenten (SuSe IS=

sure select in solution; NiGe IS= NimbleGen in solution; NiGe AB= NimbleGen array based; Img AB= imagene array based; bgi=?) der Crohnpatientin Amb132 sowie die Ergebnisse der Validierung mit Hilfe der Sanger-Sequenzierung.

Die Sanger-Validierung ergab sehr zufriedenstellende Ergebnisse. Nur vier der in der Genomsequenzierung detektierten SNPs aus dieser Liste konnten mit Hilfe der Sanger-Sequenzierung nicht validiert werden, bei einem SNP war das Ergebnis nicht eindeutig. Die ausgewählten SNPs konnten außerdem immer in mindestens einem der Exom-Sequenzierungen gefunden werden. Der Vergleich der verschiedenen Exom-Enrichment-Strategien und Sequenzierungen zeigt aber an sich große Differenzen. Hier gibt es untereinander große Abweichungen in den Ergebnissen der SNP-Detektion. In keinem der Exom-Enrichment-Experimente konnten alle SNPs, die mit Hilfe der Gesamt-Genom-Sequenzierung gefunden worden, detektiert werden. Da die Ergebnisse der Exom-Sequenzierung im Rahmen der hier vorliegenden Arbeit ausschließlich für diesen Vergleich fungieren, werden die Ergebnisse der Exomsequenzierung an dieser Stelle nicht näher erläutert .

**Tabelle 5-8.** Die Tabelle zeigt ausgewählte SNPs der Patientin. Diese SNPs wurden einer Sanger-Validierung unterzogen, außerdem zeigt die Tabelle den Vergleich mit den am Institut durchgeführten Exom-Sequenzierungen der Patienten-DNA mit verschiedenen Technologien. Ausgewählt wurden die SNPs nach einem Filterprozess nach der Annotations mit SnpActs. Unbekannte SNPs, exonische SNPs, nsSNPs waren die für diese Auswahl entscheidende Kriterien.

| Chr | Pos | RS number | mut Typ | Gene symbol | Effect | 1000G | HGMD disease | Coverage | Alt. allele frequency (%) | SuSe IS | NiGe IS | NiGen AB | Img AB | BGI | Sanger validation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 63,654,140 | 4630153 | het | ALG6 | mis | 0,716667 | Congenital disorder of glycosylation 1c, mild | 69 | 38 | + | | | | | + |
| 1 | 245,655,481 | 35829419 | het | NLRP3 | mis | 0,091667 | Cryopyrin-associated periodic syndrome | 109 | 36 | + | | + | + | + | + |
| 2 | 74,970,035 | 2229629 | het | HK2 | mis | | Diabetes, NIDDM | 74 | 40 | | | | | | + |
| 3 | 4,679,816 | 41289628 | het | NULL | mis | | Spinocerebellar ataxia 15 | 101 | 44 | | | | | | + |
| 5 | 131,439,359 | 25882 | het | CSF2 | mis | 0,241667 | Pulmonary alveolar proteinosis | 66 | 40 | + | | + | | | + |
| 6 | 18,230,485 | 10949483 | het | NHLRC1 | mis | 0,441667 | Myoclonic epilepsy of Lafora | 20 | 40 | | | | | | + |
| 6 | 25,958,824 | 56027330 | het | SLC17A3 | mis | 0,116667 | Glycogen storage disease 1c | 121 | 18 | | + | | | + | + |
| 10 | 13,380,242 | 28938169 | het | PHYH | mis | 0,15 | Refsum disease | 71 | 23 | + | | | | + | + |
| 10 | 101,819,504 | 61751507 | het | CPN1 | mis | 0,025 | Carboxypeptidase N deficiency | 39 | 39 | + | | + | | | N.A. |
| 11 | 36,552,776 | 4151031 | het | RAG1 | mis | | Omenn syndrome | 68 | 34 | | + | + | | + | + |
| 11 | 36,554,446 | - | het | RAG1 | mis | | Omenn syndrom with aniridia | 100 | 47 | | | + | | + | + |
| 1 | 99,760,1 | 357193 | het | PCCA | mis | 0,025 | Propionic acidaemia | 172 | 23 | | | | + | + | + |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 57 | 59 | | | | | | | | | | | | | | |
| 14 | 20,859,880 | 10151259 | het | RPGRIP1 | mis | 0,175 | Cone-rod dystrophy | 70 | 47 | | | | | | | + |
| 15 | 50,430,856 | 1058219 | het | MYO5A | mis | 0,183333 | Griscelli syndrome | 49 | 41 | + | | | | | | + |
| 16 | 52,250,195 | - | het | RPGRIP1L | mis | | Bardet-Biedl syndrome | 56 | 0 | + | + | + | | + | + | + |
| 16 | 88,361,077 | 17233141 | het | FANCA | mis | | Fanconi anaemia | 60 | 47 | + | | + | + | + | + | + |
| 16 | 88,513,279 | - | het | MC1R | mis | | Melanoma | 50 | 14 | | | | | | | ? |
| 19 | 5,783,209 | 778805 | het | FUT6 | mis | 0,35 | Fucosyltransferase deficiency | 89 | 36 | | | | | + | + | |
| 19 | 5,795,537 | 778986 | het | FUT3 | mis | 0,791667 | Lewis antigen, absence | 82 | 46 | + | | + | | + | + | |
| 19 | 5,795,649 | 812936 | het | FUT3 | mis | 0,791667 | Lewis antigen, absence | 51 | 26 | + | | + | + | | N.A. | |
| 19 | 5,795,804 | - | het | FUT3 | mis | | Lewis antigen, absence | 42 | 36 | | | + | + | | - | |
| 19 | 41,034,052 | 3814995 | het | NPHS1 | mis | 0,383333 | Congenital nephrotic syndrome, Finnish type | 23 | 26 | | | + | + | | + | |
| 20 | 5,242,762 | - | het | PROKR2 | mis | | Kallmann syndrome | 57 | 21 | + | + | + | + | + | + | |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 31,464,181 | 56157422 | het | SNTA1 | mis | | Long QT syndrome | 46 | 45 | + | | + | | | + |
| 20 | 60,933,967 | 751557 | het | COL9A3 | mis | 0,216667 | Pseudoachondroplasia | 25 | 38 | | | | + | | + |
| 22 | 40,787,002 | - | het | NAGA | mis | | Neuroaxonal dystrophy, infantile | 51 | 48 | | | + | | + | + |
| 22 | 49,311,121 | - | het | TYMP | mis | | Mitochondrial neurogastrointestinal encephalopathy | 17 | 46 | | | | | | N.A. |
| 1 | 98,121,473 | 1801265 | hom | DPYD | mis | 0,85 | Dihydropyrimidine dehydrogenase deficiency | 114 | 2 | + | + | + | | | + |
| 1 | 100,444,648 | 12021720 | hom | DBT | mis | 0,933333 | Maple syrup urine disease | 48 | 9 | + | + | + | | + | + |
| 1 | 115,377,546 | 10776792 | hom | TSHB | mis | 0,958333 | Hypothyroidism | 120 | 15 | + | + | | + | + | + |
| 3 | 33,030,725 | 4302331 | hom | GLB1 | mis | 1 | Gangliosidosis GM1 | 58 | 8 | + | | | | | + |
| 3 | 33,113,553 | 7637099 | hom | GLB1 | mis | 0,658333 | Gangliosidosis GM1 | 25 | 11 | + | | | | | + |
| 4 | 187,395,028 | 3733402 | hom | KLKB1 | mis | 0,491667 | Prekallikrein deficiency | 68 | 21 | + | + | + | | + | + |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 74,017,026 | 820878 | hom | HEXB | mis | 0,983333 | Sandhoff disease | 21 | 0 | + | | + | | | + |
| 5 | 137,234,459 | 66488826 | hom | MYOT | mis | 0,991667 | Myotilinopathy | 97 | 2 | + | | + | + | + | + |
| 5 | 149,341,070 | 30832 | hom | SLC26A2 | mis | 0,991667 | Diastrophic dysplasia | 80 | 7 | | + | + | + | + | + |
| 6 | 49,511,241 | 8589 | hom | MUT | mis | 0,675 | Methylmalonic aciduria | 90 | 1 | + | + | + | | + | + |
| 8 | 145,611,219 | 2977838 | hom | SLC39A4 | mis | 0,958333 | Acrodermatitis enteropathica | 13 | 14 | + | | + | | | N.A. |
| 10 | 13,206,082 | 523747 | hom | OPTN | mis | 0,991667 | Glaucoma 1, open angle | 44 | 5 | + | + | + | + | + | + |
| 11 | 68,462,250 | 17612126 | hom | IGHMBP2 | mis | 0,333333 | Spinal muscular atrophy with resp. distress 1 | 98 | 3 | + | | | + | | + |
| 11 | 87,685,231 | 217086 | hom | CTSC | mis | 0,833333 | Papillon-Lefevre syndrome | 146 | 7 | + | + | + | + | + | + |
| 12 | 120,779,718 | 1154510 | hom | HPD | mis | 0,875 | Hawkinsinuria | 79 | 5 | + | + | | | + | + |
| 13 | 31,827,387 | 169547 | hom | BRCA2 | mis | 0,991667 | Breast cancer | 100 | 6 | + | + | + | | + | + |
| 13 | 51,413,355 | 1801249 | hom | ATP7B | mis | 0,483333 | Wilson disease | 84 | 8 | + | | | | | + |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 36,205,504 | 4904210 | hom | PAX9 | mis | 0,35 | Oligodontia | 10 | 0 | | | + | | | + |
| 14 | 87,470,966 | 421262 | hom | GALC | mis | 0,991667 | Krabbe disease | 88 | 3 | + | | | | | + |
| 15 | 43,179,367 | 269868 | hom | DUOX2 | mis | 0,966667 | Hypothyroidism | 65 | 6 | + | + | | | | + |
| 15 | 56,640,371 | 3829462 | hom | LIPC | mis | 0,983333 | Hepatic lipase deficiency | 49 | 2 | | | + | | + | + |
| 16 | 55,106,002 | 4784677 | hom | BBS2 | mis | 0,991667 | Bardet-Biedl syndrome | 77 | 22 | + | + | + | + | + | + |
| 16 | 55,462,088 | 1529927 | hom | SLC12A3 | mis | 0,925 | Gitelman syndrome | 44 | 0 | + | | + | + | + | + |

## 5.8 Risikovarianten für Morbus Crohn

Für die Analyse der detektierten Varianten der Crohn-Patientin wurden die 71 Risikoloci, die in der Meta-Analyse Franke et al.[70] beschrieben sind, abgeglichen und ermittelt, welche der Riskioallele die Patientin trägt. Die Ergebnisse sind in der Tabelle 5-9 dargestellt. Die in rot eingefärbten Genotypen geben an, dass die Patientin zwei Risikoallele trägt, orange gibt an, dass die Patientin ein Risikoallel trägt. Einige Genotypen waren in der Sequenzierung nicht eindeutig bestimmbar, diese sind mit einem Fragezeichen versehen. Die Tabelle zeigt, dass die Patientin für 48 der Risikoloci mindestens ein Risikoallel trägt. Für 14 von diesen ist die Patientin sogar homozygot.

**Tabelle 5-9.** Die Tabelle zeigt, für welche der in der Publikation Franke et al. (Nature Genetics 2010) 71 beschriebenen Risikoloci die Crohn Patientin Risikoallele trägt. Der Farbcode gibt an, ob die Patientin für den entsprechenden SNP ein, zwei oder kein Risikoallel trägt. (Farbcode: Rot=zwei Risikoallele; Orange= ein Risikoallel; ?= nicht eindeutig)

| SOLiD | dbSNP ID | Chr. | Risiko-allel | Allel Frequenz in Kontroll populion | OR (95% CI); *Loci mit >1 unabhängige Assoziation | Assoziation mit anderen Phänotypen | Kantidatengene, die von Interesse sind: fettgedruckt sind die, die durch zusätzliche *in silico* Analysen interessant sind |
|---|---|---|---|---|---|---|---|
| CT | rs2797685 | 1p36 | A | 0,190 | 1.05 (1.01-1.10) | Celiac | *VAMP3* |
| CC | rs3180018 | 1q22 | A | 0,250 | 1.13 (1.06-1.19)* | T2D, Asthma, PD | *SCAMP3, MUC1* |
| AG | rs1998598 | 1q31 | G | 0,302 | 1.04 (1.00-1.09) | Asthma | *DENND1B* |
| AG | rs3024505 | 1q32 | T | 0,157 | 1.12 (1.07-1.17) | T1D, UC, SLE, BD, Hep. C, | ***IL10, IL19*** |
| AG | rs13428812 | 2p23 | G | 0,326 | 1.06 (1.03-1.10) | | *DNMT3A* |
| CT | rs780093 | 2p23 | T | 0,418 | 1.15 (1.10-1.21) | CRP, Glucose, TGs | ***GCKR*** |
| CT | rs10495903 | 2p21 | T | 0,129 | 1.14 (1.09-1.20)* | T2D, PC | ***THADA*** |
| CT | rs10181042 | 2p16§ | T | 0,420 | 1.14 (1.09-1.19) | RA, UC, Celiac | ***C2orf74*** *REL* |
| AG? | rs2058660 | 2q12† | G | 0,231 | 1.19 (1.14-1.26) | Celiac, Asthma, T1D, HSV | ***IL18RAP, IL12RL2, IL18R1, IL1RL1*** |
| AA | rs6738825 | 2q33 | A | 0,473 | 1.06 (1.02-1.11) | CAD | ***PLCL1*** |
| CT | rs7423615 | 2q37 | T | 0,187 | 1.12 (1.07-1.18) | CLL | ***SP140*** |
| GG | rs13073817 | 3p24 | A | 0,322 | 1.08 (1.03-1.13) | | |
| AC | rs7702331 | 5q13 | A | 0,600 | 1.12 (1.07-1.17) | | |
| TT? | rs2549794 | 5q15 | C | 0,409 | 1.05 (1.02-1.09) | AS, PD, T1D, PET | ***ERAP2, LRAP*** |
| CC | rs11167764 | 5q31 | C | 0,796 | 1.06 (1.02-1.11) | | *NDFIP1* |
| TT | rs359457 | 5q35 | T | 0,571 | 1.08 (1.04-1.12) | | ***CPEB4*** |
| GT | rs17309827 | 6p25 | T | 0,639 | 1.10 (1.05-1.16) | | |
| CC | rs1847472 | 6q15 | G | 0,658 | 1.07 (1.03-1.11) | T1D, Celiac | ***BACH2*** |
| TT | rs212388 | 6q25 | G | 0,393 | 1.10 (1.05-1.14) | RA, Celiac, T1D↕ | *TAGAP* |
| TT | rs6651252 | 8q24 | T | 0,865 | 1.23 (1.17-1.30) | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CT | rs4077515 | 9q34† | T | 0,411 | 1.18 (1.13-1.22) | UC, AS | ***CARD9, SNAPC4*** |
| CT | rs12722489 | 10p15 | C | 0,852 | 1.11 (1.05-1.16) | MS, T1D, Vitiligo, RA, AA, Asthma, AITD | ***IL2RA*** |
| CC | rs1819658 | 10q21 | C | 0,774 | 1.19 (1.13-1.25) | AD | *UBE2D1* |
| CC | rs1250550 | 10q22‡ | G | 0,669 | 1.19 (1.15-1.23) | Celiac, MS, Vitiligo, BC | *ZMIZ1* |
| CT | rs102275 | 11q12 | C | 0,341 | 1.08 (1.04-1.12) | CAD; Dyslipidemia | ***FADS1*** |
| AA | rs694739 | 11q13 | A | 0,626 | 1.10 (1.05-1.16) | AA | *PRDX5, ESRRA* |
| GG | rs2062305 | 13q14 | G | 0,346 | 1.10 (1.05-1.15) | BD, RA | ***TNFSF11*** |
| AG | rs4902642 | 14q24 | G | 0,584 | 1.07 (1.11-1.04)* | Celiac, T1D | *ZFP36L1* |
| CT | rs8005161 | 14q35 | T | 0,119 | 1.23 (1.16-1.31)* | | ***GALC, GPR65*** |
| CC | rs17293632 | 15q22 | T | 0,233 | 1.12 (1.07-1.16) | CAD, T2D | *SMAD3* |
| TT | rs151181 | 16p11‡ | G | 0,386 | 1.07 (1.03-1.12) | T1D, obesity, Asthma, CRC, SLE, RA, IBD | ***IL27, SH2B1, EIF3C, LAT, CD19*** |
| AG | rs3091315 | 17q12§ | A | 0,723 | 1.20 (1.14-1.26) | HIV resistance | ***CCL2, CCL7*** |
| AA | rs12720356 | 19p13 | G | 0,084 | 1.12 (1.06-1.19)* | T1D, SLE, MS, HIES | ***TYK2, ICAM1, ICAM3*** |
| CT | rs736289 | 19q13‡ | T | 0,612 | 1.06 (1.02-1.11) | | |
| GG | rs281379 | 19q13‡ | A | 0,487 | 1.07 (1.04-1.11) | B12, Norovirus, HP | ***FUT2, RASIP1*** |
| AG | rs4809330 | 20q13 | G | 0,709 | 1.12 (1.06-1.18) | Glioma | ***RTEL1, TNFRSF6B, SLC2A4RG*** |
| GG | rs181359 | 22q11 | T | 0,203 | 1.10 (1.06-1.15) | RA, Celiac, SLE, MCV | ***YDJC*** |
| GG? | rs713875 | 22q12‡ | C | 0,471 | 1.08 (1.04-1.13) | IBD, T1D | *MTMR3* |
| CT | rs2413583 | 22q13 | C | 0,830 | 1.23 (1.17-1.29) | | *MAP3K7IP1* |
| GG | rs11209026 | 1p31 | G | 0,932 | 2.66 (2.36-3.00) | UC, AS, Ps, PBC, GC, BD | ***IL23R*** |
| GG? | rs2476601 | 1p13 | G | 0,907 | 1.26 (1.17-1.37) | T1D↕, RA, SLE, Ps, Vitiligo↕, AITD | ***PTPN22*** |
| AG | rs4656940 | 1q23 | A | 0,801 | 1.15 (1.09-1.21) | SLE, RA | ***CD244,*** *ITLN1* |
| CT | rs7517810 | 1q24 | T | 0,246 | 1.22 (1.16-1.28) | Hep.C, SLE, SSc, T2D | ***TNFSF18, TNFSF4, FASLG*** |
| AC | rs7554511 | 1q32 | C | 0,726 | 1.14 (1.08-1.19) | UC, celiac, MS | ***C1orf106,*** *KIF21B* |
| GG | rs3792109 | 2q37 | A | 0,529 | 1.34 (1.29-1.40) | UC | ***ATG16L1*** |
| AG? | rs3197999 | 3p21 | A | 0,297 | 1.22 (1.16-1.27) | UC | ***MST1, GPX1, BSN*** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CT | rs11742570 | 5p13 | C | 0,606 | 1.33 (1.27-1.39) | MS | ***PTGER4*** |
| GG | rs12521868 | 5q31 | T | 0,422 | 1.23 (1.18-1.28) | Ps, Fibrinogen, Asthma,TB, UC | ***SLC22A4, SLC22A5, IRF1, IL3*** |
| AG | rs7714584 | 5q33 | G | 0,088 | 1.37 (1.28-1.47) | TB | ***IRGM*** |
| GG | rs6556412 | 5q33 | A | 0,332 | 1.18 (1.13-1.24) | Ps, SLE, Malaria, Asthma | ***IL12B*** |
| CT | rs6908425 | 6p22 | C | 0,784 | 1.17 (1.11-1.23) | T2D, Ps, UC | *CDKAL1* |
| TT | rs1799964 | 6p21 | C | 0,209 | 1.19 (1.13-1.25) | Multiple including UC | ***LTA,*** *HLA-DQA2,* ***TNF, LST1, LTB*** |
| AG | rs6568421 | 6q21 | G | 0,301 | 1.13 (1.07-1.18)* | SLE, RA | ***PRDM1*** |
| CG | rs415890 | 6q27 | C | 0,522 | 1.17 (1.12-1.22) | RA, Graves | ***CCR6*** |
| CT | rs1456896 | 7p12 | T | 0,690 | 1.14 (1.09-1.20) | AD, SLE, MCV, ALL | ***IKZF1, ZPBP, FIGNL1*** |
| AA | rs4871611 | 8q24 | A | 0,609 | 1.17 (1.12-1.23) | | |
| AC | rs10758669 | 9p24 | C | 0,349 | 1.18 (1.13-1.23) | UC, Myelo. | *JAK2* |
| CT | rs3810936 | 9q32 | C | 0,682 | 1.21 (1.15-1.27) | UC, Leprosy, SpA | ***TNFSF15, TNFSF8*** |
| AA? | rs12242110 | 10p11 | G | 0,315 | 1.15 (1.10-1.20) | UC | ***CREM*** |
| AG | rs10761659 | 10q21 | G | 0,538 | 1.23 (1.18-1.29) | BC | *ZNF365* |
| GT | rs4409764 | 10q24 | T | 0,492 | 1.22 (1.17-1.27) | UC | *NKX2-3* |
| CT | rs7927997 | 11q13 | T | 0,389 | 1.17 (1.12-1.22) | Atopy↕ | *C11orf30* |
| GG | rs11564258 | 12q12 | A | 0,025 | 1.74 (1.55-1.95) | PD, Leprosy | ***MUC19****, LRRK2* |
| AG | rs3764147 | 13q14 | G | 0,245 | 1.17 (1.12-1.23) | Leprosy | ***C13orf31*** |
| AG | rs2076756 | 16q12 | G | 0,260 | 1.53 (1.46-1.60) | Leprosy, Atopy, Blau, GvHD | ***NOD2*** |
| AG | rs2872507 | 17q21 | A | 0,458 | 1.14 (1.09-1.19) | Asthma, UC, PBC, T1D, RA, WBC | ***GSMDL, ZPBP2, ORMDL3, IKZF3*** |
| AA | rs11871801 | 17q21 | A | 0,756 | 1.15 (1.10-1.21) | MS↕, obesity, HIES | ***MLX,*** *STAT3* |
| AA | rs1893217 | 18p11 | G | 0,153 | 1.25 (1.18-1.32) | T1D↕, celiac | *PTPN2* |
| AG | rs740495 | 19p13 | G | 0,247 | 1.16 (1.10-1.21) | | *GPX4, SBNO2* |
| CC | rs1736020 | 21q21 | C | 0,579 | 1.16 (1.11-1.21) | UC | |
| GG | rs2838519 | 21q22 | G | 0,391 | 1.18 (1.13-1.23) | Celiac, UC | ***ICOSLG*** |

## 5.9 *In silico* Vorhersage-Programme

Nicht-synonyme SNPs (nsSNPs) sind kodierende Varianten, die einen Effekt auf die Aminosäuresequenz in dem korrespondierenden Protein haben. Weil ncSNPs die Proteinfunktion beeinflussen können, glaubt man, dass sie den stärksten Einfluss auf die menschliche Gesundheit verglichen mit anderen SNPs in anderen Regionen des Genoms haben[125]. Deshalb ist es wichtig, diese nsSNPs, die einen (schädigenden) Einfluss auf die Proteinfunktion haben von denen, die funktionell neutral sind zu unterscheiden. Einen Anhaltspunkt geben Vorhersageprogramme, die auf Grund von strukturellen Eigenschaften, die eine Aminosäureänderung mit sich bringt oder basierend auf Sequenzhomologien Konservierungsgrade bestimmen, einen sog. Score errechnen. Diese Scores sind ein Maß für die potentielle Wirkung eines nsSNP. Die Abbildung 5-5 zeigt die errechneten Scores für das in in diesem Projekt sequenzierte Genom im Vergleich zu den „1000-Genomes-Project“ Daten und den SNPs aus der HGDM-Datenbank. Das *in silico* Vorhersageprogramm SIFT berechnet Scores zwischen 0 und 1; 0 bedeutet „schädigend“, 1 bedeutet neutral[126]. Diese Abbildung zeigt in drei Kategorien (alle SNPs („all“); alle mit potentieller schädlicher Wirkung („all damaging“), neue mit schädlicher Wirkung („novel damaging“)), dass das sequenzierte Crohn-Genom kein typisch krankes Genom ist. Deutlicher wird dies noch in der Abbildung, in der die Grantham-Scores vergleichend aufgetragen sind. Das Grantham-Vorhersageprogramm berechnet die Effekte von Aminosäureaustausch basierend auf chemischen Eigenschaften der resultierenden Proteine, einschließlich Polarität und Molekulargewicht[127]. An dieser Abbildung sieht man sehr schön, dass sich das Crohn-Genom genauso verhält wie die Kontrollgenome des „1000-Genomes-Project“, sie sind in den Kategorien „all“ und „all damaging“ fast auf einer Linie. Die HGMD-SNPs zeigen jedoch eine Verschiebung nach rechts und jenseits des eingezeichneten Grenzwertes (62) in der Kategorie „all“, was eine schädigende Wirkung der meisten in diese Berechnung eingeflossenen SNPs anzeigt.

Das gleiche zeigt das im Internet frei zugängliche Vorhersageprogramm PhyloP (http://hyperbrowser.uio.no/test/tool_runner?tool_id=hgv_add_scores), dass den Grad der Konservierung berechnet, auch hier sieht man eine Rechtsverschiebung der HGMD-SNPs in der Kategorie „all“, hier zeigt sich erneut, dass sich das sequenzierte Crohn-Genom sehr ähnlich wie Kontrollgenome verhält.

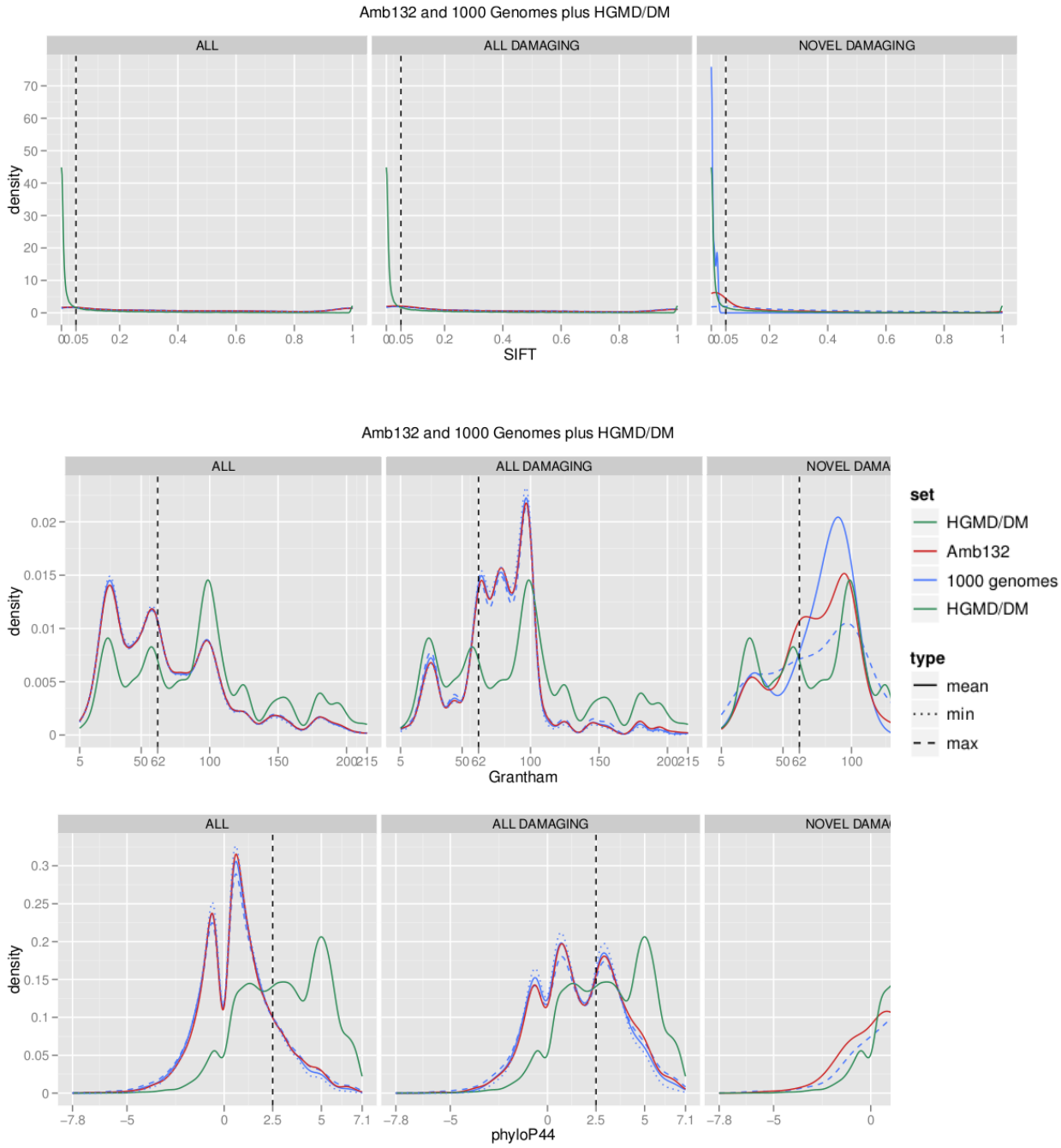



**Abbildung 5-5.** Die Abbildung zeigt die mit verschiedenen Vorhersageprogrammen errechneten Scores für das sequenzierte Crohn-Genom, die HGMD-SNPS und die SNPs aus dem „1000-Genomes-Project“. Das Crohn-Genom (rote Linie) verhält sich in den *in silico* Vorhersageprogrammen wie die Kontrollgenome (blaue Linie).

# 6 Diskussion zur Resequenzierung der GWAS-Loci

Der Fokus der vorliegenden Arbeit im ersten Teil ist weniger auf die Identifizierung der zur Aufklärung des erblichen Risikos für chronisch entzündliche Darmerkrankungen beitragenden genetischen Variationen gerichtet. Vielmehr standen die technischen und logistischen Herausforderungen eines groß angelegten Enrichment-Experiments in Vorbereitung und Verbindung mit einer Sequenzierung mittels „Next Generation Sequencing“ im Vordergrund. Da diese Arbeit keinen endgültigen Beitrag zur Aufklärung des genetischen Risikos für chronisch entzündliche Darmerkrankungen liefern konnte, bezieht sich auch die Diskussion deshalb vor allem auf die technischen und logistischen Aspekte der Studie und der damit einhergehenden Herausforderungen.

Diese Arbeit ist die erste größer angelegte Sequenzierungsstudie von aus GWAS identifizierten Suszeptibilitäts-Loci für verschiedene Phänotypen. Die DNA für diese Loci wurden in 50 Einzelindividuen und zwei HapMap-Trios für die Ziel-Regionen angereichert und anschließend sequenziert. Zwar wurde bereits eine größere Resequenzierungsstudie zu GWAS identifizierten CED-Loci veröffentlicht, aber in dieser Arbeit wurde sich auf die Sequenzierung der Exone beschränkt[88]. Außerdem basiert diese Arbeit auf einem Pooling-Ansatz, in dem die Allelfrequenzen nur geschätzt werden können. In der vorliegenden Arbeit wurden nicht nur die Exone der GWAS- identifizierten Gene für die selektierten Individuen sequenziert, sondern es wurden auch die intronischen Bereiche der entsprechenden Gene sowie intergenetische bzw. Promotorbereiche miteingeschlossen und das für 56 Einzelindividuen für jeden Locus. Für die Sequenzierung der Einzelindividuen (insgesamt 556 Individuen, je 50 für jeden Loci und zwei HapMap Trios) wurde nach dem Enrichment-Schritt ein komplexer Pooling-Ansatz entwickelt, der es unnötig machte, die verschiedenen Individuenproben mit Multiplex-Adaptoren zu versehen, um die Sequenzen nach der Sequenzierung für die Analysen den Individuen und Loci wieder zuordnen zu können.

Es ist offensichtlich, dass ein sicheres SNV-calling von einer angemessenen Coverage abhängt, da durch die vielen Amplifikationsschritte während des gesamten Arbeitsprozesses viele Fehler eingeschleppt werden können, so dass die Coverage und die Mappingergebnisse ebenso kritisch betrachtete werden müssen, wie auch die SNP-calling-Algorithmen. Deshalb werden diese Punkte als erstes und vor den eigentlichen inhaltlichen Ergebnissen diskutiert.

## 6.1 Konkordanz mit HapMap Trios

Um technische Kontrollen vorliegen zu haben, wurden die 2 HapMap-Trios, die LR-PCR basiert angereichert wurden, zusammen mit den eigentlichen Versuchsindividuen mit der SOLiD Technologie sequenziert. Der Vorteil bei der Verwendung von HapMap-Proben ist, dass Daten aus Genotypisierungen im Rahmen des internationalen HapMap-Projekt (http://hapmap.ncbi.nlm.nih.gov/) für diese Proben zur Verfügung stehen und diese auch für die Konkordanzberechnungen miteinbezogen werden konnten. Mit Hilfe von Konkordanzberechnungen, die mit den detektierten SNVs durchgeführt werden, lassen sich dann hinterher Aussagen über die Qualität und die weitere Verwendbarkeit der Sequenzierungsdaten machen.

Die Konkordanzberechnungen fallen in dieser Arbeit sehr unterschiedlich aus und weisen somit auf einige Probleme hin, die innerhalb dieses Projektes aufgetreten sind. So zeigen die Konkordanzraten vor und nach dem Filtern falsch-positiver SNV-Detektionssignale erhebliche Unterschiede. Die Konkordanzberechnungen für die HapMap-Proben der SOLiD-Sequenzierdaten vor dem Filtern potentieller falsch-positiver SNVs zeigen, verglichen mit den SOLiD- Datensätzen, die hinsichtlich falsch-positiver Signale gefiltert worden sind, eine deutliche Erhöhung der Konkordanzraten auf jeweils weit über 90%. Vor dem Filtern ergaben sich Konkordanzen zwischen 85% und 92%. Dies zeigt die Wichtigkeit der genaueren Inspektion der Daten. Ein Problem der gängigen SNP-Detektionsprogramme ist nämlich, dass diese Sequenzierungsfehler oder Information zur Coverage rund um die detektierte Variante oder allgemein nicht mitdokumentiert werden, so dass keinerlei Qualitätsaussage über die detektierte Variante gemacht werden kann. Allerdings sind die Konkordanzraten auch nach dem Filterprozess nicht optimal. Die fehlende Konkordanz auch nach dem Filterprozess lässt die Befürchtung zu, dass Kontamination und/oder Probenvertauschung stattgefunden hat/haben. Beides ist innerhalb des Enrichment-Prozesses wahrscheinlich, da bei der Komplexität und Verschachtelung innerhalb des Experiments menschliche Fehler nicht ausgeschlossen werden können. (Siehe Herausforderung Targeted Enrichment) .

Ein weiteres Problem, das durch die Konkordanzberechnungen sichtbar wird, sind die falsch negativen Signale, d. h. es wurden Variationen nicht detektiert, obwohl diese zu erwarten gewesen wären. Dies kann einerseits ein Problem fehlender Coverage oder ein „Detektionsfehler“ der Variationsdetektionssoftware sein. Die Ursachen der mangelnden Konkordanzen und Datenqualität wird in den nächsten Unterkapiteln diskutiert.

## 6.2 Herausforderung Targeted Enrichment

Mit Hilfe Genomweiter-Assoziationsstudien sind bereits Hunderte von genetischen Variationen, die mit komplexen Erkrankungen assoziiert sind, identifiziert worden. Da diese aber nur minimal das familiäre Clustern verschiedener Erkrankungen erklären und als potentielle Marker für die kausativen Variationen dienen, war es nun Ziel dieses Projektes, neue, seltene Varianten durch Sequenzierung der assoziierten Loci zu identifizieren. Mit Hilfe der „Next Generation Sequencing" -Technologie ist man nun in der Lage, große Bereiche des Genoms schnell und parallel zu sequenzieren. Die Sanger-Technologie eignet sich nur bedingt für solche Vorhaben, da die Probenbearbeitung und Datenproduktion ein zu teures und langwieriges Projekt ab einer gewissen Größe des zu sequenzierenden Bereiches darstellt. Allerdings muss man auch bei der Verwendung der „Next Generation Sequencing"-Methode die Bereiche, die von Interesse sind, aus dem Genom anreichern. Das stellte sich in diesem Projekt im Nachhinein als die größte Herausforderung dar. In dieser Arbeit wurde das Anreichern der DNA für die zu sequenzierenden Loci mit Hilfe von Long-Range-PCR realisiert. Diese Methode war im *NOD2*-Pilot- und Resequenzierungsprojekt erfolgreich[120] und ist auch allgemein etabliert[85]. Zum Startpunkt des Experiments waren kommerziell erhältliche Enrichment-Kits zudem teilweise noch in der Entwicklungs- bzw. Etablierungsphase und nicht kompatibel mit der für uns für dieses Experiments zur Verfügung stehenden SOLiD-Technologie. Das Enrichment auf LR-PCR zu basieren hat auch einige Vorteile. Zum einen ist es völlig unabhängig von einer Sequenzierungs-Technologie, d.h. der Anwender ist völlig frei, was die nachfolgende Sequenzierung angeht, da in keinem der Enrichment-Schritte Adaptoren für die anschließende Sequenzierung ligiert werden, die an eine bestimmte Sequenzierungs-Technologie gebunden sind. Außerdem ist es möglich, auch über gewisse *repeat*-reiche Sequenzen zu kommen, die bei kommerziellen Methoden, wie bsp. bei der Chip-basierten Methode der Firma NimbleGen, ausgespart werden. Hinzu kommt, dass diese Methode recht praktikabel in der Durchführung ist und in jedem Standardlabor ohne zusätzliche Anschaffungen durchführbar ist. Allerdings hat diese Methode bei allen theoretischen Vorteilen natürlich auch ihre praktischen Nachteile. Gerade die theoretische Simplizität der Durchführbarkeit der Methode stellte sich im Nachhinein und in der praktischen Umsetzung als Fehleinschätzung heraus; zumindest bei dem hier angedachten Versuchsaufbau, der sich für auf LR-PCR basierendes Enrichment als sehr komplex und logistisch sehr aufwendig herausstellte. Obwohl versucht worden ist, wenigstens einige der Schritte im Experiment zu automatisieren, um einerseits zeiteffektiver zu arbeiten, und zudem weitere Möglichkeiten für Probenvertauschungen kleinzuhalten, zeigte sich u.a. in

den Konkordanzraten, dass es höchstwahrscheinlich zu Vertauschungen gekommen ist. Für die weiteren Analysen muss die Annahme der Kontamination und vor allem der Probenvertauschung berücksichtigt werden. So ist es beispielsweise nicht mehr ohne weiteres möglich, die detektierten Varianten eindeutig auf eine DNA-Probe zurückzuführen. Selbst die Zuordnung der Varianten zu Fällen und Kontrollen kann nicht mehr mit Sicherheit gemacht werden. Bei welchem Schritt genau Vertauschungen stattgefunden haben, ist im Nachhinein nicht mehr eindeutig nachvollziehbar. Unzählige Möglichkeiten für Vertauschungen von DNAs vor und nach der PCR, LR-PCR Produkte vor und nach dem Aufreinigen, vor und nach dem Zusammenführen für die Sequenzierung sind theoretisch denkbar. Trotz dessen, dass versucht wurde, so viele Schritte wie möglich zu automatisieren, mussten entscheidende Dinge per Hand bearbeitet und pipettiert werden. Angefangen bei dem Raussuchen der DNA-Proben aus dem Probenlager, dass zum Zeitpunkt der experimentellen Durchführung durch Laborangestellte gemacht wurde, würde besser und sicherer in einem vollautomatisierten computergestützten Probenlager durchgeführt werden. Die unzähligen PCR-Reaktionen, bei denen die richtigen DNAs auf die korrekt beschrifteten Platten mit den richtigen Primern zusammenpipettiert werden mussten, stellen die nächsten großen Fehlerquellen dar. Dieses ließe sich bei unbegrenzten personellen Kapazitäten minimieren, indem man mit dem „vier Augen Prinzip“ als Qualitätsstandard arbeitet. Barcodierung von Proben und LIMS-gestütztes Arbeiten im Labor würden weitere innovative und qualitätssichernde Beiträge zu einem positiven Gelingen eines solchen Enrichtment- und Poolingansatzes liefern. Das gleiche gilt für die Umpipettierschritte der PCR-Produkte vor und nach dem Aufreinigen, vor und nach der Konzentrationsmessung und für das Herstellen der PCR-Produkte-Mixe. Um solche Fehlerquellen zukünftig auszuschließen und auch um zeiteffizienter zu arbeiten, würde man nach dem heutigen Stand der Forschung einen Versuch in dieser Größenordnung nicht mehr LR-PCR basiert durchführen, sondern auf kommerzielle Anbieter für Enrichment-Kits zurückgreifen, die in den letzten Jahren große Fortschritte gemacht haben. Mittlerweile sind einige Anbieter mit Produkten auf dem Markt, die sowohl zeit- und kostengünstiger als LR-PCR basierte Ansätze sind[128]. Die Gefahr für Probenvertauschungen können bei effizienten Kitlösungen auch minimiert werden.

Der ursprüngliche Ansatz für das Experiment, nach dem die DNAs auch nach gezielten Haplotypen ausgesucht worden sind um nachher selten detektierte Varianten mit den entsprechenden Haplotypen in Verbindung zu setzen, musste nach dieser Erkenntnis verworfen werden. Das Projekt wurde nach dem Erkennen dieses Problems der Vertauschung als reines Mutationsdetektionsprojekt angesehen, d. h. es wurde auf Grund der

angesprochenen Komplexizität nicht sehr viel Aufwand in die Aufklärung der angenommenen Vertauschungen investiert. Wären diese Gefahren zum Anfang dieses Experiments besser bedacht und realistischer eingeschätzt worden, wäre in diesem Versuch von Anfang an mit gepoolten DNAs gearbeitet worden, so wie es in Rivas et al. 2011[88] beschrieben ist. So hätte sich auch die Anzahl der Proben problemlos erhöhen lassen.

## 6.3 Mapping gegen die Ziel-Region und SNP-Detektion

Bei einem Resequenzierungsexperiment mit vorhergehendem Anreichern der Ziel-Region ist zu überlegen, wie die erhaltenen Sequenzen gemappt werden. Es gibt hier zwei Möglichkeiten. Beim sog. „whole genome approach“ mappt man die erhaltenen Sequenzen gegen das gesamte Genom, beim „target region approach“ gegen die definierte Ziel-Region. In dem vorliegenden Experiment wurden die erhaltenen Sequenzen zunächst gegen das gesamte Genom gemappt und anschließend gegen die vordefinierte Ziel-Region. Beide Ansätze haben Vor- und Nachteile. Zwar ist die Gefahr falsch-positiver Signale beim Mapping gegen das gesamte Genom geringer als beim Mappen gegen die Ziel-Region, da die Sequenzen nicht „gezwungen“ werden, innerhalb der definierten Ziel-Region zu mappen, dafür ist allerdings die Gefahr falsch-negativer Signale beim anschließenden SNP-calling höher, da viele Sequenzen mehrfach im Genom landen könnten, die dann verworfen werden, weil unsicher ist, ob sie wirklich von diesem Ort kommen. Allerdings schwindet gleichzeitig die Gefahr falsch-positiver Signale in der Detektion von SNP[129]. Beim Sichten der Ergebnisse vom Mappen gegen das gesamte Genom wurde festgestellt, dass erhebliche Lücken in der eigentlichen Ziel-Region, die nicht zu erwarten waren, weil für diese Regionen PCR-Produkte existierten, die in die Sequenzierung gegangen sind, zu verzeichnen waren. So wurde für das weitere Vorgehen für das Mappen gegen die Ziel-Region entschieden. Gerade bei LR-PCR basierten Ansätzen weiß man, woher die Sequenzen kommen, weil vorher die PCR-Produkte generiert wurden und über Gelelekrophorese überprüft wurden. Auf diese Weise wurde es möglich, Coverages für die Ziel-Region in das für das SNP-calling sehr gute Bereiche zu bekommen. Außerdem ist es bei einem Sequenzierungsexperiment, was darauf ausgelegt ist, unbekannte Varianten zu identifizieren, wichtig, möglichst alle Varianten zu detektieren. Die Gefahr falsch-positiver Signale ist dabei geringer einzuschätzen als die Gefahr echte SNPs zu übersehen. Allerdings ist eine genaue Inspektion der Detektionssignale erforderlich, um echte SNVs von Sequenzierungsfehlern bzw. Mappingfehlern zu unterscheiden.

## 6.4 Coverage

Das gesamte Experiment betrachtend sind sehr zufriedenstellende Coverages für alle Proben und die Ziel-Regionen erzielt worden, was bedeutet, das eine weitere Erhöhung der Coverage die Qualität des SNP-callings in den meisten Fällen nicht verbessert hätte. Eine Coverage-Simulation in einem vorherigen Experiment zeigte, dass die durchschnittliche Coverage bei LR-PCR basierten Enrichmentexperimenten nicht unter 40x liegen sollte[120]. Die ungleiche Verteilung der gemappten Reads impliziert, dass bestimmte Bereiche in der Ziel-Region übersequenziert sind. Durch genaues equimolares Mixen der einzelnen Fragmente kann man dieses Problem nicht gänzlich umgehen. Besonders die Enden der LR-PCR Produkte lassen regelrechte Peaks in der Coverage entstehen und treiben die durchschnittlichen Werte für die Coverage der Sequenzierung hoch. Diese Peaks sind methodisch bedingt und resultieren aus dem Schritt der Fragmentierung aus der Library-Präperation. In diesem Schritt werden die LR-PCR-Produkte in kleinere Fragmente auf eine Länge zwischen 100 und 150 bp geschert. Wo genau die Fragmente geschert werden geschieht völlig zufällig, die Enden von den generierten PCR-Produkten bleiben aber bei vielen Scherprodukten als Anfang oder Ende erhalten, so dass diese später in der Sequenzierung überrepräsentiert sind. Man könnte dieses Problem umgehen, indem man blockierende Primer einsetzt[130].

## 6.5 Sequenzierstrategie und Multiplexansatz

Es wurde für diese Arbeit eine Sequenzierstrategie und ein Multiplexansatz erarbeitet, der es möglich machte, mehrere Individuen gleichzeitig und parallel auf einem Spot des Sequenzierungs-Slides zu sequenzieren. Dieser Ansatz ließ zu, dass man in der Analyse die Sequenzen den ursprünglichen Individuen und Loci zuordnen konnte. Dieser Multiplexansatz wurde aus verschiedenen Gründen so gewählt: Es sollte die volle Kapazität des Sequenzierungs-Slides pro Lauf ausgeschöpft werden, d.h. finanzielle Ressourcen sollten geschont bleiben. Dass auf diese Weise alle 56 Individuen für alle Loci in nur 7 Sequenzierungsläufen sequenziert werden konnten, hat aber auch zeitlich große Vorteile. Ein Sequenzierungslauf benötigte zu der Zeit 14 Tage, wäre dieser Multiplexansatz nicht gewählt worden, hätten die 10 Loci in je 56 Individuen 70 (560 Individuen, 8/Lauf) Sequenzierungsläufe in Anspruch genommen. Die reine Sequenzierungszeit hätte sich somit verzehnfacht und annährend ein ganzes Jahr in Anspruch genommen. So war dieser Versuch mit diesem Multiplexansatz auch ein Pilotprojekt und eine Art „Versuchsballon“ für noch größere Sequenzierungsstudien, in denen viele einzelne Individuen für kleinere Ziel-Regionen

sequenziert werden könnten, zu diagnostischen Zwecken etc. beispielsweise. Die Problematik der Probenvertauschung beim Enrichment-Vorgang wurde bereits im Unterkapitel 6.2 diskutiert. Bei kleineren Regionen, die angereichert werden müssten, würde sich die Komplexität allerdings verringern. Die Multiplexstrategie im Sequenzieren birgt aber nicht nur die Gefahr der Probenvertauschung. Die Daten den einzelnen Individuen nach dem Sequenzierungsvorgang zuzuordnen ist ein weiteres Problem, dass sich als sehr zeitaufwendig und kompliziert in der praktischen Umsetzung gestaltete, so dass sich zumindest der zeitliche Vorteil, der sich durch das parallele Sequenzieren auf einem Sequenzierungsspot ergab, schnell auflöste. Da die Originalsequenzdaten von 10 Individuen immer in einem Verzeichnis bleiben, mussten viele bioinformatorische Schritte eingebaut werden, um einzelne Informationen zu den einzelnen Individuen zu erhalten. Das betraf vor allem das SNP-calling aber z.B. auch die Berechnungen der Coverages. Das verkomplizierte sämtliches Arbeiten mit den Sequenzdaten. Deshalb ist zu überlegen, ob dieser Multiplex-Ansatz sinnvoll ist; in dieser Form und in diesem Ausmaß würde man einen Versuch dieser Art nach den gemachten Erfahrungen wohl nicht mehr gestalten.

## 6.6 SNP-Validierung und Assoziationsexperiment

Als Folge des technischen Fortschritts wird großes Bemühen unternommen um die genetischen Variationen als Basis für viele häufige Krankheiten zu entschlüsseln[131,132]. Wenig überraschend sind die meisten Variationen in nicht-kodierenden Regionen beobachtet worden, wo sie regulative Interaktionen beeinflussen könnten. Deren funktionelle Konsequenzen sind natürlich viel schwieriger vorherzusagen und zu validieren, weil der sog. regulatorische Code sehr viel komplexer und flexibler ist als der genetische Code[113]. Ein Großteil der in diesem Projekt detektierten Varianten ist ebenfalls in nicht-kodierenden Bereichen lokalisiert. Die Herausforderung aus den vielen Sequenzinformationen und vielen detektierten Varianten die kausalen Varianten herauszulösen, gewinnt an Wichtigkeit- die größte Hürde ist nicht mehr wie einst die Datenerzeugung, sondern die sinnvolle Interpretation dieser Sequenzinformation[133]. Um die Varianten erst einmal zu priorisieren und als Anhaltspunkt für weiterführende Analysen, wurden alle detektierten Varianten wie im Material und Methoden Teil (Kapitel 2.11.1) beschrieben einer Analyse mit sTRAP zur Quantifizierung der regulativen Interaktion unterzogen.

Ziel dieses Experimentes war es, die GWAS-Loci durch das Sequenzierungsexperiment für die einzelnen Erkrankungen sehr fein aufzulösen und sogar evtl. neue unabhängige seltenere

Mutationen zu entdecken, die gemäß der „Common Disease/Rare Variant Hypothese“ in den sich anschließenden Fall/Kontroll-Studien assoziieren, die durch kommerziell erhältliche SNP-Chips nicht abgedeckt sind. Im ersten Validierungsschritt wurden 88 SNVs über die Sequenom-Technologie getestet. Die Auswahl der Variationen geschah basierend auf Transkriptionsfaktorbindeanalyse und den zusammengefassten Ergebnissen aus dem SNP-Kategorisierungstool SnpActs. Viele SNPs bzw. SNVs waren in der Genotypisierung monomorph, d.h. das alternative Allel konnte in keinem der Individuen der Analysepopulation detektiert werden. Da es sein kann, dass die getesteten Variationen seltene bzw. sehr seltene Varianten sind, heißt das nicht, dass sie in einer größeren/anderen Analysepopulation nicht zu finden wären. In diesem Experiment traten sie außer in der Sequenzierung selbst nicht noch einmal in Erscheinung. Vielleicht wäre es hilfreich gewesen, das Individuum, in dem die Variation in der Sequenzierung gefunden worden war, mit in die Analysepopulation als „Kontrollindividuum“ mithineinzunehmen. Allerdings wäre man im aktuellen Experiment an mehrere Probleme gestoßen. Zum einen wäre es ein Problem gewesen, das Individuum auf Grund der im vorangegangenen Kapitel Probenvertauschungsproblematik auszumachen, in dem die Variation tatsächlich zu finden wäre, zum anderen wäre der logistische Aufwand, die verschiedenen DNA Proben auf die die Sequenom Platten zu bekommen, größer gewesen. Das gleiche gilt für den zweiten Validierungsschritt mit 27 SNVs, auch hier zeigten sich einige monomorph.

Allerdings zeigte eine bisher unbekannte kodierende Variante im Exon von *NOD2*, ein Gen, das hinlänglich als Suszeptibilitätsgen für Morbus Crohn bekannt ist, eine Frequenz in der Genotypisierung in den Fällen von 0,0019. In den Kontrollen tauchte diese Variante in den 2500 Kontrollen überhaupt nicht auf. Diese exonische Variante wurde auch in einer größeren Untersuchung der kodierenden Regionen des Gens *NOD2*, eine Studie, in der 612 an Morbus Crohn Erkrankte und 112 Kontrollindividuen resequenziert wurden, nicht detektiert (Lesage et al., 2002)[134]. Dieser Befund demonstriert, dass das Ultratief-Resequenzieren von sogar sehr gut untersuchten Krankheitsgenen in sorgfältig ausgesuchten Individuen sehr seltene Varianten oder sogar sog. “private mutations” mit funktionellem Potential, die nicht in den GWAS abgedeckt sind, hervorbringen kann[135]. Diese Variante liegt wie die Crohn-assoziierten *NOD2*-Varianten in der Leucinrich- repeat (LRR)- Domäne[22] und spricht auch aus diesem Grund für eine mögliche Verbindung zu Morbus Crohn. Diese sehr seltene Variante in den Fällen muss nun genauer untersucht werden. Der erste Schritt wäre, diese in einer unabhängigen noch größeren Studienpopulation erneut zu testen. Experimentelle

funktionelle Studien könnten sich anschließen, um zu testen, ob diese Variante einen Einfluss z.B. auf die Expression hat.

Da die bisherigen genetischen Befunde zur Heritabilität von Morbus Crohn nicht einmal ein Viertel des erblichen Risikos erklären, ist ein Ansatz zur weiteren Aufklärung, dass neue Varianten in Genen, die als mit der Krankheit assoziiert identifiziert wurden, gefunden werden. Die andere Möglichkeit, die für dieses Experiment aber nicht in Frage kommt, ist, nach Varianten zu suchen, die in Genen, die noch nicht mit der Krankheit assoziiert sind aber die über Stoffwechselwege an die Krankheit gekoppelt sein könnten, liegen. Damit wäre auch die Limitation von Resequenzierungsstudien von identifizierten assoziierten Loci gegeben. Denn um neue Gene mit komplexen Krankheiten zu assoziieren ist man auf regionale Vorbefunde nicht mehr angewiesen. Die Zukunft wird wohl mit sinkenden Sequenzierungskosten in der Gesamt-Genom-Sequenzierung liegen.

# 7 Diskussion Gesamt-Genom-Sequenzierung

Ziel des Projektes war es, das Genom einer Crohn-Patientin aufzulösen und einen genauen Eindruck von einem Genom einer schwer an Morbus Crohn erkrankten Person zu bekommen. Das Ziel war also eher deskriptiver Natur als das Auffinden kausaler Ursachen für die komplexe Erkrankung Morbus Crohn. Für das Institut war dieses Projekt ein Pilotprojekt, ein sog. „proof-of-principle“, das auch viel Potential zum Lernen im Umgang mit Daten kommend aus einer Gesamt-Genom-Sequenzierung besitzt.

## 7.1 Datenerzeugung und Detektion von genetischen Variationen

Für dieses Projekt wurden acht Sequenzierlibraries von einer an Morbus Crohn erkrankten Patientin hergestellt, die in insgesamt 18 Sequenzierungsläufen sequenziert wurden, um eine genomweit hohe Coverage zu erhalten. Unter den acht Libraries waren sechs Mate-Paired-Libraries mit unterschiedlichen Insertgrößen (zwischen 0,4 und 6 kb, siehe Tabelle 5-1) und zwei Fragment-Libraries. Die Herstellung und Sequenzierung verschiedener Typen von Libraries (Long-Mate-Paired, Fragment) ist gegenüber der Sequenzierung der immer wieder gleichen Library vorteilig, da jeder Library-Typ Schwächen und Stärken hat. So ist die Sequenzierung der Long-Mate-Paired-Libraries sehr gut dazu geeignet, große strukturelle Varianten zu detektieren. Der Nachteil in der Präparation von Long-Mate-Paired-Libraries ist allerdings die immens große Menge an genomischer DNA, die für die Herstellung benötigt wird, während für die Konstruktion von Fragment-Libraries wenig Eingangs-DNA benötigt wird.

Das Crohn-Genom konnte beinahe vollständig mit einer zufriedenstellenden Coverage mittels SOLiD-Technologie sequenziert werden. Das SNP-calling zeigte sich recht robust, dies zeigten die Konkordanzberechnungen mit den Validierungsdaten der Sanger-Sequenzierung und der Exom-Sequenzierungsdaten. Auch die für Säugetiere gut dokumentierte und typische Transition/Transversion-Ratio von 2:1[136,137] , die für alle detektierten SNPs errechnet wurde, spricht für einen guten Gesamteindruck der SNP-Detektionsergebnisse. Allerdings zeigte das SNP-calling Schwächen bei der Detektion bisher unbekannter Varianten. Es empfiehlt sich daher, potentielle Varianten, die für die Weiterverfolgung in große Validierungs- oder Assoziationsexperimenten einfließen sollen, erst mit Hilfe einer standardisierten Methode zu überprüfen oder den SNP-Kontext in den Originaldaten zu betrachten. Das ist einerseits möglich durch Visualisierungsprogramme wie IGV (http://www.broadinstitute.org/igv/) oder

andererseits durch spezielle Programme, z.B. pipase[138], die eine Großzahl einzelner Positionen abfragen können und den Kontext der Varianten dokumentieren und somit eine Qualitätsaussage möglich machen.

## 7.2 Charakterisierung des Crohn-Genoms

Der Vergleich der Genomsequenzierung dieses Projektes mit verschiedenen Referenzdatensätzen zeigt, dass dieses Genom kein auf den ersten Blick „krankes" Genom ist. In den groben Charakteristika unterscheidet sich das Genom dieser kranken Person nicht von dem eines gesunden Kontrollindividuums. So verhält sich dieses Genom in der Analyse mit *in silico* Vorhersagetools wie Kontrolldatensätze, was die potentiell schädigende Auswirkung von genetischen Varianten angeht. Auch SNPs, die ausschließlich in diesem Genom detektiert wurden, waren prozentual nicht mehr als die individuell detektierten SNPs für andere publizierte Genome. Dies zeigte der Vergleich mit den Genomen von Venter, Quake und Watson (Ergebnisteil Abbildung 5-3).

Viele verschiedene assoziierte Risikovarianten für Morbus Crohn sind aus anderen Studien bekannt und eine mögliche Hypothese wäre, dass die sequenzierte Crohn-Patientin besonders viele dieser Risikovarianten trägt, die zusammengenommen ein erhöhtes Risiko und/oder die vergleichsweise schwere Verlaufsform erklären würden. Das kumulative Risiko dieser Frau müsste also deutlich höher sein als das der Allgemeinbevölkerung. In dieser Arbeit sind die 71 Risikoloci aus der Meta-Analyse Franke et al.[70] mit den detektierten Varianten dieser Patientin abgeglichen wurden. Für 48 der 71 überprüften Risikoloci trägt die Patientin mindestens eins der Risikoallele. Außerdem zeigt die Patientin beispielsweise die 20 kb große Deletion (homozygot) vor dem Gen *IRGM*[139], die mit Morbus Crohn assoziiert ist. Nur 5% der Bevölkerung tragen diese Risikovariante. Der Anteil der Bevölkerung, die homozygot für diese Variante sind, beläuft sich auf nur 0,25%, was sehr selten ist. Allerdings erklärt dieses weder die Erkrankung selbst noch den Verlauf. Trotz der Risikovarianten für Morbus Crohn im Genom der Patientin zeigt das kombinierte Risiko keine signifikanten Unterschiede zu dem Risiko der gesunden Kontrollindividuen. Das ist auch an sich keine überraschende Feststellung, da die bekannten Risikoloci für Morbus Crohn zusammengenommen nur ungefähr 20% des erblichen Risikos erklären. Deshalb ist es wichtig, in den Genen, die mit Morbus Crohn assoziiert sind, zusätzlich nach neuen Varianten zu suchen. Oder aber auch neue Gene zu identifizieren, die mit dem Phänotyp assoziiert sind.

Die SNPs, die durch das intelligente Filtern aus der Gesamtheit der erhaltenen SNPs bzw. SNVs nach einer erst einmal relativ naiven Vorgehensweise nach unbekannten, potentiell schädigenden SNPs herausgelöst werden, sind als erstes interessant für weitere Analysen. Dafür werden sog. Vorhersageprogramme benutzt, um die erhaltenen Variationen zu priorisieren. Diese *in silico* Vorhersageprogramme sind mit SNPs aus den Datenbanken HGMD[140] und OMIM[141] getestet und „trainiert" worden, die in Krankheitsgenen identifiziert worden sind. Die meisten SNPs in HGMD und OMIM betreffen mendelnde Krankheiten. 70-90% der in den Datenbanken HGMD und OMIM katalogisierten SNPs sind laut Aminsäureaustausch-Vorhersage schädigend, nur 10-20% werden dagegen in neutralen Datensätzen als schädigend vorhergesagt. Das zeigt, dass diese Vorhersageprogramme für mendelnde Krankheiten zwischen SNPs, die einen Aminosäureaustausch bewirken, der eine neutrale oder schädigende Wirkung auf die Proteinfunktion hat, unterscheiden können[125]. Für komplexe Erkrankungen wie Diabetes, Bluthochdruck oder Morbus Crohn sind diese Vorhersageprogramme allerdings noch in der Phase des Austestens. Die genetische Grundlage komplexer Erkrankungen kann nicht auf einen einzigen Locus zurückgeführt werden; die Interaktion mehrerer krankheitsassoziierter Loci und/oder das Zusammenspiel von genetischer Disposition und Umwelt charakterisieren die Natur komplexer Erkrankungen. Die neuen technologischen Möglichkeiten auf dem Sektor der Sequenzierung erlauben eine schnelle und genaue Auflösung ganzer Genome, es müssen aber auch Möglichkeiten gefunden werden, diese Datenmengen nach neutraler und krankheitsrelevanter Information zu filtern. Deshalb werden diese Vorhersageprogramme mit Zunahme der sequenzierten Genome und der daraus resultierender Erfahrungen in näherer Zukunft eine immer größere Bedeutung bekommen. Das gilt nicht nur für Varianten, die einen Aminosäureaustausch bewirken, sondern vor allem auch für Varianten in nicht kodierenden Bereichen, die die Genfunktion regulieren oder splicing-Varianten, da sie auch einen Großteil von neuen bisher unbekannten Varianten ausmachen.

Die Assoziation von genetischen Variationen mit Krankheiten und mit dem Ansprechen oder Nicht-Ansprechen auf bestimmte Medikamente sowie Verbesserungen in den Sequenzierungstechnologien, vor allem das Sequenzieren kompletter Genome, haben großen Optimismus geweckt für die sog. „genomic medicine"[142]. Einige Erkenntnisse aus Untersuchungen am humanen Genom lassen sich auch schon konkret nutzen, so werden Patienten vor der Gabe des immunsupressiven Medikaments Azathioprin auf bestimmte Mutationen untersucht, die mit der Enzymtätigkeit der Thiopurin-S-Methyltransferase

korreliert sind, um die Patienten so vor schwerwiegenden Komplikationen und Nebenwirkungen zu schützen[143,144]. Um allerdings das volle Potential der Genom-Sequenzierung für die menschliche Gesundheit auszuschöpfen, müssen noch einige Limitationen überwunden werden.

## 7.3 Risikovarianten, die mit Morbus Crohn assoziiert sind und potentielle Risikovarianten

Neue u.U. vielversprechende Varianten fanden sich in den Genen *MUC2* und *MUC6.* Diese Gene kodieren für die sog. Mucine. Mucine sind der strukurgebende Teil des Schleims im Organismus und agieren im Darm als protektive Barriere zwischen Mucosa und Darmlumen. Da die Gene, die für die Mucine kodieren zu den hoch polymorphen Genen gehören, ist die Auflösung der genetischen Variationen für diese Gene technisch und bioinformatorisch sehr schwierig und mit einem großen Potential für falsch-positive Signale behaftet. Aus diesem Grund werden für die Detektion von möglichen kausativen Varianten in Sequenzierungsprojekten solche Gene durch Filterschritte gewöhnlich erst einmal für weitere Analysen ausgenommen[145]. Für die vorliegende Arbeit wurden die genetischen Varianten in *Muc2* und *Muc6* aus diesen Gründen auch nicht näher analysiert. Da der Mucusschicht aber in der Integrität der Barrierefunktion des Darmes eine hohe Bedeutung zukommt, werden diese Varianten an dieser Stelle kurz diskutiert.

Mucine sind Glykoproteine, mit einer zentralen Proteinkette und langen Seitenketten aus Polysacchariden, die durch die Becherzellen sezerniert werden. Der Proteinkern enthält mehrere sog. „tandem repeat"- Domänen, variable Areale, an die die Polysaccharide kovalent gebunden sind. Der Grad der Glykolysierung hat eine zentrale Rolle in der Barrierefunktion des Mucus[146]. In Fällen von chronisch entzündlichen Darmerkrankungen werden schon seit längerer Zeit ein Zusammenhang zwischen Veränderungen in der Zusammensetzung des Mucus und der Erkrankung diskutiert. So sind Veränderungen in der Länge der Polysaccharidketten im Kontext mit chronisch entzündlichen Darmerkrankungen beschrieben[147]. Auch andere biochemische Veränderungen wie unterschiedliche Sulfatierung rufen eine Änderung in der Viskosität der Mucusschicht hervor, was mit einer gestörten Barrierefunktion einhergeht, d.h. die Barriere wird durchlässiger für Bakterien, die sich sich leichter an das Epithelium haften können und so das empfindliche Gleichgewicht im Darm negativ beeinflussen. In Experimenten an *Muc2*-defizienten Mäusen konnte das spontane Auftreten von Colitis nachgewiesen werden[148]. Veränderungen der Mucinproduktion und der

Glykolysierung sind also in Zusammenhang mit chronisch entzündlichen Darmerkrankungen wahrscheinlich, ob sie aber die Ursache für die Entstehung der Erkrankung sind oder aus den entzündlichen Prozessen resultieren, bleibt bisher unklar. Dass genetische Varianten in den für Mucine kodierenden Genen, die die Expression der Mucine und deren Glykolysierung beeinflussen, einen Einfluss auf Morbus Crohn haben, ist zumindest denkbar. Allerdings werden weitere Untersuchungen nötig sein, um die funktionelle Bedeutung der biochemischen Veränderungen der verschiedenen Komponenten der Mucusschicht und deren komplexe Interaktion mit Mikrobiota, Epithelzellen sowie dem angeborenen und adaptiven Immunsystem aufzuklären[149].

Eine große Menge von Varianten sind in der Morbus Crohn Patientin identifiziert wurden, auch solche, die als Risikovarianten für Morbus Crohn bekannt sind. Außerdem viele Varianten, die eine potentiell schädigende Wirkung haben. Es sind zusätzlich viele komplett unbekannte Varianten identifiziert worden. Hinzu kommen identifizierte strukturelle Varianten, die teilweise Gene betreffen, die in die Immunantwort eingebunden sind. Trotz dieser enormen Information zum genetischen Hintergrund der Patientin, kann diese Arbeit nur einen minimalen Beitrag zur Aufklärung der Ursachen des Morbus Crohn leisten. Morbus Crohn ist und bleibt eine komplexe Krankheit, bei der sich der Einfluss der Umgebungsfaktoren sehr schwer abschätzen lässt.

## 7.4 Strukturelle Varianten

Die Detektion und Analyse von strukturellen Varianten im Rahmen einer Gesamt-Genom-Sequenzierung stellt eine weitere große Herausforderung dar. Im Rahmen dieses Projektes wurden erste Analysen zur Identifizierung von CNVs (eng. Copy number variations) in Kooperation mit der Arbeitsgruppe um Jan Korbel durchgeführt. Die ersten „groben“ Ergebnisse mussten mühevoll „per Hand“ nachbearbeitet werden. Einige Deletionen (unbekannte und bekannte) konnten identifiziert werden, die direkt Gene betreffen, die auch im Zusammenhang mit Morbus Crohn stehen bzw. stehen könnten. Die 20 kb große Deletion vor dem Gen *IRGM*, die mit Morbus Crohn assoziiert ist, für die die Patientin homozygot ist, ist einer der Befunde, der auch mittels Nimblegen Chip-Technologie validiert werden konnte (Daten sind nicht gezeigt). Eine Deletion (heterozygot) auf Chromosom 1, die eine Größe von beinahe 350 kb aufweist, betrifft die Gene *FCGR1C* und *FCGR1B*. Die FC- Gamma-Rezeptor Familie besteht aus drei engverwandten Genen, die mit A, B, und C bezeichnet werden. Diese drei verschiedenen aber sehr ähnlichen Gene kodieren für sechs mRNA Transkripte, (durch

alternatives Splicen). Die FC-Gamma-RIB Rezeptor ist auf der Oberfläche von B-Lymphozyten und Macrophagen zu finden und spielt daher eine wichtige Rolle in der Immunantwort[150,151] und könnte aus diesem Grund auch mit der Krankheit Morbus Crohn assoziiert sein.

Eine homozygote über 30 kb große Deletion, die Gene *LCE3B* und *LCE3C* betreffend, könnte eventuell ebenfalls in Zusammenhang mit Morbus Crohn gebracht werden. Zumindestens ist diese Deletion mit chronisch entzündlichen Erkrankungen der Haut (Psoriasis) assoziiert[152].

## 7.5 Nachweis von *Mycobacterium avium* und Suszeptibilität

Ein sehr interessanter, als mitverursachender Faktor des Morbus Crohn aber wissenschaftlich sehr kontrovers diskutierter Befund sind die im Blut und Darmbiopsien (durchgeführt am Referenzzentrum Borstel) der Patientin nachgewiesenen Mycobakterien (*Mycobacterium avium* subsp. *paratuberculosis*; MAP). Der wissenschaftliche Diskurs um die Ätiologie des Morbus Crohn und die Diskussion um die Hypothese mit Morbus Crohn als Infektionskrankheit bleiben bis heute kontrovers. Die dominante Hypothese zur Ätiologie von Morbus Crohn besagt, dass die Krankheit auf einer Dysregulation des Immunsystems in der Mucosa in genetisch prädisponierten Individuen beruht, die zu einer übertriebenen und weiterführenden Aktivierung der Immunantwort gegen die eigene normale Mikroflora führt[153]. Und obwohl genetische Befunde existieren und die genetische Disposition eine Rolle zu spielen scheint, erklären sie nicht die weltweite rapide Zunahme der Inzidenz von Morbus Crohn[154]. Epidemiologische Studien haben substantielle Überlappungen von Morbus Crohn in Regionen gezeigt, die hohe Level von MAP in der Umgebung aufweisen, ältere Studien aus Japan fanden eine statistische Korrelation zwischen erhöhter Inzidenz von Morbus Crohn und erhöhtem Konsum von tierischen Proteinen, insbesondere Milch, eine bekannte Quelle von MAP[155]. Ein weiterer signifikanter, zwar seltener Befund ist die isolierte Manifestation eines Morbus Crohn im Duodenum[156], der gegen die These einer übersteuerten Immunantwort gegen die eigene Darmflora spricht, da dieser Bereich frei von Darmbakterien ist.

Der Nachweis von Mycobakterien in Morbus Crohn Patienten scheinen zumindest einen Zusammenhang von Mycobakterien und der Krankheit zu untermauern. So konnten Bull et al. in einer Studie die Bakterien-DNA in bis zu 92% der Morbus Crohn nachweisen, in den Kontrollen fand sich die DNA nur bei 26%[157]. Eine andere Untersuchung konnte zeigen, dass in frisch resektomierten Darmabschnitten von Morbus Crohn Patienten 52% Mycobakterien

aufwiesen, in den Darmgeweben von nicht an Morbus Crohn Erkrankten wurden nur in 2% die Mycobakterien gefunden[158].

Allerdings bleibt eben weiterhin die genetische Komponente der Krankheit zu berücksichtigen. Immerhin sind 50% des Krankheitsrisikos auf genetische Faktoren zurückzuführen[159]. Verschiedene genetischen Faktoren, die für Morbus Crohn prädisponieren, mit hochsignifikanten und replizierten Assoziationen einschließlich von Genen, die für intrazelluläre Rezeptoren kodieren, die Bestandteile von Bakterienzellwänden erkennen, *(NOD2/CARD15)* und für die Beseitigung von Bakterien über Autophagie *(ATG16L1* und *IRGM)* verantwortlich sind, sind hinlänglich beschrieben worden. Eine theoretische Verbindung zwischen den Risikogenen NOD2, ATG16L1 und IRGM könnte sein, dass Morbus Crohn erstmalig durch eine nicht intakte Immunantwort auf eine hartnäckige Infektion durch intrazelluläre Pathogene wie Mycobacterium avium subsp. Paratuberculosis[160] auftritt. Interessanterweise konnte eine Studie in jüngerer Vergangenheit zeigen dass Crohn-Patienten, die die *NOD2*-Mutationen tragen, eine nicht-effiziente Erkennung von MAP aufweisen[161]. Diese Resultate zeigen, dass *NOD2* in der Erkennung von *M. paratuberculosis* durch das angeborene Immunsystem miteingebunden ist und die Bakterien somit in die Pathogenese von Morbus Crohn involviert sind[162]. Die Crohn-Patientin in diesem Experiment weist neben den prädisponierenden *NOD2*-Mutationen, die, wie oben beschrieben, 20 kb große Deletion homozygot vor dem *IRGM*-Gen auf, was ebenfalls für eine Verbindung zwischen ihrer persönlichen genetischen Veranlagung und ihrer Infektion mit MAP spricht. Allerdings ist die Frage ungeklärt, ob die Infektion ursächlich für Morbus Crohn bei der Patientin ist oder die Krankheit die Ursache für die Infektion mit MAP.

Einen weiteren Beitrag zur Aufklärung des genetischen Risikos für komplexe Erkrankungen könnten sogenannte Pathway-Analysen leisten. Bei diesem Ansatz wird überprüft, ob Gen-Gen-Interaktionen mit der Krankheit assoziiert sein könnten. Es könnten mehrere SNPs, die jeder für sich genommen keinen oder nur einen geringen messbaren Einfluss in funktionellen Studien haben, gemeinsam die Ätiologie der Krankheit beeinflussen, weil sie durch Stoffwechselprozesse aneinander gekoppelt sind. Diese auf Stoffwechselprozesse basierenden Analysen wurden auf GWAS-Daten für verschiedene komplexe Erkrankungen angewendet, und es wurden einige neue krankheitsrelevante Stoffwechselwege entdeckt[163–167]. Auch für Morbus Crohn waren solche Analysen schon erfolgreich. So konnte basierend auf GWAS-Daten und Pathway-Analysen der *IL12/IL23*-Pathway aufgedeckt werden, der 20 Gene beherbergt[164].

# 8 Schlussfolgerung und Ausblick

Die Aufklärung der genetischen Ursachen für komplexe Erkrankungen erweist sich auch mit dem Fortschritt der technischen Möglichkeiten als weiterhin sehr schwierig und kompliziert. Den tatsächlichen Einfluss verschiedener Umweltfaktoren und deren Interaktion mit der individuellen genetischen Disposition auf die Ätiologie von Morbus Crohn zu bestimmen, ist eines der Hauptprobleme. Ausschließlich die Aufklärung der genetischen Ursache betreffend macht das Fehlen von Information über die Funktion von einer Vielzahl von Genen, die Unzugänglichkeit mancher genetischer Regionen mittels „Next Generation Sequencing" – Methoden und vor allem das Fehlen von zuverlässigen Vorhersageprogrammen für nicht kodierende Varianten das Auffinden kausativer Varianten schwierig und limitiert so bisher den Nutzen der Sequenzierung von bekannten krankheitsassoziierten Loci oder die Sequenzierung gesamter Genome. Verbesserungen der Sequenzierungstechnologie (z. B. längere Leselängen) und bessere Algorithmen für das Identifizieren der genetischen Varianten können die Probleme nur teilweise beheben. Da die Kosten für die Sequenzierung voraussichtlich weiter fallen werden, wird es immer mehr möglich sein, Exome oder gar komplette Genome, z. B. großer Familien oder viele Patienten mit einer bestimmten Krankheit, zu sequenzieren, was die Identifizierung von kausativen Varianten oder krankheitsassoziierter Gene vereinfachen würde. Der Fokus wird sich aber nicht ausschließlich auf den genetischen Code richten, sondern sich auch auf strukturelle Varianten und epigenetische Modifikation erweitern. Die Auflösung des menschlichen Genoms im großen Umfang auf Einzelbasenniveau wird unser Wissen über die genetische Grundlage vieler komplexer Erkrankung und Phänotypen drastisch erweitern und in fernerer Zukunft in sog. Personalisierter Medizin resultieren. Das Erfolgspotential von Sequenzierstudien in der genetischen Aufklärung und somit zum Verständnis von schweren Krankheiten ist schon an einigen Beispielen sichtbar. So konnte mit Hilfe von einer Exom-Sequenzierung und nachfolgendem intelligenten Filtern der detektierten Varianten eine sehr seltene Mutation im *XIAP*-Gen aufgespürt werden, das eine endgültige Diagnose möglich machte und einen Therapiedurchbruch durch eine hämatopoetische Stammzelltransplantation eines schwer an einer monogenetischen Form einer chronisch entzündlichen Darmerkrankung erkrankten Jungen zur Folge hatte[168]. In manchen Fällen treten chronisch-entzündliche Darmerkrankungen als eine spezielle Form einer autosomal-rezessiv, potentiell monogenetischen Erkrankung in Erscheinung. Diese monogenetische mendelnde Vererbung für CED ist selten, aber die Aufklärung dieser seltenen Form könnte helfen, die

pathophysiologischen Mechanismen in der häufigeren, komplexen Form der Erkrankung zu finden und zu verstehen. Glocker et al. sequenzierten Kandidatengene in zwei blutsverwandten Familien mit Kindern, bei denen die chronisch-entzündliche Darmerkrankung sehr früh auftrat. Ursächlich in diesen Patienten waren Mutationen in Genen für Untereinheit der IL10R Proteine[169]. Diese und viele weitere positive Beispiele zeigen das enorme Potential, welches in der Sequenzierung von Exomen und Genomen für die Aufklärung des genetischen Hintergrundes komplexer Erkrankungen liegt und somit zukünftig therapeutische oder gar prophylaktische Maßnahmen verbessert werden können.

# 9 Zusammenfassung

Die Entdeckung von SNPs, die mit komplexen Krankheiten assoziiert sind, sind in der Aufklärung der genetischen Prädisposition wichtige Meilensteine und führten zu einem neuen Krankheitsverständnis. Da sie aber nur einen Teil des genetischen Risikos erklären, ist man weiterhin auf der Suche nach kausativen Varianten, die die „fehlende Heritabilität“ erklären sollen. Stark vorangetrieben wird diese sowohl genomweite- als auch Loci-gezielte Mutationssuche durch neue technische Möglichkeiten wie „Next Generation Sequencing“.

Die vorliegende Arbeit bestand aus zwei Teilprojekten, an die jeweils die Etablierung der SOLiD- Sequenzierung für den jeweiligen Anwendungsbereich geknüpft war.

Das erste Teilprojekt befasste sich mit der Auflösung von durch GWAS identifizierte Risikoloci bis auf Einzelbasenniveau mit dem Ziel bisher unbekannte Varianten zu detektieren. In einem aufwendigen Verfahren wurden die anvisierten Ziel-Regionen in jeweils 56 Individuen über LR-PCR angereichert um in einem späteren Schritt mit Hilfe der SOLiD-Technologie sequenziert zu werden. Für Analysen, die über die SNP-Detektion hinausgingen, wurde sich im folgenden auf die CED-Loci *NOD2*, *IL23R*, *ATG16L1*, *IL10*, *NKX2-3*, *STAT3* und *IRGM* beschränkt. In einer Assoziations-Genotypisierungsstudie wurden in einem ersten Verfahren 88 SNPs, die in der Sequenzierung detektiert wurden, in 364 Morbus Crohn Patienten und 368 gesunden Kontrollen genotypisiert. In einem zweiten Ansatz wurden aus diesem ersten Experiment 26 SNPs erneut in einer größeren Studienpopulation (2800 Kontrollen, 2500 Fälle) genotypisiert. Eine bisher unbekannte Variante in dem mit Morbus Crohn assoziierten Gen *NOD2* ist in den Fällen mit einer Allelfrequenz von 0,0019 genotypisiert wurden. In den Kontrollen trat diese Variante nicht Erscheinung. Diese Variante ist ein hochinteressanter Kandidat für eine sehr seltene Mutation, die mit dem Phänotyp Morbus Crohn assoziiert sein könnte. Sie wird Gegenstand weiterer Untersuchungen sein.

Im zweiten Teil der Arbeit wurde ein vollständiges Genom einer an Morbus Crohn erkrankten Patientin ebenfalls mittels SOLiD-Technologie sequenziert. Das Genom konnte annährend vollständig mit einer sehr guten Coverage sequenziert werden. Die Validierung mittels etablierter Sanger-Technologie zeigte, dass die SNP-Detektion robust, in der Identifizierung bisher unbekannter Varianten aber mit Schwächen behaftet ist. Der genetische Hintergrund der erkrankten Person konnte gut charakterisiert werden. Dabei stellte sich heraus, dass in dem Genom der sequenzierten Patientin nicht überdurchschnittlich viele SNPs mit potentiell

schädigender Wirkung angereichert sind. Vielmehr ähnelt es „gesunden“ Genomen. Es konnten über drei Millionen SNPs detektiert werden. Es wurden Varianten identifiziert, die schon mit dem Phänotyp Morbus Crohn assoziiert sind, und so einen Teil des Krankheitsbildes erklären. Unter anderem wurde die sehr seltene Deletion vor dem Gen *IRGM* in der Patientin homozygot identifiziert. Eine bestehende Infektion mit Mycobacterien konnte ebenfalls mit der genetischen Konstitution in Verbindung gebracht werden.

Das Identifizieren von neuen Varianten, die die weitere Aufklärung der Ätiologie des Morbus Crohn vorantreiben, gestaltet sich u. a. aufgrund fehlender Vorhersageprogramme für nicht-kodierende Varianten und nicht zuletzt durch die Komplexität der Krankheit und der Interaktion mit nicht berechenbaren nicht-genetischen Ursachen weiterhin als schwierig.

## Summary

The identification of an impressive number of disease-associated single nucleotide polymorphisms in the last years represents a milestone in the dissection of genetic predisposition and has led to new insights into disease etiology. However, as only a small part of the disease heritability can be explained by these polymorphisms, the search for causative variants is still ongoing in order to explain the “missing heritability”. “Next Generation Sequencing” is an appropriate tool to detect rare and potentially causative variants in whole genome seqencing approches or in target resequencing studies.

This thesis consisted of two subprojects, both of which involved the establishment of SOLiD „Next Generation Sequencing“-technology according to its special application.

The first project aimed at in depth-sequencing of ten riskloci identified in genome-wide association studies for several phenotypes (IBD, sarcodosis, psoriasis, PSC) in order to detect unknown variations. In an elaborate procedure the target regions were enricht by LR-PCR in 56 individuals for subsequent sequencing with the SOLiD „Next Generation Sequencing“-technology. SNP-calling was performed for all loci. For further analysis beyond SNP detection this study concentrated on the IBD loci *NOD2*, *IL23R*, *ATG16L1*, *IL10*, *NKX2-3*, *STAT3* and *IRGM*. In the first follow-up case/control association study 88 potential SNPs detected by sequencing were genotyped in 364 German Crohn`s disease patients and 368 healthy German controls. In a second approach 26 SNPs from the first association study were genotyped in a large independent study population (2800 German controls, 2500 German cases). A novel variation with an allele frequency of 0,0019 in the cases could be identified in the Crohn`s disease risk gene *NOD2*. Since this variation was not found in the controls, it

represents a very promising candidate for association with Crohn`s disease will be tested in further investigations.

In the second part of this thesis whole genome sequencing of a female Crohn`s disease patient with SOLiD „Next Generation Sequencing“-Technology was performed. The whole genome was sequenced with a satisfactorily coverage. The validation of identified polymorphisms with Sanger-Sequencing technology revealed a robust SNP detection, but shows some weakness in the de-novo identification of unknown variants. However, the genetic background of the patient was comprehensively characterized. It was shown that potentially damaging SNPs are not enriched in this genome, resembling a „healthy“ genome. A total of more than three million SNPs were detected, including established Crohn`s disease risk variations, which explain part of the phenotypic variation. Amongst others the patient revealed to be a homozygous carrier of the rare deletion upstream of the *IRGM*. A present infection with Mycobacterium paratuberculosis could be connected to the genetic background of the patient.

# 10 Literaturverzeichnis


1. Baumgart, D. C. & Carding, S. R. Inflammatory bowel disease: cause and immunobiology. *Lancet* **369**, 1627–40 (2007).

2. Baumgart, D. C. & Sandborn, W. J. Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet* **369**, 1641–57 (2007).

3. Vatn, M. H. Natural history and complications of IBD. *Current gastroenterology reports* **11**, 481–7 (2009).

4. Bernstein, C. N. Extraintestinal manifestations of inflammatory bowel disease. *Current gastroenterology reports* **3**, 477–83 (2001).

5. Rosenstiel, P., Sina, C., Franke, A. & Schreiber, S. Towards a molecular risk map--recent advances on the etiology of inflammatory bowel disease. *Seminars in immunology* **21**, 334–45 (2009).

6. Houlston, R. S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics* **40**, 1426–35 (2008).

7. Yang, H., Taylor, K. D. & Rotter, J. I. Inflammatory bowel disease. I. Genetic epidemiology. *Mol Genet Metab* **74**, 1–21 (2001).

8. Odes, S. *et al.* Cost analysis and cost determinants in a European inflammatory bowel disease inception cohort with 10 years of follow-up evaluation. *Gastroenterology* **131**, 719–28 (2006).

9. Loftus, E. V Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology* **126**, 1504–17 (2004).

10. Shivananda, S. *et al.* Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD). *Gut* **39**, 690–7 (1996).

11. Roth, M. P. *et al.* Familial empiric risk estimates of inflammatory bowel disease in Ashkenazi Jews. *Gastroenterology* **96**, 1016–20 (1989).

12. Kurata, J. H., Kantor-Fish, S., Frankl, H., Godby, P. & Vadheim, C. M. Crohn's disease among ethnic groups in a large health maintenance organization. *Gastroenterology* **102**, 1940–8 (1992).

13. Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474**, 307–317 (2011).

14. Gitlin, L., Borody, T. J., Chamberlin, W. & Campbell, J. Mycobacterium avium ss paratuberculosis-associated Diseases Piecing the Crohn ' s Puzzle Together. **46**, 649–655 (2012).

15. Russell, R. K. & Satsangi, J. IBD: a family affair. *Best practice & research. Clinical gastroenterology* **18**, 525–39 (2004).

16. Thompson, N. P., Driscoll, R., Pounder, R. E. & Wakefield, A. J. Genetics versus environment in inflammatory bowel disease: results of a British twin study. *BMJ (Clinical research ed.)* **312**, 95–6 (1996).

17. Tysk, C., Lindberg, E., Järnerot, G. & Flodérus-Myrhed, B. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* **29**, 990–6 (1988).

18. Brant, S. R. Exposed: the genetic underpinnings of ulcerative colitis relative to Crohn's disease. *Gastroenterology* **136**, 396–9 (2009).

19. Küster, W., Pascoe, L., Purrmann, J., Funk, S. & Majewski, F. The genetics of Crohn disease: complex segregation analysis of a family study with 265 patients with Crohn disease and 5,387 relatives. *American journal of medical genetics* **32**, 105–8 (1989).

20. Satsangi, J., Jewell, D. P., Rosenberg, W. M. & Bell, J. I. Genetics of inflammatory bowel disease. *Gut* **35**, 696–700 (1994).

21. Orholm, M. *et al.* Investigation of inheritance of chronic inflammatory bowel diseases by complex segregation analysis. *BMJ (Clinical research ed.)* **306**, 20–4 (1993).

22. Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J. & Krawczak, M. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet* **6**, 376–388 (2005).

23. Gaya, D. R., Russell, R. K., Nimmo, E. R. & Satsangi, J. New genes in inflammatory bowel disease: lessons for complex diseases? *Lancet* **367**, 1271–84 (2006).

24. Zheng, J. J. *et al.* Crohn's disease in mainland China: a systematic analysis of 50 years of research. *Chinese journal of digestive diseases* **6**, 175–81 (2005).

25. Lakatos, P. L. Environmental factors affecting inflammatory bowel disease: have we made progress? *Digestive diseases (Basel, Switzerland)* **27**, 215–25 (2009).

26. Guarner, F. *et al.* Mechanisms of disease: the hygiene hypothesis revisited. *Nature clinical practice. Gastroenterology & hepatology* **3**, 275–84 (2006).

27. Riordan, A. M., Ruxton, C. H. & Hunter, J. O. A review of associations between Crohn's disease and consumption of sugars. *European journal of clinical nutrition* **52**, 229–38 (1998).

28. Geerling, B. J. *et al.* Diet as a risk factor for the development of ulcerative colitis. *The American journal of gastroenterology* **95**, 1008–13 (2000).

29. Desai, H. G. & Gupte, P. A. Increasing incidence of Crohn's disease in India: is it related to improved sanitation? *Indian journal of gastroenterology : official journal of the Indian Society of Gastroenterology* **24**, 23–4

30. Hampe, J., Heymann, K., Krawczak, M. & Schreiber, S. Association of inflammatory bowel disease with indicators for childhood antigen and infection exposure. *International journal of colorectal disease* **18**, 413–7 (2003).

31. Gent, A. E., Hellier, M. D., Grace, R. H., Swarbrick, E. T. & Coggon, D. Inflammatory bowel disease and domestic hygiene in infancy. *Lancet* **343**, 766–7 (1994).

32. Cosnes, J. Tobacco and IBD: relevance in the understanding of disease mechanisms and clinical practice. *Best practice & research. Clinical gastroenterology* **18**, 481–96 (2004).

33. Mahid, S. S., Minor, K. S., Stromberg, A. J. & Galandiuk, S. Active and passive smoking in childhood is related to the development of inflammatory bowel disease. *Inflammatory bowel diseases* **13**, 431–8 (2007).

34. Jones, D. T., Osterman, M. T., Bewtra, M. & Lewis, J. D. Passive smoking and inflammatory bowel disease: a meta-analysis. *The American journal of gastroenterology* **103**, 2382–93 (2008).

35. Taurog, J. D. *et al.* The germfree state prevents development of gut and joint inflammatory disease in HLA-B27 transgenic rats. *The Journal of experimental medicine* **180**, 2359–64 (1994).

36. Dieleman, L. A., Rath, H. C. & Schultz, M. Different Subsets of Enteric Bacteria Induce and Perpetuate Experimental Colitis in Rats and Mice. **69**, 2277–2285 (2001).

37. Umesaki, Y., Setoyama, H., Matsumoto, S. & Okada, Y. Expansion of alpha beta T-cell receptor-bearing intestinal intraepithelial lymphocytes after microbial colonization in germ-free mice and its independence from thymus. *Immunology* **79**, 32–7 (1993).

38. Vaishnava, S., Behrendt, C. L., Ismail, A. S., Eckmann, L. & Hooper, L. V Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 20858–63 (2008).

39. Ley, R. E., Peterson, D. a & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837–48 (2006).

40. Guarner, F. & Malagelada, J.-R. Gut flora in health and disease. *Lancet* **361**, 512–9 (2003).

41. Swidsinski, A. *et al.* MUCOSAL FLORA IN CROHN ' S DISEASE AND ULCERATIVE COLITIS - AN OVERVIEW. 61–71 (2009).

42. Chiodini, R. J., Kruiningen, H. J. V. A. N., Merkal, R. S., Thayer, W. R. & Coutu, J. A. Characteristics of an Unclassified Mycobacterium Species Isolated from Patients with Crohn ' s Disease. **20**, 966–971 (1984).

43. Chiodini, R. J., Van Kruiningen, H. J., Thayer, W. R., Merkal, R. S. & Coutu, J. A. Possible role of mycobacteria in inflammatory bowel disease. I. An unclassified Mycobacterium species isolated from patients with Crohn's disease. *Digestive diseases and sciences* **29**, 1073–9 (1984).

44. Sechi, L. A. *et al.* Identification of Mycobacterium avium subsp . paratuberculosis in Biopsy Specimens from Patients with Crohn ' s Disease Identified by In Situ Hybridization Identification of Mycobacterium avium subsp . paratuberculosis in Biopsy Specimens from Patients wi. (2001).doi:10.1128/JCM.39.12.4514

45. Barta, Z., Mekkel, G. & Zeher, M. Seroprevalence of Mycobacterium paratuberculosis in Patients with Crohn ' s Disease. **42**, 5432–5433 (2004).

46. Hulten, K. *et al.* Detection of Mycobacterium avium subspecies paratuberculosis in Crohn's diseased tissues by in situ hybridization. *The American journal of gastroenterology* **96**, 1529–35 (2001).

47. Ryan, P. *et al.* PCR detection of Mycobacterium paratuberculosis in Crohn's disease granulomas isolated by laser capture microdissection. *Gut* **51**, 665–70 (2002).

48. Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y. & Chia, K. S. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* **55**, 403–415 (2010).

49. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature genetics* **36**, 949–51 (2004).

50. Tamaki, K. & Jeffreys, A. J. Human tandem repeat sequences in forensic DNA typing. *Legal medicine (Tokyo, Japan)* **7**, 244–50 (2005).

51. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature reviews. Genetics* **6**, 109–18 (2005).

52. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–33 (2001).

53. Day, I. N. M. dbSNP in the detail and copy number complexities. *Human Mutation* **31**, 2–4 (2010).

54. Hugot, J. P. *et al.* Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* **379**, 821–3 (1996).

55. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).

56. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).

57. Hampe, J. *et al.* Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* **357**, 1925–8 (2001).

58. Van Limbergen, J., Wilson, D. C. & Satsangi, J. The genetics of Crohn's disease. *Annual review of genomics and human genetics* **10**, 89–116 (2009).

59. Van Heel, D. A., McGovern, D. P. & Jewell, D. P. Crohn's disease: genetic susceptibility, bacteria, and innate immunity. *Lancet* **357**, 1902–4 (2001).

60. Bonen, D. K. *et al.* Crohn's disease-associated NOD2 variants share a signaling defect in response to lipopolysaccharide and peptidoglycan. *Gastroenterology* **124**, 140–6 (2003).

61. Rosenstiel, P. *et al.* Regulation of DMBT1 via NOD2 and TLR4 in intestinal epithelial cells modulates bacterial recognition and invasion. *Journal of immunology (Baltimore, Md. : 1950)* **178**, 8203–11 (2007).

62. Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).

63. Tremelling, M. *et al.* IL23R variation determines susceptibility but not disease phenotype in inflammatory bowel disease. *Gastroenterology* **132**, 1657–64 (2007).

64. Parkes, M. *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* **39**, 830–832 (2007).

65. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature genetics* **39**, 207–11 (2007).

66. Parkes, M. *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* **39**, 830–832 (2007).

67. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature genetics* **39**, 596–604 (2007).

68. Raelson, J. V *et al.* Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc Natl Acad Sci U S A* **104**, 14747–14752 (2007).

69. Franke, A. *et al.* Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. *Nat Genet* **40**, 713–715 (2008).

70. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118–1125 (2010).

71. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* **43**, 246–252 (2011).

72. Zhou, L. *et al.* IL-6 programs T(H)-17 cell differentiation by promoting sequential engagement of the IL-21 and IL-23 pathways. *Nature immunology* **8**, 967–74 (2007).

73. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955–962 (2008).

74. Estivill, X. *et al.* A candidate for the cystic fibrosis locus isolated by selection for methylation-free islands. *Nature* **326**, 840–5

75. Marian, A. J. Molecular genetic studies of complex phenotypes. *Translational research : the journal of laboratory and clinical medicine* **159**, 64–79 (2012).

76. Manolio, T. A. news and views Cohort studies and the genetics of complex disease. **41**, 5–6 (2009).

77. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124–137 (2001).

78. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311–321 (2008).

79. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (New York, N.Y.)* **261**, 921–3 (1993).

80. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362–7 (2009).

81. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* **19**, 212–219 (2009).

82. Romeo, S. *et al.* Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature genetics* **39**, 513–6 (2007).

83. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.)* **324**, 387–9 (2009).

84. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* **24**, 133–41 (2008).

85. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–73 (2010).

86. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–6 (2008).

87. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).

88. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics* **43**, 1066–73 (2011).

89. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nature genetics* **44**, 623–30 (2012).

90. Kato, K. Review Article Impact of the next generation DNA sequencers. *Clinical and Experimental Medicine* 193–202 (2009).

91. Peloso, G. M. *et al.* Exome Sequencing, ANGPTL3 Mutations, and Familiar Combined Hypolipidemia. 2220–2227 (2010).at <http://www.ncbi.nlm.nih.gov/pubmed/20942659>

92. Christodoulou, K. *et al.* Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes. *Gut* (2012).doi:10.1136/gutjnl-2011-301833

93. Siu, H., Zhu, Y., Jin, L. & Xiong, M. Implication of next-generation sequencing on association studies. *BMC genomics* **12**, 322 (2011).

94. Sobreira, N. L. M. *et al.* Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS genetics* **6**, e1000991 (2010).

95. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).

96. Holm, H. *et al.* syndrome. **43**, 316–320 (2011).

97. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–95 (1977).

98. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–65 (1981).

99. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

100. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133–141 (2008).

101. Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**, 1622–1629 (2009).

102. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

103. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England)* **21**, 263–5 (2005).

104. Hill, W. G. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**, 229–39 (1974).

105. Flachsbart, F. Kandidatengen-Studien zur Identifizierung genetischer Suszeptibilitätsfaktoren für Langlebigkeit beim Menschen. (2007).

106. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)* **132**, 365–86 (2000).

107. Mitra, R. D., Shendure, J., Olejnik, J. & Church, G. M. Fluorescent in situ sequencing on polymerase colonies. *Analytical Biochemistry* **320**, 55–65 (2003).

108. Biosystems, A. Applied Biosystems SOLiD ™ 4 System Library Preparation Guide. (2010).

109. Biosystems, A. Applied Biosystems SOLiD ™ System BioScope ™ Software for Scientists Guide. (2010).

110. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–9 (2009).

111. Forster, M. *et al.* From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. *Nucleic acids research* **41**, e16 (2013).

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

113. Manke, T., Heinig, M. & Vingron, M. Quantifying the effect of sequence variation on regulatory interactions. *Hum Mutat* **31**, 477–483 (2010).

114. Gabriel, S., Ziaugra, L. & Tabbaa, D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* **Chapter 2**, Unit 2.12 (2009).

115. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *American journal of human genetics* **76**, 887–93 (2005).

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

117. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308–11 (2001).

118. Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome medicine* **1**, 13 (2009).

119. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164 (2010).

120. Melum, E. *et al.* SNP discovery performance of two second-generation sequencing platforms in the NOD2 gene region. *Human mutation* **31**, 875–85 (2010).

121. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome research* **19**, 1124–32 (2009).

122. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

123. Huang, X. *et al.* High-throughput genotyping by whole-genome resequencing. *Genome Res* **19**, 1068–1076 (2009).

124. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–73 (2010).

125. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics* **7**, 61–80 (2006).

126. Shihab, H. a *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation* **34**, 57–65 (2013).

127. Rudd, M. F. *et al.* The predicted impact of coding single nucleotide polymorphisms database. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **14**, 2598–604 (2005).

128. Ku, C.-S. *et al.* Technological advances in DNA sequence enrichment and sequencing for germline genetic diagnosis. *Expert review of molecular diagnostics* **12**, 159–73 (2012).

129. Elsharawy, A. *et al.* Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing. *BMC genomics* **13**, 417 (2012).

130. Tewhey, R. *et al.* Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol* **10**, R116 (2009).

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

132. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature genetics* **27**, 234–6 (2001).

133. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature reviews. Genetics* **12**, 628–40 (2011).

134. Lesage, S. *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70**, 845–857 (2002).

135. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**, 695–701 (2008).

136. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature genetics* **22**, 231–8 (1999).

137. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature genetics* **22**, 239–47 (1999).

138. Forster, M. *et al.* From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. *Nucleic Acids Research* 1–12 (2012).doi:10.1093/nar/gks836

139. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**, 1107–1112 (2008).

140. Stenson, P. D. *et al.* Human Gene Mutation Database ( HGMD ): 2003 Update. **581**, 577–581 (2003).

141. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & Mckusick, V. A. Online Mendelian Inheritance in Man ( OMIM ), a knowledgebase of human genes and genetic disorders. **33**, 514–517 (2005).

142. Genomics, C. Finishing the euchromatic sequence of the human genome. 931–945 (2004).

143. Booth, R. A. *et al.* Assessment of thiopurine S-methyltransferase activity in patients prescribed thiopurines: a systematic review. *Annals of internal medicine* **154**, 814–23, W–295–8 (2011).

144. Jackson, A. P., Hall, A. G. & McLelland, J. Thiopurine methyltransferase levels should be measured before commencing patients on azathioprine. *The British journal of dermatology* **136**, 133–4 (1997).

145. Fuentes Fajardo, K. V *et al.* Detecting false-positive signals in exome sequencing. *Human mutation* **33**, 609–13 (2012).

146. Shirazi, T., Longman, R. J., Corfield, a P. & Probert, C. S. Mucins and inflammatory bowel disease. *Postgraduate medical journal* **76**, 473–8 (2000).

147. Morita, H., Kettlewell, M. G., Jewell, D. P. & Kent, P. W. Glycosylation and sulphation of colonic mucus glycoproteins in patients with ulcerative colitis and in healthy subjects. *Gut* **34**, 926–32 (1993).

148. Van der Sluis, M. *et al.* Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology* **131**, 117–29 (2006).

149. Kim, Y. S. & Ho, S. B. Intestinal goblet cells and mucins in health and disease: recent insights and progress. *Current gastroenterology reports* **12**, 319–30 (2010).

150. Indik, Z. K., Park, J. G., Hunter, S. & Schreiber, A. D. The molecular dissection of Fc gamma receptor mediated phagocytosis. *Blood* **86**, 4389–99 (1995).

151. Fridman, W. H. Fc receptors and immunoglobulin binding factors. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **5**, 2684–90 (1991).

152. Li, M. *et al.* Deletion of the late cornified envelope genes LCE3C and LCE3B is associated with psoriasis in a Chinese population. *The Journal of investigative dermatology* **131**, 1639–43 (2011).

153. Podolsky, D. K. Inflammatory bowel disease. *The New England journal of medicine* **7**, 417–29 (2002).

154. Russell, R. K., Wilson, D. C. & Satsangi, J. Unravelling the complex genetics of inflammatory bowel disease. *Archives of disease in childhood* **89**, 598–603 (2004).

155. Shoda, R., Matsueda, K., Yamato, S. & Umeda, N. Epidemiologic analysis of Crohn disease in Japan: increased dietary intake of n-6 polyunsaturated fatty acids and animal protein relates to the increased incidence of Crohn disease in Japan. *The American journal of clinical nutrition* **63**, 741–5 (1996).

156. Nugent, F. W., Richmond, M. & Park, S. K. Crohn's disease of the duodenum. *Gut* **18**, 115–20 (1977).

157. Bull, T. J. *et al.* Detection and Verification of Mycobacterium avium subsp. paratuberculosis in Fresh Ileocolonic Mucosal Biopsy Specimens from Individuals with and without Crohn's Disease. *Journal of Clinical Microbiology* **41**, 2915–2923 (2003).

158. Autschbach, F. *et al.* High prevalence of Mycobacterium avium subspecies paratuberculosis IS900 DNA in gut tissues from individuals with Crohn's disease. *Gut* **54**, 944–9 (2005).

159. Kaser, A. & Blumberg, R. S. Autophagy, microbial sensing, endoplasmic reticulum stress, and epithelial function in inflammatory bowel disease. *Gastroenterology* **140**, 1738–47 (2011).

160. Glasser, A.-L. & Darfeuille-Michaud, A. Abnormalities in the handling of intracellular bacteria in Crohn's disease: a link between infectious etiology and host genetic susceptibility. *Archivum immunologiae et therapiae experimentalis* **56**, 237–44 (2008).

161. Friswell, M., Campbell, B. & Rhodes, J. The role of bacteria in the pathogenesis of inflammatory bowel disease. *Gut and liver* **4**, 295–306 (2010).

162. Ferwerda, G. *et al.* Mycobacterium paratuberculosis is recognized by Toll-like receptors and NOD2. *Journal of leukocyte biology* **82**, 1011–8 (2007).

163. Menashe, I. *et al.* NIH Public Access. **70**, 4453–4459 (2011).

164. Wang, K. *et al.* Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *American journal of human genetics* **84**, 399–405 (2009).

165. Baranzini, S. E. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human molecular genetics* **18**, 2078–90 (2009).

166. Menashe, I. *et al.* Large-scale pathway-based analysis of bladder cancer genome-wide association data from five studies of European background. *PloS one* **7**, e29396 (2012).

167. Zhang, M., Liang, L., Xu, M., Qureshi, A. a & Han, J. Pathway analysis for genome-wide association study of basal cell carcinoma of the skin. *PloS one* **6**, e22760 (2011).

168. Worthey, E. a *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in medicine : official journal of the American College of Medical Genetics* **13**, 255–62 (2011).

169. Glocker, E.-O. *et al.* Infant colitis--it's in the genes. *Lancet* **376**, 1272 (2010).

# 11 Erklärung

Hiermit erkläre ich, dass die vorliegende Arbeit - abgesehen von den Beratungen durch meine akademischen Lehrer - nach Inhalt und Form meine eigene Arbeit ist. Die Arbeit wurde bis jetzt weder vollständig noch in Teilen einer anderen Stelle im Rahmen eines Prüfungsverfahren vorgelegt. Ferner erkläre ich, dass ich noch keine früheren Promotionsversuche unternommen habe.

Lübeck, den

# 12 Lebenlauf

## Persönliche Daten

| | |
|---|---|
| Name | Sandra May |
| Geburtsdatum | 29.07.1979 |
| Geburtsort | Halberstadt |
| Familienstand | verheiratet |
| Staatsangehörigkeit | deutsch |

## Schulischer Werdegang

| | |
|---|---|
| 1986-1988 | POS Ernst Thälmann |
| 1988-1991 | POS Friedenschule mit erweitertem Russischunterricht |
| 1991-1998 | Gymnasium Martineum Halberstadt |
| 1998 | Abitur |

## Studium

| | |
|---|---|
| 1999-2002 | Grundstudium der Biologie an der CAU Kiel |
| 2002-2006 | Hauptstudium Biologie<br>Hauptfach: Zoologie<br>Nebenfächer: Zellbiologie, Zellbiologie |
| 04/2005 | Hauptdiplomprüfungen |
| 05/2005-04/2006 | Diplomarbeit am Zoologischen Institut der CAU Kiel „Charakterisierung des Hämocyanins bei *Porcellio scaber* und Versuche zur Aktivierung in ein phenyl-oxidaseähnliches Enzym“ |
| seit 2008 | Promotion am Institut für Klinische Molekularbiologie (Direktor: Prof. Dr. med. S. Schreiber)<br>„"Next Generation Sequencing" basierte Prozessentwicklung zur systematischen Patientengenomanalyse am Beispiel der chronisch entzündlichen Darmerkrankungen“ |

**Publikationen**

Franke, A., Balschun, T., Karlsen, T. H., Hedderich, J., **May, S.**, Lu, T., Schuldt, D., et al. (2008). Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. *Nat Genet*, *40*(6), 713–715. doi:ng.148 [pii] 10.1038/ng.148

Akil, I., Ozguven, A., Canda, E., Yilmaz, O., Nese, N., Ozkol, M., **May, S.**, et al. (2010). Co-existence of chronic renal failure, renal clear cell carcinoma, and Blau syndrome. *Pediatric Nephrology*, *25*(5), 977–981. doi:10.1007/s00467-009-1413-5

Melum, E., **May, S.**, Schilhabel, M. B., Thomsen, I., Karlsen, T. H., Rosenstiel, P., Schreiber, S., et al. (2010). SNP discovery performance of two second-generation sequencing platforms in the NOD2 gene region. *Human mutation*, *31*(7), 875–85. doi:10.1002/humu.21276

Forster, M., Forster, P., Elsharawy, A., Hemmrich, G., Kreck, B., Wittig, M., Thomsen, I., Stade, B., Barann, M., Ellinghaus, D., Petersen, B.-S., **May, S.**, Melum, E., Schilhabel, M., Keller, A., Schreiber, S., et al. (2013). From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. *Nucleic acids research*, *41*(1), e16. doi:10.1093/nar/gks836

# 13 Danksagung

Ich danke Herrn Prof. Stefan Schreiber, Prof. Philip Rosenstiel und Prof. Andre Franke für die Möglichkeit mich eines solchen herausfordernden Themas zu stellen zu dürfen sowie für die exzellenten Arbeitsbedingungen am IKMB. Prof. Andre Franke möchte ich besonders für die immer konstruktive Unterstützung und stetige Geduld während meiner Arbeit danken.

Ein ganz besonders herzlicher Dank gilt Catharina Fürstenau, die hochmotiviert und diszipliniert unzählige PCRs pipettiert, Gele gegossen, Platten beschriftet hat und immer fröhlich bei der Stange geblieben ist. Danke für Deinen persönlichen Einsatz und für die schöne vertrauensvolle Zusammenarbeit.

Dem NextGen-Labor möchte ich für die Herstellung der SOLiD-Libraries und für die Sequenzierung derselben danken. Casi und Melli für das selbstlose Aufreinigen der vielen „projektfernen“ PCR-Produkte, für die Treue zum „Pinguin“ (und deren Chefin), für die vielen Tassen Kaffee und Kekse, die es für und manchmal auch gegen die Motivation gab. Danke für die schöne Laborstimmung.

Allen MTAs des IKMBs, besonders denen des Sanger-Labors und Intakes ein großes Dankeschön für die große Unterstützung dieses Mega-Labor-Projektes.

Markus Schilhabel danke ich für seine unbedingte Offenheit, mit der er all seine Erfahrungen in Sachen Sequenzierung, von denen ich stark profitiert habe, weitergibt. Unsere gemeinsamen Bürozeiten (jeder mit Kopfhörer und immer bei offener Bürotür) mit ergiebigen Gesprächen bleiben unvergessen.

Michael Forster, Ingo Thomsen und Björn Stade möchte ich danken für die bioinformatorische Realisierung dieses Projektes. Ohne Euch wäre vieles nicht möglich gewesen.

Ein sehr persönlicher Dank geht an meine Kollegin und Freundin Heidi Schaarschmidt, mit der zusammen ich viele berufliche und private Durststrecken durchlebt habe. Danke für Deine Freundschaft.

Meiner Familie und Freunde außerhalb des IKMBs möchte ich dafür danken, dass sie mir in jeder Lebensphase eine Anlaufstelle bieten, Kraft, Verständnis, Freundschaft und Liebe zu tanken, was mir gute Rahmenbedingungen für das Durchhalten von schwierigen Phasen gab.

# 14 Anhang

**Tabelle 14-1:** GWAS Katalog für Morbus Crohn Risikovarianten mit bekannten OR. Die Gene sind sortiert nach höchsten OR. Datenbankabfrage war der 15.01.2012.

| Reported Gene(s) | Strongest SNP-Risk Allele | OR or beta | 95% CI (text) |
|---|---|---|---|
| **NOD2,ADCY7** | rs2066847-T | 3.10 | [1.497-1.618] |
| **IL23R** | rs11209026-G | 2.84 | [NR] |
| **HLA-DQA2, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DOB, PSMB9** | rs7765379-G | 1.93 | [1.78-2.09] |
| **IL23R** | rs11465804-? | 1.89 | [1.47-2.44] |
| **MUC19,LRRK2** | rs11564258-A | 1.74 | [1.55-1.95] |
| **TNFSF15, LOC100129633, LOC645266, TNFSF8** | rs6478106-T | 1.73 | [1.60-1.86] |
| **Intergenic** | rs224136-? | 1.67 | [NR] |
| **PSORS1C3** | rs3094188-C | 1.61 | [1.33-1.94] |
| **Intergenic** | rs6596075-C | 1.55 | [1.00-2.39] |
| **IRGM** | rs1000113-T | 1.54 | [1.31-1.82] |
| **LRRK2, MUC19** | rs11175593-T | 1.54 | [NR] |
| **Intergenic** | rs17234657-G | 1.54 | [1.34-1.76] |
| **PBX2, NOTCH4** | rs9267911-T | 1.50 | [1.29-1.76] |
| **SLCO6A1** | rs7705924-G | 1.48 | [1.17-1.87] |
| **Intergenic** | rs10801047-? | 1.47 | [1.22-1.76] |
| **NOD2** | rs2076756-G | 1.46 | [NR] |
| **Intergenic** | rs1373692-? | 1.46 | [NR] |
| **IL12B** | rs10045431-? | 1.45 | [1.27-1.64] |
| **HLA-F, MOG, HLA-G, GABBR1, HLA-H, UBD, HLA-A** | rs9258260-T | 1.45 | [1.21-1.68] |
| **PSMB10** | rs11574514-A | 1.44 | [1.35-1.52] |
| **PTGER4** | rs1992660-? | 1.42 | [1.24-1.67] |
| **AIF1** | rs9348876-T | 1.41 | [1.22-1.63] |
| **IL23R** | rs11805303-T | 1.39 | [1.22-1.58] |
| **NR** | rs7807268-G | 1.38 | [1.20-1.60] |
| **IL23R** | 17 marker haplotype-1 | 1.38 | [1.23-1.53] |
| **ATG16L1** | rs3792109-A | 1.38 | [NR] |
| **IRGM** | rs13361189-? | 1.38 | [1.15-1.66] |
| **CCR6,FGFR10P,RNASE2** | rs2301436-? | 1.37 | [1.22-1.53] |
| **PTGER4** | rs9292777-T | 1.37 | [1.28-1.48] |
| **STAT3** | rs9891119-A | 1.37 | [1.27-1.48] |
| **IRGM** | rs7714584-G | 1.37 | [1.28-1.47] |
| **IBD5** | rs2188962-? | 1.36 | [1.21-1.52] |
| **PTPN2** | rs2542151-G | 1.35 | [NR] |
| **NKX2-3, SLC25A28, GOT1, ENTPD7, CNNM1, COX15, CUTC** | rs11190141-C | 1.34 | [1.25-1.43] |
| **ATG16L1** | rs3792109-A | 1.34 | [1.29-1.40] |
| **TBC1D1** | rs1487630-T | 1.33 | [1.22-1.44] |
| **IRGM** | rs11747270-G | 1.33 | [NR] |

| | | | |
|---|---|---|---|
| **PTGER4** | rs11742570-C | 1.32 | [NR] |
| **ATG16L1, SAG, DGKD, INPP5D, USP40** | rs2241880-G | 1.32 | [1.24-1.41] |
| **PTGER4** | rs4613763-C | 1.32 | [NR] |
| **NELL1** | rs1793004-? | 1.30 | [1.12-1.52] |
| **CARD9, SNAPC4** | rs4077515-T | 1.29 | [NR] |
| **NOD2** | rs17221417-G | 1.29 | [1.13-1.46] |
| **ZNF365** | rs10761659-G | 1.28 | [NR] |
| **ATG16L2, FCHSD2** | rs72981516-T | 1.28 | [1.27-1.29] |
| **JAK2** | rs2274471-A | 1.27 | [1.15-1.41] |
| **RBX1, EP300** | rs4820425-A | 1.27 | [1.17-1.38] |
| **TCERG1L** | rs10734105-G | 1.27 | [1.10-1.43] |
| **ELF1, microRNA2276, SLC25A15, WBP4** | rs7329174-G | 1.27 | [1.17-1.38] |
| **IL12B** | rs6887695-? | 1.26 | [1.12-1.41] |
| **PTPN22** | rs2476601-G | 1.26 | [1.17-1.37] |
| **PTPN2** | rs1893217-G | 1.25 | [1.18-1.32] |
| **ZNF365** | rs10995271-C | 1.25 | [NR] |
| **ATG16L1** | rs3828309-G | 1.25 | [NR] |
| **C10orf67** | rs1398024-A | 1.23 | [1.04-1.45] |
| **IL3, ACSL6, P4HA2, PDLIM4, SLC22A4** | rs3091338-T | 1.23 | [1.08-1.42] |
| **GALC,GPR65** | rs8005161-T | 1.23 | [1.16-1.31] |
| **SLC22A4,SLC22A5,IRF1,IL3** | rs12521868-T | 1.23 | [1.18-1.28] |
| **MAP3K7IP1** | rs2413583-C | 1.23 | [1.17-1.29] |
| **ATG16L1, INPP5D** | rs12994997-A | 1.23 | [1.193-1.274] |
| **IFNGR2** | rs2834215-? | 1.22 | [1.12-1.32] |
| **SOX11** | rs11894081-T | 1.22 | [1.2-1.22] |
| **TNFSF15** | rs4263839-G | 1.22 | [NR] |
| **TNFSF18,TNFSF4,FASLG** | rs7517810-T | 1.22 | [1.16-1.28] |
| **MST1,GPX1,BSN** | rs3197999-A | 1.22 | [1.16-1.27] |
| **NKX2-3** | rs4409764-T | 1.22 | [1.17-1.27] |
| **TNFSF15,TNFSF8** | rs3810936-C | 1.21 | [1.15-1.27] |
| **FOXP2** | rs1869839-? | 1.20 | [1.11-1.3] |
| **NKX2-3** | rs10883365-G | 1.20 | [1.03-1.39] |
| **Intergenic** | rs1456893-A | 1.20 | [NR] |
| **CCL2,CCL7** | rs3091315-A | 1.20 | [1.14-1.26] |
| **PTPN22,DCLRE1B** | rs6679677-C | 1.20 | [1.129-1.268] |
| **NKX2-3** | rs11190140-T | 1.20 | [NR] |
| **Intergenic** | rs1906493-A | 1.19 | [1.09-1.28] |
| **ZNF365, ERG2, ADO** | rs7076156-G | 1.19 | [1.10-1.30] |
| **LTA,HLA-DQA2,TNF,LST1,LTB** | rs1799964-C | 1.19 | [1.13-1.25] |
| **IL18RAP,IL12RL2,IL18R1,IL1RL1** | rs2058660-G | 1.19 | [1.14-1.26] |
| **ATG16L1** | rs10210302-T | 1.19 | [1.01-1.41] |
| **Intergenic** | rs6651252-T | 1.19 | [1.128-1.246] |
| **UBE2D1** | rs1819658-C | 1.19 | [1.13-1.25] |

| | | | |
|---|---|---|---|
| **ZMIZ1** | rs1250550-G | 1.19 | [1.15-1.23] |
| **RUNX3** | rs7551188-T | 1.18 | [1.10-1.28] |
| **Intergenic** | rs1736135-T | 1.18 | [NR] |
| **Intergenic** | rs11584383-T | 1.18 | [NR] |
| **STAT3** | rs744166-A | 1.18 | [NR] |
| **JAK2** | rs10758669-C | 1.18 | [1.13-1.23] |
| **IL12B** | rs6556412-A | 1.18 | [1.13-1.24] |
| **ICOSLG** | rs2838519-G | 1.18 | [1.13-1.23] |
| **CDKAL1** | rs6908425-C | 1.17 | [1.11-1.23] |
| **Unknown** | rs7746082-C | 1.17 | [NR] |
| **CCR6** | rs415890-C | 1.17 | [1.12-1.22] |
| **Intergenic** | rs13126505-A | 1.17 | [1.10-1.248] |
| **intergenic** | rs4871611-A | 1.17 | [1.12-1.23] |
| **C11orf30** | rs7927997-T | 1.17 | [1.12-1.22] |
| **Intergenic** | rs13204742-T | 1.17 | [1.118-1.23] |
| **NR** | rs6601764-C | 1.16 | [1.01-1.33] |
| **TMEM17, EHBP1, CPAMD8, AK3** | rs6545946-C | 1.16 | [1.06-1.27] |
| **Intergenic** | rs17582416-G | 1.16 | [NR] |
| **C11orf30** | rs7927894-T | 1.16 | [NR] |
| **ADAM30** | rs3897478-T | 1.16 | [1.101-1.224] |
| **intergenic** | rs1736020-C | 1.16 | [1.11-1.21] |
| **GPX4,SBNO2** | rs740495-G | 1.16 | [1.10-1.21] |
| **SOCS1** | rs4780355-T | 1.16 | [NR] |
| **ZMIZ1** | rs1250544-G | 1.16 | [NR] |
| **LACC1** | rs3764147-G | 1.16 | [1.112-1.199] |
| **GPX4,HMHA1** | rs2024092-A | 1.16 | [1.112-1.201] |
| **CD244,ITLN1** | rs4656940-A | 1.15 | [1.09-1.21] |
| **Intergenic** | rs2836754-? | 1.15 | [1.03-1.28] |
| **MLX,STAT3** | rs11871801-A | 1.15 | [1.10-1.21] |
| **C8orf84, TERF1, RPL7, RDH10, KCNB2** | rs12677663-T | 1.15 | [1.04-1.28] |
| **SLC43A3, PRG2, PRG3** | rs11229030-C | 1.15 | [1.10-1.39] |
| **CREM** | rs12242110-G | 1.15 | [1.10-1.20] |
| **GCKR** | rs780093-T | 1.15 | [1.10-1.21] |
| **HLA-C,PSORS1C1,NFKBIL1,MICB** | rs9264942-C | 1.15 | [1.107-1.184] |
| **NR** | rs9469220-A | 1.14 | [0.98-1.32] |
| **C1orf106,KIF21B** | rs7554511-C | 1.14 | [1.08-1.19] |
| **Intergenic** | rs12035082-? | 1.14 | [1.02-1.27] |
| **CCL7, CCL2, CCL11, CCL8, CCL13, CCL1** | rs3091316-G | 1.14 | [1.03-1.27] |
| **IKZF1,ZPBP,FIGNL1** | rs1456896-T | 1.14 | [1.09-1.20] |
| **C2orf74,REL** | rs10181042-T | 1.14 | [1.09-1.19] |
| **GSMDL,ZPBP2,ORMDL3,IKZF3** | rs2872507-A | 1.14 | [1.09-1.19] |
| **ITLN1** | rs2274910-C | 1.14 | [NR] |
| **THADA** | rs10495903-T | 1.14 | [1.09-1.20] |

| | | | |
|---|---|---|---|
| **LGALS9,NOS2** | rs2945412-A | 1.14 | [1.10-1.175] |
| **PRDM1** | rs6568421-G | 1.13 | [1.07-1.18] |
| **Intergenic** | rs17391694-C | 1.13 | [1.077-1.194] |
| **ICOSLG** | rs762421-G | 1.13 | [NR] |
| **SCAMP3,MUC1** | rs3180018-A | 1.13 | [1.06-1.19] |
| **SP140** | rs6716753-C | 1.13 | [1.089-1.18] |
| **FASLG,TNFSF18** | rs9286879-G | 1.13 | [1.083-1.167] |
| **ZDHHC23** | rs1386478-A | 1.12 | [NR] |
| **IL6ST,IL31RA** | rs10065637-C | 1.12 | [1.079-1.17] |
| **TYK2,ICAM1,ICAM3** | rs12720356-G | 1.12 | [1.06-1.19] |
| **SP140** | rs7423615-T | 1.12 | [1.07-1.18] |
| **IL10,IL19** | rs3024505-T | 1.12 | [1.07-1.17] |
| **RTEL1,TNFRS-F6B,SLC2A4RG** | rs4809330-G | 1.12 | [1.06-1.18] |
| **UCN** | rs1728918-A | 1.12 | [1.086-1.16] |
| **IFNGR2,IFNAR1,IFNAR2,IL10RB,GART,TMEM50B** | rs2284553-G | 1.12 | [1.086-1.162] |
| **SMAD3** | rs17293632-T | 1.12 | [1.07-1.16] |
| **CAPZB** | rs7667-? | 1.11 | [NR] |
| **IL2RA** | rs12722489-C | 1.11 | [1.05-1.16] |
| **TAGAP** | rs212388-C | 1.11 | [1.069-1.141] |
| **DBP,SPHK2,IZUMO1,FUT2** | rs516246-T | 1.11 | [1.071-1.143] |
| **intergenic** | rs17309827-T | 1.10 | [1.05-1.16] |
| **CPEB4** | rs17695092-T | 1.10 | [1.055-1.136] |
| **Intergenic** | rs10865331-A | 1.10 | [1.062-1.134] |
| **PRDX5,ESRRA** | rs694739-A | 1.10 | [1.05-1.16] |
| **TNFSF11** | rs2062305-G | 1.10 | [1.05-1.15] |
| **Intergenic** | rs4802307-G | 1.10 | [1.06-1.139] |
| **Intergenic** | rs12663356-C | 1.10 | [1.06-1.131] |
| **YDJC** | rs181359-T | 1.10 | [1.06-1.15] |
| **BSN, MST1** | rs9858542-A | 1.09 | [0.96-1.24] |
| **Intergenic** | rs10486483-A | 1.09 | [1.048-1.13] |
| **TXK,TEC,SLC10A4** | rs6837335-G | 1.09 | [1.049-1.123] |
| **CREB5,JAZF1** | rs864745-T | 1.09 | [1.052-1.123] |
| **RASGRP1,SPRED1** | rs16967103-C | 1.09 | [1.045-1.132] |
| **Intergenic** | rs7702331-A | 1.09 | [1.05-1.126] |
| **RIPK2** | rs7015630-T | 1.08 | [1.035-1.116] |
| **intergenic** | rs13073817-A | 1.08 | [1.03-1.13] |
| **Intergenic** | rs1551398-A | 1.08 | [NR] |
| **Intergenic** | rs9491697-G | 1.08 | [1.042-1.112] |
| **FADS1** | rs102275-C | 1.08 | [1.04-1.12] |
| **MTMR3** | rs713875-C | 1.08 | [1.04-1.13] |
| **CPEB4** | rs359457-T | 1.08 | [1.04-1.12] |

| | | | |
|---|---|---|---|
| **BACH2** | rs1847472-G | 1.07 | [1.03-1.11] |
| **ZFP36L1** | rs4902642-G | 1.07 | [1.11-1.04] |
| **IL27,SH2B1,EIF3C,LAT,CD19** | rs151181-G | 1.07 | [1.03-1.12] |
| **FUT2,RASIP1** | rs281379-A | 1.07 | [1.04-1.11] |
| **intergenic** | rs736289-T | 1.06 | [1.02-1.11] |
| **PLCL1** | rs6738825-A | 1.06 | [1.02-1.11] |
| **NDFIP1** | rs11167764-C | 1.06 | [1.02-1.11] |
| **DNMT3A** | rs13428812-G | 1.06 | [1.03-1.10] |
| **VAMP3** | rs2797685-A | 1.05 | [1.01-1.10] |
| **PUS10, PEX13, REL, KIAA1841, C2orf74, PAPOLG, USP34** | rs13003464-G | 1.05 | [1.00-1.40] |
| **ERAP2,LRAP** | rs2549794-C | 1.05 | [1.02-1.09] |
| **DENND1B** | rs1998598-G | 1.04 | [1.00-1.09] |
| **FUT2** | rs504963-A | NR | NR |
| **PUS10** | rs10188217-C | NR | NR |
| **IL23R** | rs7517847-? | NR | NR |
| **NOD2** | rs5743289-? | NR | NR |

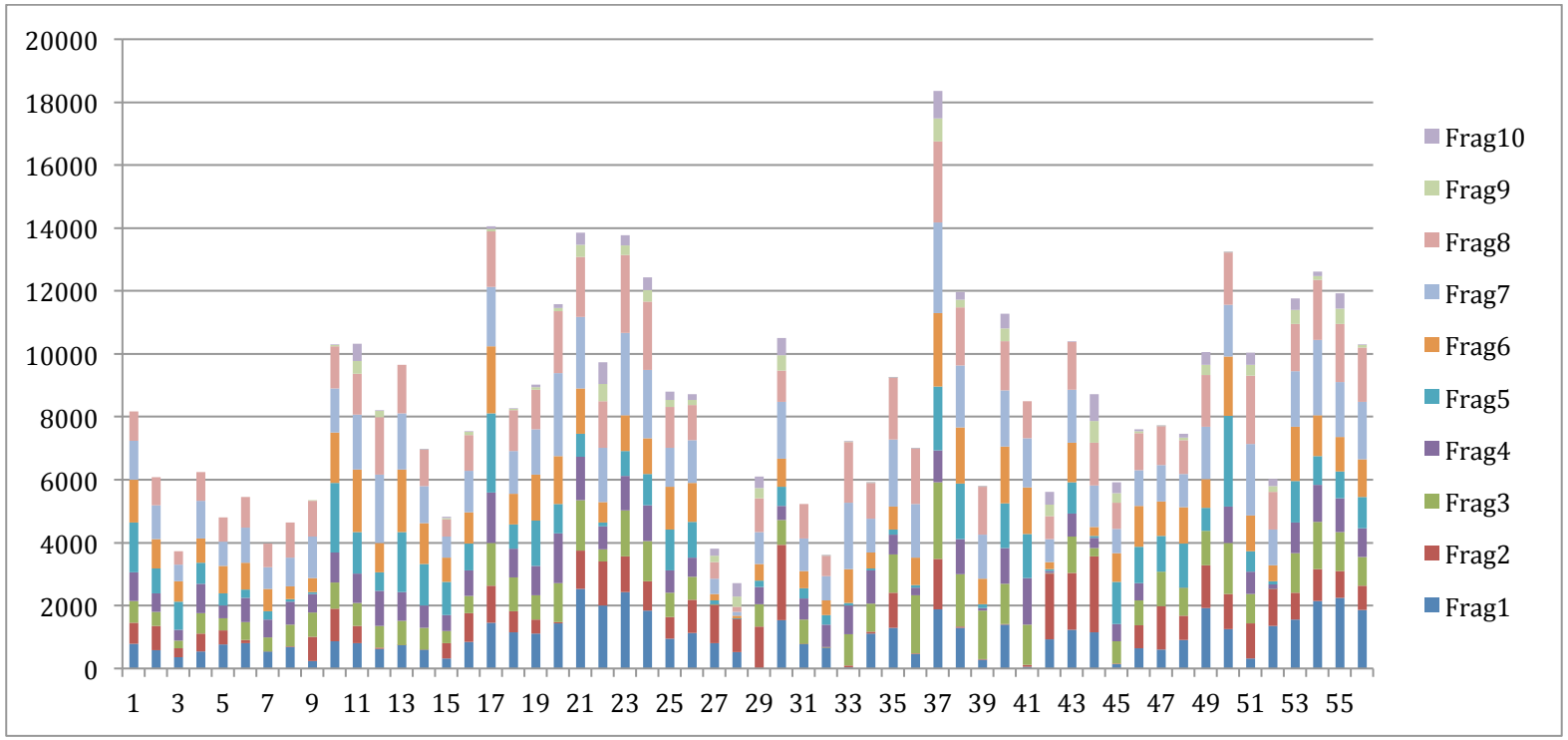



**Abbildung 14-1:** Coverage-Diagramm für den CED-Locus IL23R. Das Diagramm zeigt die Coverages fragmentweise für jede der 56 sequenzierten Libraries.

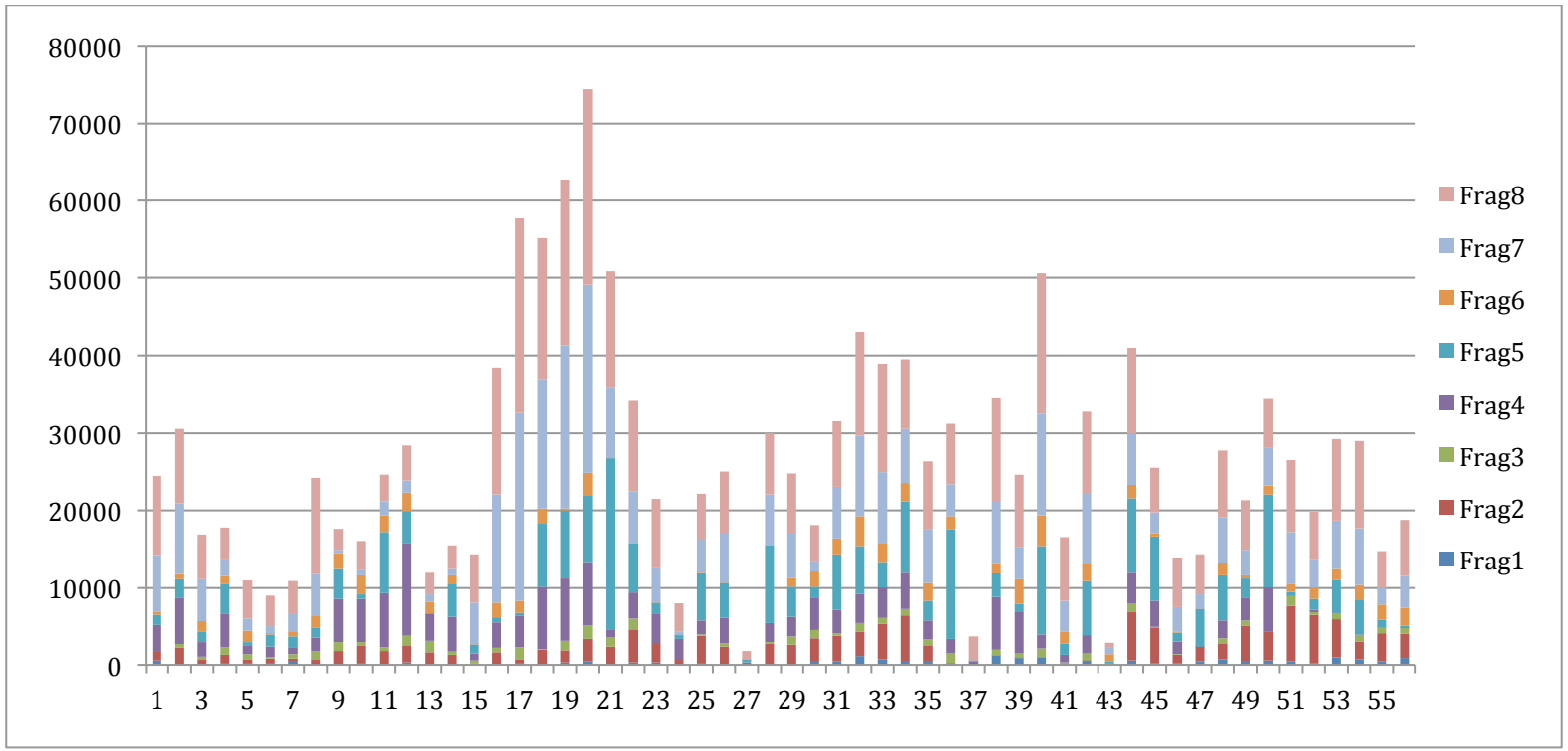



**Abbildung 14-2:** Coverage-Diagramm für den CED-Locus ATG16L1. Das Diagramm zeigt die Coverages fragmentweise für jede der 56 sequenzierten Libraries.

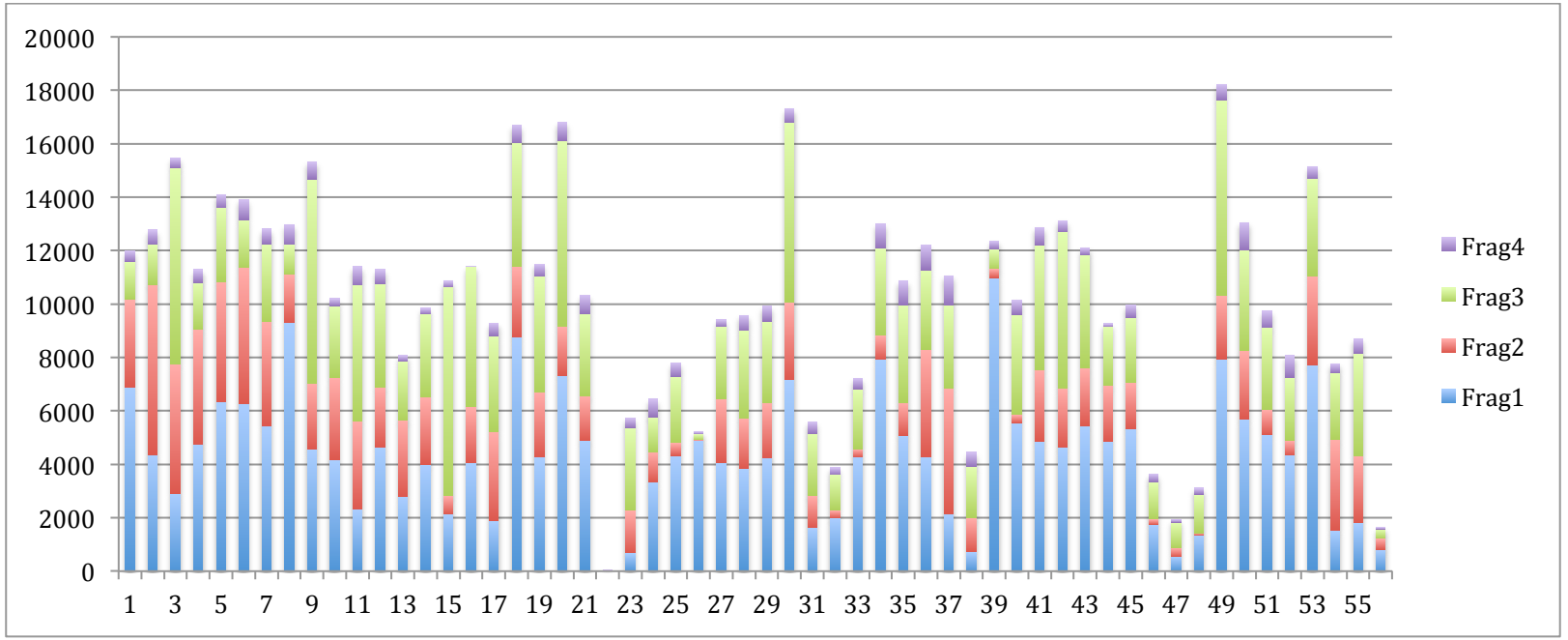



**Abbildung 14-3:** Coverage-Diagramm für den CED-Locus NKX2-3. Das Diagramm zeigt die Coverages fragmentweise für jede der 56 sequenzierten Libraries.

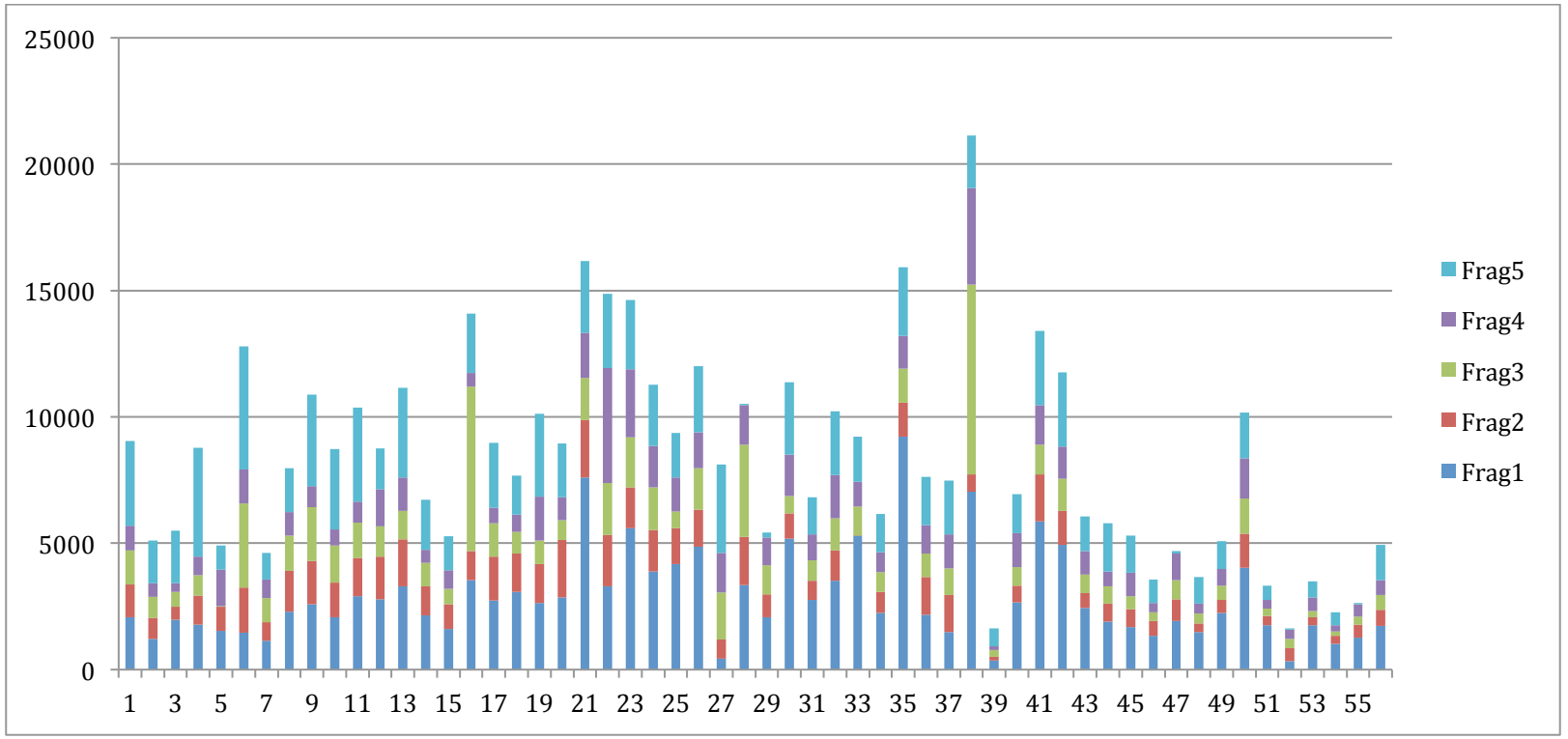



**Abbildung 14-4:** Coverage-Diagramm für den CED-Locus NOD2. Das Diagramm zeigt die Coverages fragmentweise für jede der 56 sequenzierten Libraries.

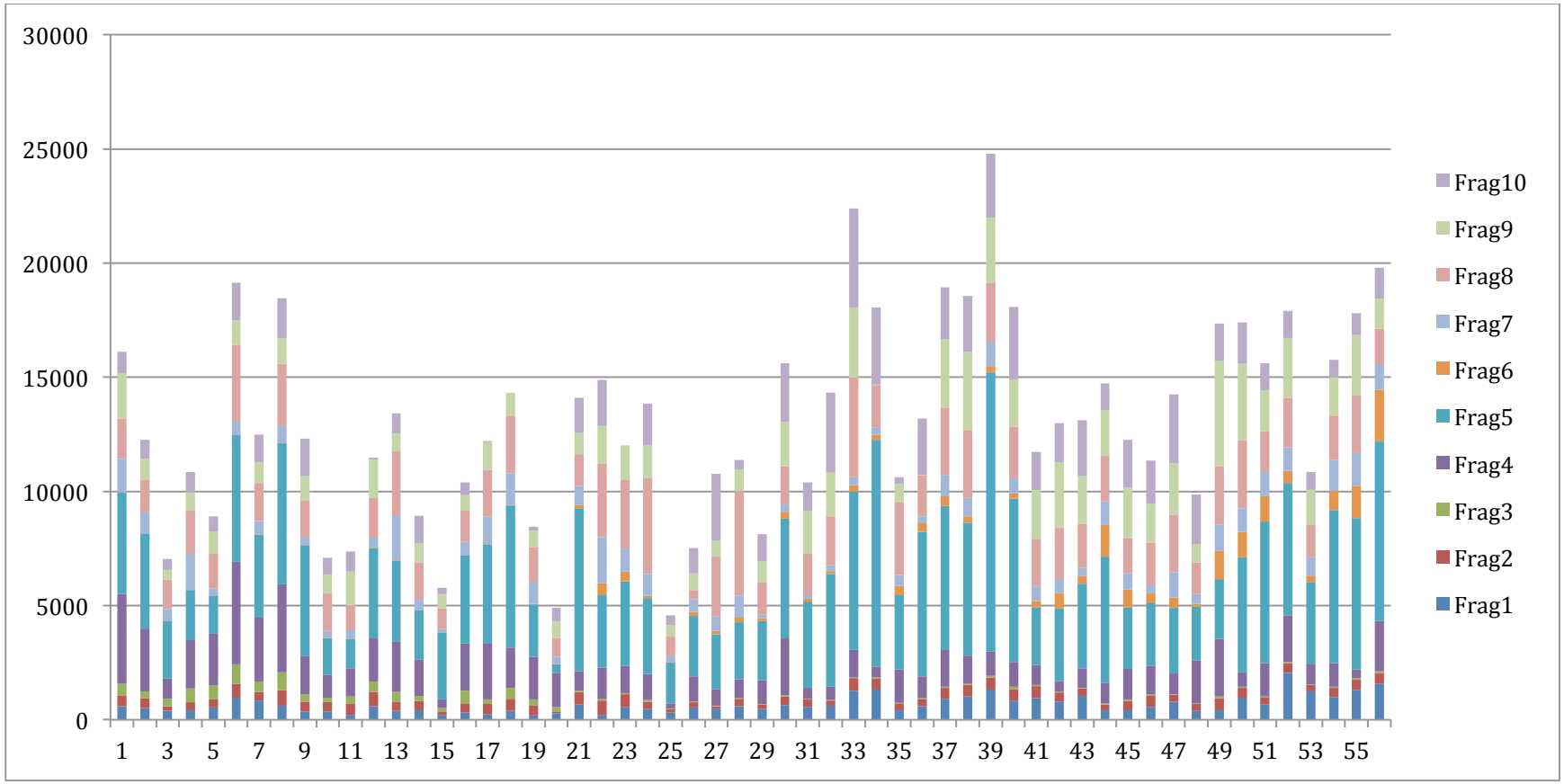



**Abbildung 14-5:** Coverage-Diagramm für den CED-Locus STAT3. Das Diagramm zeigt die Coverages fragmentweise für jede der 56 sequenzierten Libraries.

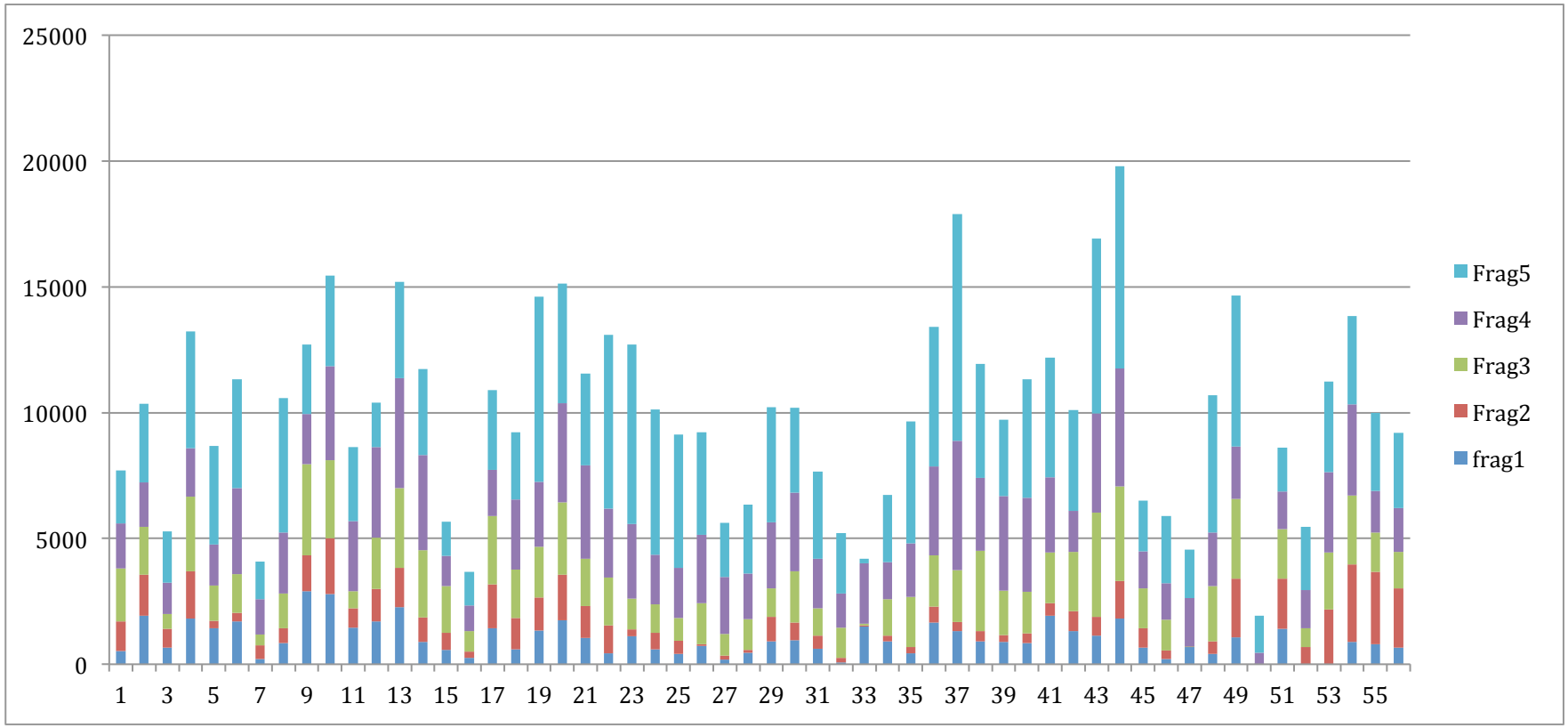



**Abbildung 14-6:** Coverage-Diagramm für den CED-Locus IL10. Das Diagramm zeigt die Coverages fragmentweise für jede der 56 sequenzierten Libraries.

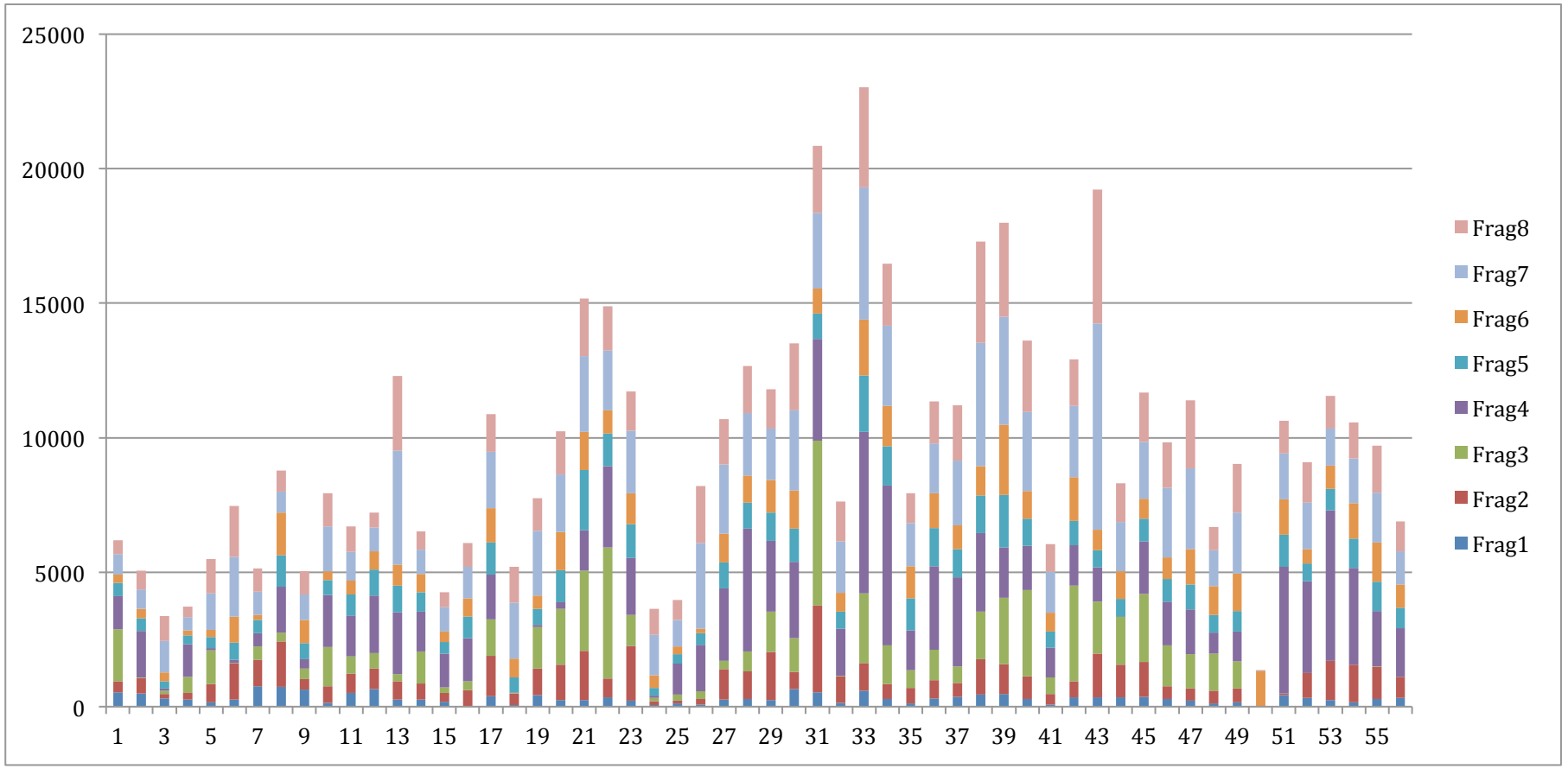



**Abbildung 14-7:** Coverage-Diagramm für den CED-Locus IRGM. Das Diagramm zeigt die Coverages fragmentweise für jede der 56 sequenzierten Libraries.

**Tabelle 14-2:** Die Tabelle zeigt die Sequenzen für die in der LR-PCR verwendeten Primer im Enrichmentprozess für die CED-Loci an.

| Primer_ID | Forward | revers | Länge |
|---|---|---|---|
| **ATG16L1_1** | TGTTAGTTGGTTTCTTGACTATCCTTT | TAGGTGATTCTAATGTGGAGCTAAGTT | 6053 |
| **ATG16L1_2** | CCTAAATTATAATGGTCCTAAGGTGGT | TAAGAGTCTGAACCTCAAGAAGAAGTC | 9157 |
| **ATG16L1_3** | GATTTAGGATGGTTTCATGTAAAGTGT | TTTAAACACTATGCATTTTTGACTCTG | 10974 |
| **ATG16L1_4** | ATTTGGTTTCTTCAGAATAACAGCTTA | CTTGGGGAAGAGAATATGTGATATTTA | 5119 |
| **ATG16L1_5** | AAGTTTTCAATCTGAGACCTCCTAAG | ATACATTAAAACAGGGCACTGAAATAC | 5070 |
| **ATG16L1_6** | ATTTAATCAAATCATTGAGATTTTTGC | AGGAAACATCCTTGTGCTATCTTAGTA | 6848 |
| **ATG16L1_7** | ATGAGCAATTTTGTGTTTAGAAATAGG | AAGTCCTTAACCATCTACAGACTTCAA | 6811 |
| **ATG16L1_8** | TTAGTTTGCTTTTAAAACCTGAAATTG | GATGATCTGAGACTACTCGAACACTTA | 5116 |
| **IL10_1** | AAGAGCGGTGTCTAGTTTAGGTTTTAG | GTTATACGTGGACTAACCAGATTTCATA | 7930 |
| **IL10_2** | ATTATTTTTCCCAATTCTAATAGCACAC | CTCCTCTGAATAATAACAGGAAGAAAGT | 11769 |
| **IL10_3** | TTATAGAACAAGCCATAATCTCAACAGT | TTAAGCCATTTGGTATGTATTATGTAGC | 8448 |
| **IL10_4** | AAAACACATATGTACCTTCATACCTGAG | GTGAAGACTTTCTTTGTGAGTATGATTC | 8576 |
| **IL10_5** | TACCTGGAGGAATTACTCATAGACACTA | ATCTTCAAAAACTTGAGAAATAGGAGTC | 5285 |
| **IRGM_1** | TGAAAGAATAAACATTATTGGGTAAGG | TCATTCCTTGTGGTCAGATAAAATACT | 8960 |
| **IRGM_2** | AAGTATTTTATCTGACCACAAGGAATG | TGTGTGATAAGTGCTTAGTTAAGATGG | 8099 |
| **IRGM_3** | ACTAAAAGCCACCATATCTAAGTAGCA | CTTTACGGATTCACTATAGATCTCAGG | 9439 |
| **IRGM_4** | TTACTTCACAGCATTCTACATTCAAAC | AGCTGATCCTCTTACAACTATACGAAA | 8459 |
| **IRGM_5** | CCTTTCATATTCTGATTACCCACTAAA | GTAAAACTAAGAGATGCCCTAACTCCT | 9803 |
| **IRGM_6** | GTAAAAGTTCTAGGAGTTCAGTGGTTG | GCGGTAGAGAAGTCCTTTACTATTTTT | 8374 |
| **IRGM_7** | CATCACTGCAATAATTTAAGGACTACA | ACTTAATAGCAGAGCTTCACCATACAT | 9455 |
| **IRGM_8** | CTAAAGAAGCTTTTGAGCTGAGACTA | TTCATATACTTACATAAACCGTGGTGA | 8639 |
| **NKX2-3_1** | AGAAAAGCTTATACCTTGATTTTAGCC | CTCCAATTTTGAAGGAAGTTACTGTAG | 6101 |
| **NKX2-3_2** | TAGAATCTTCTAACAAACACCCCTAGA | CCTAGTGCTTTTAGTCTCCACATAATC | 6149 |
| **NKX2-3_3** | GAAGTTGTCTTTAGATAAACCGTTCAC | TTAGTAGTGGGGTAACAGAGAAAGAAA | 7675 |
| **NKX2-3_4** | ACAATAGAAGGAAATGGAGAAGGAT | TGTCCTCTTAAATACACTTCCTTGG | 7840 |
| **STAT3_1** | GACAGCAGCTTATAAACCACCTTATAG | ATTTGAGAGGAGAATCAGAATCACTTA | 8645 |
| **STAT3_2** | AAAGAACCCAGAGAGTCAAAACAG | CTTTTCATTGATTCAGATTATTGGATT | 8051 |
| **STAT3_3** | CTTCTAAGCATTCAAGCAAGACTTTAT | AACTCTACATGAGAATGCCTGTCTATT | 10233 |
| **STAT3_4** | TAAAATTTATACCACTAGGAGGCATTG | GTGTATTCTTGGGAGAAGTGTAGAAAG | 8054 |
| **STAT3_5** | GAGCCTCTATCCAAATTGACTTAAAC | CAATTTCTACATTTTGGCTACAGTACA | 7470 |
| **STAT3_6** | ACTATGAGCCGATTAAACCTCTTTT | TAGCGTGAATTGCTTTCTCTTACTT | 9411 |
| **STAT3_7** | AAATCTATTTCAGAATGTGTTTGTGTG | TTATGGACTTAGCTAGATTCAAAGGAA | 8647 |
| **STAT3_8** | AAATTCCTCAGATGATGCTAAAGTG | TATCTAGGGAAAGAGCCTATCAGGT | 6087 |
| **STAT3_9** | ATCAAGTAGGTGTTATTCCCATTTACA | TCTTGGAGACATATTTATTCCAAATTC | 7055 |
| **STAT3_10** | ACGGTTTGAATCTTGTTAACTTCAG | AGGAAAGAGACATATCTGCATAACG | 7021 |
| **NOD2_1** | GGAGTGGGCCTTGGAGTC | GTCCAGGACACTCTCGAAGC | 5963 |
| **NOD2_2** | CTGGCACTCTTTAGGGCTTG | GCTGCAACTGAATCCAGACA | 10722 |
| **NOD2_3** | GTTCTCCTAGCTGCCACACC | GTCCACACAACCGCTCCTAT | 4276 |
| **NOD2_4** | CCTGGTGGGGAACAACATT | AGAGCAGGGCTTTCAAACAA | 8402 |
| **NOD2_5** | GCAGCGAATGCAGATATCAA | GCGAAAGGAGACTCAACACC | 3053 |
| **IL23R_1** | TAACAGGTCCTCTCTGATCACTACC | TTAGCAGATAATCGTGAAGATAGGG | 8600 |

| | | | |
|---|---|---|---|
| **IL23R_2** | TTTATCAGTGCAGGATACCTCCCC | CAGCTCCTGTAAACTTTTATCAAGG | 8912 |
| **IL23R_3** | AGTACCGACAAAACAGTTCACTTTC | TGCACATTTTCCACTACTCATTTTA | 9470 |
| **IL23R_4** | GTGAGGTCAGGTAATGAATGTATCC | TCTTTTACCACACATACTGAGACCA | 9486 |
| **IL23R_5** | CTTCATATCCACCCCCATTG | CCTGGCCAACATAGCAAAAT | 3498 |
| **IL23R_6** | ATCACAGAGGTAGGGAACATTTAATTC | CATTTTTACTCGGATTATCATCATCTT | 7501 |
| **IL23R_7** | TTCAGAGACAGTATTTGATTATGT | CTTTATATATCTTTTCTGCTGAGCACAGTGG | 396 |
| **IL23R_8** | GTCATCATAAAAGACAACATAGGGAAT | GCTAGTTTAATCAAATCCACTCAAAAG | 7606 |
| **IL23R_9** | AAAAACCCACTGACATCAGATACAT | GAGAAACAACTGGTATTCAAGAGGA | 3249 |
| **IL23R_10** | CAGGAATATCTTTGTGGTGTTCTCTAT | TTTCCTCCTCTACTGAGTTTTGTATGT | 10842 |