**Technische Universität Braunschweig**

Fakultät für Lebenswissenschaften

# Untersuchungen zur Verwendung der Next Generation Sequenziertechnologie für die Analyse von Antikörpergenen bei Gesunden und Autoimmunpatienten

**Diplomarbeit**

zur Erlangung des akademischen Grades
Diplom-Biotechnologin

angefertigt in Kooperation mit dem
**Max-Planck-Institut für molekulare Genetik**

vorgelegt von
**Svetlana Mollova**

Braunschweig/Berlin, November 2009

Referent: Prof. Dr. Stefan Dübel
Korreferent: Prof. Dr. Hans Lehrach

*Man muss das Unmögliche versuchen, um das Mögliche zu erreichen.*

Hermann Hesse

# Inhaltsverzeichnis

# Abkürzungsverzeichnis

## I. Abkürzungen

| | | |
|---|---|---|
| **A:** | a | Atto |
| **B:** | bp | Basenpaar(e) (aus *"base pair(s)"*) |
| **C:** | C | konstante Region eines Antikörpers (aus *"constant region"*), Cytosin |
| | °C | Grad Celsius (Einheit der Temperatur) |
| | cDNA | komplementäre Desoxyribonukleinsäure (aus *"complementary deoxyribonucleic acid"*) |
| | CDR | Antigenerkennungs- und -bindungsstelle des Antikörpers, Komplementarität-bestimmende Region (aus *"complementarity-determining region"*) |
| | CH1 | konstante Region 1 der schweren Immunglobulin-Kette |
| | CH2 | konstante Region 2 der schweren Immunglobulin-Kette |
| | CH3 | konstante Region 3 der schweren Immunglobulin-Kette |
| | csv | Dateiformat mit Komma-getrennten Werten (aus *"comma-separated values file format"*) |
| **D:** | D | Diversitätsgensegment (aus *"diversity gene segment"*) |
| | dATP | Desoxyadenosintriphosphat |
| | dCTP | Desoxycytidintriphosphat |
| | $ddH_2O$ | doppelt destilliertes Wasser |
| | dGTP | Desoxyguanosintriphosphat |
| | DNA | Desoxyribonukleinsäure (aus *"deoxyribonucleic acid"*) |
| | DNase I | Desoxyribonuklease I |
| | dNTPs | Desoxynukleosidtriphosphate (Desoxynukleotide) |
| | ds | doppelsträngig |
| | dTTP | Desoxythymidintriphosphat |
| **E:** | EDTA | Ethylendiamintetraessigsäure (aus *"ethylene-diamine-tetraacetic acid"*) |
| | et al. | et alii |

| | | |
|---|---|---|
| **F:** | _f, Fw | Forward-Primer |
| | Fab | Antigen-bindendes Fragment eines Antikörpers (aus *"fragment antigen binding"*) |
| | FR | Gerüstregion (aus *"framework region"*) |
| | Fv | variables Fragment eines Antikörpers (aus *"fragment variable"*) |
| **G:** | g | Gramm; Erdbeschleunigung |
| **H:** | H, HC | schwere Kette (aus *"heavy chain"*) |
| | hum | human |
| **I:** | IG, Ig | Immunglobulin |
| | IgA | Immunglobulin A |
| | IgD | Immunglobulin D |
| | IgE | Immunglobulin E |
| | IgG | Immunglobulin G |
| | IgM | Immunglobulin M |
| | IGH | schwere Immunglobulin-Kette (s. IG, H) |
| | IGHD | Diversitätsgensegment der schweren Immunglobulin-Kette (s. IG, H, D) |
| | IGHJ | Verbindungs-Gensegment der schweren Immunglobulin-Kette (s. IG, H, J) |
| | IGHV | variable Region oder variables Gensegment der schweren Immunglobulin-Kette (s. IG, H, V) |
| | IGK | leichte Kappa Immunglobulin-Kette (s. IG, K) |
| | IGKJ | Verbindungs-Gensegment der leichten Kappa Immunglobulin-Kette (s. IG, K, J) |
| | IGKV | variable Region oder variables Gensegment der leichten Kappa Immunglobulin-Kette (s. IG, K, V) |
| | IGL | leichte Lambda Immunglobulin-Kette (s. IG, L) |
| | IGLJ | Verbindungs-Gensegment der leichten Lambda Immunglobulin-Kette (s. IG, L, J) |
| | IGLV | variable Region der leichten Lambda Immunglobulin-Kette (s. IG, L, V) |
| **J:** | J | Verbindungs-Gensegment (aus *"joining gene segment"*) |
| **K:** | κ | leichte Kappa Kette |

| | | |
|---|---|---|
| **L:** | λ | leichte Lambda Kette |
| | L | leichte Kette; Liter |
| | LC | leichte Kette |
| **M:** | µ | Mikro |
| | m | Milli |
| | M | Molarität |
| | min | Minute |
| | mol | Mol, Einheit für Stoffmenge |
| | MPIMG | Max-Planck-Institut für molekulare Genetik |
| | mRNA | Boten-Ribonukleinsäure (aus *"messenger ribonucleic acid"*) |
| **N:** | n | Nano |
| | N | nicht in der DNA-Matrize kodiertes Nukleotid (aus „*non-templated nucleotide*“) |
| | ND | gesunde Probanden |
| | NTC | Negativkontrolle (Ansatz ohne Template-DNA) (aus *„non-template control"*) |
| **P:** | p | Piko |
| | P | palindromisches Nukleotid |
| | PCR | Polymerase-Kettenreaktion (aus *"polymerase chain reaction"*) |
| | PDGF | aus „*platelet-derived growth factor*“ |
| **Q:** | qRT-PCR | quantitative *Real-Time*-Polymerase-Kettenreaktion (aus *"quantitative real-time polymerase chain reaction"*) |
| **R:** | RA | rheumatoide Arthritis |
| | RNA | Ribonukleinsäure (aus *"ribonucleic acid"*) |
| | RNase | Ribonuklease |
| | rpm | Umdrehungen pro Minute (aus *"revolutions per minute"*) |
| | RT | reverse Transkription |
| | Rv | Reverse-Primer |
| **S:** | s | Sekunde |
| | SA | Standardabweichung |
| | sc | Einzelkette (aus *"single chain"*) |

| | | |
|---|---|---|
| **S:** | scFv | einzelkettiges variables Fragment eines Antikörpers (aus *"single chain fragment variable"*) |
| | SD | Sklerodermie |
| | SLE | systemischer Lupus erythematodes |
| | snRNA | kleine nukleare RNA (aus *"small nuclear RNA"*) |
| | ss | einzelsträngig (aus *"single stranded"*) |
| **T:** | TBE | Tris-Borat-EDTA |
| | TD-PCR | Touchdown Polymerase-Kettenreaktion (aus *"touchdown polymerase chain reaction"*) |
| | Tris | Tris(hydroxymethyl)-aminomethan |
| | TU | Technische Universität |
| **U:** | U | Einheit der Enzymaktivität (aus *"unit"*) |
| **V:** | V | variable Domäne oder Region eines Antikörpers; variables Gensegment |
| **W:** | % w/v | Gewichtsprozent (aus *"weight per volume"*) |

II. Nukleotidbezeichnungen [1]

| | |
|---|---|
| A | Adenin |
| C | Cytosin |
| G | Guanin |
| T | Thymin |
| B | C oder G oder T |
| D | A oder G oder T |
| H | A oder C oder T |
| K | G oder T |
| M | A oder C |
| N | A oder C oder G oder T |
| R | A oder G |
| S | C oder G |
| V | A oder C oder G |
| W | A oder T |
| Y | C oder T |

# 1. Einleitung

## 1.1. Antikörper: Bildung, Struktur, Funktion

Antikörper (Immunglobuline, Ig) sind Glykoproteine der adaptiven Immunantwort. Sie werden von B-Lymphozyten (B-Zellen) hergestellt und kommen in Körperflüssigkeiten wie Blutserum und Sekreten vor, wo sie Antigene (Bakterien, Viren sowie Fremdstoffe) spezifisch erkennen und binden. Je nach Antigenart können folgende Antikörper-Effektormechanismen ausgelöst werden: Neutralisation (Verhinderung der Wechselwirkung zwischen Antigenen und Zellen), Opsonisierung (Umhüllung mit Antikörpern) oder Aktivierung des Komplementsystems (System von Plasmaproteinen). Unabhängig vom Effektormechanismus werden die Antigene durch Phagozytose abgewehrt. [2, 3, 4].

Grundstruktur von Immunglobulinen

IgG Antikörper bestehen aus insgesamt vier Ketten: zwei identische leichte Ketten (LC) des Isotyps Kappa (κ) oder Lambda (λ) und zwei identische, durch Disulfidbrücken verbundene schwere Ketten (HC). Die LC enthält eine variable (VL) und eine konstante (CL) Domäne, während die HC aus einer variablen und drei konstanten Domänen aufgebaut ist (VH, CH1, CH2 und CH3). Zwischen den CH1 und CH2 Domänen befindet sich die Gelenkregion („Hinge“) des IgG Antikörpers (Abbildung 1A) [2, 3].
Für die Antigenerkennung und -bindung sind im nativen Antikörper je drei komplementaritätsbestimmende Regionen (CDRs) der HC und der LC zuständig. Sie werden durch je vier für die Grundstruktur zuständige Gerüstregionen (FRs) flankiert und bilden zusammen die variablen (V) Domänen (Abbildung 1B) [2, 3].
Neben kompletten IgG Antikörpern werden für zahlreiche Anwendungen (z.B. Phage Display) Antikörper-Fragmente eingesetzt. Ein verbreitetes rekombinantes Format ist das einzelkettige, variable Fragment eines Antikörpers (scFv). Es besteht aus VH und VL Regionen, die mit einem flexiblen Peptidlinker verbunden sind. Das scFv Format besitzt die gleiche Antigenbindungsspezifität wie native Antikörper [5, 6].

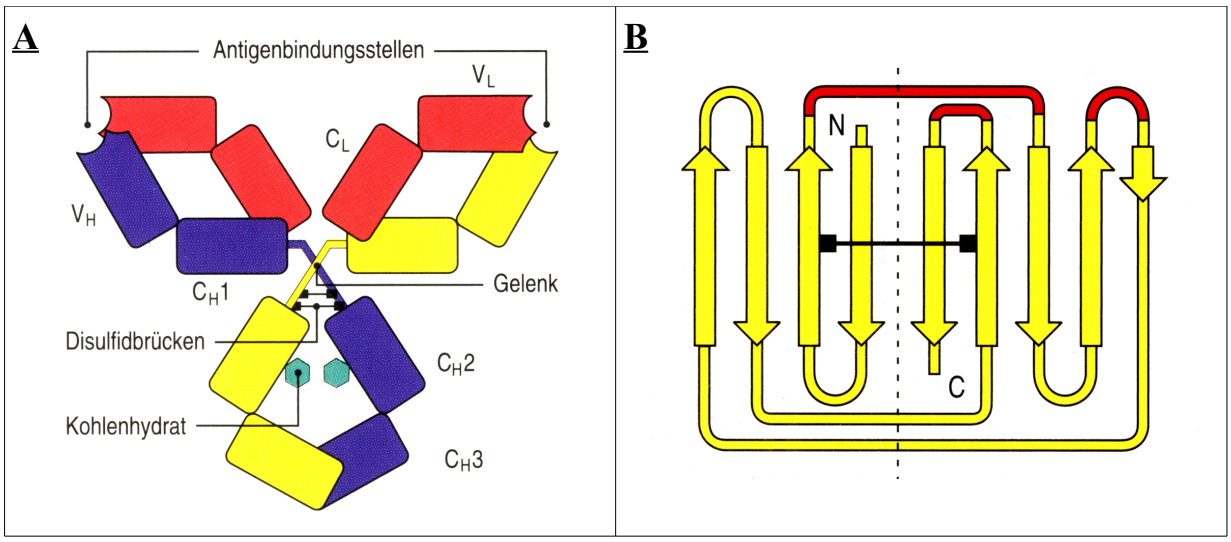



**Abbildung 1:** Grundstruktur eines Antikörpers (A) sowie einer Ig V Domäne (B). B: die CDRs sind rot gekennzeichnet, während die FRs gelb dargestellt sind (adaptiert von [2]).

## Antikörper-Isotypen

Es existieren fünf Ig-Isotypen. Sie unterscheiden sich in den CH Regionen, welche für die Interaktion mit anderen in der Immunantwort agierenden Molekülen und Zellen zuständig sind (Abbildung 2). *IgM* wird als erster Isotyp bei einer Primärantwort hergestellt und bewirkt keine längerfristige Immunität. Seine monomere Form liegt membrangebunden bei unreifen und reifen B-Lymphozyten vor, während die pentamere Form etwa 10 % der Antikörper im Blutserum beträgt. *IgD* ist ein Monomer, das entweder gebunden auf der Oberfläche von reifen B-Zellen oder löslich im Blutserum vorhanden ist. Es zeichnet sich durch eine verkürzte Halbwertszeit aus, weniger als 1 % der Immunglobuline im Serum sind IgD [2, 3, 4]. Seine genaue Funktion ist unbekannt, obwohl er bei vielen Erkrankungen eine Rolle spielen könnte [7]. *IgG* ist ein Monomer und der häufigste Isotyp im Blutserum (70-75 %). Es existieren vier IgG Subklassen (IgG1 – IgG4), die nach Häufigkeit ihres Vorkommens durchnummeriert sind. *IgE* ist ein Monomer, das nach der Sekretion ins Blutserum an Mastzellen gebunden wird. Seine Konzentration steigt bei Parasitenabwehr und bei Allergien. *IgA* kommt einerseits als Monomer im Blutserum und andererseits in dimerer Form als häufigster Ig in Sekreten vor. Dementsprechend schützt es als erstes Ig gegen Antigene im respiratorischen und gastrointestinalen Trakt [2, 3, 4].

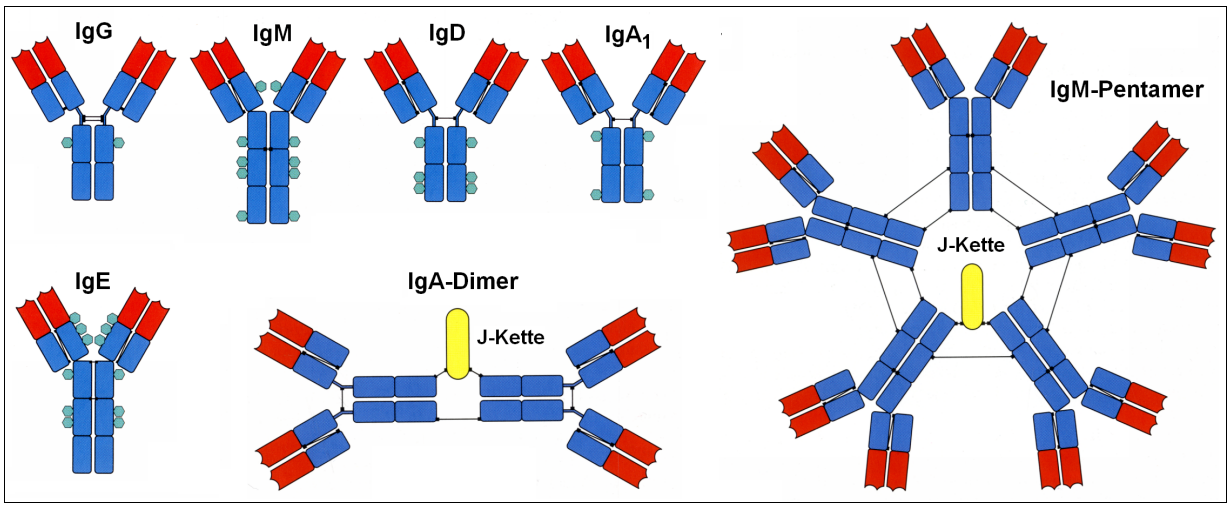



**Abbildung 2:** Monomere und polymere Ig Formen. Dargestellt sind VH bzw. VL (rot), CH bzw. CL (blau) und Disulfidbrücken (schwarze Linien). Die Kohlenhydratseitenketten (türkis) sind nur in den monomeren Formen veranschaulicht. Die J-Kette (gelb) kommt in den polymeren Ig Formen vor. IgM sowie IgE besitzen statt Hinge Region noch eine zusätzliche CH Domäne (adaptiert von [2]).

Genetische Grundlagen der Antikörperbildung *in vivo*

Die V(D)J Rekombination bildet die Basis der Antikörperbildung (Abbildung 3): variable (V), Diversitäts- (D) und Verbindungs-Gensegmente (J) werden bei der HC kombiniert (V und J bei der LC), um funktionelle Ig-Gene durch Deletion oder Inversion der DNA-Abschnitte zwischen den Genesegmenten zu bilden. Im Genom befinden sich mehrere V, D und J Gensegmente, die aufgrund Homologien in Genfamilien aufgeteilt sind. Einige der V Gene sind Pseudogene – sie ähneln funktionellen Genen, können aber aufgrund von Defekten (Auftreten von Stopcodons) keine funktionellen VDJ *Rearrangements* bilden [2, 3, 8]. Andere V Gene sind die so genannten Orphans (Weisen), die aufgrund räumlicher Trennung nicht rekombinieren können (Locus auf einem anderen Chromosom) [8]. Somit entsteht bei Berücksichtigung der funktionellen Gene in der VBASE2 Datenbank [8] eine theoretische kombinatorische Antikörperdiversität von über $10^6$. Die Diversität wird durch das Ausschneiden von bestehenden Nukleotiden an den Enden der Gensegmente sowie durch das Hinzufügen von zusätzlichen palindromischen (P) oder nicht in der DNA-Matrize kodierten (N) Nukleotiden zwischen den Gensegmenten weiter erhöht. Sie nimmt durch die somatische Hypermutation (Auftreten von zufälligen Punktmutationen in den VH und VL Ketten) bei der Affinitätsreifung weiter zu [2, 3].

Nicht funktionelle *Rearrangements* entstehen einerseits, indem V Pseudogene mit D und J rekombinieren. Andererseits kann durch die Deletion und/oder Insertion von Nukleotiden bei der V(D)J Rekombination eine Leserasterverschiebung (*Frameshift*) im J Segment auftreten

(*Out-of-Frame Rearrangement*). Zusätzlich könnten während der VDJ Rekombination oder der somatischen Hypermutation Stopcodons entstehen [2, 3, 9, 10].

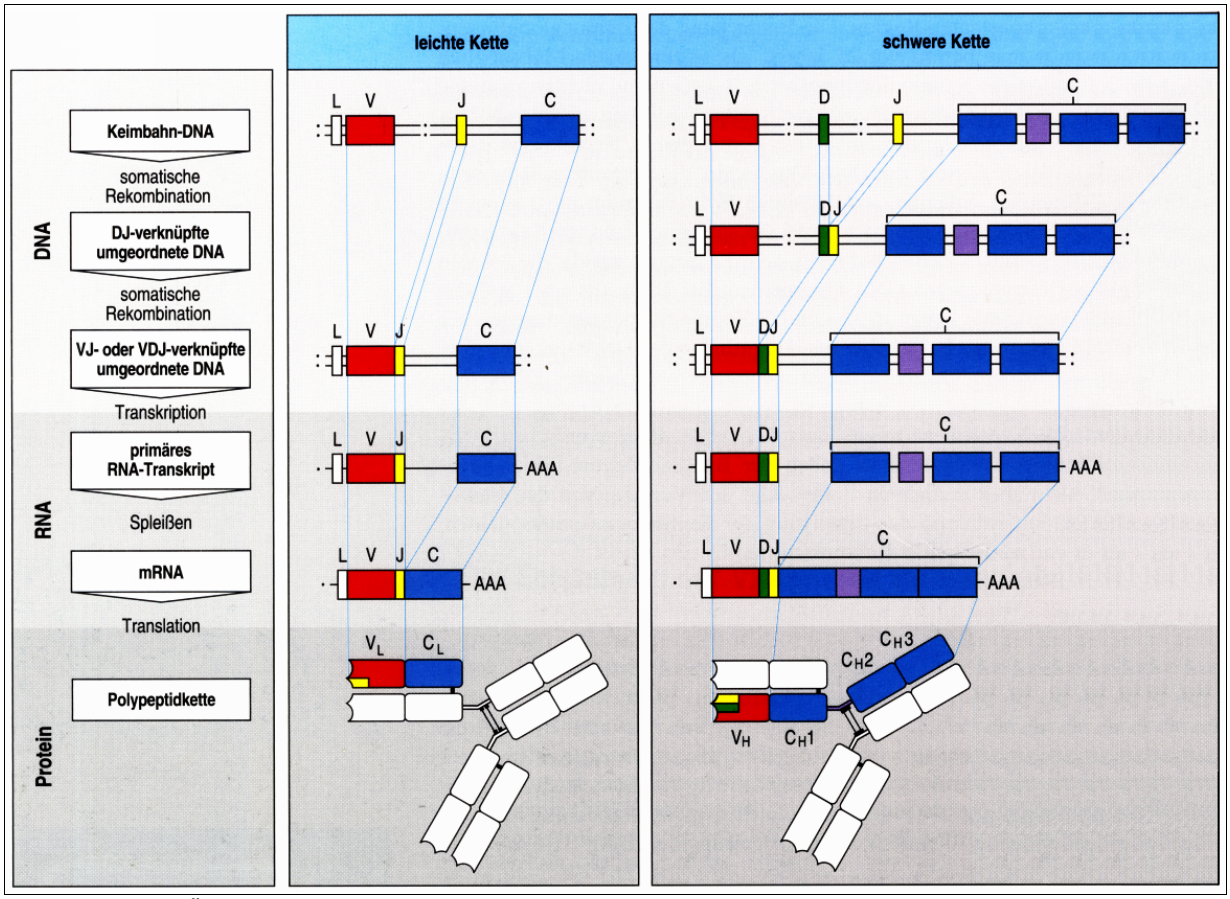



**Abbildung 3:** Übersicht der V(D)J Rekombination sowie der Transkription und der Translation eines Antikörpers [2]. L: Leader-Sequenz.

## Ablauf der Immunantwort

Die monomere Form von IgM besitzt eine niedrige Antigen-Affinität und wird zusammen mit IgD durch alternatives Spleißen von mRNA auf der Oberfläche der naiven B-Lymphozyten produziert. Beim ersten Kontakt mit einem Antigen werden die B-Zellen aktiviert und entwickeln sich zu Gedächtniszellen oder zu Antikörper-produzierenden Plasmazellen (primäre Immunantwort). Die Affinität zum Antigen wird durch die somatische Hypermutation erhöht. Außerdem findet ein Wechsel des Isotyps zu IgG, IgA oder IgE („Class Switch“) durch somatische Rekombination statt. Beim erneuten Antigen-Kontakt differenzieren die Gedächtniszellen zu Plasmazellen und Antikörper können somit schneller produziert werden (sekundäre Immunantwort) [2, 3, 4].

## 1.2. Antikörper und ihre Rolle bei Autoimmunerkrankungen

Die Hauptfunktion der Immunantwort ist die Abwehr von Fremdkörpern, z.B. Krankheitserregern. In der Regel besteht Autotoleranz aufgrund der klonalen Deletion: in unreifen Lymphozyten, die Ig gegen körpereigene Moleküle (Autoantigene) bilden, wird vor der Reifung der programmierte Zelltod (Apoptose) eingeleitet. Eventuelle Störungen der Regulation können zu Autoimmunerkrankungen führen [2, 3, 11].

Bei vielen Krankheiten fehlt allerdings der endgültige Nachweis, dass es sich eindeutig um eine Autoimmunerkrankung handelt, z.B. bei der rheumatoiden Arthritis (RA), bei dem systemischen Lupus erythematodes (SLE) und bei der Sklerodermie (SD) [11].

RA ist eine chronische entzündliche Gelenkerkrankung, die bei ca. 1% der Erwachsenen auftritt (dreifache Häufigkeit bei Frauen). Die Entzündung beginnt in der Synovialmembran der Gelenke und kann zu Gelenkknorpel-Abbau und Gelenkversteifung in den späteren Krankheitsstadien führen [11]. Hinweise, dass es sich bei der RA um eine Autoimmunerkrankung handelt, sind die Bildung von Rheumafaktoren (große Immunkomplexe aus IgG und IgM anti-IgG) [11] sowie von ACPA Autoantikörpern (Antikörper gegen citrullinierte Peptide) [12, 13, 14] während des Krankheitsverlaufs.

SLE ist eine Multiorganerkrankung und betrifft neunfach häufiger Frauen. Der Name der Krankheit leitet sich von dem oft auftretenden schmetterlingsförmigen Erythem (Hautrötung) auf dem Gesicht ab. In vielen Fällen sind die Symptome Haut-, Gelenk- und Nierenprobleme. In den späten Stadien können schwere Organschäden die Folge sein [11]. SLE ist charakterisiert durch das Auftreten antinukleärer Autoantikörper, die gegen doppelsträngige (ds) DNA und Ribonukleoproteine (Sm-Antigene) gerichtet sind [15, 16, 17, 18].

SD manifestiert sich durch progressive Sklerose der Haut und der inneren Organe. Es existieren zwei Formen: bei der ersten ist vorwiegend die Haut, bei der zweiten – auch systemische Sklerose genannt – sind hauptsächlich andere Organe betroffen. Typische Folgen sind Nierenversagen, Lungen- und Herzprobleme [11]. Hinweise für die autoimmune Ursache der Erkrankung ist die Anwesenheit von Autoantikörpern [19, 20, 21].

## 1.3. Next Generation Sequenzierung

Sequenzierung ist die Entschlüsselung der Reihenfolge der vier Nukleotide C, G, A und T in DNA- oder C, G, A und U in RNA-Molekülen. Hierfür werden unterschiedliche Methoden eingesetzt. Der Standard ist die Didesoxymethode nach Sanger (Kettenabbruch-Synthese). Die Länge der so sequenzierten DNA-Abschnitte (Reads) kann mehr als 1000 Basenpaare (bp) erreichen, dies ist jedoch mit hohen Kosten und Zeitaufwand verbunden [22, 23]. Eine Alternative zum Standard bietet die Hochdurchsatz-Sequenzierung (Next Generation Sequenzierung): dank der Detektion von vielen parallel ablaufenden und räumlich getrennten Sequenzierungsreaktionen auf kleinster Fläche wird im Vergleich zum Standard schneller ein deutlich höherer Durchsatz bei kürzeren Read-Längen erreicht. Abhängig vom verwendeten Next Generation Sequenzierungssystem können pro Lauf bis zu 30 Gigabasen abgelesen werden [24].

Next Generation Sequenzierungsgeräte werden zur Untersuchung vieler Fragestellungen eingesetzt [25, 26]. Auf dem Markt sind vier konkurrierende Systeme erhältlich: Helicos *HeliScope™ Single Molecule Sequencer*, Illumina *Genome Analyzer*, Applied Biosystems *SOLiD™ System* und 454 Life Sciences *Genome Sequencer FLX Instrument*. Die Systeme von Illumina, Helicos und Applied Biosystems können nur kurze Reads aufschlüsseln, die zur Zeit meist deutlich kürzer als 100 bp lang sind [26, 27]. Diese Länge ist für die geplante Sequenzierung nicht ausreichend. Einerseits weisen die V Gene Homologien auf, daher könnten sie bei so kurzen Reads möglicherweise nicht eindeutig identifiziert werden. Andererseits können die für die Antigenbindung zuständigen CDR-Sequenzen nicht analysiert werden.

Aus diesem Grund wurde in der vorliegenden Arbeit das *Genome Sequencer FLX Instrument* [28, 29] verwendet, mittels welches Antikörpergene erfolgreich sequenziert wurden [30]. Das Gerät ist für Sequenzierung von längeren Reads geeignet: mit den Standard-Reagenzien ist eine Read-Länge von 200-300 bp möglich, während die neuen Titanium-Reagenzien eine Länge von über 400 bp erlauben. Ein Nachteil des Systems ist jedoch der im Vergleich zu den konkurrierenden Systemen kleinere Durchsatz (bis zu 400 Megabasen pro Lauf mit dem Titanium-Kit) [31].

Die Technologie von 454 Life Sciences, Pyrosequenzierung in Picotiterplatten (Abbildung 4), basiert auf „Sequenzierung durch Synthese“ [26, 27]. Zunächst wird eine einzelsträngige (ss) Matrizen- (Template-) DNA (sstDNA) Bibliothek vorbereitet: die Bindestellen für die

Sequenzierprimer, die so genannten Primer A und B, werden an den zu sequenzierenden DNA-Fragmenten (in diesem Fall Antikörpergene) angefügt. Als Nächstes werden die Fragmente einzeln auf die Trägersubstanz (kleine Kügelchen, Beads) immobilisiert, so dass an einem Bead nur ein Fragment gebunden ist. Die DNA wird mittels Emulsion-PCR (emPCR) in einem getrennten Reaktionsraum für jedes Bead (Wasser-in-Öl Mikroreaktor) amplifiziert. Darauffolgend werden die Beads auf eine Pikotiterplatte aufgetragen, wobei eine Pore nur ein Bead enthalten kann. Danach erfolgt die Sequenzierung-durch-Synthese: zunächst werden die zu sequenzierenden einzelsträngigen (ss) DNA-Moleküle mit Sequenzierprimern hybridisiert, darauffolgend werden die vier Desoxynukleosidtriphosphate (dNTPs) einzeln (z.B. nur Adenintriphosphat) zyklisch nacheinander zugegeben. Die Detektion wird auf Basis von zwei parallel zu dem DNA-Nukleotideinbau laufenden Enzymreaktionen realisiert. Bei dem Einbau von dNTPs (komplementär zur ssDNA-Matrix) durch die DNA-Polymerase wird das abgespaltene Pyrophosphat durch das Enzym Sulfurylase zu Adenosintriphosphat umgesetzt, welches seinerseits von dem Enzym Luziferase für die Oxidation von Luziferin zu Oxiluziferin verbraucht wird. Diese zweite Reaktion ist mit einem Lichtsignal gekoppelt, das durch eine im Gerät integrierte CCD-Kamera detektiert wird [27, 28, 29]. Die Intensität des Signals ist proportional zu der Anzahl eingebauter Nukleotide, sofern die Homopolymer-Länge kürzer als 8 Nukleotide ist [28]. Laut anderer Untersuchungen ist dies jedoch nicht immer der Fall [32, 33].

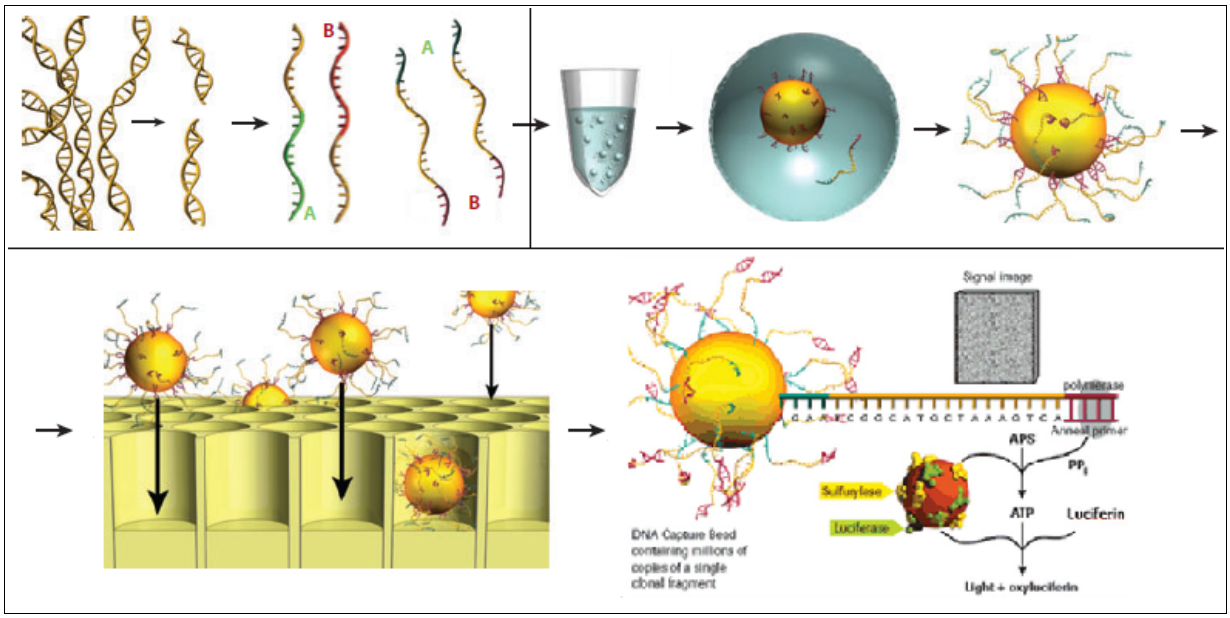



**Abbildung 4:** Ablauf der Next Generation Sequenzierung mittels *Genome Sequencer FLX Instrument*: Vorbereitung und Titration der DNA Bibliothek (oben links), Emulsion-PCR (emPCR, oben rechts) und Sequenzierung (unten) [adaptiert von 27 und 29].

Next Generation Sequenzierung und (Bio-)Informatik

Die Herausforderung für Next Generation Sequenzierungssysteme liegt in der Bewältigung der Datenmengen, die (bio)informatisch ausgewertet werden müssen. Es existieren mehrere (bio)informatische Programme, die jeweils die beste Lösung für ein bestimmtes System und Einsatzgebiet anbieten. Beispielweise stellt 454 Life Sciences die *GS Amplicon Variant Analyzer Software* zur Identifizierung von PCR-Produktvarianten mit dem *GS FLX System* zur Verfügung [29]. Jedoch existiert bisher keine Next-Generation-Sequenzierung-Software, die auf die Analyse von Antikörpergenen spezialisiert ist.

## 1.4. Bioinformatische Analyse von Antikörpergensequenzen

Ig Keimbahngen-Sequenzen sind über folgende Datenbanken verfügbar: VBASE [34], VBASE2 [8], IMGT/LIGM-DB [35] und IMGT/GENE-DB [36]. Für die Identifizierung von Gensegmenten (V, D, J) in Antikörper-Sequenzen sind spezielle bioinformatische Programme erforderlich. Einige davon werden nicht mehr aktualisiert (VBASE [34], IgBLAST [37]), während andere regelmäßig verbessert werden (VBASE2 [8, 38, 39, 40], IMGT/V-QUEST [41], IMGT/JunctionAnalysis [42]). Alternativen für die VDJ Identifizierung bieten JOINSOLVER [43], JointML [44] und iHMMune-align [45], die auf innovativen Analyse-Algorithmen basieren.

In der vorliegenden Arbeit wurden die VBASE2 Datenbank [8] sowie die VBASE2 Perl-Skripte verwendet. VBASE2 ist eine Datenbank, die humane und murine Ig V Keimbahngene enthält. Momentan sind 576 humane V Gene der H und L Kette annotiert. VBASE2 enthält umfangreiche Informationen über die V Gene wie z.B. triviale Namen der Gene, Nukleotid- und Aminosäuresequenzen, Anfang und Ende der CDR und FR Regionen, Referenzen zu den Sequenzquellen, Referenzen zu anderen Ig-Gen-Datenbanken wie z.B. VBASE, IMGT/LIGM-DB und weitere. Eine wichtige Option ist die Bewertung der Qualität und der Funktionalität der V Gen Sequenzen, die auf Grundlage der spezifischen VBASE2 Klassifikation erfolgt. Klasse 1 V Gene weisen Referenzen für genomische sowie für rearrangierte Sequenzen auf und sind somit funktionelle Gene. Klasse 2 V Gene basieren nur

auf genomischen Referenzen und sind dementsprechend Pseudogene, Orphans oder V Gene, deren Funktionalität noch nicht bewiesen worden ist, da sie in keinen *Rearrangements* identifiziert wurden. Klasse 3 V Gene stützen sich nur auf rearrangierte Sequenzen, bisher wurden sie auf genomischer Ebene nicht nachgewiesen [8].
VBASE2 ermöglicht nicht nur die Suche nach V-Gen-Informationen, sondern auch die Verwendung von DNAPLOT [38] basierten Anwendungen wie „*DNAPLOT Query*“ [39] oder *„Fab Analysis“* [40]. Sie dienen der Identifizierung von V(D)J Genen und dem Auffinden anderer relevanter Informationen, wie z.B. Vorhandensein von P und N Nukleotiden; Anzahl der aufgetretenen Mutationen in Sequenzen oder Einteilung von Nukleotid- und Aminosäuresequenz in CDRs und FRs. Das Programm weist darauf hin, falls die untersuchte Sequenz ein nicht funktionelles *Rearrangement* ist oder sein könnte. Dies ist z.B. der Fall beim Auftreten von Stopcodons, bei der Benutzung von Pseudogenen oder wenn sich bestimmte konservierte Aminosäuren nicht an ihrer Position laut der IMGT Nummerierung [46] befinden (z.B. Cystein an Position 104). Dies ermöglicht eine schnelle Einschätzung der Funktionalität von Antikörpersequenzen [39].

Die Verwendung der VBASE2-Perl-Skripte ermöglicht eine Optimierung der Analyse-Algorithmen. Außerdem können mehrere Sequenzen auf einmal analysiert werden [39]. Diese Option, kombiniert mit den standardmäßig ausgegebenen Tabellen für schnellen Ergebnis-Export in Tabellenkalkulationsprogramme oder Datenbanken [39], bietet die erfolgversprechendste Grundlage für die Analyse von Next Generation Sequenzierdaten.

## 1.5. Zielsetzung

Ziel des Projekts ist das Erhalten und Analysieren von Antikörper-Gensequenzen von gesunden Probanden und Autoimmunpatienten mittels Next Generation Sequenzierung (Amplicon-Sequenzierung [29] auf dem 454 Life Sciences *Genome Sequencer FLX Instrument*). Durch anschließende bioinformatische Analysen sollten krankheitsspezifische Muster entdeckt werden, wie z.B. Genverteilung und bevorzugte V(D)J Rekombinationen.

Das Ziel dieser Diplomarbeit ist die experimentelle Etablierung der Next Generation Technologie für Antikörpergensequenzierung und bioinformatisch die Auswertung der Datenmengen zu ermöglichen. Zunächst soll mittels quantitativer *Real-Time*-Polymerase-Kettenreaktion (qRT-PCR) analysiert werden, ob mehrere Proben für die Sequenzierung zusammen gemischt werden können, ohne die Ergebnisse zu verfälschen. Damit wäre in wenigen Sequenzierungsläufen ein repräsentativer Vergleich zwischen Gesunden und Kranken möglich. Außerdem sollen mittels qRT-PCR Differenzen in der IgG- und der IgD-Expression zwischen Gesunden und Autoimmunpatienten aufgedeckt werden. Weiterhin sollen IgG bzw. IgD Antikörpersequenzen von Gesunden und RA-Patienten, die ein V Gen der VH4 Familie und speziell das V Gen VH4(DP63) (humIGHV201, IGHV4-34*01) enthalten, für die Amplicon-Sequenzierung vorbereitet werden. Darüber hinaus sollte eine neue, für die statistische Auswertung der Genverteilung geeignete, VBASE2 Anwendung programmiert werden, mit der Next Generation Sequenzierdaten ausgewertet werden sollen. Die erlangten Daten sollen genutzt werden, um eventuell bestehende Korrelationen zum Gesundheitszustand der Probanden zu prüfen.

# 2. Material

## 2.1. Experimenteller Teil

### 2.1.1. Laborausstattung

- *Applied Biosystems* ABI Prism 7900HT
- *Bio-Rad* Gel Doc 2000
- *Eppendorf* Bio Photometer
- *Eppendorf* epMotion 5075
- *Eppendorf* Centrifuge 5414 S und Centrifuge 5415 C
- MJ Research Peltier Thermal Cycler PCT-200
- *Ohaus* Adventurer™ Pro AV812
- *Peqlab Biotechnologie GmbH* NanoDrop Spectrometer
- *Pharma Biotech* EPS 200, EPS 300, EPS 301, EPS 600
- *Roth* Micro Centrifuge SD

### 2.1.2. Verbrauchsmaterial

- *Applied Biosystems* Reaktionsgefäße für die PCR
- *Applied Biosystems* Reaktionsgefäße für die qRT-PCR (96- / 384-Well)
- *Eppendorf* Reaktionsgefäße 0,5 / 1,5 / 2,0 mL
- *Roth* Einweg-Küvetten für Photometer

### 2.1.3. Chemikalien, Enzyme, Kit-Systeme, Lösungen, Puffer

Falls nicht anders ausgewiesen, stammen alle Chemikalien von den Firmen *New England Biolabs*, *Fermentas*, PEQLAB Biotechnogie GmbH, *QUIAGEN*, *Jena Bioscience*, *Sigma-Aldrich*, *Merck*, *Roche* und *Roth*. Doppelt destilliertes Wasser (dd$H_2O$) wurde mittels *Barnstead* NANOpure Diamond, *ELGA LabWater* PURELAB Plus UV/UF und *Heraeus* Destamat Bi-Dest gewonnen.

Agarosegelelektrophorese:

- *Fermentas* 6X DNA Loading Dye
- 10X TBE-Puffer
  - 107,8 g Tris (0,89 M)
  - 55,03 g Borsäure (0,89 M)
  - 20 ml EDTA, pH 8,0 (Stock: 500 mM)
  - 863,3 ml $ddH_2O$
- Agarosegel
  - 1 – 1,8 g *Invitrogen™* UltraPure™ Agarose
  - 0,5X TBE bis auf 100 ml
  - Ethidiumbromid (0,00005% in TBE)

DNA-Verdau und Reverse Transkription (RT):

- *Sigma®* Amplification Grade Deoxyribonuclease I (DNase I)
  - Amplification Grade DNase I 1 U/µL
  - 10X Reaction Buffer
  - Stop Solution (50 mM EDTA)
- *Invitrogen™* SuperScriptTM II Reverse Transcriptase
  - SuperScript™ II RT 200 U/µL
  - 5X First-Strand Buffer
  - 0.1 M DTT
- I*nvitrogen™* RNaseOUT™ Recombinant Ribonuclease Inhibitor 40 U/µL

PCR:

- *Finnzymes* Phusion™ Hot Start High-Fidelity DNA Polymerase
  - Phusion™ Hot Start DNA Polymerase 500 U (250 µL)
  - 5X Phusion™ HF Buffer
- *New England BioLabs* ET SSB (500 µg/ml) (Extreme Thermostable Single-Stranded DNA Binding Protein)
- *Invitrogen*™ dNTPs (dATP, dCTP, dGTP, dTTP in $ddH_2O$, jeweils 2,5 mM)
- *PEQLAB Biotechnogie GmbH* Cycle Pure Kit

qRT-PCR:

- *Applied Biosystems* SYBR® Green Master Mix
- *Invitrogen*™ dNTPs (dATP, dCTP, dGTP, dTTP in $ddH_2O$, jeweils 10 mM)

## 2.1.4. Primer

**Tabelle 1:** Verwendete Oligonukleotide (Fw: Forward-Primer, Rv: Reverse-Primer; Hersteller *Eurofins*).

| Methode | | Name | DNA-Sequenz (5' – 3') | Hybridisierungsstelle |
|---|---|---|---|---|
| RT | | Random Primers# und Oligo(dT)$_{12-18}$ Primer## | | zufällige Hybridisierungsstellen bzw. Poly-A-Schwanz (als Reverse-Primer für IgG benutzt) |
| | | IgD-Hinge-Rv | GGTTAGCAGGTAGACGCCAAGAGGC | IgD-Hinge Region |
| | PCR1 | VH4-Fw | CAGGTGCAGCTGCAGGAGTCSG | Anfang der V Gene der VH4-Genfamilie |
| | | VH4(DP63)-Fw | CAGGTGCAGCTACAGCAGTGGG | Anfang des VH4(DP63)-Gens sowie Anfang der anderen Genallele des VH4(DP63)-Gens |
| | | IgG-CH1-Rv | ACTCTCTTGTCCACCTTGGTGTTGC | CH1 der schweren IgG-Kette |
| | | IgD-CH1-Rv | AGATCTCCTTCTTACTCTTGCTG | CH1 der schweren IgD-Kette |
| PCR2, TD-PCR2, Gradient-PCR2 | | IgG-MID1-Rv | TCAGACGAGTGCGTAGTC AGGGCGCCAGGGGGAAGAC | CH1 der schweren IgG-Kette nahe an der variablen Region |
| | | IgD-MID1-Rv | TCAGACGAGTGCGTAGTC TCTGCACCCTGATATGAT | CH1 der schweren IgD-Kette nahe an der variablen Region |
| | PCR3, Gradient-PCR3 | FusionA-VH4-Fw | GCCTCCCTCGCGCCATCAG CAGGTGCAGCTGCAGGAGTCSG | wie VH4-Fw |
| | | FusionA-VH4 (DP63)-Fw | GCCTCCCTCGCGCCATCAG CAGGTGCAGCTACAGCAGTGGG | wie VH4(DP63)-Fw |
| | | FusionB-MID1-Rv | GCCTTGCCAGCCCGC TCAGACGAGTGCGTAGTC | MID1-Sequenz des IgG-MID1-Rv bzw. IgD-MID1-Rv Primers |

\# *Invitrogen*™ Random Primers 3 µg/µL

\## *Invitrogen*™ Oligo(dT)$_{12-18}$ Primer 0,5 µg/µL

**Tabelle 1 (Fortsetzung):** Verwendete Oligonukleotide (Fw: Forward-Primer, Rv: Reverse-Primer; Hersteller *Eurofins*).

| Methode | Name | DNA-Sequenz (5' – 3') | Hybridisierungsstelle |
|---|---|---|---|
| qRT-PCR | qRT-PCR-IgG-Fw | TTCCCGGCTGTCCTACAGTC | CH1 der schweren IgG-Kette, 145 bp Fragment |
| | qRT-PCR-IgG-Rv | TGGGCTCAACTTTCTTGTCCA | |
| | qRT-PCR-IgD-Fw | CACCGCCAGCAAGAGTAAGAA | CH1 und Hinge Region der schweren IgD-Kette, 142 bp Fragment |
| | qRT-PCR-IgD-Rv | TGTGTTACGGGTGGTGGCT | |
| | GAPDH-Fw | CTGGTAAAGTGGATATTGTTGCCAT | Glycerinaldehyd - 3 - phosphat-Dehydrogenase (GAPDH), 81 bp Fragment |
| | GAPDH-Rv | TGGAATCATATTGGAACATGTAAACC | |
| | HPRT-Fw | GTAATTGGTGGAGATGATCTCTCAACT | Hypoxanthin - Phosphoribosyl-Transferase 1 (HPRT1), 81 bp Fragment |
| | HPRT-Rv | TGTTTTGCCAGTGTCAATTATATCTTC | |
| | U6-Fw | CTCGCTTCGGCAGCACA | U6 kleine nukleare RNA (snRNA), 94 bp Fragment |
| | U6-Rv | AACGCTTCACGAATTTGCGT | |
| | U5-Fw | TGGTTTCTCTTCAGATCGCATAAA | U5 snRNA, 102 bp Fragment |
| | U5-Rv | CCAAGGCAAGGCTCAAAAAAT | |
| | 7SK-Fw | CCCCTGCTAGAACCTCCAAAC | 7SK snRNA, 106 bp Fragment |
| | 7SK-Rv | CACATGCAGCGCCTCATTT | |

## 2.1.5. RNA- und cDNA-Proben

DNA-Verdau und RT

Die Lymphozyten-Gesamt-RNA stammt aus Proben von gesunden Testpersonen. In Tabelle 2 sind die jeweiligen Gesamt-RNA-Konzentrationen der Proben aufgelistet.

**Tabelle 2:** RNA-Konzentration der Proben aus den gesunden Probanden.

| Bezeichnung der Probe | RNA-Konzentration [ng/µL] |
|---|---|
| **C1330** | 191 |
| **C1331** | 400 |
| **C1332** | 462 |
| **C1333** | 529 |
| **C1334** | 367 |

PCR

Für die PCR wurden aus Gel isolierte 700 bp Fragmente eingesetzt, die aus cDNA mit VH4-Fw bzw. VH4(DP63)-Fw als Fw-Primern und IgG-CH1-Rv bzw. IgD-CH1-Rv als Rv-Primern amplifiziert wurden.

qRT-PCR

Für die qRT-PCR wurde bei den Patienten vorhandene oder bei den gesunden Testpersonen selbst synthetisierte cDNA benutzt.

Für die Kalibrierung wurden vier Proben eingesetzt (Tabelle 3), um die Konzentration von IgG und IgD cDNA in der cDNA nach der RT zu bestimmen. VH ist die Bezeichnung für Mischungen von 700 bp Fragmenten (aus Gel isoliert), die mit Fw-Primern aller VH-Familien und IgG-CH1-Rv bzw. IgD-CH1-Rv als Rv-Primern hergestellt wurden. VH4 ist die Bezeichnung für Proben, die mit den gleichen Rv-Primer und VH4-Fw amplifiziert wurden.

**Tabelle 3:** Konzentrationen der Proben für die Kalibrierung mittels qRT-PCR.

| Bezeichnung | Konzentration [ng/µL] |
|---|---|
| VH IgG | 0,82 |
| VH IgD | 1,23 |
| VH4 IgG | 10 |
| VH4 IgD | 10 |

## 2.1.6. DNA-Längenstandards

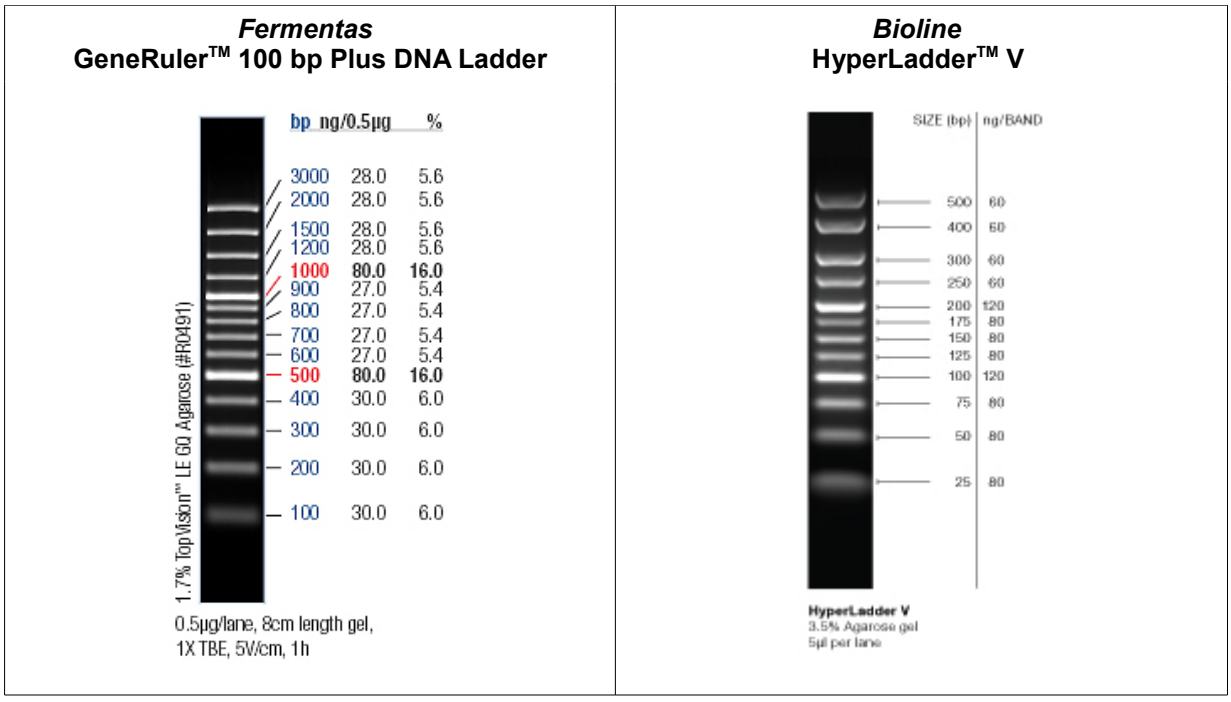



**Abbildung 5:** Verwendete DNA-Längenstandards.

## 2.2. Bioinformatischer Teil

### 2.2.1. Hardware

- Fujitsu Siemens Computers Amilo M 6450

### 2.2.2. Software

- *Applied Biosystems* Sequence Detection System® (SDS®) Software Version 2.1
- *Invitrogen™* VectorNTI Advance 10 mit Align X-Modul

Operationssystem: Canonical Ltd. Ubuntu 8.04 (*Hardy Heron*):

- Geany 0.13 „Vensell“
- Eclipse SDK 3.2.2
- EPIC (Eclipse Perl Integration) 0.5.37
- Perl 5.8.8
- PHP5 5.2.4
- PostgreSQL 8.2.9
- phppgadmin 4.1.3
- Apache2 2.2.8
- Mozilla Firefox Version 3.0.10
- GIMP 2.4.5
- Evince Dokumentenbetrachter 2.22.2
- OpenOffice.org 3.0.0
- Außerdem wurden die VBASE2-Datenbank [8], die VBASE2 Perl-Skripte sowie das DNAPLOT-Programm [38] eingesetzt.

Operationssystem: Microsoft® Windows XP Home Edition Version 5.1 (Service Pack 3):

- GraphPad Prism Version 5.00
- Adobe® Reader® Version 9.2
- OpenOffice.org 3.1.0

### 2.2.3. Internetquellen: Anwendungen und Datenbanken

| | |
|---|---|
| VBASE2: | http://www.vbase2.org/ |
| ClustalW2: | http://www.ebi.ac.uk/Tools/clustalw2/ |
| UCSC In-Silico PCR: | http://genome.ucsc.edu/cgi-bin/hgPcr |
| Entrez Nucleotide: | http://www.ncbi.nlm.nih.gov/nuccore/ |
| Uniprot: | http://www.uniprot.org/ |
| ExPASy: | http://au.expasy.org/ |
| IMGT: | http://imgt.cines.fr/ |
| SRS: | http://srs.ebi.ac.uk/ |

### 2.2.4. Antikörper-Sequenzen

scFv Sequenzen: Um Vortests durchführen zu können, wurden von Michael Hust (Technische Universität Braunschweig) 48 Sequenzen aus den naiven humanen scFv Antikörperbibliotheken HAL7 und HAL4 [47,48] zur Verfügung gestellt, davon enthalten 8 κ und 40 λ leichte Ketten.

Sequenzen aus Klonierung: Um Gesunde mit Patienten zu vergleichen, wurden von Florian Rubelt (Max-Planck-Institut für molekulare Genetik, MPIMG) Antikörpersequenzen zur Verfügung gestellt. Die Next Generation Sequenzierung erfolgte nach der vorgeschlagenen „S*hotgun*“-Methode mit dem Titanium-Kit (s. Kapitel 5.), jedoch mit Modifikationen. Die Sequenzen wurden von Volker Sievert (MPIMG) mit dem EMBOSS Programm „fuzznuc“ nach MIDs und nach spezifischen Primersequenzen in Gruppen aufgeteilt.

Sequenzen der Antikörper-Isotypen IgG und IgD: Um die Bindungsstellen und das Design der Primer-Sequenzen für die qRT-PCR zu überprüfen, wurden von Theam Soon Lim (MPIMG) die von ihm für Primerdesign benutzten IgG bzw. IgD Sequenzen zur Verfügung gestellt.

## 2.2.5. Primer

**Tabelle 4:** Genspezifische Oligonukleotide, die an der Technischen Universität (TU) Braunschweig und am MPIMG benutzt werden und für die bioinformatischen Tests eingesetzt wurden.

| IG-Kette | Oligonukleotide, die an der TU Braunschweig benutzt werden | | Oligonukleotide, die an dem MPIMG benutzt werden | |
|---|---|---|---|---|
| | Name | DNA-Sequenz (5' – 3') | Name | DNA-Sequenz (5' – 3') |
| variabel schwer | MHVH1_f | CAGGTBCAGCTGGTGCAGTCTGG | VH1 | CAGGTCCAGCTKGTRCAGTCTGG |
| | MHVH1/7 f | CARRTSCAGCTGGTRCARTCTGG | VH157 | CAGGTGCAGCTGGTGSARTCTGG |
| | MHVH2 f | CAGRTCACCTTGAAGGAGTCTGG | VH2 | CAGRTCACCTTGAAGGAGTCTG |
| | MHVH3 f1 | SARGTGCAGCTGGTGGAGTCTGG | VH3 | GAGGTGCAGCTGKTGGAGWCY |
| | MHVH3 f2 | GAGGTGCAGCTGKTGGAGWCYSG | | |
| | MHVH4 f1 | CAGGTGCARCTGCAGGAGTCGGG | VH4 | CAGGTGCAGCTGCAGGAGTCSG |
| | MHVH4 f2 | CAGSTGCAGCTRCAGSAGTSSGG | VH4(DP63) | CAGGTGCAGCTACAGCAGTGGG |
| | MHVH5 f | GARGTGCAGCTGGTGCAGTCTGG | | |
| | MHVH6 f | CAGGTACAGCTGCAGCAGTCAGG | VH6 | CAGGTACAGCTGCAGCAGTCA |
| variabel leicht kappa | MHVK1 f1 | GACATCCAGATGACCCAGTCTCC | VK1 | GACATCCRGDTGACCCAGTCTCC |
| | MHVK1 f2 | GMCATCCRGWTGACCCAGTCTCC | VK246 | GATATTGTGMTGACBCAGWCTCC |
| | MHVK2 f | GATRTTGTGATGACYCAGWCTCC | | |
| | MHVK3 f | GAAATWGTGWTGACRCAGTCTCC | VK3 | GAAATTGTRWTGACRCAGTCTCC |
| | MHVK4 f | GACATCGTGATGACCCAGTCTCC | | |
| | MHVK5_f | GAAACGACACTCACGCAGTCTCC | VK5 | GAAACGACACTCACGCAGTCTC |
| | MHVK6 f | GAWRTTGTGMTGACWCAGTCTCC | | |
| variabel leicht lambda | MHVL1 f1 | CAGTCTGTGCTGACTCAGCCACC | VL1 | CAGTCTGTSBTGACGCAGCCGCC |
| | MHVL1 f2 | CAGTCTGTGYTGACGCAGCCGCC | VL1459 | CAGCCTGTGCTGACTCARYC |
| | | | VL15910 | CAGCCWGKGCTGACTCAGCCMCC |
| | MHVL2 f | CAGTCTGCCCTGACTCAGCCT | VL2 | CAGTCTGYYCTGAYTCAGCCT |
| | MHVL3 f1 | TCCTATGWGCTGACWCAGCCACC | VL3 | TCCTATGWGCTGACWCAGCCAC |
| | MHVL3 f2 | TCTTCTGAGCTGACTCAGGACCC | VL3(DP16) | TCCTCTGAGCTGASTCAGGASCC |
| | | | VL3(38) | TCCTATGAGCTGAYRCAGCYACC |
| | MHVL4 f1 | CTGCCTGTGCTGACTCAGCCC | | |
| | MHVL4 f2 | CAGCYTGTGCTGACTCAATCRYC | | |
| | MHVL5 f | CAGSCTGTGCTGACTCAGCC | | |
| | MHVL6 f | AATTTTATGCTGACTCAGCCCCA | VL6 | AATTTTATGCTGACTCAGCCCC |
| | MHVL7/8 f | CAGRCTGTGGTGACYCAGGAGCC | VL78 | CAGDCTGTGGTGACYCAGGAGCC |
| | MHVL9/10 f | CAGSCWGKGCTGACTCAGCCACC | | |

# 3. Methoden

## 3.1. Experimenteller Teil

### 3.1.1. DNA-Verdau und Reverse Transkription

Die RT ist der Prozess der DNA-Synthese mit Hilfe des Enzyms reverse Transkriptase (RNA-abhängige DNA-Polymerase), wobei RNA als Matrize für die Herstellung der so genannten komplementären DNA (cDNA) dient. Um eventuellen DNA-Verunreinigungen der zu transkribierenden mRNA zu entfernen, wurden die Proben vor der eigentlichen RT mit dem Enzym DNase I behandelt. Das eingesetzte Protokoll orientiert sich an dem von Lim [49] entwickelten.

Protokoll für den DNA-Abbau

Die folgende Beschreibung des eingesetzten Protokolls basiert auf Auszügen aus dem Originalprotokoll [50]:

- Zugabe in ein RNase-freies Reaktionsgefäß von
  - RNA in 8 µL Wasser, 1 µL 10X Reaction Buffer und 1 µL Amplification Grade DNase I (1 U/µL) bei den Proben C1331, C1332, C1333 und C1334;
  - RNA in 10,5 µL Wasser, 1,31 µL 10x Reaction Buffer und 1,31 µL Amplification Grade DNase I bei der Probe C1330.
- Vorsichtiges Mischen, Inkubation: 15 Minuten (min) bei Raumtemperatur.
- Hinzufügen von 1 µL Stop Solution zur DNase I-Inaktivierung vor dem Erhitzen.
- Erhitzen (10 min bei 70 °C) zur Denaturierung der DNase I und der RNA.
- Abkühlen auf Eis.

Protokoll für die reverse Transkription

Die folgende Beschreibung des eingesetzten Protokolls basiert auf Auszügen aus dem Originalprotokoll [51]:

- Zugabe in ein Nuklease-freies Reaktionsgefäß von
  - VH4-Fw bzw. VH4(DP63)-Fw als Fw-Primer; als Rv-Primer Random Rrimers und Oligo(dT)$_{12-18}$ Primer für IgG bzw. IgD-Hinge-Rv für IgD (insgesamt 1 µL);
  - 400 ng Gesamt-RNA bei den Proben C1330, C1332, C1333 und C1334 bzw. 200 ng Gesamt-RNA bei der Probe C1331;
  - 1 µL dNTPs (10 mM) und
  - dd$H_2O$ bis auf 12 µL.
- Erhitzen (5 min bei 65 ºC) und schnelles Abkühlen auf Eis.
- Kurze Zentrifugation und Hinzufügen von
  - 4 µL 5X First-Strand Buffer,
  - 2 µL 0.1 M DTT,
  - 1 µL RnaseOUT™.
- Vorsichtiges Mischen, Inkubation (2 min bei 42 ºC).
- Zugabe von 1 µL SuperScript™ II RT, Mischen durch sorgfältiges Pipettieren.
- Inkubation der IgG-Proben (10 min bei 25 ºC).
- Inkubation aller Proben (50 min bei 42 ºC).
- Abbruch der Reaktion durch Erhitzen (15 min bei 70 ºC).

### 3.1.2. qRT-PCR

Die qRT-PCR bietet die Möglichkeit, eine zeitliche Quantifizierung von DNA-Vervielfältigung vorzunehmen. Der im Ansatz verwendete Fluoreszenzfarbstoff (SYBR® Green) interkaliert nur mit dsDNA und erlaubt somit eine zeitliche Erfassung der dsDNA-Menge während der PCR-Zyklen. Dabei wird für jede Probe der Anfang der exponentiellen Phase der Amplifizierung ermittelt und der entsprechende Ct-Wert („*threshold cycle*“, Schwellenwert-Zyklus) gemessen. Ct ist die Nummer des Zyklus (angegeben als Dezimalbruchzahl), bei der eine bestimmte

Menge amplifizierter DNA vorliegt – d.h. die gemessene Fluoreszenz erreicht einen Wert über der Hintergrund-Fluoreszenz [52].

Primer-Effizienz-Test

Für den Primer-Effizienz-Test wurde als Template eine Mischung aus jeweils 20 ng cDNA bei den gesunden Probanden C1330, C1331, C1333 und C1334 erstellt, die mit $ddH_2O$ auf ein Gesamtvolumen von 50 µL aufgefüllt wurde. Aus der so hergestellten 1:1 Verdünnung (1,6 ng/µL) wurden weitere angesetzt (1:10, 1:100 und 1:1000). Als Kontrolle für die Normierung wurden fünf konstitutiv exprimierte Haushaltsgene amplifiziert (U5, U6, 7SK, HPRT und GAPDH). Die Auftragung der Proben auf die 96-Well Platte ist in Tabelle 5 dargestellt:

**Tabelle 5:** Schema der Probenauftragung für den Primer-Effizienz-Test.

| | | | Verdünnungsreihe Template-DNA | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1:1 | | | 1:10 | | | 1:100 | | | 1:1000 | | |
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Primer | GAPDH | A | | | | | | | | | | | | |
| | HPRT | B | | | | | | | | | | | | |
| | U6 | C | | | | | | | | | | | | |
| | U5 | D | | | | | | | | | | | | |
| | 7SK | E | | | | | | | | | | | | |
| | qRT-PCR-IgG | F | | | | | | | | | | | | |
| | qRT-PCR-IgD | G | | | | | | | | | | | | |
| | NTC* | H | GAPDH | | | HPRT | | | U5 | | | 7SK | | |
| | | | qRT-PCR-IgG | | | qRT-PCR-IgD | | | | | | | | |

* NTC: Negativkontrolle (Ansatz ohne Template-DNA)

Für die Auswertung wird die Standardkurve-Methode benutzt [53]. Zunächst werden die Ct-Durchschnittswerte der Mehrfachbestimmungen für die einzelnen Primer bzw. Verdünnungen berechnet. Die Werte werden in einer Grafik mit logarithmischer X-Achse (Logarithmus der DNA-Menge in ng) und linearer Y-Achse (Ct) aufgetragen. Die Steigungen der Geraden wurde mit GraphPad Prism ermittelt (Funktion „XY analyses“ – „Linear regression“). Die erhaltene Grundgleichung kann Gleichung 1 entnommen werden:

**Gleichung 1:** Grundgleichung für die grafische Auswertung von Primereffizienz-Tests mittels qRT-PCR.

$$f(x) = a * x + b$$

*x:* Logarithmus der DNA-Menge in ng; *f(x):* Ct-Wert; *a:* Steigung; *b:* Ordinaten-Abschnitt

Mit Hilfe der Steigungswerte kann die jeweilige Primereffizienz anhand Gleichung 2 kalkuliert werden:

**Gleichung 2:** Gleichung für die Berechnung von Primereffizienz bei qRT-PCR-Experimenten.

$$Primereffizienz = 10^{-\frac{1}{Steigung}}$$

Experiment mit gesunden Probanden und Patienten

Als Template wurde einerseits die cDNA der gesunden Probanden bzw. der Patienten (einzeln sowie als Pool in Konzentrationen von 25 ng cDNA in 50 µL) und andererseits zwei Verdünnungsreihen von 700 bp PCR-Produkten (s. Kapitel 2.1.5.) eingesetzt. Als Kontrolle wurden U6 und HPRT Primer eingesetzt.
Die Probenauftragung auf die 384-Well Platte ist in Tabelle 6 dargestellt. Die Verdünnungsreihen für die Konzentrationsbestimmung von IgG bzw. IgD cDNA sowie die Proben für die Kreuzreaktivitätsüberprüfung wurden per Hand pipettiert. Das automatische Pipettiergerät *Eppendorf* epMotion 5075 wurde für die Primer, für die NTC sowie für die cDNA-Proben aus Gesunden und Patienten benutzt.
Die Auswertung erfolgt nach der „*Efficiency adjusted* ΔΔCt"-Methode [54]. Berechnet wird die Differenz in der Expression von zwei Genen, bezogen auf die Expression von einem Kontrollgen unter Berücksichtigung der unterschiedlichen Primereffizienzen ($2^{(-\Delta\Delta Ct_{adjusted})}$). Da zwei Gene als Kontrollen eingesetzt wurden, wurde die Normalisierung unter Verwendung des geometrischen Mittelwertes der $Ct_{adjusted}$ Werte der Kontrollgene durchgeführt [55].

Um die Ig cDNA Stoffmengen zu berechnen, wurden Kalibriergeraden nach der Standardkurve-Methode [53] erstellt. Dabei wurden die $Ct_{adjusted}$ Werte [54] gegen den Logarithmus des aus den Konzentrationen der 1:1 Verdünnungen (s. Tabelle 3) kalkulierten Gewichts aufgetragen. Damit die erhaltenen Werte miteinander vergleichbar sind, wurde folgende Normalisierungskorrektur der $Ct_{adjusted}$ Werte vorgenommen:
1. Bildung der $Ct_{adjusted}$-Durchschnittswerte aus den Mehrfachbestimmungen für die einzelnen Gene [54].
2. Bildung des jeweiligen geometrischen Mittelwertes des $Ct_{adjusted}$ bzw. der Standard-abweichung des $Ct_{adjusted}$ Wertes (SA) zwischen den Kontrollen U6 und HPRT (Normalisierung der Kontrollen) [55].

**Tabelle 6:** Probenauftragung für das Experiment mit den gesunden Probanden und Patienten. VH IgG, VH IgD, VH4 IgG, VH4 IgD: Bezeichnungen der Verdünnungsreihen, wobei die verwendeten Primer für die Herstellung der Template für die qRT-PCR als Abkürzung für die Reihe benutzt werden: Fw-Primer: VH (alle VH-Familien) bzw. VH4, Rv-Primer: IgG (IgG-CH1-Rv) bzw. IgD (IgD-CH1-Rv).

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| C1330 | | | | | | | | SLE66 | | | |
| C1331 | | | | | | | | SLE67 | | | |
| C1332 | | | | | | | | SLE68 | | | |
| C1333 | | | | | | | | SLE75 | | | |
| C1334 | | | | | | | | SLE76 | | | |
| Pool gesunde Kontrolle | | | | | | | | Pool SLE | | | |
| RA26 | | | | | | | | SD63 | | | |
| RA27 | | | | | | | | SD64 | | | |
| RA28 | | | | | | | | SD69 | | | |
| RA29 | | | | | | | | SD70 | | | |
| RA30 | | | | | | | | SD71 | | | |
| Pool RA | | | | | | | | Pool SD | | | |
| VH | 1:1 | VH | 1:1 | VH4 | 1:1 | VH4 | 1:1 | NTC | NTC | NTC | NTC |
| IgG | 1:10 | IgD | 1:10 | IgG | 1:10 | IgD | 1:10 | | | | |
| | 1:100 | | 1:100 | | 1:100 | | 1:100 | VH IgD 1:1 | VH IgG 1:1 | | |
| | 1:1000 | | 1:1000 | | 1:1000 | | 1:1000 | VH4 IgD 1:1 | VH4 IgG 1:1 | | |

| Primer-Kennzeichnung | qRT-PCR-IgG | qRT-PCR-IgD | U6 | HPRT |
|---|---|---|---|---|

3. Kalkulation der Korrekturen für IgG $Ct_{adjusted}$ nach Gleichung 3:

**Gleichung 3:** Kalkulation der $Ct_{adjusted}$-Korrekturen für IgG.

$$Korrektur_{(Probe)} = \overline{Ct_{adjusted}} - Ct_{adjusted\,(Probe)}$$

$\overline{Ct\ adjusted}$: Durchschnittlicher $Ct_{adjusted}$-Wert der normalisierten Kontrollen

$Ct_{adjusted(Probe)}$: $Ct_{adjusted}$-Wert der normalisierten Kontrolle der jeweiligen Probe

4. Berechnung der korrigierten $Ct_{adjusted}$-Werte für IgG nach Gleichung 4:

**Gleichung 4:** Kalkulation der korrigierten $Ct_{adjusted}$-Werte für IgG.

$$Ct_{adjusted\,(korrigiert)} = Ct_{adjusted\,(Probe)} - Korrektur_{(Probe)}$$

Nach der Normalisierungskorrektur werden die erhaltenen Grundgleichungen laut Gleichung 1 für die Interpolation mittels Kalibrierungsreihen umformuliert (Gleichung 5):

**Gleichung 5:** Umformulierung der Grundgleichung für die Berechnung der Verdünnung aus dem $Ct_{adjusted}$-Wert bei qRT-PCR-Experimenten.

$$x = \frac{f(x) - b}{a}$$

*x:* Logarithmus der DNA-Menge in ng
*f(x):* korrigierter $Ct_{adjusted}$-Wert
*a:* Steigung
*b:* Ordinaten-Abschnitt

Mit den auf dieser Weise ermittelten $Ct_{adjusted}$-Werten der Proben können die dazugehörigen Konzentrationen (in ng) durch Interpolation berechnet werden. Daraus werden die Ig cDNA Stoffmengen der jeweiligen Proben laut Gleichung 6 kalkuliert:

**Gleichung 6:** Gleichung für die Berechnung der Ig cDNA Stoffmenge bei qRT-PCR-Experimenten.

$$n_{Ig\,cDNA} = \frac{10^{x}}{L * 2 * \overline{MG}}$$

$n_{Ig\,cDNA}$: Ig cDNA Stoffmenge in nmol
*x:* Logarithmus der DNA-Menge in ng, interpoliert aus den Kalibriergeraden
*L*: Länge der cDNA-Kontrolle (700 bp)
$\overline{MG}$: Durchschnittliches Molekulargewicht eines Nukleotids (333 g/mol)

Experiment mit gesunden Probanden

Als Template wurde die cDNA aus allen gesunden Probanden benutzt (einzeln sowie als Pool in Konzentrationen von 50 ng cDNA in 50 µL). Als Kontrollen wurden U6 und HPRT Primer eingesetzt. Das Auftragungsschema der Proben auf die 96-Well Platte ist in Tabelle 7 dargestellt. Die Auswertung für die Berechnung von $2^{(-\Delta\Delta Ct\ adjusted)}$ erfolgt wie beim Experiment mit den gesunden Probanden und den Patienten nach der „Efficiency adjusted ΔΔCt"-Methode [54] unter Verwendung des geometrischen $Ct_{adjusted}$ Mittelwertes der Kontrollgene [55].

**Tabelle 7:** Probenauftragung für das Experiment mit den gesunden Probanden.

| | | Primer | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | qRT-PCR-IgG | | | qRT-PCR-IgD | | | U6 | | | HPRT | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Template-DNA | C1330 | A | | | | | | | | | | | |
| | C1331 | B | | | | | | | | | | | |
| | C1332 | C | | | | | | | | | | | |
| | C1333 | D | | | | | | | | | | | |
| | C1334 | E | | | | | | | | | | | |
| | Pool* | F | | | | | | | | | | | |
| | NTC** | G | | | | | | | | | | | |
| | | H | | | | | | | | | | | |

* Pool: Mischung mit gleichen DNA-Mengen aller gesunden Kontrollen
** NTC: Negativkontrolle (Ansatz ohne Template-DNA)

Protokoll

Die Programmierung von ABI Prism 7900HT sowie der Ansatz für die drei qRT-PCR-Experimente sind Tabelle 8 zu entnehmen.

**Tabelle 8:** PCR-Ansatz und Standard-Programmierung von ABI Prism 7900HT für die qRT-PCR.

| Ansatz (10 µL) | |
|---|---|
| Template-DNA | 2 µL |
| Fw-Primer | 0,25 µL |
| Rv-Primer | 0,25 µL |
| SYBR® Green | 5 µL |
| $ddH_2O$ | 2,5 µL |

| Schritte | N# | min | °C |
|---|---|---|---|
| UNG Aktivierung* | | 2:00 | 50 |
| Denaturierung | | 10:00 | 95 |
| Denaturierung | 40 | 0:15 | 95 |
| Annealing und Extention | | 1:00 | 60 |
| Dissoziationskurve | | 0:15 | 95 |
| | | 0:15 | 60 |
| | | 0:15 | 95 |

\# *N*: Anzahl Zyklen
* Standardschritt, Aktivierung von Uracyl-N-Glycosylase (nicht in SYBR® Green Master Mix enthalten)

### 3.1.3. Sequenzierungsvorbereitung

Um die Antikörper-Gene der H Kette für die Amplicon-Sequenzierung vorzubereiten, müssen die Bindestellen der Sequenzierprimer (vom Hersteller als Primer A und Primer B bezeichnet [56]) an den zu amplifizierenden DNA-Fragmenten angefügt werden (Abbildung 6). Beide Primer sind jeweils 19 bp lang und besitzen am 3' Ende eine sich wiederholende 4 bp lange Sequenz („Schlüssel“) [56].

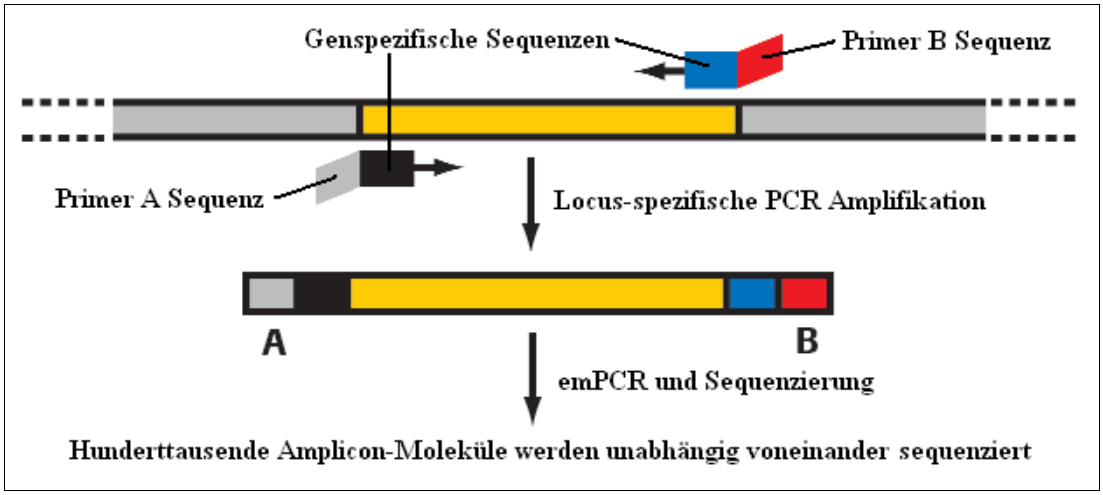



**Abbildung 6:** Vorbereitung der Sequenzen für die Amplicon-Sequenzierung: Anfügen von Primer A und B (Bindestellen der Sequenzierprimer) (adaptiert von [56]).

Um eine eindeutige Erkennung der sequenzierten PCR-Fragmente bei Mischungen aus individuellen Proben (Pools) bei den hohen Durchsätzen der Next Generation Sequenzierung zu ermöglichen, können zusätzlich 10 bp lange Multiplex Identifiers (MIDs) eingeführt werden, ähnlich einem „Barcode“ [29]. Die Sequenzen aller MIDs können dem Anhang entnommen werden (Tabelle A1). Für die Identifizierung der Sequenzen ist das Anfügen der MIDs neben dem Primer B erforderlich: so werden die MIDs zuerst sequenziert, sonst könnten sie bei längeren Antikörper-Sequenzen eventuell nicht in der Readlänge enthalten sein. Außerdem werden MID und Primer B aufgrund der bestrebten Ig-Isotyp-Bestimmung am 3' Ende der Antikörper-Sequenzen angehängt: somit werden die revers-komplementären Ig-Sequenzen gewonnen.

Um die Antikörpergene für die Sequenzierung vorzubereiten, werden sie zunächst mittels genfamilienspezifischer Fw-Primer und Ig-Isotyp-Gelenkregion-spezifischer Rv-Primer von mRNA in komplementäre Desoxyribonukleinsäure (cDNA) durch RT umgeschrieben. In der darauffolgenden ersten PCR-Phase (PCR1) werden die Gene mit einem Ig-Isotyp-CH1-spezifischen Rv-Primer und dem gleichen Fw-Primer amplifiziert. In der zweiten PCR-Phase (PCR2) werden Primer A und MID mit Hilfe von zusammengesetzten Fusion-Primern eingeführt (Tabelle 9). Für die PCR2 wird der Primer A (mit „Schlüssel“) an die 5'-Seite des spezifischen Fw-Primers angefügt, während die Ig-Isotyp spezifischen Rv-Primer durch die vorgesehenen MID-Sequenzen (MID1 für gesunde Probanden und MID2 für RA-Patienten) auf der 5'-Seite modifiziert werden. Dabei wurde bei dem Design der spezifischen Rv-Primer beachtet, dass sie möglichst nah an der VH Region binden (ca. 20-30 bp davon entfernt) (Abbildung 7). Da bei dem geplanten PCR-Ablauf die MID-Sequenzen als Primerbindungsstelle für die darauffolgende PCR dienen sollen, sind sie mit ihrer Länge von 10 bp zu kurz. Daher werden der in den Bindestellen der Sequenzierprimer enthaltene „Schlüssel“ und vier zusätzliche Nukleotide zu jeder MID Sequenz hinzugefügt. In der letzten PCR-Phase (PCR3) wird der Primer B angehängt, wobei er mit dem „Schlüssel“, der MID-Sequenz und den zusätzlichen Nukleotiden den zusammengesetzten Rv-Primer bildet (Tabelle 5). Somit sind die PCR-Produkte mit beiden Bindestellen der Sequenzierprimer (Primer A und B) sowie einer Erkennungssequenz (MID) ausgestattet und können für die Sequenzierung eingesetzt werden.

Der gesamte Ablauf ist in Abbildung 8 dargestellt. Für die L Ketten ist der Ablauf analog, jedoch müssen die Rv-Primer-Sequenzen angepasst werden.

Nach der vorbereitenden PCR-Amplifizierung sollte die Anzahl der DNA Moleküle für die Amplicon Sequenzierung $10^{11}$ nicht überschreiten. Außerdem ist es empfehlenswert, PCR Produkte mit einer maximalen Länge von 500 bp einzusetzen [57].

**Tabelle 9:** Bausteine für das Primerdesign der zusammengesetzten Primer für die Vorbereitung der Amplicon-Sequenzierung und einige Beispiele für Fusion-Primer.

| Verwendung, Beispiele | | Name | Sequenz in 5'-3' Richtung |
|---|---|---|---|
| Bausteine | | Primer A | GCCTCCCTCGCGCCATCAG |
| | | Primer B | GCCTTGCCAGCCCGCTCAG |
| | | „Schlüssel“ | TCAG |
| | | MID, hier MID1 | ACGAGTGCGT |
| | | Zusätzliche Nukleotide | AGTC |
| Bausteine (spezifische Primer) | | IgD-Rv | TCTGCACCCTGATATGAT |
| | | VH4-Fw | CAGGTGCAGCTGCAGGAGTCSG |
| Fusion Primer | PCR2 | IgD-MID1-Rv | TCAGACGAGTGCGTAGTCTCTGCACCCTGATATGAT |
| | PCR2, PCR3 | FusionA-VH4-Fw | GCCTCCCTCGCGCCATCAGCAGGTGCAGCTGCAGGAGTCSG |
| | PCR3 | FusionB-MID1-Rv | GCCTTGCCAGCCCGCTCAGACGAGTGCGTAGTC |

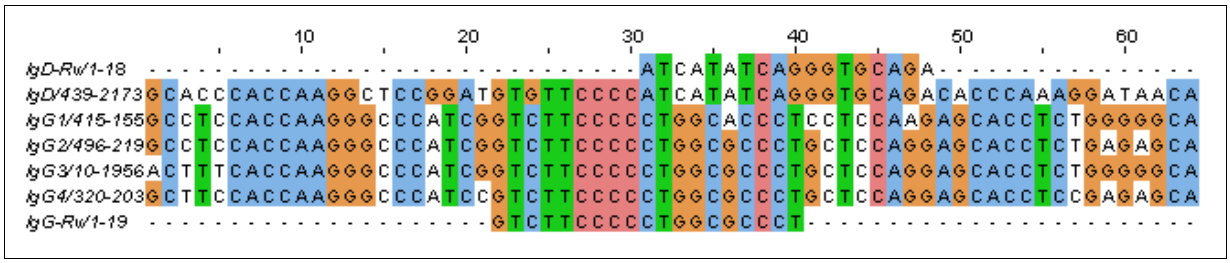



**Abbildung 7:** Darstellung der Bindestellen der CH1-spezifischen Primer IgD-Rv und IgG-Rv am Anfang der CH1 Domänen. *IgD, IgG1-IgG4:* Sequenzen der Ig-Isotypen. Zu beachten ist, dass die revers-komplementären Primersequenzen abgebildet sind, um den Vergleich mit den IgD und IgG Sequenzen zu erleichtern.

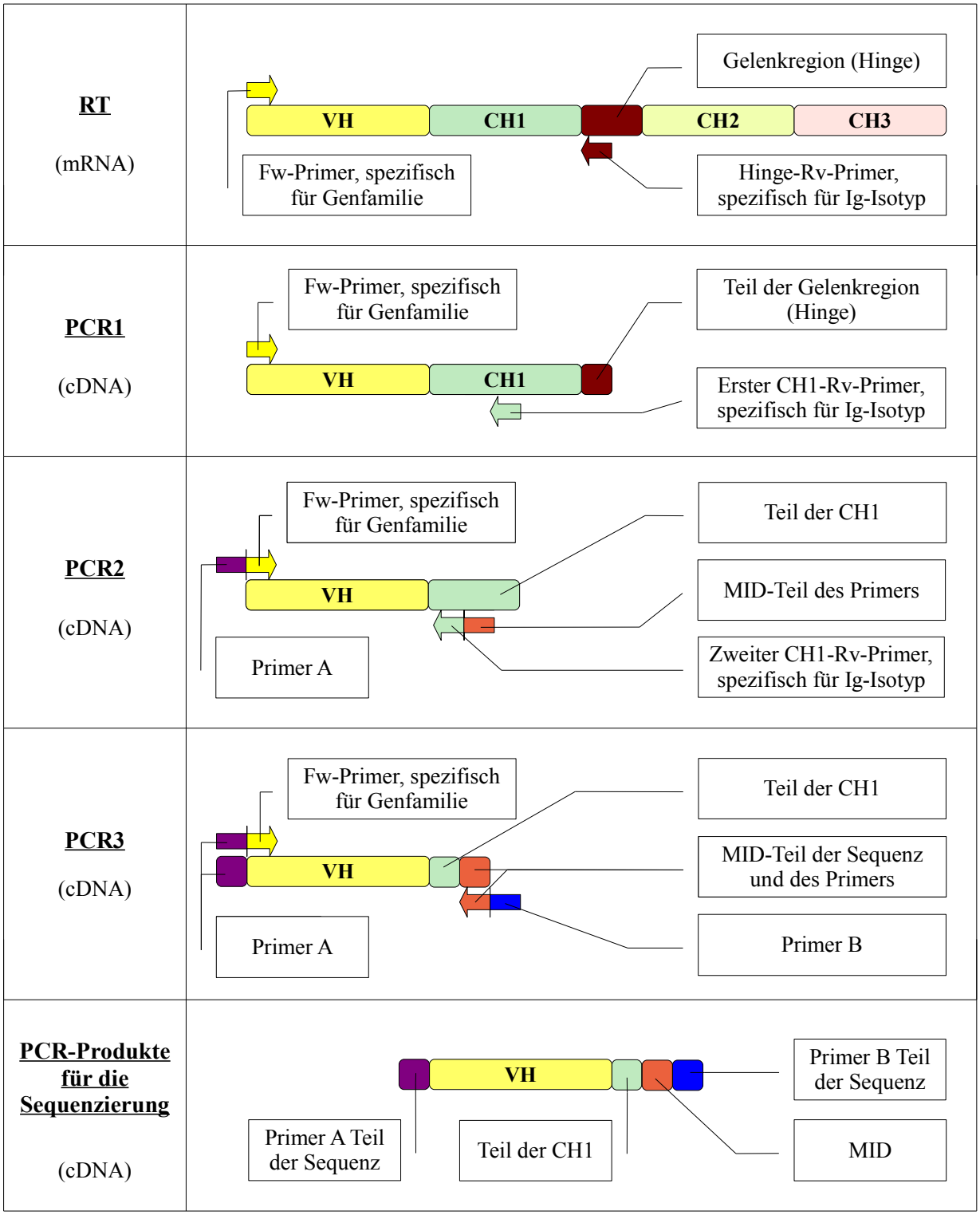



**Abbildung 8:** Schematische Darstellung der Vorbereitung für die Amplicon-Sequenzierung.

### 3.1.4. Polymerase-Kettenreaktion

Die Polymerase-Kettenreaktion (PCR, *Polymerase Chain Reaction*) ist eine Methode für *in vitro* DNA-Amplifizierung. Basis für die technische Realisierung der Methode ist ein Gerät (Thermocycler), das eine zyklische Wiederholung von drei molekularbiologischen Schritten ermöglicht. Im ersten Schritt (Denaturierung) wird die zu vervielfältigende DNA (Template) zwecks Strangtrennung auf 95 °C erhitzt. Im zweiten Schritt (Annealing) hybridisieren spezifische Oligonukleotide (Primer) bei reduzierter Temperatur mit der einzelsträngigen Template-DNA. Im dritten Schritt (Extension) werden dNTPs ausgehend von den freien 3' OH-Enden der gebundenen Primer mittels thermostabiler DNA-Polymerase eingefügt, bis der komplementäre Strang synthetisiert worden ist [58].
Die Qualität der PCR-Produkte hängt von vielen Faktoren ab – wie z.B. gewählte Polymerase, Reaktionsadditive, Puffer oder Reaktionsbedingungen (Dauer und insbesondere Temperatur des Annealing-Schritts), die ihrerseits durch Art der Polymerase, Länge bzw. Sequenz der Primer und Länge der PCR-Produkte beeinflusst werden. Für die Experimente wurde die Phusion™ Hot Start DNA Polymerase verwendet, da sie einerseits weniger Fehler bei der DNA-Amplifizierung aufgrund ihrer Proofreading-Aktivität (“Korrekturlesen”: 3' – 5' Exonukleaseaktivität im Falle einer Fehlpaarung) verursacht, andererseits die Amplifizierung erst beim Erreichen einer bestimmten Temperatur (72 °C) erlaubt [59]. Außerdem wurde der Einfluss von ET SSB Zugabe untersucht, welche die Stabilisierung der einzelsträngigen DNA bewirken soll und somit die Zugänglichkeit der Template durch Verhindern der Ausbildung von DNA-Sekundärstrukturen sichern soll [60].

Weiterhin wurden Touchdown-PCR (TD-PCR) [61] und Gradient-PCR [62] angewendet, um die Nebenproduktamplifizierung zu reduzieren. TD-PCR ist eine PCR-Methode zur Minimierung der Konzentration von möglichen Nebenprodukten. Die TD-PCR-Reaktion erfolgt in zwei Phasen: in der ersten Phase wird das gewünschte PCR-Produkt durch schrittweise Erniedrigung der Annealing-Temperatur bevorzugt amplifiziert, wobei die erhöhte Spezifität der Primer bei höheren Temperaturen von Vorteil ist; während in der zweiten Phase die PCR bei Standardbedingungen weitergeführt wird [61]. Gradient-PCR ist eine PCR Methode zur Bestimmung der optimalen Temperatur im Annealing-Schritt und dient somit zur Reduzierung unerwünschter Nebenproduktamplifizierung. Dabei werden die gleichen PCR-Ansätze bei unterschiedlichen Hybridisierungstemperaturen behandelt. Die optimale Annealing-Temperatur ist diejenige, bei der die Ziel-DNA ohne Nebenprodukte in möglichst großer Menge amplifiziert worden ist [62].

Protokoll

Die Ansätze für die unterschiedlichen PCR-Methoden sind in Tabelle 10 dargestellt, während die Programmierung des Thermocyclers Tabelle 11 entnommen werden kann.

**Tabelle 10:** PCR-Ansätze (m: mit ET SSB, o: ohne ET SSB).

| | PCR1 | PCR2m, PCR2o, TD-PCR2, Gradient-PCR2 | PCR3, Gradient-PCR3 |
|---|---|---|---|
| **Template** | ca. 700 bp PCR-Produkt (aus Gel extrahiert) | ca. 700 bp PCR-Produkt aus PCR1 | ca. 450 bp PCR-Produkt aus PCR2 |
| **PCR-Produkt** | ca. 700 bp | ca. 450 bp | ca. 465 bp |
| **Fw-Primer** | VH4-Fw, VH4(DP63)-Fw | FusionA-VH4-Fw, FusionA-VH4(DP63)-Fw | FusionA-VH4-Fw, FusionA-VH4(DP63)-Fw |
| **Rv-Primer** | IgG-CH1-Rv, IgD-CH1-Rv | IgG-MID1-Rv, IgD-MID1-Rv | FusionB-MID1-Rv |
| **Beispielansatz (50 µL)** | • 1 µL Template-DNA<br>• 4 µL dNTPs<br>• 10 µL 5X HF-Buffer | • 1 µL Fw-Primer<br>• 1 µL Rv-Primer<br>• 0,5 µL Phusion | • 0,5* (0**) µL ET SSB<br>• dd$H_2O$ bis auf 50 µL |

* bei den Methoden PCR1 und PCR2m
* bei den Methoden PCR2o, TD-PCR2, Gradient-PCR2, PCR3, Gradient-PCR3

**Tabelle 11:** Programmierung des Thermocyclers für die verschiedenen PCR-Methoden.

| PCR-Schritt | Standard-PCR | | |
|---|---|---|---|
| | N# | min | °C |
| Denaturierung | | 1:30 | 95 |
| Denaturierung | | 1:00* | 95 |
| Annealing | | | |
| *PCR1* | 31 *(PCR1)* | 0:45 | 58 |
| *PCR2* | 21 *(PCR2)* | 0:30 | 55 |
| *PCR3* | 30 *(PCR3)* | 0:50 | 58 |
| Extension | | 1:00* | 72 |
| End-Extension | | 5:00 | 72 |
| Kühlen | | -## | 4 |

| | PCR-Schritt | TD-PCR2 | | |
|---|---|---|---|---|
| | | N# | min | °C |
| Denaturierung | | | 0:30 | 95 |
| Phase 1 | Denaturierung | 10 | 0:30 | 95 |
| | Annealing | | 0:45 | 65** |
| | Extension | | 1:00 | 72 |
| Phase 2 | Denaturierung | 20 | 0:30 | 95 |
| | Annealing | | 0:45 | 55 |
| | Extension | | 1:00 | 72 |
| End-Extension | | | 5:00 | 72 |
| Kühlen | | | -## | 4 |

| PCR-Schritt | Gradient-PCR | | |
|---|---|---|---|
| | N# | min | °C |
| Denaturierung | | 1:30 | 95 |
| Denaturierung | 30 | 0:30 | 95 |
| Annealing | | 0:45 | |
| *Gradient-PCR2* | | | *** |
| *Gradient-PCR3* | | | **** |
| Extension | | 1:00 | 72 |
| End-Extension | | 5:00 | 72 |
| Kühlen | | -## | 4 |

\# *N*: Anzahl Zyklen
## bis der Thermocycler ausgeschaltet wurde
* Beispieldauer für PCR1, für PCR2 bzw. PCR3 wurde die Dauer je nach PCR-Produktlänge angepasst
** Anfangstemperatur, jeden Zyklus Temperaturerniedrigung um 1°C
*** Temperaturen: 55,7; 56,8; 58,3; 60,0; 61,4; 62,5; 63,3; 63,8
**** Temperaturen: 55,2; 56,6; 58,5; 60,8; 63,5; 65,8; 67,5; 68,8

Zusätzlich wurde eine modifizierte PCR-Methode eingesetzt (PCRqrt), wobei die Bedingungen möglichst ähnlich der qRT-PCR unter Berücksichtigung der unterschiedlichen Polymerase gewählt wurden. Die Konzentrationen der Reagenzien im Ansatz und das Thermocycler-Programm wurden geändert (Tabelle 12).

**Tabelle 12:** PCR-Ansatz und Programmierung des Thermocyclers für die PCRqrt-Methode.

| Template | cDNA aus allen Gesunden sowie ca. 700 bp PCR-Produkte (aus Gel extrahiert) |
|---|---|
| PCR-Produkt | 145 bp (IgG) bzw. 142 (IgD) |
| Fw-Primer | qRT-PCR-IgG-Fw bzw. qRT-PCR-IgD-Fw |
| Rv-Primer | qRT-PCR-IgG-Rv bzw. qRT-PCR-IgD-Rv |

| Ansatz | |
|---|---|
| Template-DNA | 2 µL |
| dNTPs | 1,6 µL |
| 5X HF-Buffer | 4 µL |
| Fw-Primer | 1 µL |
| Rv-Primer | 1 µL |
| Phusion | 0,2 µL |
| $ddH_2O$ | 10,2 µL |

| PCR-Schritt | N# | min | °C |
|---|---|---|---|
| Vorschritt | | 2:30 | 50 |
| Denaturierung | | 3:00 | 95 |
| Denaturierung | 30 | 0:15 | 95 |
| Annealing | | 0:30 | 60 |
| Extension | | 0:30 | 72 |
| End-Extension | | 5:00 | 72 |
| Kühlen | | -## | 4 |

\# *N*: Anzahl Zyklen
## bis der Thermocycler ausgeschaltet wird

### 3.1.5. Aufreinigung von PCR-Produkten

Die Aufreinigung der PCR-Produkte von dNTPs, Primer, Polymerase usw. erfolgte mit dem Cycle Pure Kit. Die folgende Beschreibung des eingesetzten Protokolls basiert auf Auszügen aus dem Originalprotokoll [63]:

- Versetzen der PCR-Ansätze mit dem gleichen Volumen XP1-Puffer.
- Vortexen (sorgfältiges Mischen).
- Beladen der HiBind DNA-Säule und Bindung der DNA an die Silikamembran; Zentrifugation (1 min bei 10.000g).
  - *Abweichung vom Originalprotokoll:* erneutes Beladen der Säule mit dem Durchfluss zwecks besserer Bindung der DNA an der Membran, Zentrifugation (1 min bei 10.000g).
- Waschen mit 750 µL SPW-Puffer (komplettiert mit Ethanol), Zentrifugation (1 min bei 10.000g).
- Wiederholung des vorherigen Schrittes.
- Trockenzentrifugation der Säule (2 min bei 10.000g).
- Elution der DNA mit Elutionspuffer (40 µL), Zentrifugation (1 min bei 5.000g).

### 3.1.6. DNA-Gelelektrophorese

Die Gelelektrophorese ist eine Methode zur Trennung von DNA Molekülen mittels Anlegen eines elektrischen Feldes, wobei die Wanderungsgeschwindigkeit der negativ geladenen DNA-Moleküle zu der positiven Elektrode (Anode) antiproportional zu ihrer Größe ist. Das Agarosegel ist die Trägersubstanz für die Molekültrennung. Je nach Gewichtsprozent Agarose weisen die Gele eine unterschiedliche Porengrößenverteilung auf und legen somit den Auftrennungsbereich fest. Aufgrund der angestrebten feineren Auftrennung im Bereich der kleinen DNA-Fragmente wurde TBE-Puffer als mobile Phase gewählt. Für den Nachweis der DNA wird Ethidiumbromid eingesetzt, das als DNA-interkalierende Substanz die Sichtbarkeit der DNA unter UV-Licht (306 nm) gewährleistet [64].

Protokoll

- Aufkochen der Agarose (1% bzw. 1,8 % w/v) mit 0,5X TBE-Puffer.
- Hinzufügen von Ethidiumbromidlösung nach kurzer Abkühlung der Lösung.
- Gießen des Gels in einer Elektrophoresekammer, Einsetzen des Kamms.
- Nach Erstarren des Gels Entfernung des Kamms und Füllen der Kammer mit TBE-Puffer.
- Auftragung der mit 6X DNA Loading Dye vermischten Proben sowie der DNA-Längenstandards.
- Anlegen eines elektrischen Feldes (300 V); Laufzeit (30 bis 60 min) ist von der Größe des Gels abhängig.
- Fotografieren der Gele unter UV-Licht (306 nm).

### 3.1.7. Überblick der durchgeführten Experimente

Der Ablauf der durchgeführten qRT-PCR Experimente ist in der Abbildung 9 veranschaulicht. Zur Etablierung der PCR-Methoden 1-3 für die Vorbereitung der Sequenzierung wurden Proben der gesunden Probanden eingesetzt (Abbildung 10). Dabei wurden statt cDNA gelaufgereinigte Produkte der PCR1 benutzt.

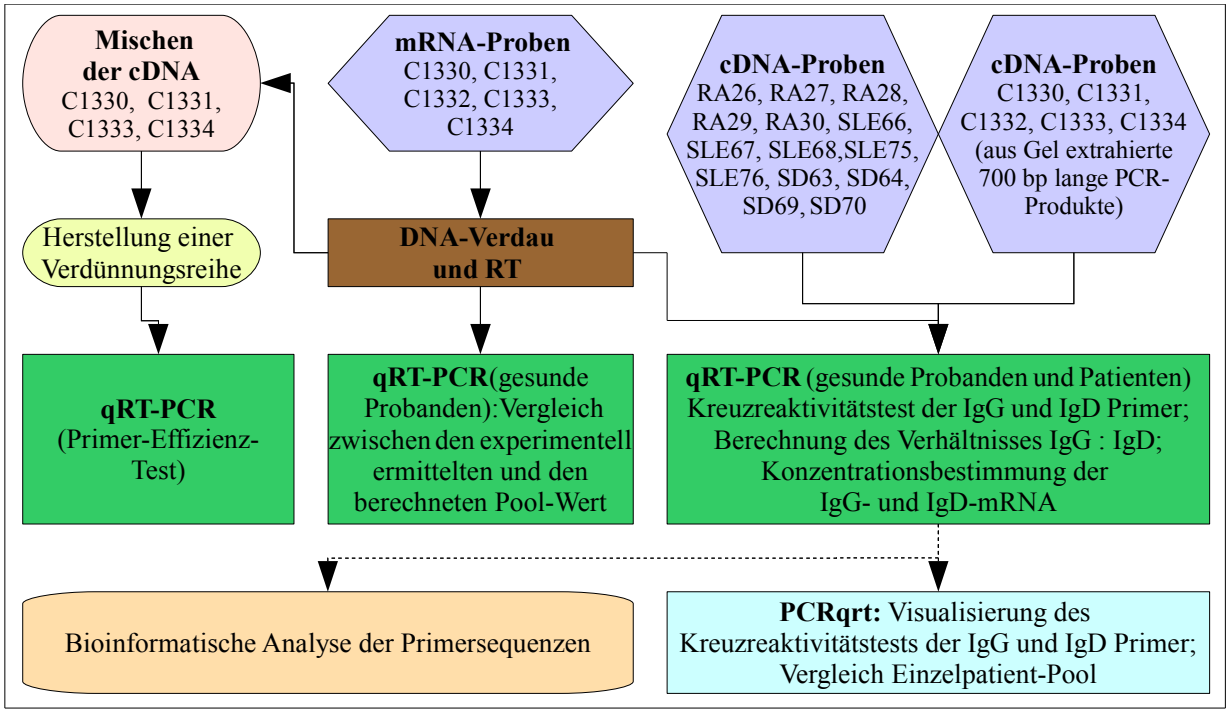



**Abbildung 9:** Überblick der durchgeführten qRT-PCR-Experimente; mit gestrichelter Linie sind die damit verbundene bioinformatische Analyse bzw. PCRqrt-Experiment dargestellt.

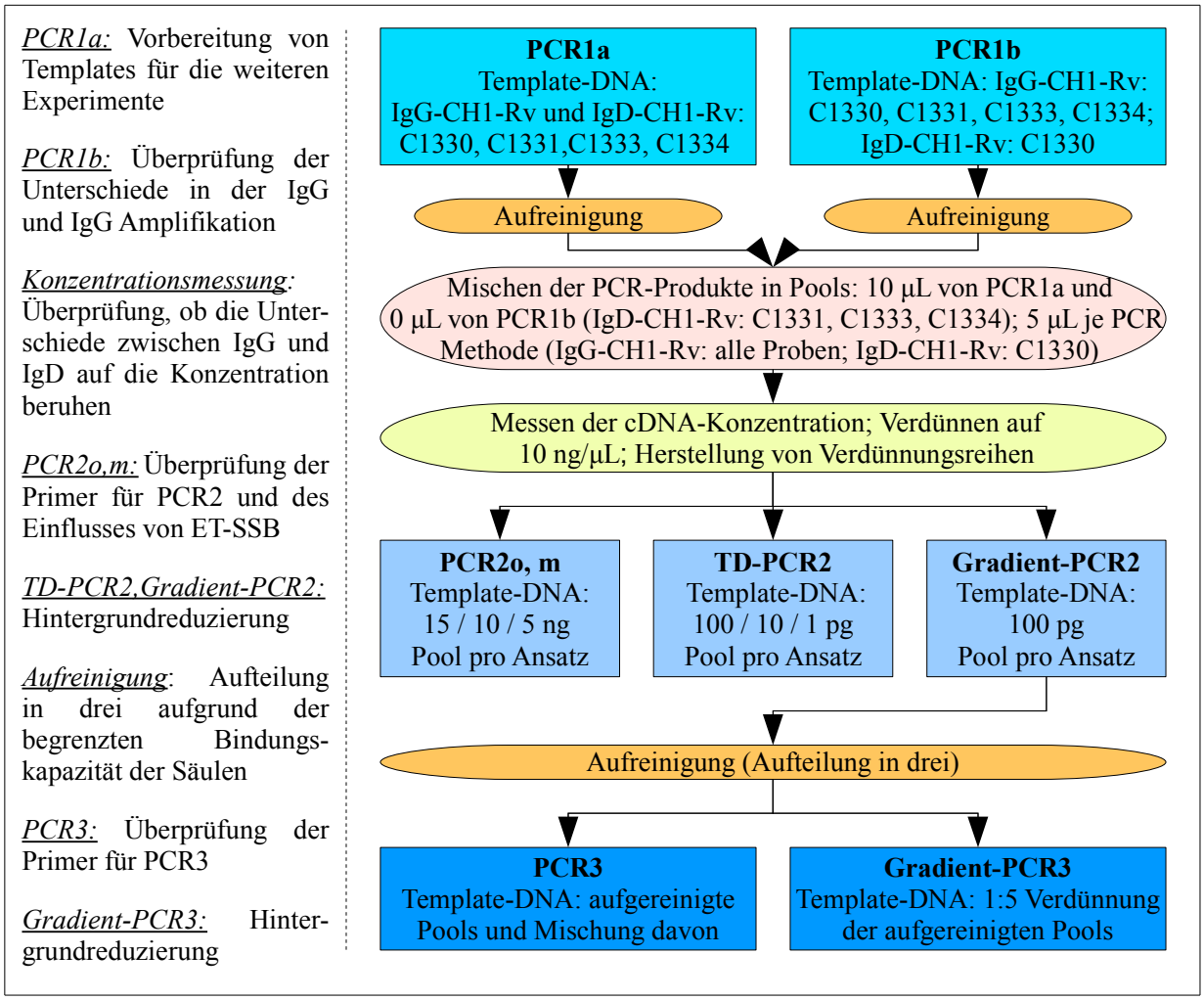



**Abbildung 10:** Überblick der durchgeführten PCR-Experimente für die Sequenzierungsvorbereitung.

# 3.2. Bioinformatischer Teil

## 3.2.1. Unterschiedliche Vortests mittels Perl-Programme

Für die Untersuchung folgender Fragestellungen wurden Perl-Skripte programmiert, mit deren Hilfe die Auswertungen durchgeführt wurden:

Gekürzte Sequenzen

Die Sequenzierung der Antikörper vom 3' Ende mittels *Genome Sequencer FLX Instrument* führt dazu, dass der 5' Anfang des V-Gens z.T. nicht mehr sequenziert wird. Um auszuschließen, dass fälschlicherweise ein anderes V Gen identifiziert wird, wurde eine bioinformatische Analyse mittels *„DNAPLOT Query“* der ungekürzten und der jeweils in 10 bp Schritten (bis 100 bp) gekürzten humanen V Genen aus der VBASE2 sowie scFv-Testsequenzen aus den naiven humanen scFv Antikörperbibliotheken HAL7 und HAL4 [47, 48] vorgenommen.

Länge der scFv-Nukleotidsequenzen

Um die optimale sowie die minimal erforderliche Read-Länge der Next Generation Sequenzierdaten zu ermitteln, wurde die durchschnittliche Länge der scFv-Testsequenzen in Nukleotiden berechnet. Nur Sequenzen mit vollständiger Abdeckung aller Gensegmente wurden für die Analyse eingesetzt.

Homopolymere

Die scFv-Testsequenzen sowie die humanen V Gene aus der VBASE2 wurden hinsichtlich der Länge und Anzahl der Homopolymere analysiert, da das *Genome Sequencer FLX Instrument* möglicherweise die Anzahl der Nukleotide in Homopolymersequenzen nicht korrekt detektiert (s. Kapitel 1.3).

Mutationen

Die Anzahl der Mutationen in den scFv-Testsequenzen wurde ermittelt, um den zu erwartenden Mutation-Durchschnitt in Prozent zu berechnen. Dieser Mittelwert ist für die Berechnung des „*Cut Off*"-Wertes für die folgenden Analysen erforderlich (Gleichung 7). Der Parameter *Cut Off* entspricht demnach dem Prozentsatz Identität zwischen der getesteten Sequenz und dem Keimbahngen. Er bestimmt, ob eine Sequenz für die weiteren Analysen beibehalten wird oder aufgrund der vielen auftretenden Mutationen verworfen wird.

**Gleichung 7:** Gleichung für die Berechnung des *Cut Off* Wertes.

$$Cut\,Off = (100 - \overline{Mutationen}) \quad [\%]$$

$\overline{\text{Mutationen}}$: Mittelwert der Mutationen in Prozent

### 3.2.2. Primer-Analysen

Design der qRT-PCR Primer: Das Design der qRT-PCR Primer wurde mittels ClustalW2 überprüft.

VectorNTI Align X Analyse der an dem MPIMG eingesetzten Fw-Primer [49]: Die humanen V Gensequenzen sind auf der Internetseite der VBASE2 Datenbank zum Download bereitgestellt. Sie wurden nach Familien sortiert und die an dem MPIMG benutzten Primer wurden mittels des AlignX® Moduls für Vector NTI Advance™ 10 (Invitrogen) mit den zu amplifizierenden Genfamilien aligniert. Die Mutationen (*Mismatches)* zwischen Primer und Sequenz wurden manuell gezählt. Ein Gen wurde als amplifiziert bezeichnet, sofern die Anzahl der *Mismatches* nicht mehr als vier betrug und sowie die überlappte Region am 3'-Ende des Primers eine Länge von 18 bp nicht unterschritt. Daraufhin wurde die Abdeckung (*Coverage*) der Primer kalkuliert.

Mutationsanalyse der an der TU-Braunschweig eingesetzten Fw-Primer: Die Ergebnisse der Mutationsanalyse für die *Cut Off* Berechnung wurden erweitert, indem die Mutationen in FR1 nicht ab dem 1., sondern ab dem 24. Nukleotid ermittelt wurden. Diese Zahl richtet sich nach der Länge der verwendeten Primer – bis zu 23 bp. Somit kann kalkuliert werden, wie viele Mutationen in FR1 durch die Primer verursacht werden, wobei ein Ansatz mit Beibehaltung der Mutationsrate in FR1 gewählt wurde (Gleichung 8 und 9).

**Gleichung 8:** Gleichung für die Berechnung der Anzahl erwarteter Mutationen in FR1 bei Beibehaltung der Mutationsrate in FR1.

$$Mutationen_{erwartet} = \frac{Mutationen_{ab\,dem\,24.\,Nukleotid} * Länge\ der\ FR1}{Länge\ der\ FR1 - 23}$$

**Gleichung 9:** Gleichung für die Berechnung der Anzahl der durch Primer verursachten Mutationen in FR1 bei Beibehaltung der Mutationsrate in FR1.

$$Mutationen_{durch\ Primer\ verursacht} = Mutationen_{ab\,dem\,1.\,Nukleotid} - Mutationen_{erwartet}$$

### 3.2.3. VBASE2 „*Statistic Analysis*“

Die VBASE2 „*Statistic Analysis*“ ist eine neue VBASE2 Anwendung, die auf „*DNAPLOT Query*“ basiert. Die entsprechenden PHP-HTML-Eingabeformulare sowie Perl-Skripte wurden an die Anforderungen (Vergleich der Gen-Nutzung zwischen zwei Sequenzen-Sets) angepasst.

### 3.2.4. *next*IGbase Datenbank

Die *next*IGbase Datenbank ist eine relationale PostgreSQL-Datenbank, die zur Vereinfachung der Auswertung programmiert wurde. Das Füllen der Datenbank mit Daten sowie die Auswertung der Daten wurde mit Hilfe von Perl-Skripten über das DBI-Modul realisiert. Zusätzlich wurde ein VBASE2-ähnliches Webinterface mit PHP und HTML programmiert.

### 3.2.5. Statistische Auswertung

Die statistischen Auswertungen beim Vergleich zwischen Gesunden und Autoimmunpatienten sowie bei der Ermittlung von bioinformatischen Parametern (Unterschiede zwischen Mittelwerten, Medianen, Verteilungen) wurden mit GraphPad Prism vorgenommen. Die im Programm eingebauten Analysen wurden eingesetzt; ein Ergebnis mit einer Wahrscheinlichkeit von $P < 0.05$ wird für signifikant erachtet. Der jeweils benutzte Test wird angegeben.

# 4. Ergebnisse

## 4.1. Experimentelle Untersuchungen

### 4.1.1. Effizienz-Test und Kreuzreaktivitätstest der qRT-PCR-Primer

Um die Menge an IgG bzw. IgD Antikörpergentranskripte mittels qRT-PCR zu quantifizieren bzw. das IgG:IgD Mengenverhältnis zu ermitteln, muss sichergestellt werden, dass mit allen qRT-PCR-Primern eine gleichmäßige Amplifizierung der Ziel-Gene möglich ist. Hierzu wurde ein Primer-Effizienz-Test durchgeführt. Außerdem müssen die ermittelten Werte für IgG bzw. IgD Mengen mit Hilfe von einer Kontrolle normiert werden. Daher wurden fünf konstitutiv exprimierte Haushaltsgene amplifiziert (U5, U6, 7SK, HPRT und GAPDH), um Kontrollprimer mit ähnlichen Primereffizienzen wie die IgG- und IgD-Primer auszuwählen.
Für den Test wurde eine cDNA-Template-Verdünnungsreihe aus 700 bp PCR1 Produkten eingesetzt und für jedes qRT-PCR-Primer-Paar wurden die Ct Werte (Schwellenwert-Zyklen, s. Kapitel 3.1.6) gemessen. Aus den Messdaten wurde die Effizienz der Primer ermittelt, basierend auf den signifikant voneinander abweichenden Geradensteigungen in einem Ct gegen log ng cDNA Diagramm (Wahrscheinlichkeit für gleiche Steigung $P < 0{,}0001$) (Abbildung 11).

Die berechnete Effizienz der meisten Primerpaare liegt um 2,00 (Tabelle 13); wobei die IgD bzw. U5 Werte davon abweichen (1,84 bzw. 2,28). Eine Primereffizienz von 2,0 ist optimal und bedeutet, dass bei jedem PCR Zyklus die Anzahl an DNA Molekülen verdoppelt wird [54]. Als Kontrollen für die darauffolgenden qRT-PCR-Tests wurden zwei Primer-Paare gewählt: U6 ist eine snRNA, die am mRNA-Spleißen beteiligt ist; während HPRT ein am Purin-Stoffwechsel beteiligtes Enzym ist.

**Tabelle 13:** Berechnete Primer-Effizienzen

| Primer | Effizienz | Effizienz [%] |
|---|---|---|
| GAPDH | 1,98 | 98,9 |
| HPRT | 2,00 | 99,8 |
| U6 | 2,02 | 101,3 |
| U5 | 2,28 | 118,6 |
| 7SK | 1,93 | 95,2 |
| IgG | 1,97 | 97,7 |
| IgD | 1,84 | 87,9 |

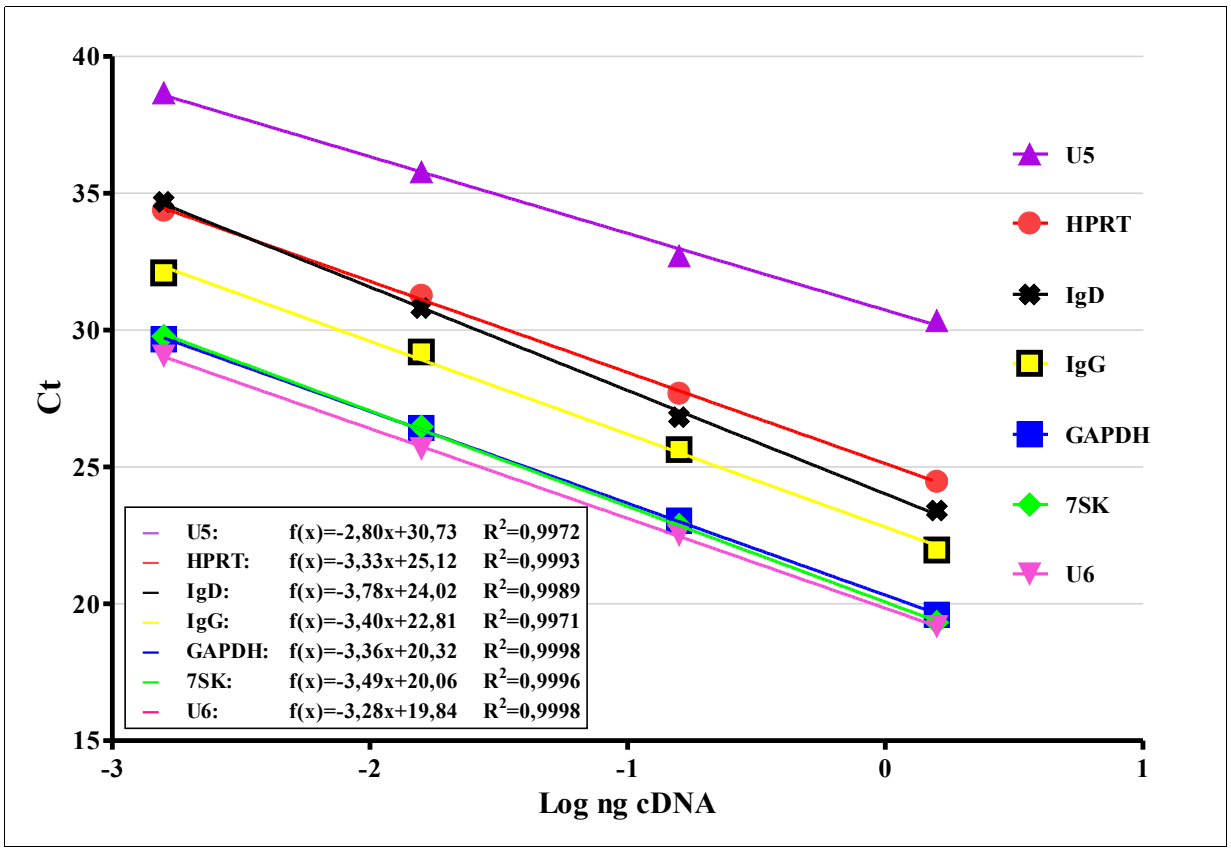



**Abbildung 11:** Darstellung der Ergebnisse des Primer-Effizienz-Tests. Die Gleichungen der Linearen Regression sowie die dazugehörigen Regressionskoeffizienten sind abgebildet.

Um die IgG bzw. IgD Antikörpergentranskripte mittels qRT-PCR zu quantifizieren, muss ausgeschlossen werden, dass die qRT-PCR-Primer kreuzreagieren. Diese möglichen Kreuzreaktionen wurden mittels qRT-PCR überprüft (nicht dargestellt). Als Template wurden 700 bp Fragmente eingesetzt, die V sowie die CH1 Region von IgG enthalten und aus cDNA mit VH4-Fw bzw. Fw-Primern aller VH-Familien und IgG-CH1-Rv bzw. IgD-CH1-Rv als Rv-Primern amplifiziert wurden. Mittels ANOVA Tests mit anschließendem Tukey-Test der erhaltenen $Ct_{adjusted}$-Werte wurde eine Kreuzreaktivität des qRT-PCR-IgG-Primers ausgeschlossen, die vom qRT-PCR-IgD-Primer jedoch nicht. Um die Ergebnisse des qRT-PCR-Kreuzreaktivitätstests zu analysieren, wurden ein PCR-Kreuzreaktivitätstest nach der Methode PCRqrt und eine anschließende Agarose-Gelelektrophorese durchgeführt: das Ergebnis des qRT-PCR-Tests wurde bestätigt (nicht dargestellt). Daher wurde eine bioinformatische Analyse der Primer durchgeführt. Es wurde festgestellt, dass die 700 bp IgD PCR1 Produkte als Kalibrierung-Template für die IgD-Mengenermittlung nicht einsetzbar sind, da der qRT-PCR-IgD-Rv Primer statt in der IgD CH1 Region in der IgD Hinge Region bindet (Abbildung 12). Weiterhin wurde eine Kreuzreaktivität des qRT-PCR-IgD-Primers ausgeschlossen, da der qRT-PCR-IgD-Rv Primer an keiner Stelle der gesamten IgG Sequenzen

binden kann (nicht dargestellt). Daher kann das qRT-PCR-IgD Primerpaar für die Quantifizierung der IgD Transkripte aus cDNA verwendet werden.

| Primer | qRT-PCR-IgD-Fw | IgD-CH1-Rv | qRT-PCR-IgD-Rv | IgD-Hinge-Rv |
|---|---|---|---|---|

IgD Sequenz

```
661 cagcagtggc gccaaggcga gtacaaatgc gtggtccagc acaccgcCAG CAAGAGTAAG
721 AAGGAGATCT tccgctggcc agagtctcca aaggcacagg cctcctcagt gcccactgca
781 caaccccaag cagagggcag cctcgccaag gcaaccacag ccccagccac cacccgtaac
841 acaggaagag gaggagaaga gaagaagaag gagaaggaga aagaggaaca agaagagaga
901 gagacaaaga caccagagtg tccgagccac acccagcctc ttggcgtcta cctgctaacc
961 cctgcagtgc aggacctgtg gctccgggac aaagccacct tcacctgctt cgtggtgggc
```

**Abbildung 12:** Teil der IgD-Sequenz (EMBL [65] BC063384). Der Teil der CH1 Region ist unterstrichen und endet mit dem 724. Nukleotid, während die Hinge Region mit dem 725. Nukleotid beginnt (IMGT/LIGM-DB [35] K02875 und K02876). Die Bindungsstelle des qRT-PCR-IgD-Fw Primers ist durch gelben Hintergrund dargestellt, die des qRT-PCR-IgD-Rv durch grünen und die des IgD-Hinge-Rv durch roten Hintergrund. Die Bindungsstelle des IgD-CH1-Rv ist durch groß und fett geschriebene Nukleotide gekennzeichnet. Zu beachten ist, dass die IgD Sequenz hier in 5' – 3' Richtung abgebildet ist, dementsprechend wäre die Bindungsstelle des qRT-PCR-IgD-Fw Primers auf der nicht gezeigten komplementären Sequenz in 3' – 5' Richtung.

### 4.1.2. Mengenverhältnis zwischen IgG und IgD bei Gesunden und Patienten

Das Experiment dient einerseits des Vergleichs zwischen Gesunden und Patienten und andererseits der Ermittlung des Mengenverhältnisses zwischen IgG und IgD Transkripten bei den einzelnen Probanden. Die Ergebnisse der SLE Proben konnten nicht analysiert werden, da die erhaltenen Ct Werte um ca. 10 höher als bei RA bzw. bei den gesunden Kontrollen (ND) waren und das Kontrollgen HPRT in der Hälfte der Proben nicht nachweisbar war.

Während die gesunden Kontrollen vergleichbar kleine IgG:IgD Verhältnisse zeigen (ca. 0,2 – 0,5), enthalten vier von fünf RA Patientenproben ca. 1,8- bis 2,9-fache IgG Menge. Die SD Proben sind heterogen mit einem IgG:IgD Verhältnis von 0,22 bis 1,14 (Abbildung 13). Die RA und SD Patienten weisen somit deutlich höhere IgG:IgD Verhältnisse als Gesunden auf.

Die IgG:IgD-Verhältnis-Mittelwerte betragen 0,33 bei den Gesunden (ND), 1,93 bei RA und 0,80 bei SD (Abbildung 13). Um festzustellen, ob die Mittelwert-Differenzen relevant sind, wurde ein ANOVA-Test mit darauffolgendem Dunnett-Test durchgeführt. Ein signifikanter Unterschied existiert zwischen RA-Patienten und Gesunden, während dies für SD und ND Kontrolle statistisch nicht bewiesen wurde.

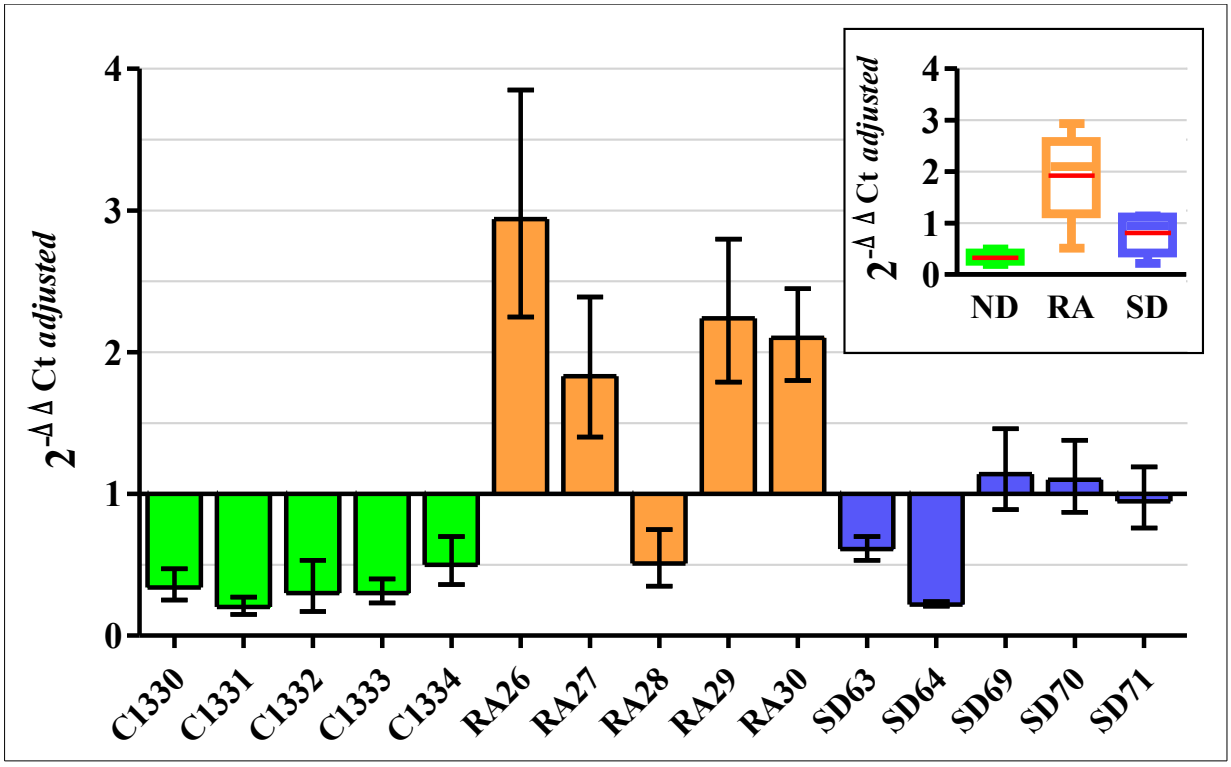



**Abbildung 13:** Verhältnisse zwischen IgG und IgD bei gesunden Probanden und RA sowie SD Patienten (Mittelwert und Standard-abweichung der Dreifachbestimmung). Die Verhältnisse bei den SLE Proben sind nicht abgebildet. Oben rechts sind die gleichen Daten mit Median, Quartile, Minimum und Maximum dargestellt, wobei die Mittelwerte mit roten Linien veranschaulicht sind.

### 4.1.3. Einsatz von cDNA Pools

Mit diesem Experiment sollte festgestellt werden, ob eine Mischung von mehreren Patientenproben für die Sequenzierung eingesetzt werden kann. Andererseits kann dadurch ein eventueller Einfluss der cDNA-Anfangskonzentration auf die Ergebnisse ausgeschlossen werden.

Dafür wurden die Werte für $2^{(-\Delta\Delta Ct\ adjusted)}$ und der $2^{(-\Delta\Delta Ct\ adjusted)}$-SA für den Pool der cDNA-Proben einerseits durch das Experiment ermittelt und andererseits theoretisch berechnet. Für die theoretische Berechnung wurde für jeden Pool der Durchschnitt der $Ct_{adjusted}$-Werte der fünf einzelnen Proben berechnet, analog wurde der Durchschnitt der SA der $Ct_{adjusted}$-Werte kalkuliert. Daraufhin wurde die weitere Auswertung wie für die einzelnen Proben bzw. für den experimentell ermittelten Pool durchgeführt.

Die IgG:IgD Verhältnisse (Abbildung 14) sind mit den ND Werten beim vorherigen Experiment (s. Kapitel 4.1.2) vergleichbar (IgG:IgD von ca. 0,18 – 0,52), somit kann ein Einfluss der Konzentration ausgeschlossen werden.

Der Unterschied zwischen dem experimentell ermittelten Pool (Pool-Experimentell, IgG:IgD = 0,36) und dem aus den Werten für die einzelnen Proben kalkulierten Pool-Mittelwert (Pool-Berechnung, IgG:IgD = 0,32) fällt gering aus. Dies gilt auch für die Standardabweichung bei einer maximalen Differenz von 0,09. Somit können unproblematisch Pools für die Sequenzierung eingesetzt werden.

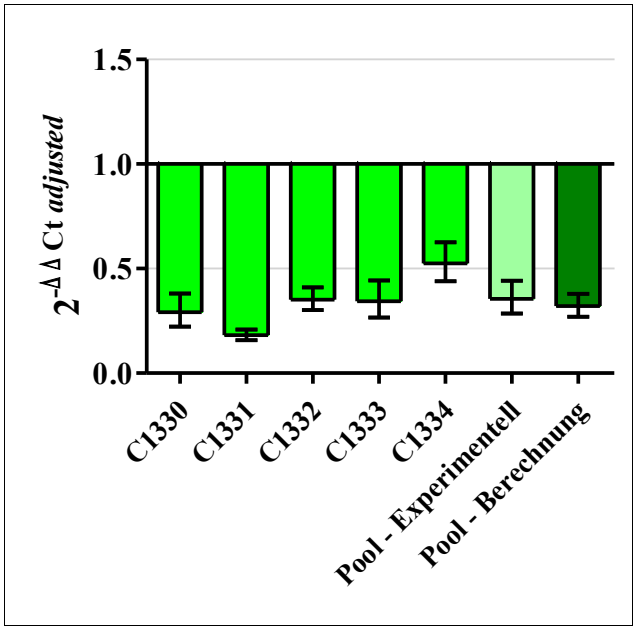



**Abbildung 14:** Verhältnis zwischen IgG und IgD bei den einzelnen Gesunden (Mittelwert und Standardabweichung der Dreifachbestimmung dargestellt).

### 4.1.4. Ermittlung der cDNA Stoffmenge in den Proben

Eine Quantifizierung der IgD bzw. IgG Transkripte könnte Unterschiede zwischen Gesunden und Autoimmunpatienten aufdecken. Daher wurde ein qRT-PCR-Experiment durchgeführt, um die cDNA Stoffmengen in den Proben nach der RT zu ermitteln.

Zunächst wurde eine Kalibrierung durchgeführt, um aus den $Ct_{adjusted}$-Werten der Proben die cDNA-Menge zu berechnen. Eingesetzt wurden IgG-VH bzw. IgG-VH4 Verdünnungsreihen von PCR1 Produkten mit bekannter cDNA-Konzentration (s. Kapitel 2.1.5). Alle DNA-Fragmente sind ca. 700 bp lang und beinhalten die V sowie die CH1 Region von IgG.

Bei der Erstellung der Kalibriergeraden wurde der 4. $Ct_{adjusted}$-Punkt bei IgG VH aufgrund der großen $Ct_{adjusted}$ Abweichung nicht berücksichtigt. Da die Steigung der IgG VH4 Kalibriergerade (-2,8, Abbildung A1 im Anhang) zu stark von der erwarteten abweicht (-3,4 bei der Primereffizienzermittlung, Abbildung 11), wurde für die Berechnung der cDNA-Stoffmengen nur die IgG VH Kalibriergerade eingesetzt. Für die Ermittlung der IgD Werte wurden die kalkulierte IgG:IgD Mengenverhältnise bei Gesunden und Patienten (s. Kapitel 4.1.2) eingesetzt. Die berechneten cDNA-Stoffmengen (Abbildung 15) entsprechen den mRNA-Stoffmengen in den Proben.

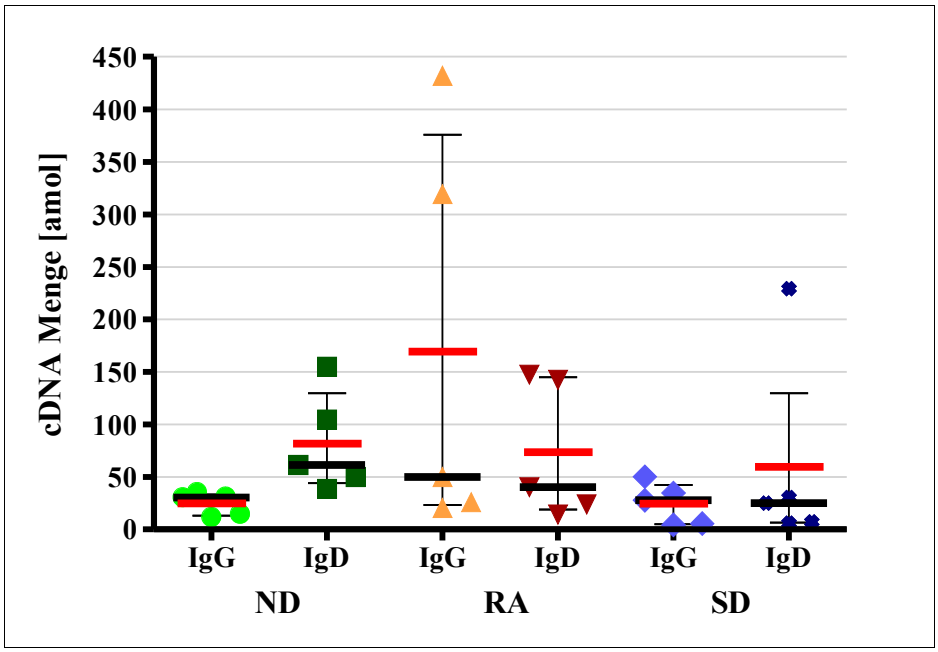



**Abbildung 15:** Ermittelte IgG Stoffmenge in den cDNA-Proben (entspricht mRNA Menge). Mit dicken schwarzen Linien sind die Mediane dargestellt. Die dünnen schwarzen Linien repräsentieren die Quartile, während die dicken roten Linien Mittelwerte abbilden.

Die ermittelten IgG Werte sind bei ND und SD in einem engen Bereich zusammengefasst (ca. 6-50 amol). Im Gegensatz dazu ist die IgG Verteilung bei RA Autoimmunpatienten wesentlich größer: 20 – 432 amol. Die IgD Werte sind bei ND und RA im Bereich von 14 bis 152 amol wesentlich breiter gestreut als bei SD: bei vier von fünf Patienten ist die IgD cDNA Stoffmenge zwischen 5 und 31 amol, während die letzte SD Probe ein Ausreißer ist (229,1 amol). Laut Two-way ANOVA existieren keine signifikante Unterschiede zwischen den Gruppen, jedoch zeigt der anschließende Bonferroni-Posttest signifikante Differenz zwischen RA und Gesunden in der IgG Stoffmenge ($P < 0,05$).

### 4.1.5. Vorbereitung der Next Generation Amplicon Sequenzierung

Für die Sequenzierung mittels *Genome Sequencer FLX Instrument* müssen die einzusetzenden DNA-Fragmente die Bindestellen der Sequenzierprimer (vom Hersteller als Primer A und Primer B bezeichnet [56]) enthalten. Zusätzlich würde das Einführen von Erkennungs-sequenzen (MIDs) [29] die Zuordnung der Sequenzen zu der jeweiligen Probe erleichtern. Das Einfügen der Primer A und B sowie der MIDs erfolgt mittels zusammengesetzter Fusion-Primer in drei vorbereitenden PCR-Schritten (PCR1 – PCR3). In PCR1 werden die Gene amplifiziert, um ausreichend Material für die weiteren PCR Phasen zu erhalten. In PCR2

werden Primer A und MID angehängt, während in PCR3 die Sequenzen mit Primer B vervollständigt werden.

Für die Sequenzierungsvorbereitung wurden stets vier Primerkombinationen eingesetzt, um Unterschiede in der Gen-Nutzung der V-Genfamilie VH4 und speziell des V Gens VH4(DP63) bei den Isotypen IgG und IgD zu untersuchen.

PCR1 Phase: Um genügend Ausgangsmaterial zu bekommen, wurden zunächst aus Gel isolierte 700 bp Fragmente (mit VH4-Fw bzw. VH4(DP63)-Fw als Fw-Primer und IgG-CH1-Rv bzw. IgD-CH1-Rv als Rv-Primer hergestellt) amplifiziert und aufgereinigt. Mittels Agarose-Gelelektrophorese wurde wie erwartet ein ca. 700 bp Fragment nachgewiesen, jedoch wurden weitere DNA-Fragmente mit anderen Größen (hauptsächlich über 700 bp) trotz Gelaufreinigung der Template sichtbar gemacht (nicht dargestellt). Um eventuelle Einflüsse der cDNA-Konzentration auf die weiteren Experimente auszuschließen, wurden die PCR-Produkte gemischt und nach Konzentrationsmessung auf 10 ng/µL verdünnt.

PCR2 Phase: Als Nächstes wurde überprüft, ob die neuen Fusion-Primer für die Einführung des Primers A und der MID-Sequenz funktionieren. Parallel wurde untersucht, ob eine unterschiedliche Template-Konzentration sowie die Zugabe von ET SSB einen positiven Einfluss auf die Amplifizierung haben (Abbildung 16). Gelelektrophoretisch wurde ein ca. 450 bp langes PCR-Produkt nachgewiesen, jedoch sind zusätzliche DNA-Fragmenten vorhanden. Weiterhin sind keine Unterschiede bezüglich des Additivs ET SSB feststellbar.

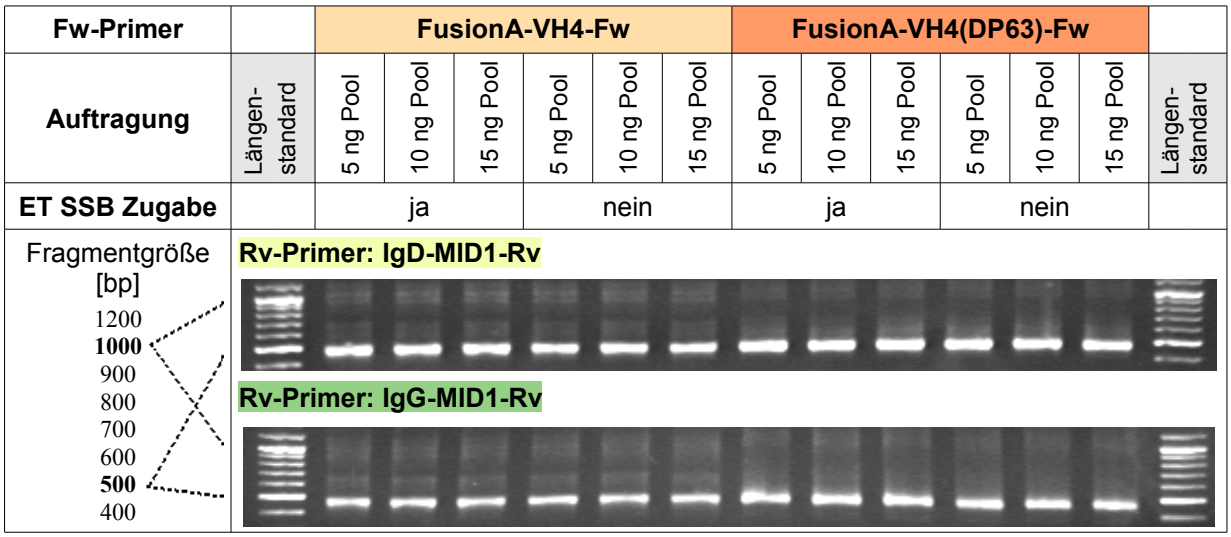



**Abbildung 16:** Vergleich der Methoden PCR2o und PCR2m.

Da der Einfluss der Template-Konzentration mittels Agarose-Gelelektrophorese nicht ermittelt werden konnte, wurde als Nächstes die notwendige Template Konzentration für die Amplicon Sequenzierung berechnet. Laut Hersteller sollte die ursprüngliche Probe 1 bis 5 ng DNA (in Form von PCR-Fragmenten) bzw. 10 – 50 ng komplexe (z.B. genomische) DNA in höchstens 2 µL enthalten sein, während die Anzahl der Moleküle mit beiden Fusion-Primern am Ende der Amplifizierung $10^{11}$ nicht überschreiten darf [57]. Für die Kalkulation wurde eine OpenOffice.org Calc Tabelle programmiert, wobei dsDNA mit einem durchschnittlichen Molekulargewicht eines Nukleotids von 333 g/mol als Berechnungsbasis diente. Zusätzlich wurde angenommen, dass DNA erst ab 10 ng mittels Agarose-Gelelektrophorese detektierbar ist.

Es besteht keine Möglichkeit, die gewünschte Anzahl an Molekülen zu erhalten und gleichzeitig das Ergebnis der Amplifizierung bei Auftragung von 5 µL auf einem Agarosegel zu verifizieren (Tabelle A2 im Anhang). Aus diesem Grund wurde die erforderliche Anzahl von PCR Zyklen für die Einhaltung des Sequenzierungskriteriums (weniger als $10^{11}$ Moleküle nach Amplifizierung) und die Anzahl der zusätzlichen PCR Zyklen für eine Detektierbarkeit auf Agarosegel für die drei PCR-Phasen bei 1 ng IgG Template und einer Gesamtlänge der Template von 1450 bp ermittelt (Tabelle 14). Eine Berechnung der Template-Konzentration kann nur gemeinsam für alle drei PCR Phasen erfolgen, da sich das Sequenzierungskriterium auf die Molekülanzahl nach der dritten PCR Phase bezieht. Durch die präzise Kalkulation ergeben sich insgesamt 11 Zyklen für die drei PCR Phasen, wobei die Reihenfolge 4-4-3, 3-4-4 sowie 4-3-4 Zyklen sein darf (Ergebnis nicht dargestellt). Für die Reihenfolge 4-4-3 ergeben sich 9, 8 bzw. 6 weitere PCR Zyklen, die für einen Nachweis mittels Gelelektrophorese notwendig sind. Die genaue Ig cDNA Konzentration in den durch RT hergestellten Proben ist jedoch aufgrund der benutzten Gesamt-RNA nicht bekannt. Daher wurde eine modifizierte OpenOffice.org Calc Tabelle programmiert, um die erforderliche Anzahl von PCR Zyklen aus den mittels qRT-PCR ermittelten Ig cDNA Stoffmengen zu berechnen. Eingesetzt wurden die niedrigsten Mittelwerte (24,51 bzw. 59,55 amol für IgG bzw. IgD). Somit sind für die Einhaltung des Sequenzierungskriteriums jeweils 5 PCR Zyklen pro Phase maximal möglich. Für den Nachweis mittels Gelelektrophorese sind bei dem niedrigsten Mittelwert (24,51 amol) jeweils zusätzliche 15-13-10 PCR Zyklen erforderlich. Da die so ermittelte Gesamtanzahl an PCR Zyklen höher ist als bei der Berechnung mit 1 ng Template-Menge, befindet sich in den Proben weniger als 1 ng Ig cDNA. Dementsprechend wurden für die darauffolgenden Experimente der PCR2 Phase deutlich niedrigere Template-Mengen eingesetzt (100, 10 bzw. 1 pg).

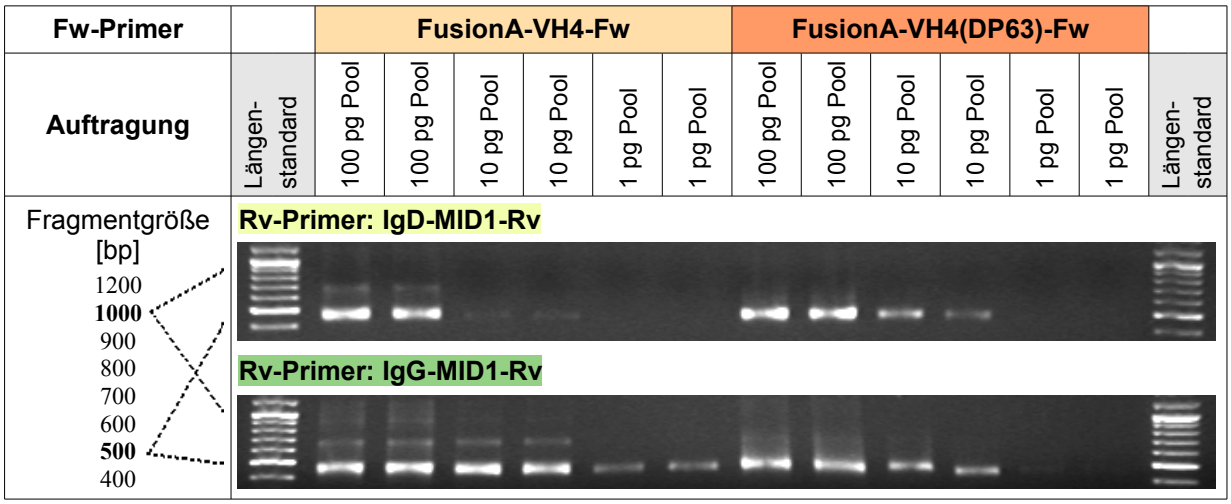
**Tabelle 14:** Berechnung der erforderlichen Anzahl von PCR Zyklen mit der programmierten OpenOffice.org Calc Tabelle. Dargestellt ist die Kalkulation der erforderlichen Entnahme (Volumen in µL) von den Ansätzen für die darauffolgende PCR sowie der erforderlichen Anzahl an Zyklen für jede PCR. Rechts ist die nötige Anzahl der PCR Zyklen für die Detektierbarkeit mittels Gelelektrophorese, wobei jeweils ein zusätzlicher Zyklus zur Sicherheit angegeben wurde. Die grau unterlegten Zahlen werden nicht eingegeben, sondern automatisch berechnet. Annahme: die IgG Template-DNA hat eine Gesamtlänge von ca. 1450 bp.

| | Eingabe | | | | | | | Ausgabe | | |
|---|---|---|---|---|---|---|---|---|---|---|
| PCR | Anzahl der PCR Zyklen | Template [ng] | Template Konzentration [ng/µL] | Länge der Template [bp] | Länge des PCR Produkts [bp] | PCR Ansatzvolumen [µL] | Von vorheriger PCR zu entnehmen [µL] | PCR Produkt & Template [Moleküle] | PCR Endkonzentration [ng/µL] | Anzahl der zusätzlichen PCR Zyklen für Detektierbarkeit |
| PCR1 | 4 | 1,00 | 0,50 | 1450 | 700 | 50 | | 9,98E+09 | 0,16 | 9 |
| PCR2 | 4 | *1,65* | *0,16* | | 450 | 50 | 10 | 3,19E+10 | 0,33 | 8 |
| PCR3 | 3 | *3,31* | *0,33* | | 465 | 50 | 10 | 5,11E+10 | 0,53 | 6 |

Um die Nebenproduktamplifizierung bei der PCR2 Phase zu reduzieren und um den Einfluss der Template-Menge zu untersuchen, wurde eine Touchdown PCR durchgeführt (Abbildung 17A). Die amplifizierten PCR-Produkte wurden mittels Agarose-Gelelektrophorese nachgewiesen, wobei sich die Intensität der DNA-Banden mit Abnahme der Template-Konzentration verringert. Bei 100 pg und teilweise bei 10 ng Template traten Nebenprodukte der PCR auf. Dementsprechend wurde eine Gradient-PCR mit 100 pg Template durchgeführt (Abbildung 17B). Die zu erwartenden DNA-Fragmente (ca. 450 bp) sind als sehr intensive Banden auf dem Agarose-Gel sichtbar, jedoch konnte die Nebenproduktkonzentration bei keiner der unterschiedlichen Annealing-Temperaturen signifikant reduziert werden. Die mit den vier unterschiedlichen (bezüglich V Gen bzw. Antikörper-Isotyp) Primerkombinationen amplifizierten PCR-Produkte unterscheiden sich nur in der Größe der amplifizierten Nebenprodukte.

PCR3 Phase: Der neue Primer für die letzte PCR Phase (Einführung des Primers B) wurde eingesetzt und seine Funktionalität wurde überprüft. Nach der Standard-PCR3 (Abbildung 18A) wurde eine Reduzierung der Nebenproduktkonzentration mittels Gradient-PCR3 angestrebt (Abbildung 18B). Bei beiden PCR-Methoden sind die Ergebnisse ähnlich: mittels Agarose-Gelelektrophorese werden die ca. 465 bp langen PCR-Produkte nachgewiesen, jedoch wurde keine Reduzierung der Nebenproduktamplifizierung bei der untersuchten Annealing-Temperaturen festgestellt. Bei der Gen-Amplifizierung mit der Primerkombination FusionA-VH4(DP63)-Fw – FusionB-MID1-Rv (auf IgG-MID1-Rv) werden zum ersten Mal kurze DNA-Fragmente mit einer Größe von ca. 180 bzw. ca. 420 bp identifiziert.

**A**




**B**

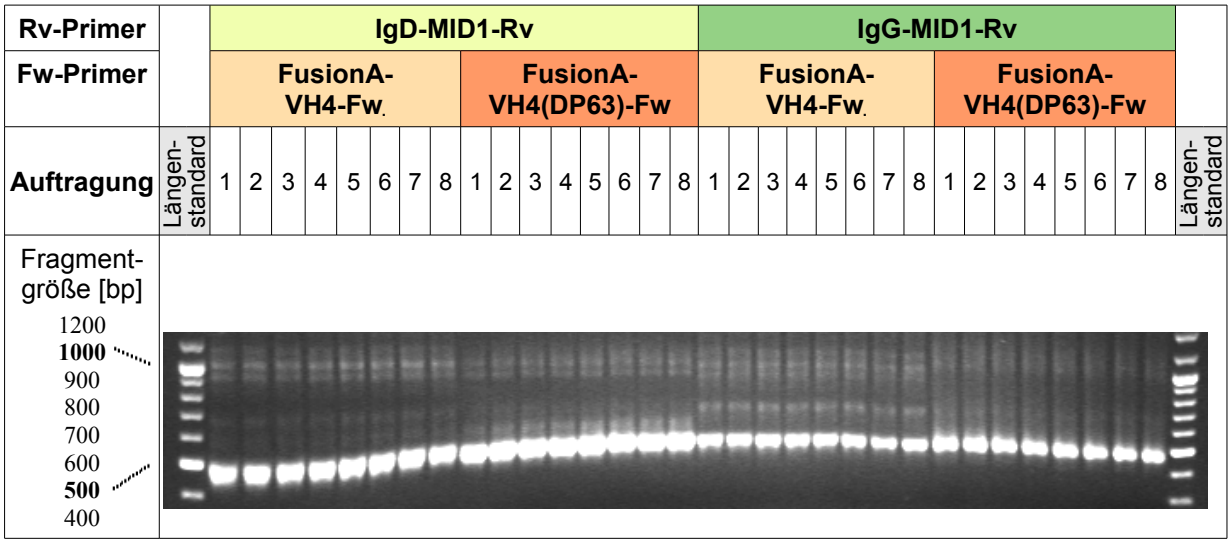
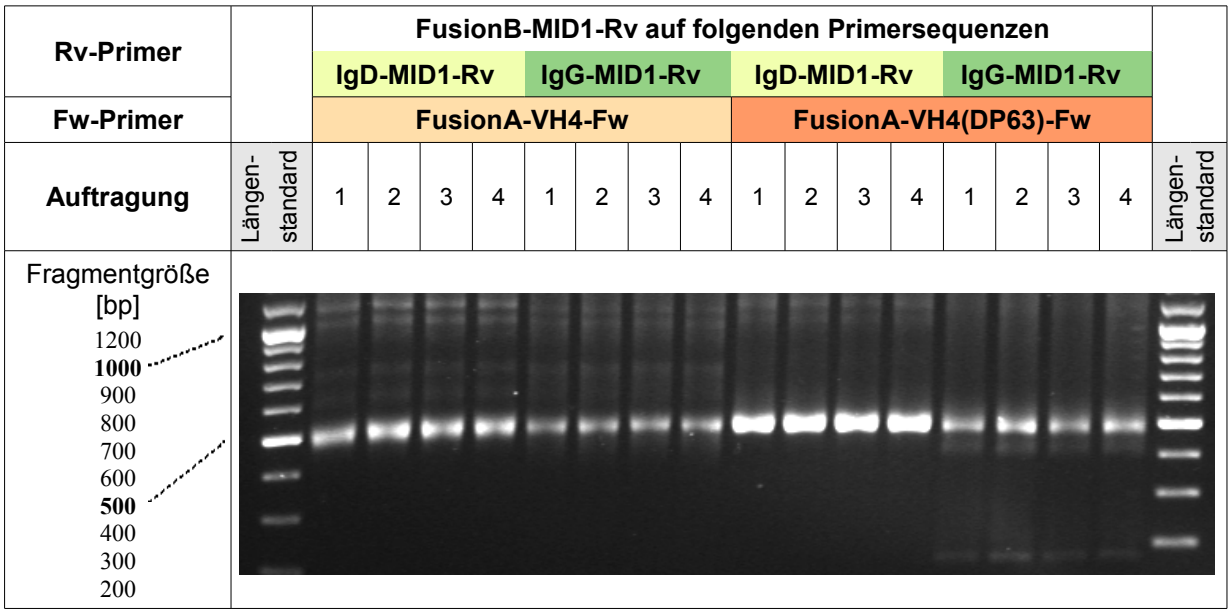



**Abbildung 17:** Darstellung der Ergebnisse der Methoden TD-PCR2 (A) bzw. Gradient-PCR2 (B). 1-8: 100 pg Template-DNA, amplifiziert bei den jeweils folgenden Annealing-Temperaturen: 55,7; 56,8; 58,3; 60,0; 61,4; 62,5; 63,3; 63,8 °C.

**A**




**B**

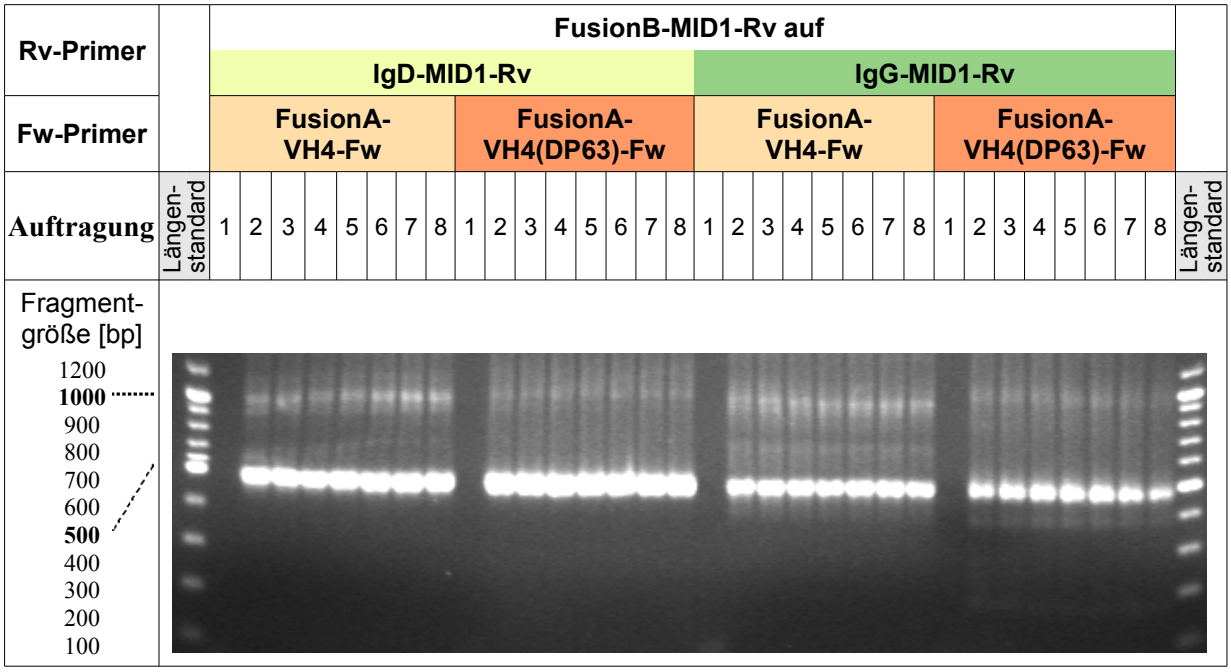



**Abbildung 18:** Darstellung der Ergebnisse der Methoden PCR3 (A) und Gradient-PCR3 (B):
A: Template-DNA: 1-3: die drei aufgereinigten PCR-Produkte der Gradient-PCR2 (einzeln), 4: Mischung aus den drei aufgereinigten PCR-Produkten der Gradient-PCR2.
B: 1: NCT, amplifiziert bei 55,2 ºC. 2-8: Template-DNA: 1:5 Verdünnung aus der Mischung der drei aufgereinigten PCR-Produkte der Gradient-PCR2; amplifiziert bei den jeweils folgenden Annealing-Temperaturen: 56,6; 58,5; 60,8; 63,5; 65,8; 67,5; 68,8 ºC.

## 4.2. Bioinformatische Untersuchungen

Die Auswertung von Next Generation Sequenzierdaten stellt eine besondere bioinformatische Herausforderung aufgrund der eingesetzten Pyrosequenzierung dar. Infolgedessen wurden bioinformatische Untersuchungen durchgeführt, um mögliche Probleme im Voraus zu erkennen und programmiertechnisch zu lösen bzw. um bestimmte Parameter empirisch zu ermitteln. Zusätzlich wurde die *Coverage* der genspezifischen Fw-Primer untersucht.

### 4.2.1. Genauigkeit der Identifizierung von V Genen in gekürzten Sequenzen und Ermittlung der minimal erforderlichen Read-Länge

Ein Test mit gekürzten Sequenzen wurde durchgeführt, um die Genauigkeit der V-Gen-Identifizierung zu untersuchen, da der 5'-Anfang des V-Gens aufgrund der Next Generation Antikörpergensequenzierung vom 3'-Ende z.T. nicht mehr sequenziert wird.
Als Erstes wurden die menschlichen Keimbahngene aus der VBASE2 in Schritten von 10 bp bis zu 100 bp in 5'-3' Richtung gekürzt und anschließend gegen sich selbst aligniert. Daraufhin wurden sie in drei Gruppen aufgeteilt. Die erste Gruppe beinhaltet alle Gene, bei denen trotz Kürzung nur das analysierte Gen gefunden worden ist. Die Gruppe 2 enthält Gene, bei denen außer dem ursprünglichen V Gen ein oder mehrere Gene den gleichen Score besitzen. Bei der Gruppe 3 wurde kein Gen oder ein anderes Gen (bzw. mehrere Gene) als das Ausgangsgen identifiziert. Demnach führt eine Kürzung in 5'-3' Richtung der Gruppe 2 bzw. 3 Gene zu einer nicht eindeutigen oder falschen V Gen Identifikation.
Als Zweites wurde das gleiche Verfahren mit den scFv-Testsequenzen aus den naiven humanen scFv Antikörperbibliotheken HAL7 und HAL4 [47,48] durchgeführt. Jedoch könnten mehrere scFv-Testsequenzen das gleiche Keimbahngen enthalten, das zusätzlich Mutationen aufweist. Daher wurde die Gruppe 4 für die ungekürzten Testsequenzen definiert, bei denen mehrere Gene mit dem gleichen Score gefunden worden sind und daher kein Vergleich mit einem einzigen, ursprünglichen Gen möglich ist.
Die Aufteilung der Gene bzw. Testsequenzen in den Gruppen ist in Abbildung 19 dargestellt.

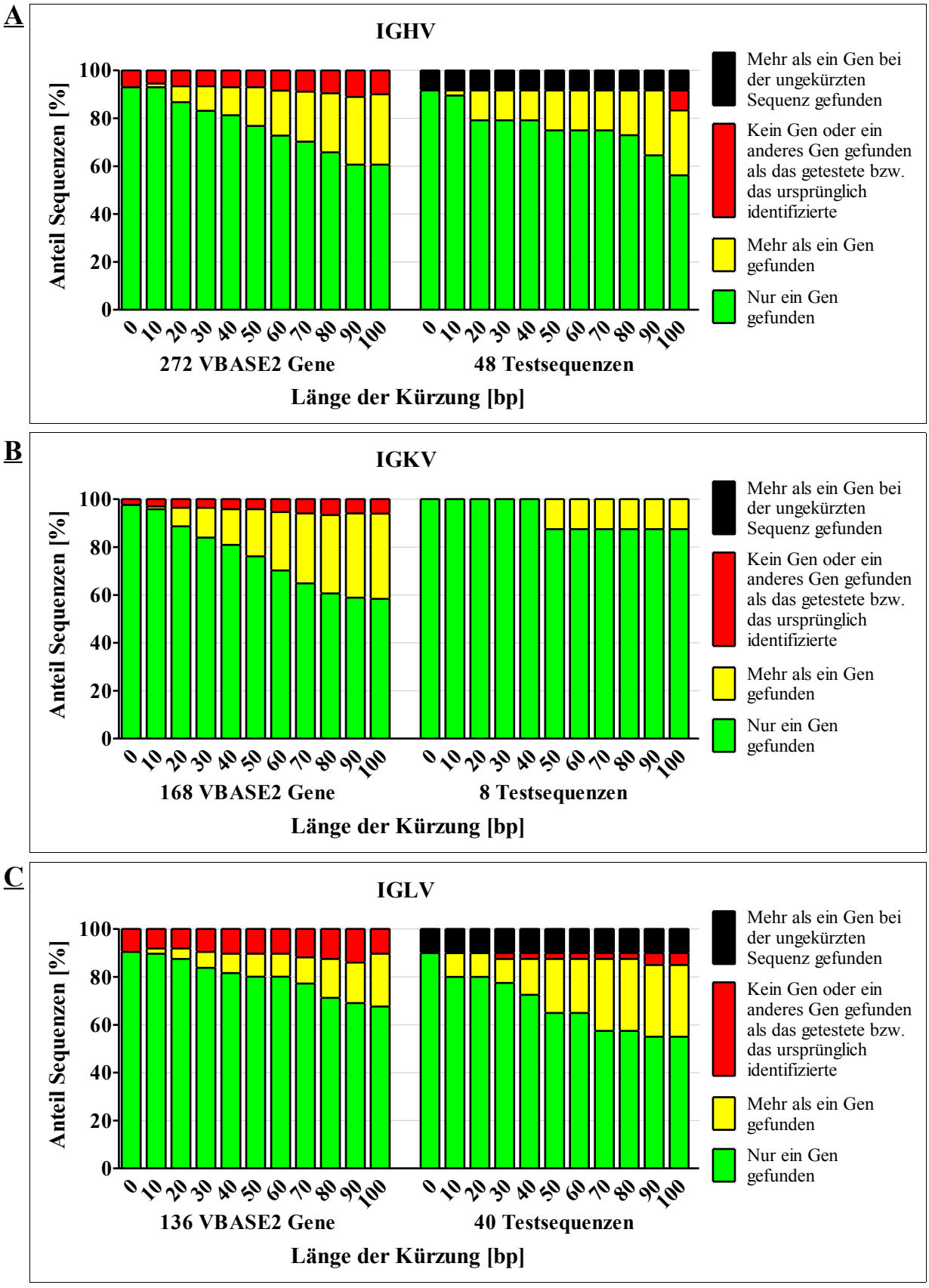



**Abbildung 19:** Identifikation von gekürzten IGHV (A), IGKV (B) und IGLV (C) Sequenzen.

Im Detail zeigt sich, dass bei den ungekürzten V Genen ein geringer Anteil der Gene (weniger als 10 %) der Gruppe 3 gehört. Alle dieser Gene sind entweder Pseudogene oder Orphans mit keiner zugeordneten Familie. Eine Ausnahme stellt das humIGHV164 (VH6-DP74, IGHV6-1*01) dar: anstelle des funktionellen humIGHV164 wurde das Pseudogen humIGHV072 aus der VH6-Familie identifiziert. Der Prozentsatz der Gruppe 1 Gene sinkt erwartungsgemäß mit der Erhöhung der Kürzungslänge, wobei diese Abnahme am höchsten bei IGKV ist, gefolgt von IGHV und IGLV (Abbildung A2 im Anhang). Bei 100 bp Kürzung sind noch ca. 60 % der V Gene in Gruppe 1.

Die Auswertung der Testsequenzen liefert ähnliche Ergebnisse. Allerdings sind 8,3 bzw. 10 % der ungekürzten IGHV und IGLV Testsequenzen in Gruppe 4 (statt in Gruppe 3), während alle IGKV Sequenzen der Gruppe 1 zugeordnet sind. Außerdem wechselt nur eine von acht IGKV Sequenzen von Gruppe 1 in Gruppe 2, und zwar ab einer Kürzungslänge von 50 bp. Somit gehören bei 100 bp Kürzung 87,5 % der IGKV bzw. ca. 55 % der IGHV und IGLV scFv-Testsequenzen der Gruppe 1 an.

Um ca. 80 % der Gene korrekt zu identifizieren, darf die Kürzung in 5'-3' Richtung des V Gens eine Länge von ca. 20 bp nicht überschreiten.

Um eine falsche V-Gen-Identifizierung aufgrund von gekürzten Sequenzen zu vermeiden, wurde die optimale Read-Länge berechnet. Dafür wurde die Länge der scFv-Testsequenzen in Nukleotiden ermittelt (Abbildung 20). Um die längste IGHV Sequenz (390 bp) vollständig mit dem *Genome Sequencer FLX Instrument* zu ermitteln, ergibt sich bei Berücksichtigung des (längeren) IgD Sequenzabschnitts (48 bp) und des MID (10 bp) eine optimale Read-Länge von 448 bp. Um ca. 60 % der V Gene korrekt zu identifizieren, wurde zusätzlich die minimal erforderlichen Read-Länge für IGHV Sequenzen ermittelt (404 bp), indem das IGHV 75 % Quartil (366 bp) eingesetzt wurde sowie eine maximale Kürzung von 20 bp angenommen wurde. Da eine Read-Länge von 400 bp mit dem Titanium-Kit erreicht werden kann, ist bei der Auswahl der Sequenzen für den Gennutzungsvergleich eine minimale Sequenzlänge von 400 bp zu empfehlen.

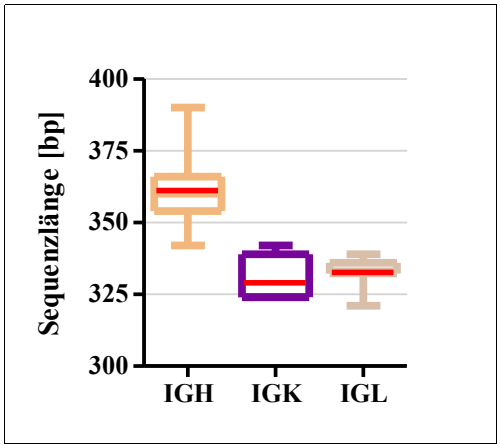



**Abbildung 20:** scFv-Sequenzlänge: Median, Quartile, Minimum und Maximum sind dargestellt, wobei auch der Mittelwert zu sehen ist (jeweils eine rote Linie).

Analog können die Read-Längen für IGKV und IGLV kalkuliert werden. Da sich die Maxima um nur 3 bp von den 75 % Quartilen bei beiden unterscheiden, könnten sie eingesetzt werden (IGKV: 342 bp, IGLV: 339 bp). Für eine Berechnung wären jedoch die Längen der amplifizierten CL Sequenzabschnitte notwendig.

### 4.2.2. V-Gen-Analyse auf Nukleotid-Homopolymeren

Das *Genome Sequencer FLX Instrument* basiert auf der Pyrosequenziertechnologie, die anfällig für Homopolymere ist. Ein Beispiel ist in Abbildung 21 dargestellt: das humIGHV178 (VH3-DP38, IGHV3-15*01) Gen hat insgesamt 7 Homopolymerregionen. Davon sind vier 4 bp, eins 5 bp und zwei 6 bp lang. Das zu untersuchende Gen VH4(DP63) (humIGHV201, IGHV4-34*01) umfasst insgesamt vier Homopolymerregionen (drei von 4 G und eins von 5 C) (Abbildung A3 im Anhang).

> humIGHV178 300 bp

GAGGTGCAGCTGGTGGAGTCTGGGGGAGGCTTGGTAAAGCCTGGGGGG
TCCCTTAGACTCTCCTGTGCAGCCTCTGGATTCACTTTCAGTAACGCC
TGGATGAGCTGGGTCCGCCAGGCTCCAGGGAAGGGGCTGGAGTGGGTT
GGCCGTATTAAAAGCAAAACTGATGGTGGGACAACAGACTACGCTGCA
CCCGTGAAAGGCAGATTCACCATCTCAAGAGATGATTCAAAAAACACG
CTGTATCTGCAAATGAACAGCCTGAAAACCGAGGACACAGCCGTGTAT
TACTGTACCACA

| Länge der Homo-polymere | Anzahl der Homo-polymere |
|---|---|
| 4 bp | 4 |
| 5 bp | 1 |
| 6 bp | 2 |
| ≥ 7 bp | 0 |

**Abbildung 21:** Nukleotid-Homopolymere am Beispiel des humIGHV178 (VH3-DP38, IGHV3-15*01) Gens. Die unterschiedlichen Längen der Homopolymere sind mit verschiedenen Farben gekennzeichnet.

Da eine falsch ermittelte Nukleotidanzahl in Homopolymerregionen Alignment-*Frameshifts* zur Folge hat, wurden die scFv-Testsequenzen sowie die humanen VBASE2 V Gene hinsichtlich der Länge und der Anzahl der Nukleotid-Homopolymere analysiert.
Nur eins von allen VBASE2 V Genen (576) sowie nur eine Testsequenz (von 96) wiesen eine Homopolymerlänge von 8 oder mehr bp auf, daher wurden vier Längenkategorien betrachtet: Homopolymere gleich lang oder länger als jeweils 4, 5, 6 oder 7 bp.

Die Ergebnisse der Homopolymer-Analyse sind in Abbildung 22 veranschaulicht:

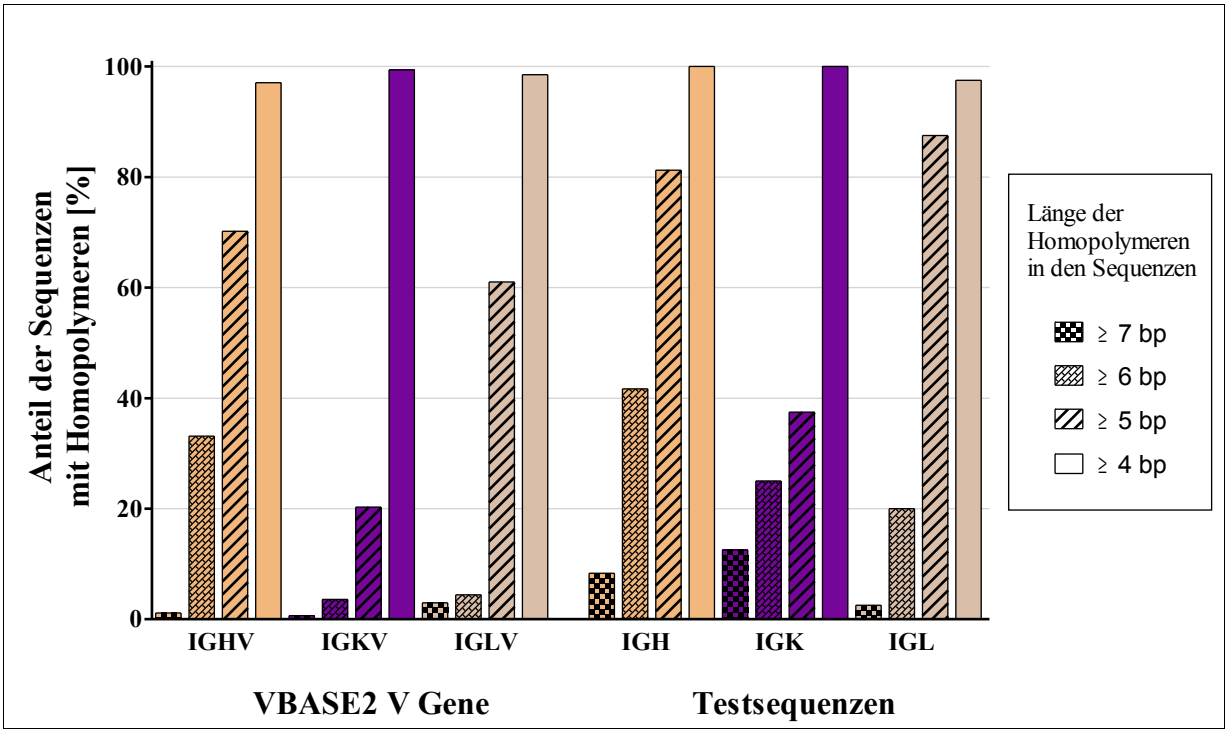



**Abbildung 22:** Anteil der Sequenzen mit Homopolymeren, sortiert nach der Homopolymerlänge.

Ca. 100 % der analysierten Testsequenzen bzw. V Gene enthalten Homopolymerregionen mit einer Länge ab 4 bp (Abbildung 22). Mit Zunahme der Länge sinkt der Anteil der Sequenzen mit Homopolymeren sowohl bei den V Genen als auch bei den Testsequenzen, jedoch ist dieser Anteil bei den Testsequenzen wesentlich höher. Bei den IGHV Genen ist der Anteil der Sequenzen mit Homopolymerlänge ab 6 bp (33,1 %) deutlich höher als bei IGKV (3,6 %) bzw. IGLV (4,4 %), während der Anteil bei den Testsequenzen auf 41,7 % (IGHV), 25 % (IGKV) bzw. 20 % (IGLV) steigt. Außerdem weist IGKV einen erheblich geringeren Anteil an Sequenzen mit Homopolymeren ab 5 bp auf (ca. 40-50 % weniger als IGHV bzw. IGLV).
Darüber hinaus wurde die Anzahl der Homopolymere in den Sequenzen in Abhängigkeit von der Homopolymerlänge analysiert (Abbildung A4 im Anhang): je kürzer die Homopolymere sind, desto höher ist die Anzahl. Bei den V Genen wurden bis zu 7 Homopolymerregionen gezählt, bei den Testsequenzen lag die maximale Anzahl bei 12.

### 4.2.3. Bestimmung der *Cut Off* Werte in den scFv-Testsequenzen

Um den programmiertechnischen Parameter *Cut Off* zu berechnen, wurde der durchschnittliche Mutationsanteil der scFv-Testsequenzen (verglichen mit den Keimbahngenen) ermittelt. Der *Cut Off* Wert entspricht dem Prozentsatz der Identität zwischen der getesteten Sequenz und dem Keimbahngen und kann einzeln für die Ig Ketten sowie für die Gensegmente berechnet werden. Der festgelegte *Cut Off* Wert wird mit der Identität zwischen einer Sequenz und dem jeweiligen Keimbahngen verglichen: falls der *Cut Off* Wert größer ist, enthält die Sequenz zu viele Mutationen. Bei dem D Segment wird kein *Cut Off* Wert aufgrund der kurzen D Segment Länge sowie der durch „*DNAPLOT Query*“ begrenzten Anzahl an Mutationen (maximal zwei) ermittelt.

Bei beiden betrachteten Gensegmenten (V und J) ist der maximale Mutationsanteil am höchsten für IGKV, gefolgt von IGLV und IGHV (Abbildung 23). Da sich die Mediane (V: 2,0-4,0; J: 0,0) und die Mittelwerte (V: 2,5-4,6; J: 1,2-3,0) deutlich voneinander unterscheiden, wurden zwei statistische Tests durchgeführt. *One-Way ANOVA* mit anschließendem *Tukey's Multiple Comparison Test* ergab, dass sich nur die IGHV Mittelwerte signifikant von IGKV bzw. IGLV unterscheiden. Dementsprechend kann ein gemeinsamer *Cut Off* Wert für die V Gene der LC sowie ein gemeinsamer *Cut Off* Wert für alle J Gene berechnet werden. *Kruskal-Wallis Test* mit anschließendem *Dunn's Multiple Comparison Test* lieferte jedoch einen signifikanten Unterschied zwischen den IGHV und IGLV Medianen und zusätzlich zwischen den IGHJ und IGLJ Medianen, obwohl beide 0,0 betrugen. Da jedoch nur 8 IGK Sequenzen ausgewertet wurden, wurden drei *Cut Off* Werte berechnet: für IGHV (97,5 %), für IGKV und IGLV (94,6 %) sowie für J Gensegmente (97,0 %).

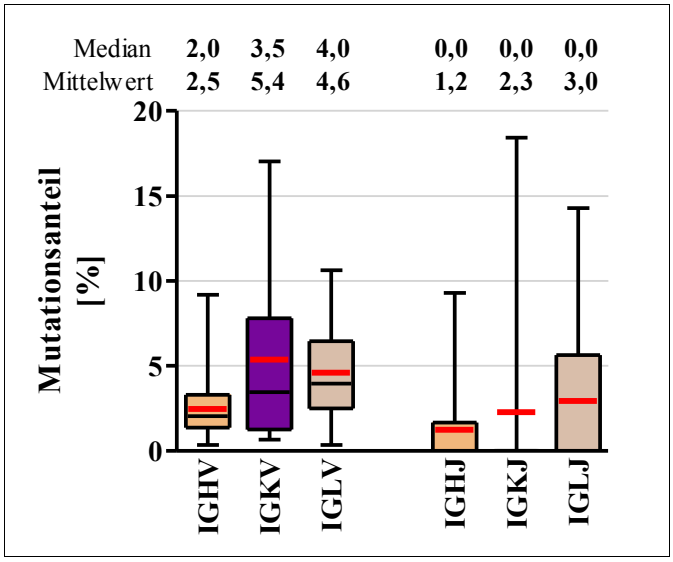



**Abbildung 23:** Mutationsanteil in den V und J Gensegmenten der scFv-Testsequenzen. Median, Quartile, Minimum und Maximum sind dargestellt, wobei auch der Mittelwert veranschaulicht ist (jeweils eine rote Linie).

## 4.2.4. Anzahl der durch Fw-Primer verursachten Mutationen in den scFv-Testsequenzen

Bei der Bestimmung des Mutationanteils in den scFv-Testsequenzen ist die hohe Anzahl an Mutationen in FR1 aufgefallen (Abbildung 24), die bei einer Histogrammdarstellung der Mutationsanteile eventuell übersehen werden könnte (Abbildung A5 im Anhang).
Da FR1 doppelt so viele Mutationen wie die um ca. 48 % längere FR3 Region enthält, wurde überprüft, ob die Mutationen in FR1 teilweise durch die Primer verursacht werden. Dies trifft bei über 70 % der IGHV bzw. IGLV Mutationen und bei ca. 42 % der IGKV Mutationen zu (Tabelle 15).

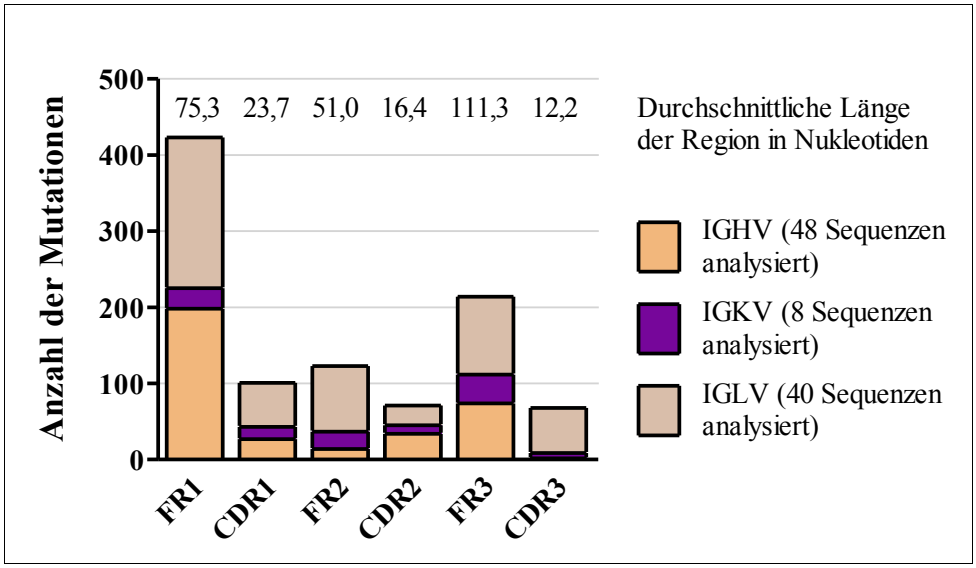



**Abbildung 24:** Anzahl der Mutationen im V Gensegment der scFv-Testsequenzen, aufgeteilt nach FRs und CDRs.

**Tabelle 15:** Anzahl und prozentualer Anteil der durch Primer verursachten Mutationen in FR1.

| Gene | Anzahl der analysierten Sequenzen | Anzahl der Mutationen in FR1 ab dem 1. Nukleotid | Anzahl der Mutationen in FR1 ab dem 24. Nukleotid | Theoretische Anzahl der Mutationen in FR1 bei Beibehaltung der Mutationsrate | Durch Primer verursachte Mutationen in FR1 | |
|---|---|---|---|---|---|---|
| | | | | | [Anzahl] | [%] |
| IGHV | 48 | 225 | 42 | 54,81 | 170,19 | 75,64 |
| IGKV | 8 | 27 | 14 | 15,6 | 11,4 | 42,22 |
| IGLV | 40 | 198 | 44 | 59,13 | 138,87 | 70,13 |

Außerdem wurde berechnet, wie hoch der durchschnittliche Mutationsanteil in den V Gen Sequenzen wäre, falls die Primer keine Mutationen verursachen würden (Tabelle 16). Demnach wären bei Beibehaltung der FR1-Mutationsrate in den IGHV Genen ca. 1%, in IGKV ca. 0,4 % und in IGLV ca. 1,2 % weniger Mutationen FR1 zu finden.

**Tabelle 16:** Anteil der Mutationen im V Gen. *Theoretisch:* bei Beibehaltung der Mutationsrate in FR1 und kein Auftreten von durch Primer verursachten Mutationen; *praktisch:* ermittelte Werte.

| | | Gene | | |
|---|---|---|---|---|
| | | IGHV | IGKV | IGLV |
| Anteil der Mutationen im V Gen, Mittelwert [%] | praktisch | 2,48 | 5,39 | 4,61 |
| | theoretisch | 1,49 | 5,02 | 3,42 |

### 4.2.5. V-Gen-*Coverage*

Die für die Durchführung des Projekts vorgesehenen Primersequenzen basieren auf bisher veröffentlichten Oligonukleotiden [66]. Um festzustellen, ob dieses Primer Set ausreichende Abdeckung (*Coverage*) der momentan 576 V Gene in VBASE2 gewährleistet, wurde eine *Coverage*-Analyse durchgeführt. Sofern die überlappte Region eine minimale Länge von 18 bp vom 3' Ende des Primers aufwies und die Anzahl der *Mismatches* (verglichen mit den Keimbahngenen) 4 nicht überschritt, wurde das Gen als von den Primern amplifiziert definiert. Da die Primer degeneriert sind, werden zum Teil V Gene unterschiedlicher V Genfamilien amplifiziert. Dies wurde bei der Benennung der V Gen Primer berücksichtigt: z.B. VL1459 amplifiziert IGLV Gene aus den VL Familien VL1, VL4, VL5 und VL9. Falls hinsichtlich der Degeneration der Oligonukleotide ein Gen von mehr als einem Primer amplifiziert wurde, wurde nur der Primer mit der niedrigeren *Mismatch*-Anzahl berücksichtigt.

Laut Sblattero und Bradbury [66] amplifiziert der Primer VH157 die Genfamilie VH2, während der Primer VK3 die V Gene der VK6 Familie abdeckt und der Primer VK246 *Coverage* der Familie VK3 leistet. Dies konnte jedoch durch die durchgeführten Analysen nicht bestätigt werden. Weiterhin ist die Nukleotidbezeichnung S an der falschen Position im Primer VH4: statt CAGGTGCAGCTGCAGGAGTC**S**G [66] sollte die Sequenz CAGGTGCAGCTGCAGGAGT**S**GG lauten. Außerdem wird durch den Primer VL15910 eine Verdopplung der Gene der Familien VL1, VL5 und VL9 verursacht, da der Primer VL1459 diese Gene genauso gut amplifiziert. Die beiden Gene, die VL15910 besser als VL1459 amplifiziert, sind Pseudogene. Bei den zwei V Genen der VL10 Familie verursacht der Primer VL1 nur ein zusätzliches *Mismatch* (3 statt 2).

Die detaillierten Ergebnisse der untersuchten Primer *Coverage* (prozentualer Anteil von den amplifizierten Genen sowie Anzahl der *Mismatches*) sind in Tabelle 17 für Genfamilien und in Tabelle 18 für Gene ohne zugewiesene Genfamilie („*not assigned*“) dargestellt und in Abbildung A6 im Anhang veranschaulicht.

**Tabelle 17:** Analyse der *Coverage* der VBASE2-Genfamilien mit dem am MPIMG eingesetzten Primer Set. Dargestellt sind die Anzahl der verursachten *Mismatches* sowie deren Anteil von den amplifizierten Genen.

| Genfamilie | Anzahl amplifizierter (nicht amplifizierter) Gene | Primer | Degene-ration | *Mismatches* | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | **0** (0 [%]) | **1** (1 [%]) | **2** (2 [%]) | **3** (3 [%]) | **4** (4 [%]) |
| *Variable Heavy Chain* | | | | | | | | |
| VH1 | 43 (1*) | VH1 | 4 | **4** (9,3) | **1** (2,3) | **5** (11,6) | | |
| | | VH157 | 4 | **16** (37,2) | **3** (7,0) | **4** (9,3) | | |
| | | VH1 & VH157 | | | **7** (16,3) | **1** (2,3) | **2** (4,7) | |
| VH2 | 10 | VH2 | 2 | **7** (70) | **3** (30) | | | |
| VH3 | 130 (5*) | VH3 | 8 | **81** (62,3) | **27** (20,8) | **15** (11,5) | **6** (4,6) | **1** (0,8) |
| VH4 | 31 | VH4 | 2 | **21** (67,7) | **6** (19,4) | | | **1** (3,2) |
| | | VH4(DP63) | 1 | **2** (6,5) | **1** (3,2) | | | |
| VH5 | 4 | VH157 | 4 | | **2** (50) | | **2** (50) | |
| VH6 | 2 | VH6 | 1 | **2** (100) | | | | |
| VH7 | 6 | VH157 | 4 | **5** (83,3) | | | | **1** (16,7) |
| *Variable Light Kappa Chain* | | | | | | | | |
| VK1 | 61 (6*) | VK1 | 6 | **45** (73,8) | **12** (19,7) | **2** (3,3) | | **2** (3,3) |
| VK2 | 29 (10*) | VK246 | 12 | **15** (51,7) | **7** (24,1) | **2** (6,9) | **4** (13,8) | **1** (3,4) |
| VK3 | 21 (4*) | VK3 | 8 | **17** (81,0) | **2** (9,5) | **1** (4,8) | | **1** (4,8) |
| VK4 | 1 | VK246 | 12 | | | **1** (100) | | |
| VK5 | 1 | VK5 | 1 | **1** (100) | | | | |
| VK6 | 4 | VK246 | 12 | | **3** (75) | **1** (25) | | |
| *Variable Light Lambda Chain* | | | | | | | | |
| VL1 | 13 (1*) | VL1 | 6 | **7** (53,8) | | | | |
| | | VL1459 | 4 | | | | | |
| | | VL15910 | 8 | | | | | |
| | | VL1459 & VL15910 | | | **6** (46,2) | | | |
| VL2 | 18 | VL2 | 8 | **16** (88,9) | **2** (11,1) | | | |
| VL3 | 22 (1*) | VL3 | 4 | | **6** (27,3) | | | |
| | | VL3(DPL16) | 4 | **2** (9,1) | | | | **1** (4,6) |
| | | VL3(38) | 8 | **8** (36,4) | **1** (4,5) | | | |
| | | VL3 & VL3(38) | | | **2** (9,1) | **1** (4,6) | | **1** (4,6) |
| VL4 | 4 | VL1459 | 4 | **1** (25) | **2** (50) | | **1** (25) | |
| VL5 | 11 | VL1459 | 4 | **5** (45,5) | **4** (36,4) | **1** (9,1) | | |
| | | VL15910 | 8 | | | | | |
| | | VL1459 & VL15910 | | **1** (9,1) | | | | |
| VL6 | 6 | VL6 | 1 | **5** (83,3) | **1** (16,7) | | | |
| VL7 | 5 | VL78 | 6 | **4** (80) | **1** (20) | | | |
| VL8 | 5 | VL78 | 6 | **5** (100) | | | | |
| VL9 | 2 | VL1459 | 4 | | | | | |
| | | VL15910 | 8 | | | | | |
| | | VL1459 & VL15910 | | **2** (100) | | | | |
| VL10 | 2 | VL15910 | 8 | | **2** (100) | | | |

* Gene, die von den Primern nicht amplifiziert werden:
*VH1:* humIGHV140;
*VH3:* humIGHV120, humIGHV144, humIGHV147, humIGHV218, humIGHV318;
*VK1:* humIGKV057, humIGKV058, humIGKV064, humIGKV111, humIGKV118, humIGKV200;
*VK2:* humIGKV038, humIGKV039, humIGKV042, humIGKV076, humIGKV100, humIGKV132, humIGKV138, humIGKV178, humIGKV195, humIGKV197;
*VK3:* humIGKV119, humIGKV122, humIGKV176, humIGKV194;
*VL1:* humIGLV163;
*VL3:* humIGLV171

**Tabelle 18:** Analyse der *Coverage* der VBASE2 Gene, die keiner Genfamilie zugeordnet sind, mit dem am MPIMG eingesetzten Primer Set. Dargestellt sind die Anzahl der verursachten *Mismatches* sowie deren Anteil von den amplifizierten Genen.

| **Anzahl amplifizierter (nicht amplifizierter) Gene** | **Primer** | **Degene-ration** | ***Mismatches*** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **0** (0 [%]) | **1** (1 [%]) | **2** (2 [%]) | **3** (3 [%]) | **4** (4 [%]) |
| *Variable Heavy Chain* | | | | | | | |
| 21 (19*) | VH1 | 4 | | | | **4** (19,0) | |
| | VH157 | 4 | **1** (4,8) | | **2** (9,5) | | |
| | VH3 | 8 | **6** (28,6) | **1** (4,8) | | | |
| | VH4 | 2 | | | | **1** (4,8) | |
| | VH1 & VH157 | | | **1** (4,8) | | | |
| | VH1, VH3 & VH157 | | | | | **5** (4,8) | |
| *Variable Light Kappa Chain* | | | | | | | |
| 25 (6*) | VK1 | 6 | **5** (20) | **4** (16) | **2** (8) | | **3** (12) |
| | VK246 | 12 | | **1** (4) | **2** (8) | | **3** (12) |
| | VK3 | 8 | | | **4** (16) | | **1** (4) |
| *Variable Light Lambda Chain* | | | | | | | |
| 36 (10*) | VL1 | 6 | **1** (2,8) | | | | |
| | VL1459 | 4 | **5** (13,9) | **3** (8,3) | **3** (8,3) | | **1** (2,8) |
| | VL15910 | 8 | | | | | **2** (5,6) |
| | VL1, VL1459 & VL15910 | | | | | **1** (2,8) | |
| | VL2 | 8 | **2** (5,6) | | | | |
| | VL3 | 4 | | **1** (2,8) | | | |
| | VL3(DPL16) | 4 | | | **1** (2,8) | **1** (2,8) | **2** (5,6) |
| | VL3(38) | 8 | | **3** (8,3) | | | |
| | VL3 & VL3(38) | | **1** (2,8) | **6** (16,7) | | | |
| | VL78 | 6 | | | | | **3** (8,3) |

* Gene, die nicht von den Primern amplifiziert werden:
*VH:* humIGHV075, humIGHV091, humIGHV113, humIGHV156, humIGHV160, humIGHV166, humIGHV177, humIGHV180, humIGHV191, humIGHV205, humIGHV224, humIGHV235, humIGHV243, humIGHV254, humIGHV265, humIGHV276, humIGHV283, humIGHV302, humIGHV310
*VK:* humIGKV020, humIGKV112, humIGKV114, humIGKV117, humIGKV142, humIGKV153
*VL:* humIGLV030, humIGLV084, humIGLV092, humIGLV114, humIGLV123, humIGLV126, humIGLV132, humIGLV133, humIGLV138, humIGLV158

Mit dem Primer Set können 431 von den 459 V Genen in der VBASE2, die einer Genfamilie zugeordnet sind, amplifiziert werden. Von den 117 Genen, die zu keiner Genfamilie gehören, werden 82 abgedeckt. Insgesamt ergibt sich eine *Coverage* von 89,1 % der 576 V Gene. Da 63 Gene nicht amplifiziert werden, wurde die Funktionalität der V Gene über ihre Klassifikation in der VBASE2 [8] geprüft. Unter Vernachlässigung der 208 Pseudogene sowie der 47 Orphans ergibt sich, dass 321 V Gene funktionell (Klasse 1) oder möglicherweise funktionell (Klasse 2 und 3) sind. Nur 19 davon (5,92 %) werden mit dem am MPIMG eingesetzten Primer Set aufgrund von Deletionen, Insertionen oder zu vielen *Mismatches* nicht amplifiziert.

Zwecks einer besseren *Coverage* wurden neue Primer (Tabelle 19) nach den folgenden Kriterien entworfen: Amplifizierung des einzigen Klasse 1 V Gens (humIGHV318) oder Amplifizierung von mindestens drei Klasse 2 V Genen. Mit den neuen Primern steigt die *Coverage* der funktionellen Gene (VBASE2 Klasse 1) auf 100 % (auf 97,8 % der funktionellen oder möglicherweise funktionellen V Gene) (Abbildung 25). Bei Betrachtung von allen 576 VBASE2 V Genen ergibt sich eine Erhöhung der Amplifizierung von 89,1 % auf 91,3 %. Eine Zusammenfassung der Ergebnisse ist Tabelle A3 im Anhang zu entnehmen.

**Tabelle 19:** Neue Primer zur Erweiterung des am MPIMG eingesetzten Primer Sets.

| Gen-familie | Primer Namen | Primer Sequenz in 5'-3' Richtung | Degene-ration | Zu amplifizierende Gene | |
|---|---|---|---|---|---|
| | | | | Anzahl | Namen |
| VH3 | VH3N | TCAACACAACGGTTCCCAGTTA | 1 | 1 | humIGHV318 |
| VK2 | VK2N1 | AGATGCTGTGTGAMCCAGCCTC | 2 | 4 | humIGKV038, humIGKV076, humIGKV138, humIGKV195 |
| | VK2N2 | TCCCTCCAAGTTCACATCCTGAG | 1 | 5 | humIGKV039, humIGKV042, humIGKV100, humIGKV132, humIGKV178 |
| VL „*not assigned*“ | VLNA | GTCCAGTTCCTCTATTATGRTAG | 2 | 3 | humIGLV084, humIGLV092, humIGLV138* |

* Pseudogen

Durch das ursprüngliche Primer Set werden alle Gene der zu untersuchenden Genfamilie VH4 amplifiziert. Mit dem Primer VH4(DP63) werden drei Allele des gleichen Gens amplifiziert: humIGHV201, humIGHV033 sowie humIGHV051. Die beiden ersten sind funktionelle Gene der Klasse 1, während das letzte Klasse 2 Gen keine bewiesene Funktionalität besitzt.

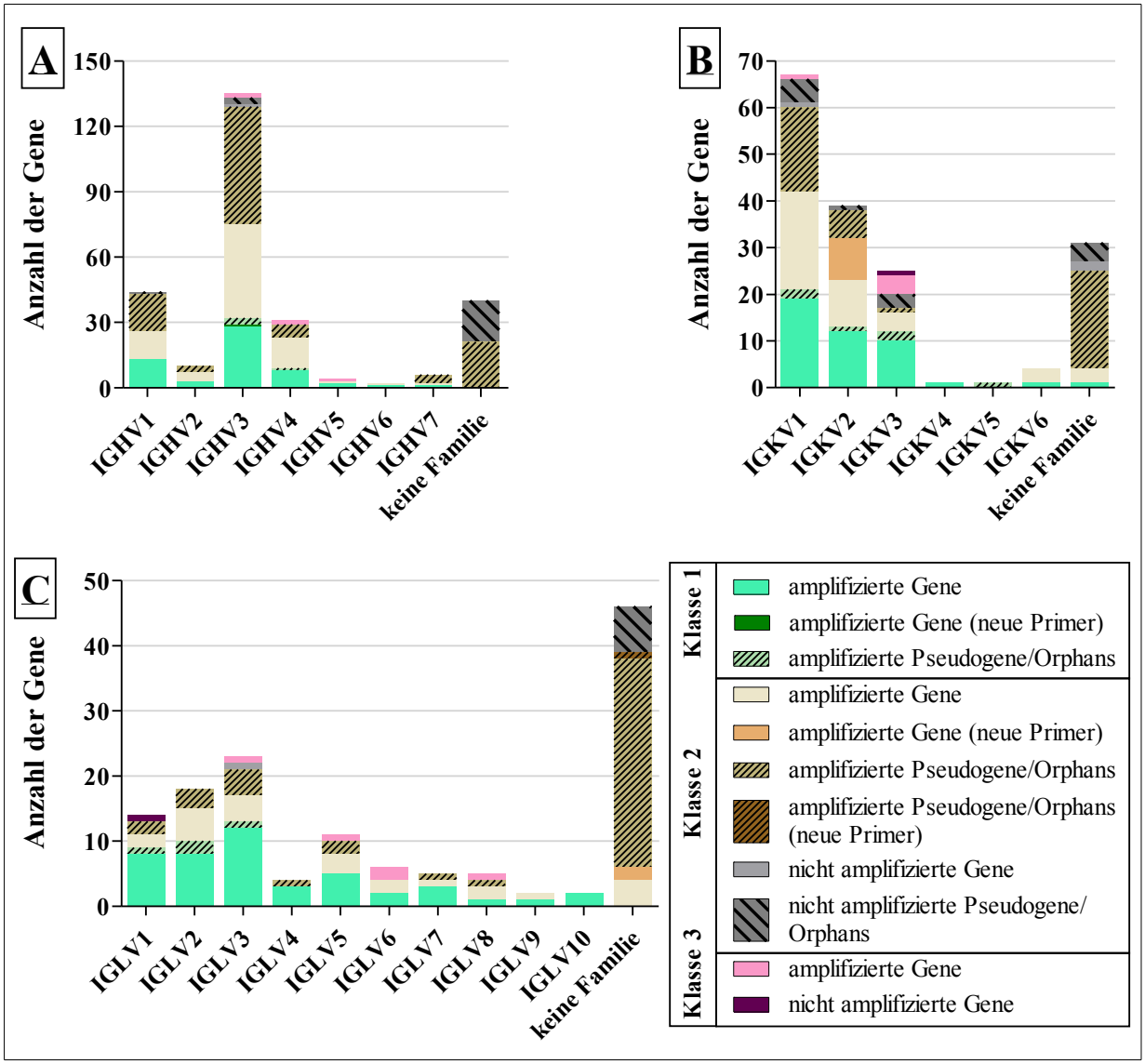



**Abbildung 25:** Amplifizierung der VBASE2 IGHV (A), IGKV (B) sowie IGLV Gene (C) durch die am MPIMG eingesetzten Primer, aufgeteilt nach Klassen.

# 4.3. Auswertung der Next Generation Sequenzierdaten

## 4.3.1. VBASE2 „*Statistic Analysis*“ und *next*IGbase

Um die Auswertung der Next Generation Sequenzierdaten vorzubereiten, wurde eine neue auf „*DNAPLOT Query*“ basierte VBASE2 Anwendung programmiert, die den Vergleich der Gen-Nutzung zwischen zwei Sequenzen-Sets ermöglicht. Das HTML Eingabeformular sowie die Ergebnisse einer Beispielanalyse sind in Abbildung 26 dargestellt. Ausgegeben werden die Anzahl sowie der Anteil an Sequenzen mit den identifizierten Genen bei beiden untersuchten Datensets.

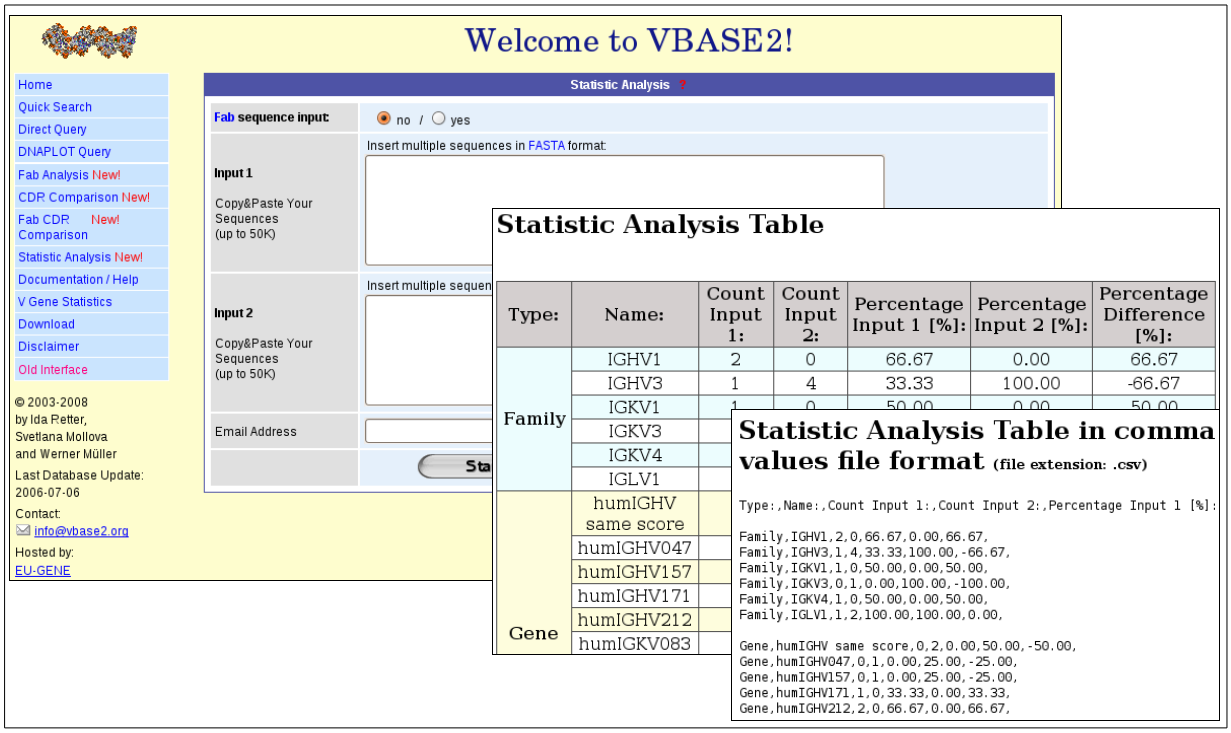



**Abbildung 26:** HTML Eingabeformular und Ergebnisse einer Beispielanalyse mittels der neuen VBASE2 „*Statistic Analysis*“ Anwendung.

Um die umfangreichen Datenmengen aus den Ergebnissen der „*DNAPLOT Query*“ zu bearbeiten sowie die Auswertung zu vereinfachen, wurde die relationale PostgreSQL-Datenbank *next*IGbase angelegt sowie ein Web-Interface programmiert (Abbildung 27). Für die Auswertung der Daten wurden Perl-Skripte eingesetzt.

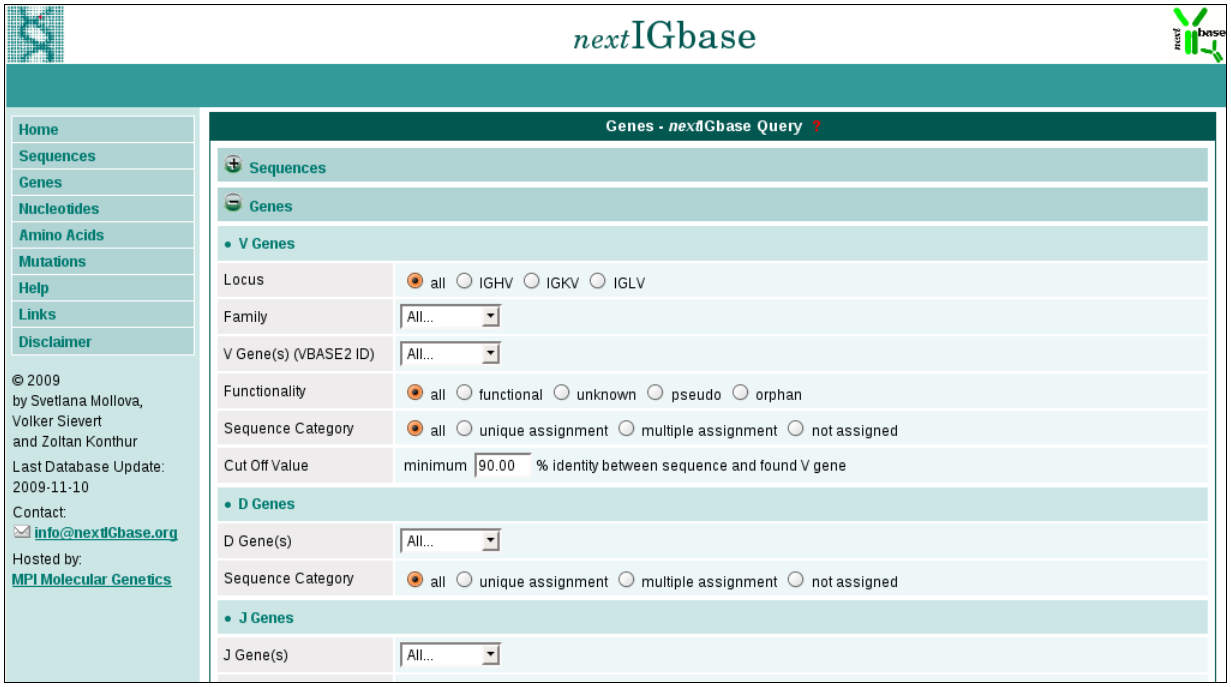



**Abbildung 27:** Webinterface der neuen Datenbank *next*IGbase am Beispiel der HTML Seite „*Genes*“.

## 4.3.2. Vergleich der Antikörpersequenzen zwischen Gesunden und Autoimmunpatienten

Im Rahmen dieser Arbeit werden die Kombination von V, D und J Gensegmenten, das Hinzufügen von nicht-kodierten (N) Nukleotiden zwischen den Gensegmenten und die Mutationen in den Genen aufgrund der somatischen Hypermutation untersucht. Dabei stehen eventuell vorhandene Unterschiede zwischen RA-Patienten und gesunden Probanden bezüglich der Genfamilie VH4 und speziell bei dem V Gen VH4(DP63) sowie Differenzen zwischen den Antikörper-Isotypen IgG und IgD im Vordergrund. Davor muss jedoch sichergestellt werden, dass eventuelle Probleme aufgrund der Sequenzierungsmethode wie z.B. kürzere Read-Längen sowie fehlerhafte Ermittlung der Homopolymer-Nukleotidanzahl im Voraus behoben werden und somit keine Schwierigkeiten bei der Genidentifizierung verursachen. Außerdem muss ausgeschlossen werden, dass bei der Auswertung Sequenzen aufgrund einer überproportionalen Überamplifizierung während der Sequenzierungs-vorbereitung mehrmals berücksichtigt werden.

**Untersuchungen über die eventuellen Probleme aufgrund der Sequenzierungsmethode**

Mittels Next Generation Sequenzierung wurden insgesamt 69847 Sequenzen ermittelt. 25601 davon sind länger als die minimal erforderliche Read-Länge (400 bp, s. Kapitel 4.2.1.) und können den MIDs sowie den spezifischen Primersequenzen zugeordnet werden. Sie wurden für Voranalysen eingesetzt, um die nicht korrekt alignierten Sequenzen zu ermitteln. Die meisten davon enthalten *Frameshifts* aufgrund der falsch ermittelten Nukleotidanzahl in Homopolymersequenzen. Dabei kann auch die Detektion der Länge von kurzen Homopolymeren (weniger als 4 bp lang) fehlerhaft sein (Abbildung 28). Teilweise sind mehrere *Frameshifts* in einer Sequenz vorhanden, wobei das ursprüngliche Leseraster wiederhergestellt werden kann (Abbildung 28).

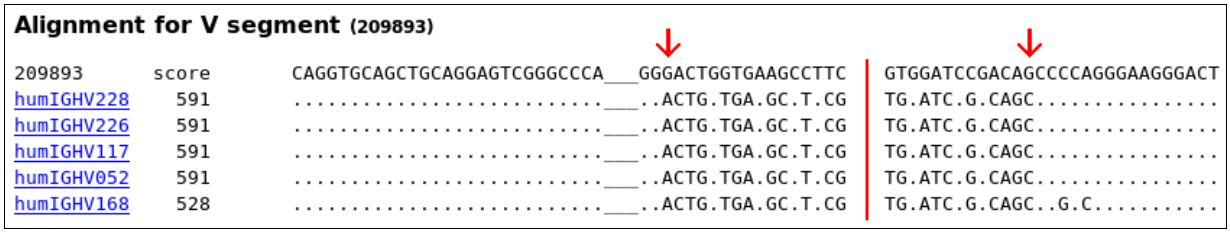



**Abbildung 28:** Interner *Frameshift*-Bereich am Beispiel der Sequenz 209893 (Screenshot des mit „*DNAPLOT Query*“ erstellten Alignments des V Gens). Nur das Alignment in der Nähe der *Frameshifts* ist dargestellt, die rote Linie kennzeichnet fehlende Nukleotide. Rote Pfeile weisen auf die Leserasterverschiebungen hin. Beim ersten *Frameshift* wurden vom *Genome Sequencer FLX Instrument* drei statt zwei G, beim zweiten *Frameshift* wurden vier statt fünf C ermittelt.

Jedoch ist aufgrund der großen Anzahl an Sequenzen eine manuelle Überprüfung der Sequenzalignments wegen des zu hohen Zeitaufwands auszuschließen. Daher wurde eine Auftrennung der Sequenzen nach CDR3 beabsichtigt, um die von Homopolymeren ausgelösten Alignment-Probleme zu lösen. Es wurde angenommen, dass bei identischen CDR3 Regionen auf Nukleotidebene auch der Rest der Sequenzen mit sehr großer Wahrscheinlichkeit identisch ist. Diese Annahme erwies sich allerdings als falsch. Zum Beispiel besitzen die acht Sequenzen 182460, 183792, 188125, 189969, 191272, 192607, 197521 und 217359 identische CDR3, dennoch unterscheiden sich die Nukleotidsequenzen der restlichen Regionen durch Anzahl und Position der Punktmutationen im V bzw. im J Gen: es sind vier unterschiedliche Nukleotidsequenzen und nicht wie vermutet nur eine (Abbildung 29, Abbildung A7 im Anhang). Schließlich wurden 15546 Sequenzen (inklusive *Out-of-Frame-Rearrangements* und Sequenzen mit Stopcodons oder Pseudogene) von der weiteren Auswertung ausgeschlossen und nur die 10055 korrekt alignierten Sequenzen weiter analysiert.

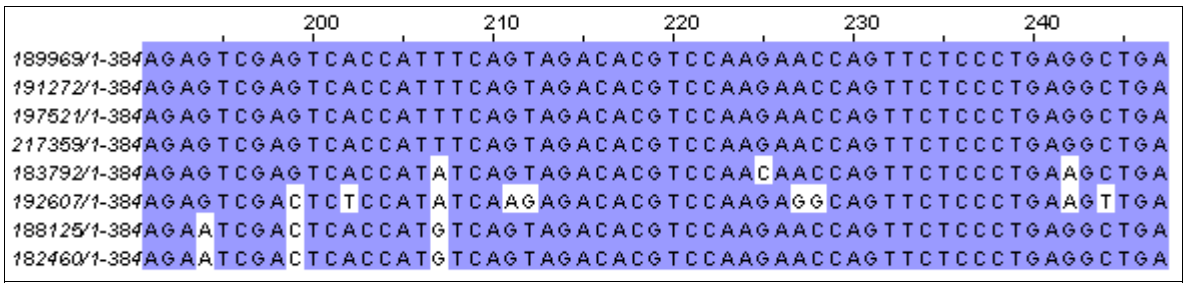



**Abbildung 29.** Screenshot eines Teils des mit ClustalW2 erstellten Alignments von acht VDJ Sequenzen mit identischer CDR3. Das Farbschema BLOSUM62 Score wurde gewählt. Die oberen vier (189969, 191272, 197521 und 217359) sowie die unteren zwei (182460 und 188125) Nukleotidsequenzen sind jeweils identisch. Das gesamte Alignment kann Abbildung A7 im Anhang entnommen werden.

Weiterhin wurde der Anteil der korrekt alignierten Zielsequenzen ermittelt (Tabelle 20), um sicherzustellen, dass die Fw-Primer ausschließlich mit ihren jeweiligen Zielsequenzen hybridisieren. Der VH4(DP63)-Fw Primer amplifiziert zu 97,5 % VH4(DP63) Zielsequenzen, wobei die restlichen der VH4 Genfamilie zuzuordnen sind. Der Anteil der VH4 Zielsequenzen des VH4-Fw Primers ist niedriger (92,1 %), weil Sequenzen aus der Genfamilie VH3 amplifiziert wurden. Somit binden die Primer hauptsächlich, jedoch nicht ausschließlich, an ihre Zielsequenzen. In den darauffolgenden Analysen wurden nur die eigentlichen Zielsequenzen berücksichtigt.

**Tabelle 20:** Gesamtanzahl der untersuchten korrekt alignierten Sequenzen sowie Aufteilung der Sequenzen nach den untersuchten V-Gen-Gruppen.

| Gesundheits-zustand | Ig Isotyp | Primer | Sequenzen [Ahzahl] | Sequenzen [%] | | |
|---|---|---|---|---|---|---|
| | | | | VH4(DP63) | VH4 | VH3 |
| ND | IgD | VH4(DP63)-Fw | 779 | 98,5 | 1,5 | 0 |
| | IgG | | 1452 | 94,4 | 5,6 | 0 |
| RA | IgD | | 1114 | 98,7 | 1,3 | 0 |
| | IgG | | 2136 | 98,6 | 1,4 | 0 |
| ND | IgD | VH4-Fw | 916 | 0,1 | 79,6 | 20,3 |
| | IgG | | 1235 | 0,1 | 94,7 | 5,3 |
| RA | IgD | | 898 | 0 | 99,7 | 0,3 |
| | IgG | | 1525 | 0,4 | 93,0 | 6,6 |

Als Nächstes wurde überprüft, ob die ermittelten Werte für den programmiertechnischen Parameter *Cut Off* (97,5 % bzw. 97 % für V bzw. J Gensegmente, s. Kapitel 4.2.3.) bei der Analyse von Next Generation Sequenzierdaten eingesetzt werden können. Demnach wurde die Anzahl an Zielsequenzen ermittelt, wobei nur komplette *Rearrangements* (Zielsequenzen mit identifizierten V und J Genen) mit korrekt aligniertem V-Gen-Anfang analysiert wurden (Tabelle 21). Die ermittelten *Cut Off* Werte erwiesen sich als zu restriktiv: nur 1212 Sequenzen entsprechen den gewählten Kriterien, davon sind dem Isotyp IgG nur 99 (8,2 %) zugeordnet.

Daher wurde ein *Cut Off* Wert von 90 % für beide Gensegmente (V und J) gewählt (Tabelle 21). Somit entsprechen 5127 Sequenzen den gewählten Kriterien.

**Tabelle 21:** Anzahl der Zielsequenzen in der erstellten Datenbank bei unterschiedlichen *Cut Off* Werten. Untersucht wurden nur komplette *Rearrangements* mit korrekt aligniertem V-Gen-Anfang.

| Gesundheits-zustand | Ig Isotyp | Primer | Gesamtanzahl an Zielsequenzen bei | |
|---|---|---|---|---|
| | | | 97,5 % (V) und 97 % (J) Cut Off | 90 % (V und J) Cut Off |
| ND | IgD | VH4(DP63)-Fw | 245 | 291 |
| | | VH4-Fw | 273 | 491 |
| | IgG | VH4(DP63)-Fw | 32 | 681 |
| | | VH4-Fw | 20 | 635 |
| RA | IgD | VH4(DP63)-Fw | 346 | 411 |
| | | VH4-Fw | 249 | 694 |
| | IgG | VH4(DP63)-Fw | 32 | 1232 |
| | | VH4-Fw | 15 | 692 |

Weiterhin wurde überprüft, ob die ursprüngliche Verteilung der *Rearrangements* erhalten geblieben ist und keine Gene bevorzugt amplifiziert worden sind. Dafür wurde die Anzahl der Zielsequenzen, die identisch sind und mehrmals in der zu untersuchenden Gruppe vorkommen (*Uniques*), in der erstellten Datenbank ermittelt (Tabelle A4 im Anhang). Als Vergleichsbasis diente entweder die vom *Genome Sequencer FLX Instrument* ermittelte Gesamtsequenz oder die von VBASE2 *„DNAPLOT Query“* ermittelte V(D)J Sequenz. Der Anteil an *Uniques* von den 9555 Zielsequenzen beträgt 67,6 % bezogen auf die Gesamtsequenz bzw. 63,5 % bezogen auf die V(D)J Sequenz. Für die weiteren *Uniques*-Analysen wurde aufgrund des niedrigeren Prozentsatzes die V(D)J Sequenz als maßgebend eingesetzt.

Da jedoch für die weiteren Analysen ein *Cut Off* Wert von 90 % geplant ist, wurde mit diesem *Cut Off* Wert die Verteilung der gleichen Nukleotidsequenzen pro *Unique* ermittelt (Abbildung 30). Von den insgesamt 2852 *Uniques* bei 90 % *Cut Off* sind 68,3 % nur ein einziges Mal bzw. 92,7 % maximal dreimal vorhanden. Nur bei IgG VH4(DP63)-Fw sind *Uniques* enthalten, die mehr als 20-Mal gefunden worden sind: bei ND kommen zwei *Uniques* 67- bzw. 29-Mal vor sowie bei RA zwei weitere *Uniques* 213- bzw. 57-Mal. Dabei besitzen diese *Uniques* eine Identität mit den V Keimbahngenen von 91,1 bis 96,6 % und wären bei dem ursprünglich ermittelten *Cut Off* Wert von 97,5 % bei den weiteren Analysen nicht berücksichtigt worden. Dies hätte die *in vivo* vorhandene Verteilung der Gene verfälscht.

Somit beruhen die mehrmals vorkommenden Sequenzen nicht auf Überamplifizierung und können für die weitere Auswertung eingesetzt werden.

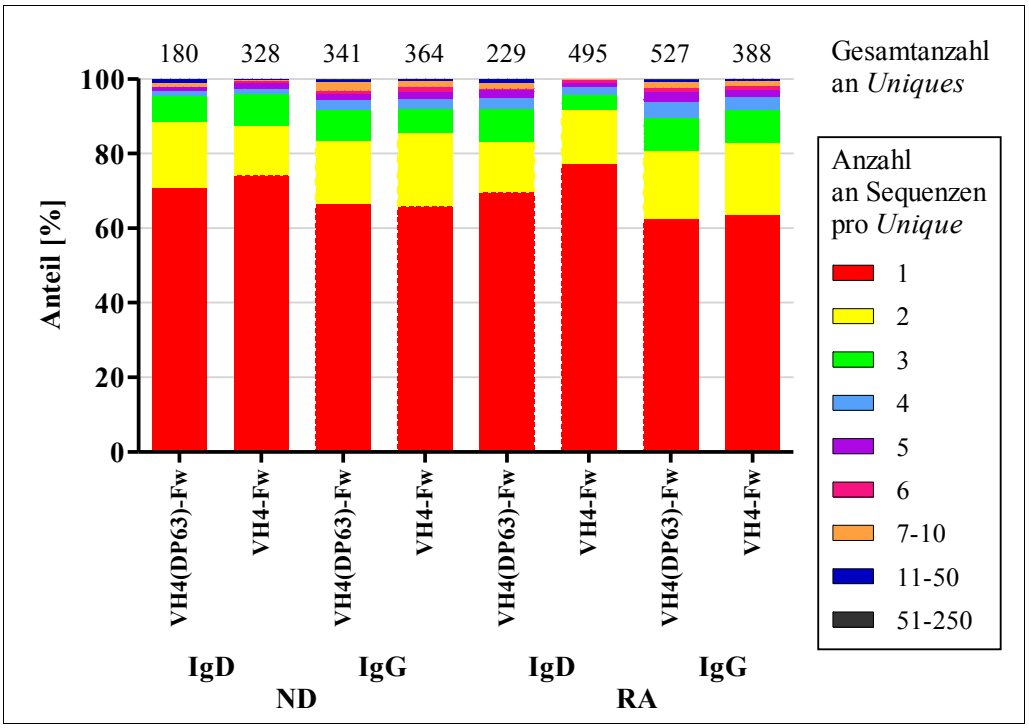



**Abbildung 30:** Verteilung der Sequenzen nach *Uniques* bezogen auf die V(D)J Sequenz. Untersucht wurden nur komplette *Rearrangements* mit korrekt aligniertem V-Gen-Anfang bei einem *Cut Off* Wert von 90 %.

Als Nächstes wurde die Genauigkeit der V-Gen-Identifizierung bei gekürzten Sequenzen überprüft. Dafür wurden nur komplette *Rearrangements* mit korrekt aligniertem V-Gen-Anfang bei einem *Cut Off* Wert von 90 % untersucht. Ermittelt wurden die Anzahl der Sequenzen mit nicht vollständig abgedeckten V Genen sowie die Anzahl an Sequenzen, bei denen mehrere V Keimbahngene den gleichen Score aufweisen (Tabelle 22). Insgesamt sind 390 Sequenzen mit nicht vollständig abgedecktem V-Gen-Anfang vorhanden. Eine multiple V-Gen-Zuordnung wurde bei 34 (8,7 %) festgestellt. Andererseits entsprechen diese 34 Sequenzen nur 25,6 % der insgesamt 133 Sequenzen mit multipler V-Gen-Zuordnung, somit sind Mutationen im V Gen die Hauptursache für Ungenauigkeit der V-Gen-Identifizierung.

**Tabelle 22:** Anzahl der Sequenzen mit nicht vollständig abgedeckten V Genen sowie Anzahl an Sequenzen mit multipler V-Gen-Zuordnung. Untersucht wurden komplette *Rearrangements* mit korrekt aligniertem V-Gen-Anfang bei einem *Cut Off* Wert von 90 %. In Klammern ist die Anzahl der *Uniques* dargestellt.

| Gesundheits-zustand | Ig Isotyp | Primer | Sequenzen mit nicht vollständig abgedeckten V Genen | Multiple V-Gen-Zuordnung, V Gen vollständig abgedeckt: | |
|---|---|---|---|---|---|
| | | | | nein | ja |
| ND | IgD | VH4(DP63)-Fw | 18 | 0 (0) | 0 (0) |
| | | VH4-Fw | 52 | 5 (5) | 6 (5) |
| | IgG | VH4(DP63)-Fw | 36 | 0 (0) | 0 (0) |
| | | VH4-Fw | 73 | 8 (5) | 35 (17) |
| RA | IgD | VH4(DP63)-Fw | 26 | 0 (0) | 0 (0) |
| | | VH4-Fw | 86 | 9 (8) | 23 (15) |
| | IgG | VH4(DP63)-Fw | 43 | 0 (0) | 0 (0) |
| | | VH4-Fw | 56 | 12 (9) | 35 (16) |

**Auswertung der Next Generation Sequenzierdaten**

Die Analyse der Next Generation Sequenzen dient der Überprüfung, ob Unterschiede bei der Antikörperentstehung sowie -reifung zwischen Gesunden und Autoimmunpatienten existieren. Dabei wurde die Gennutzung bei der VH4 Familie und speziell bei dem Gen VH4(DP63) untersucht, wobei auch Differenzen zwischen den Antikörper-Isotypen IgG und IgD erforscht wurden. Für die Datenauswertung wurden komplette *Rearrangements* mit korrekt aligniertem V-Gen-Anfang bei einem *Cut Off* Wert von 90 % eingesetzt.

Zunächst wurden die Anzahl an vorhandenen VDJ Kombinationen sowie die Verteilung der Sequenzen nach VDJ Kombinationen ermittelt (Abbildung 31), wobei Sequenzen mit multipler V, D oder J Genzuordnung nicht berücksichtigt wurden. Die Namen der benutzten V, D und J Gene sind Tabelle 23 zu entnehmen. In der VH4(DP63)-Gruppe wurde ausschließlich das Gen humIGHV201 identifiziert, während insgesamt 14 Gene der VH4 Familie (zusätzlich zum humIGHV201) identifiziert wurden. Acht davon wurden bisher nur auf genomischer Ebene gefunden (VBASE2 Klasse 2 V Gene), ihre Funktionalität wurde jedoch mittels der durchgeführten Next Generation Sequenzierung bewiesen (Tabelle 24). Die meisten der D Gene wurden sowohl durch Deletion als auch durch Inversion rekombiniert.
Bei den untersuchten Sequenzen wurden insgesamt 568 VDJ Kombinationen gezählt, davon 103 mit dem Gen VH4(DP63). Einige Kombinationen sind bei beiden Ig Isotypen sowohl bei Gesunden als auch bei Patienten im vergleichbaren Maß zu finden. Jedoch werden bestimmte D und J Gensegmente bevorzugt mit VH4(DP63) bei dem IgG Isotyp rekombiniert: bei RA-Patienten wurden zwei VDJ Kombinationen bei 84,1 % der 1139 Sequenzen gefunden, während zwei andere VDJ Kombinationen bei 65,5 % der 635 Sequenzen der gesunden Kontrollgruppe identifiziert wurden. Davon sind zwischen 19 % und 33 % auf *Uniques* zurückzuführen, die mehr als 20 Mal gefunden wurden. Somit wurden bedeutende Unterschiede bei IgG VH4(DP63) zwischen Gesunden und RA Autoimmunpatienten festgestellt. Bei den IgG VH4 Genen existiert nur eine einzige VDJ Kombination, die bei mehr als 12,5 % der Sequenzen gefunden wurde, und zwar bei 19,6 % der 501 RA IgG Sequenzen.
Im Gegensatz zu IgG wurden bei IgD weder bei ND noch bei RA bevorzugte VDJ Kombinationen bei den untersuchten V Genen festgestellt.

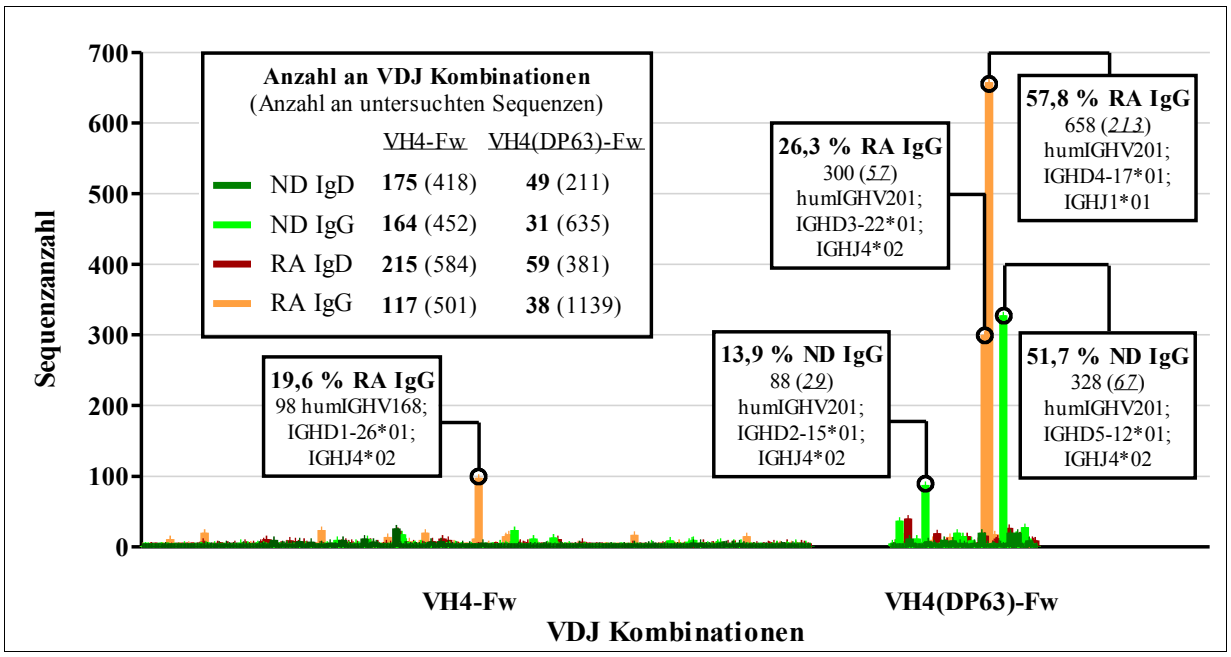



**Abbildung 31:** Anzahl an vorhandenen VDJ Kombinationen sowie deren Häufigkeit. Die Namen der Gene in den Kombinationen sind nicht dargestellt. Explizit beschrieben sind die VDJ Kombinationen, die mehr als 12,5 % der Sequenzen ausmachen, wobei die genaue Anzahl der Sequenzen sowie die Namen der V, D und J Gene genannt werden. In Klammern, kursiv und unterstrichen ist die Anzahl der identischen Sequenzen pro *Unique*, welches mehr als 20 Mal gefunden worden sind.

**Tabelle 23:** Namen der identifizierten V, D und J Gensegmente. Sequenzen mit multipler Genzuordnung wurden nicht berücksichtigt.

| V Gene | D Gene | | J Gene |
|---|---|---|---|
| humIGHV052 | IGHD1-1*01 | IGHD3-9*01 | IGHJ1*01 |
| humIGHV053 | IGHD1-7*01 | IGHD3-10*01 | IGHJ2*01 |
| humIGHV105 | IGHD1-14*01 | IGHD3-10*02 | IGHJ3*01 |
| humIGHV117 | IGHD1-20*01 | IGHD3-16*01 | IGHJ3*02 |
| humIGHV135 | IGHD1-26*01 | IGHD3-16*02 | IGHJ4*01 |
| humIGHV137 | IGHD2-2*01 | IGHD3-22*01 | IGHJ4*02 |
| humIGHV161 | IGHD2-2*02 | IGHD4-17*01 | IGHJ4*03 |
| humIGHV168 | IGHD2-2*03 | IGHD4-23*01 | IGHJ5*01 |
| humIGHV193 | IGHD2-8*01 | IGHD5-12*01 | IGHJ5*02 |
| humIGHV197 | IGHD2-8*02 | IGHD5-24*01 | IGHJ6*02 |
| humIGHV201 | IGHD2-15*01 | IGHD6-6*01 | IGHJ6*03 |
| humIGHV207 | IGHD2-21*01 | IGHD6-13*01 | IGHJ6*04 |
| humIGHV226 | IGHD2-21*02 | IGHD6-19*01 | |
| humIGHV228 | IGHD3-3*01 | IGHD6-25*01 | |
| humIGHV306 | IGHD3-3*02 | IGHD7-27*01 | |

**Tabelle 24:** VH4 Gene mit bisher unbekannter Funktionalität. Sequenzen mit multipler Genzuordnung wurden nicht berücksichtigt.

| V Gene | Anzahl der Sequenzen | |
|---|---|---|
| | insgesamt | Uniques |
| **humIGHV052** | 15 | 13 |
| **humIGHV053** | 78 | 58 |
| **humIGHV105** | 13 | 10 |
| **humIGHV137** | 1 | 1 |
| **humIGHV161** | 7 | 5 |
| **humIGHV193** | 66 | 45 |
| **humIGHV197** | 8 | 7 |
| **humIGHV228** | 176 | 114 |

Als zweiter Schwerpunkt wurde die Länge der CDR3 Region berechnet, da diese Region aus V(D)J Gensegmenten sowie P und N Nukleotiden besteht und somit Erkenntnisse über die V(D)J Rekombination liefert. Zum einen wurde die Verteilungen der Sequenzen nach CDR3 Längen ermittelt. Nur bei IgG VH4(DP63) wurden gravierende Unterschiede zwischen ND

und RA festgestellt (Tabelle 25). Zum anderen wurden die Mittelwerte sowie die Mediane der Verteilungen ermittelt (Abbildung 32).

**Tabelle 25:** Anzahl sowie Anteil an Sequenzen mit unterschiedlicher CDR3 Aminosäurenlänge. Dargestellt sind nur CDR Längen, bei denen gravierende Unterschiede zwischen Gesunden und Autoimmunpatienten bestehen.

| Gesundheits-zustand | Ig Isotyp | Primer | Anzahl sowie prozentualer Anteil an Sequenzen mit einer CDR3 Länge in Aminosäuren von: | | |
|---|---|---|---|---|---|
| | | | 12 | 15 | 20 |
| ND | IgG | VH4(DP63)-Fw | 29 (4,3%) | 41 ( 6,0 %) | 418 (61,4 %) |
| RA | | | 728 (59,1 %) | 382 (31,0 %) | 9 (0,7 %) |

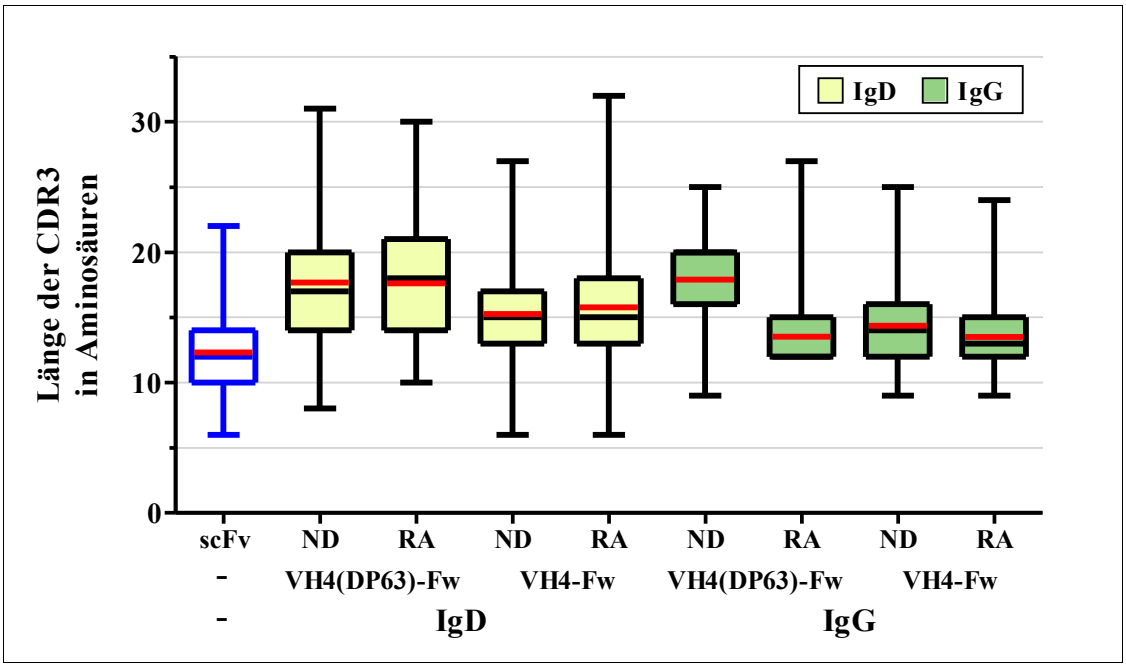



**Abbildung 32:** Länge der CDR3 Region bei den untersuchten Gruppen. Dargestellt sind Minimum, Maximum, Quartile und Mediane. Die roten Linien repräsentieren Mittelwerte.

Es fällt auf, dass bei IgD die Längen-Spannweite um durchschnittlich sieben Aminosäuren größer ist als bei IgG. Außerdem liegen die Maxima der CDR3 Längen aller IgD sowie der RA IgG VH4(DP63) Sequenzen deutlich über 25 Aminosäuren. Ein Beispiel einer solchen Sequenz ist in Abbildung 33 dargestellt.

**Junction** (238352) (V P N P D P N P J)

238352 TGTGCAAGAGGCCGGGGATATAGTGGCTACGATCCAAAAACAAAACAAGAGGACAGAAGTACACCAACAGGGGTGTTTCGATACTACTTTGACTACTGG

**Translation of the Junction** (238352)

238352 CARGRGYSGYDPKTKQEDRSTPTGVFRYYFDYW

**Abbildung 33:** CDR3 Region mit einer Länge von 31 Aminosäuren am Beispiel der korrekt alignierten Sequenz 238352: Screenshot der mit „*DNAPLOT Query*“ erstellten *Junction*-Nukleotidsequenz und ihre Translation. Die *Junction* enthält zusätzlich zu der CDR3 die letzte Aminosäure von FR3 (Cystein) sowie die erste Aminosäure von FR4 (Tryptophan).

Da die P Nukleotidanzahl gering im Vergleich zu N Nukleotidanzahl ist, könnten die auffällig langen CDR3 durch längere D Gensegmente oder durch längere N Nukleotidsequenzen verursacht worden sein. Daher wurde die Länge von diesen Regionen bestimmt. Es wurde festgestellt, dass bei dem IgD Isotyp durchschnittlich 20,6 % der Sequenzen D Segmente mit einer Länge ab 20 bp enthalten, während dieser Prozentsatz bei IgG nur 1,9 % beträgt. Mittels zweiseitigen Chi-Quadrat-Vierfeldertests wurde nachgewiesen, dass die Differenz statistisch signifikant ist ($P < 0{,}0001$).

Bezüglich der Anzahl an N Nukleotiden ist auffällig, dass die meisten (71,1 %) ND IgG VH4(DP63) Sequenzen eine N Nukleotidanzahl über 20 bp aufweisen. Dieser Prozentsatz liegt hingegen bei allen anderen untersuchten Gruppen bei maximal 12,3 % (Tabelle 26). Ein zweiseitiger Chi-Quadrat-Test bestätigte die statistische Signifikanz des Unterschieds ($P < 0{,}0001$).

Dementsprechend wurde manuell überprüft, ob D-D Fusionen [67] die Ursache für die langen N Nukleotidregionen sind. Tatsächlich wurden D-D Fusionen identifiziert (Abbildung 34), wobei sogar Beweise für D-D-Fusion von drei D Gensegmenten gefunden wurden (Abbildung A8 im Anhang).

**Tabelle 26:** Prozentualer Anteil an Sequenzen mit Länge ab 20 bp innerhalb der D Gensegmente bzw. der N Nukleotide. Die N Nukleotidanzahl bezieht sich entweder auf die N Nukleotide zwischen V und D Segment oder auf die N Nukleotide zwischen D und J Segment.

| Gesundheits-zustand | Ig Isotyp | Primer | Sequenzen [%] | |
|---|---|---|---|---|
| | | | D Gensegment ab 20 bp | N Nukleotide ab 20 bp |
| ND | IgD | VH4(DP63)-Fw | **23,4** | 4,1 |
| | | VH4-Fw | **18,1** | 2,4 |
| | IgG | VH4(DP63)-Fw | 0,3 | **71,1** |
| | | VH4-Fw | 4,9 | 3,5 |
| RA | IgD | VH4(DP63)-Fw | **23,6** | 5,6 |
| | | VH4-Fw | **19,3** | 2,3 |
| | IgG | VH4(DP63)-Fw | 1,1 | 1,6 |
| | | VH4-Fw | 2,5 | 12,3 |


**Abbildung 34:** D-D Fusion am Beispiel der Sequenz 228059: Screenshot des mit „*DNAPLOT Query*“ erstellten Alignments des D Segments.

Als Nächstes wurde der Einfluss der somatischen Hypermutation untersucht, indem die Verteilung der Sequenzen nach Anzahl der Mutationen bestimmt wurde (Abbildung 35). Sequenzen von Gesunden enthalten weniger Mutationen als jene von Autoimmunpatienten, genauso ist die Anzahl an Mutationen in VH4(DP63) Sequenzen geringer als die Anzahl an Mutationen in VH4 Sequenzen. Jedoch ist die größte Differenz zwischen IgD und IgG festzustellen: der Prozentsatz an Sequenzen mit maximal 9 Mutationen liegt bei IgD zwischen 50,6 und 93,1 %, verglichen mit 2,5 – 7,9 % bei IgG bei den untersuchten Gruppen (nach Ig Isotyp, Primer und Gesundheitszustand aufgeteilt). Mittels zweiseitigen Chi-Quadrat-Viefeldertests wurde nachgewiesen, dass der Unterschied statistisch signifikant ist ($P < 0{,}0001$). Ein weiterer zweiseitiger Chi-Quadrat-Viefeldertest wurde durchgeführt, um die statistische Signifikanz zwischen IgG und IgD bei den unmutierten Sequenzen zu überprüfen – mit dem gleichen Ergebnis. Somit hat der Ig Isotyp bedeutenden Einfluss auf die Anzahl an Mutationen.

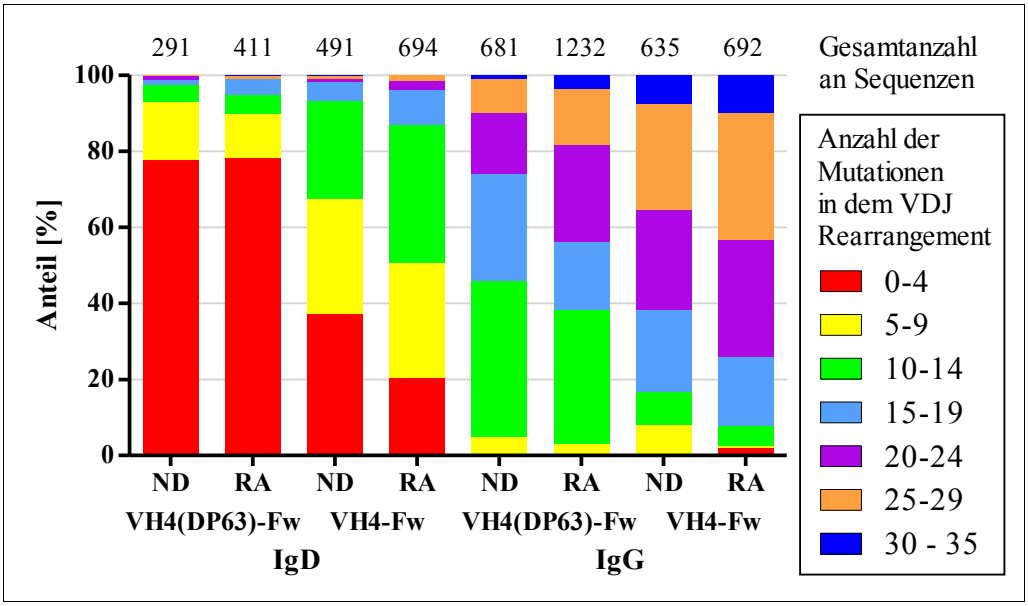



**Abbildung 35:** Verteilung der Sequenzen nach Anzahl der Mutationen im V(D)J *Rearrangement*.

Abschließend wurde untersucht, wie sich die drei Faktoren (VDJ Rekombination, Einfügen von N Nukleotiden, somatische Hypermutation) auf die Antikörperdiversität auswirken. Von den 5127 untersuchten Sequenzen sind anhand ihrer Aminosäuresequenz 2770 *Uniques*. Jedoch sind die drei CDR Regionen für die Antigenbindung zuständig und daher wurde die Anzahl der *Uniques* (auf die CDR Kombinationen bezogen) ermittelt. Insgesamt wurden 2235 Kombinationen (80,7% von 2770) nachgewiesen, wobei keine einzige Kombination bei den

anderen untersuchten Gruppen wiedergefunden wurde (Tabelle 27). Somit sind, bezogen auf die CDR Kombination, 43,6 % der untersuchten 5127 Nukleotidsequenzen *Uniques*. Die Aminosäuresequenzen der CDR Kombinationen, die über 100 Mal vorkommen, sind Tabelle 28 zu entnehmen.

**Tabelle 27:** Gesamtanzahl an *Uniques* (bezogen auf die CDR Kombination).

| Gesundheitszustand | ND | | | | RA | | | |
|---|---|---|---|---|---|---|---|---|
| Ig Isotyp | IgD | | IgG | | IgD | | IgG | |
| Primer | VH4 (DP63)-Fw | VH4-Fw | VH4 (DP63)-Fw | VH4-Fw | VH4 (DP63)-Fw | VH4-Fw | VH4 (DP63)-Fw | VH4-Fw |
| ***Uniques*** | 115 | 288 | 248 | 335 | 155 | 439 | 312 | 343 |

**Tabelle 28:** Aminosäuresequenz der CDR Kombination der *Uniques* mit über 100 Sequenzen pro *Unique*.

| Aminosäuresequenz | ND IgG VH4-(DP63)-Fw | RA IgG VH4-(DP63)-Fw |
|---|---|---|
| NGSFSGYY; ISHTGSA; ARVRITVARITTQRHRYNDY | 104 | |
| GGSFSGYY; LHHRGST; ARGRRTSSYSYYFDS | | 126 |
| GGSFSDYY; INRSGST; ARSGKTTVPPLD | | 316 |

# 5. Diskussion

Die Next Generation Sequenzierung ist eine Hochdurchsatz-Methode zur Entschlüsselung der Nukleotid-Reihenfolge in DNA Molekülen [26, 27]. Diese Technologie wird in dieser Arbeit zum ersten Mal zur Sequenzierung von humanen Antikörpergenen eingesetzt. Aus diesem Grund sind sowohl experimentelle sowie bioinformatische Untersuchungen bezüglich ihrer Verwendung bei Gesunden und Autoimmunpatienten erforderlich.

In dieser Arbeit wurde mittels qRT-PCR überprüft, ob Differenzen in der IgG- und IgD-Expression zwischen Gesunden und Autoimmunpatienten bestehen. Weiterhin wurde die Basis für eine Testsequenzierung von IgG bzw. IgD VDJ Antikörpergenen von Gesunden und RA-Patienten mittels *Genome Sequencer FLX Instrument* geschaffen, wobei nur V Gene der VH4 Familie und speziell das Gen VH4(DP63) untersucht wurden. Zusätzlich wurden bioinformatische Untersuchungen zwecks Anpassung des Analysealgorithmus an Next Generation Sequenzierdaten durchgeführt und eine neue VBASE2-Anwendung sowie eine relationale Datenbank wurden programmiert. Anschließend wurden mittels *Genome Sequencer FLX Instrument* ermittelte Antikörpersequenzen hinsichtlich der Unterschiede zwischen RA-Patienten und gesunden Probanden ausgewertet.

<u>Vergleich von Ig Isotypen zwischen Gesunden und Autoimmunpatienten mittels qRT-PCR</u>

Als erster experimenteller Teil wurden qRT-PCRs durchgeführt, um eine Quantifizierung der Ig DNA-Amplifizierung vorzunehmen. Dabei gilt für alle cDNA-Konzentrationsberechnungen die Annahme, dass bei allen Proben die cDNA mit der gleichen RT-Effizienz aus RNA hergestellt worden ist.
Bei der qRT-PCR wird zunächst ein Primer-Effizienz-Test durchgeführt, um eventuelle Effizienzunterschiede in der Amplifizierung bei der Auswertung zu berücksichtigen. Da sich die Primereffizienzen unterscheiden, wurde die „Efficiency adjusted ΔΔCt“-Methode [54] mit Einsatz des geometrischen $Ct_{adjusted}$ Mittelwertes der Kontrollgene [55] angewendet.
Als Nächstes wurde überprüft, ob die Primer kreuzreagieren. Beide Primer amplifizieren DNA Fragmente mit den erwarteten Größen von Lymphozyten cDNA. Eine Kreuzreaktivität der qRT-PCR-IgG Primer wurde mittels qRT-PCR sowie mittels PCR mit anschließender

Gelelektrophorese ausgeschlossen. Es wurde außerdem bioinformatisch gezeigt, dass die qRT-PCR-IgD Primer nur mit IgD Proben hybridisieren können. Demnach sollten die qRT-PCR-IgD Primer keine Kreuzreaktivität aufweisen. Um dies experimentell zu prüfen, wären jedoch entweder cDNA Fragmente aus der RT für die Kalibrierung oder neue qRT-PCR-IgD Primer notwendig. Eine andere Möglichkeit wäre der Einsatz von genspezifischen Primern für die IgG cDNA Synthese (statt der benutzten Random Primers sowie Oligo$(dT)_{12\text{-}18}$ Primer).

Weiterhin wurde ein Test mit Patientenproben sowie mit gesunden Kontrollen durchgeführt, um das Verhältnis zwischen IgG und IgD in Abhängigkeit des Gesundheitszustandes zu ermitteln. SLE Proben konnten diesbezüglich nicht ausgewertet werden: entweder enthielten sie nach der RT zu wenig cDNA (trotz gleicher RNA Anfangskonzentration und trotz der Anwendung des gleichen RT Protokolls) oder das Kontrollgen HPRT ist für diese Proben nicht geeignet. Außerdem wies ein RA Patient ein IgG:IgD Verhältnis auf, welches ungefähr im Bereich der gesunden Probanden liegt, während die IgG:IgD Werte bei den SD Patientenproben heterogen waren. Eine Erklärung dafür könnte der tatsächliche Gesundheitszustand der gesunden Probanden sein: sie wurden zwar als gesund im Moment der Probeentnahme eingestuft, allerdings wäre es wichtig zu wissen, ob sie sonst an Erkrankungen leiden (wie z.B. Allergien wie Heuschnupfen oder Unverträglichkeiten gegenüber Lebensmittel) oder z.B. kurz vor Probeentnahme erkältet gewesen sind. Außerdem könnte das Stadium der Autoimmunerkrankung der Patienten Einfluss auf das Ig-Level haben. Es ist zudem nicht auszuschließen, dass die Patienten zusätzlich an anderen Erkrankungen gelitten haben. Die berechneten IgG:IgD Verhältnisse sind jedoch überraschend niedrig: die Mittelwerte betragen 0,33 amol bei ND; 1,93 amol bei RA und 0,8 amol bei SD. Erwartet wurden deutlich höhere Verhältnisse, da IgG ca. 75 % der Antikörper im Blutserum entspricht, während IgD weniger als 1 % davon ausmacht [4]. Der Grund für die niedrigen Werte könnte z.B. sein, dass in den Proben mehr naive B-Lymphozyten vorhanden waren als IgG-Gedächtniszellen und IgG-produzierende Plasmazellen. Trotzdem bestätigt dieser Test die Vermutung, dass bei Autoimmunerkrankungen aufgrund der hohen Spezifität von IgG gegen Antigene das IgG Level meistens höher als bei gesunden Personen ist. Mittels ANOVA-Tests mit darauffolgendem Dunnett-Test sowie mittels Kruskal-Wallis Tests wurde ein signifikanter Unterschied zwischen RA und Gesunden nachgewiesen, während dies für SD und ND nicht zutraf. Jedoch wären deutlich mehr als je fünf Proben je Gesundheitszustand erforderlich, um repräsentative Ergebnisse für die Bevölkerung zu erhalten und eindeutige Aussagen diesbezüglich zu treffen.

Daraufhin wurde ein Test mit gesunden Probanden durchgeführt. Er dient der Überprüfung, ob mehrere Patientenproben für die Sequenzierung gemischt werden können. Dafür wurden die $2^{(-\Delta\Delta Ct\ adjusted)}$-Werte des Verhältnisses IgG:IgD für den Pool der cDNA-Proben experimentell ermittelt (IgG:IgD = 0,36) und andererseits theoretisch aus den Werten der fünf Proben berechnet (IgG:IgD = 0,32). Da der Unterschied gering ist, werden alle Proben mit vergleichbarer Effizienz amplifiziert und somit könnte für die Sequenzierung ein Pool eingesetzt werden. Außerdem wurde durch diesen Test der davor ermittelte Bereich der IgG:IgD Verhältnisse bei den Gesunden bestätigt (0,18– 0,53) und somit wurde ein Einfluss der cDNA-Anfangskonzentration auf die Ergebnisse ausgeschlossen.
Außerdem wurde die Ig cDNA Stoffmenge in den Proben anhand einer Kalibrierung bestimmt. Dabei sind die beiden höchsten IgG Werte bei RA-Patienten zu verzeichnen (319,3 bzw. 432,0 amol), während die IgG cDNA Stoffmenge bei allen anderen IgG Proben (unabhängig vom Gesundheitszustand) ca. 4 – 50 amol betrug. Die IgD Werte sind wesentlich breiter gestreut und liegen bei 5 – 229 amol. Dies liegt möglicherweise an der Berechnungsmethode, da für die Kalkulation die ermittelten IgG:IgD Verhältnisse eingesetzt wurden. Signifikante Unterschiede wurden nur zwischen RA und ND mittels eines an Two-way ANOVA anschließenden Bonferroni-Posttests nachgewiesen ($P < 0,05$). Der Grund dafür könnte wie zuvor erklärt im tatsächlichen Gesundheitszustand der Probanden liegen. Zusätzlich konnte anhand der gewonnenen Daten mittels einer OpenOffice.org Calc-Tabelle für die beiden niedrigsten Mediane die maximale Anzahl an PCR Zyklen für jede Phase berechnet werden (jeweils fünf). Falls gewünscht, könnte vor der Sequenzierungsvorbereitung bei allen Proben zunächst eine qRT-PCR durchgeführt werden, um so aus den ermittelten Stoffmengen die exakte Anzahl an PCR Zyklen mit der Calc-Tabelle zu kalkulieren. Hiermit könnte aufgrund der genaueren Daten eine Überamplifizierung vermieden werden.

Sequenzierung

Als Zweites wurden PCR Untersuchungen zur Vorbereitung der Amplicon Sequenzierung durchgeführt. Die Bindestellen der Sequenzierprimer (Primer A und B) wurden erfolgreich mittels Fusion-Primer an die Antikörpergene hinzugefügt. Die Länge der erhaltenen PCR Fragmente entspricht der empfohlenen für die Amplicon Sequenzierung (bis ca. 500 bp). Jedoch wurden PCR Nebenprodukte trotz der Verwendung von gelaufgereinigten PCR Fragmenten als Ausgangsmaterial nachgewiesen. Es konnte auch keine Verbesserung durch die

Zugabe von ET SSB festgestellt werden. Anscheinend ist ET SSB nur bei der genspezifischen Amplifizierung von cDNA von Nutzen [49]. Anschließend wurde untersucht, ob durch Änderung der PCR Bedingungen die Amplifizierung von Nebenprodukten reduziert werden kann. Als Erstes wurde eine TD-PCR mit einer um 10 ºC höheren als der standardmäßigen Anfangstemperatur durchgeführt, jedoch ohne Erfolg. Daraufhin wurde eine Gradient-PCR (Gradient-PCR2) bei der gleichen Temperaturspanne durchgeführt, um die optimale Primer Annealing-Temperatur zu ermitteln. Allerdings wurden keine Unterschiede in der Amplifizierung festgestellt. Durch Gradient-PCR3 wurden auch keine zufriedenstellenden Ergebnisse erzielt. Möglicherweise wäre ein RNA Verdau nach der RT hilfreich, um die vorhandene RNA als Matrize für Kontaminationen abzubauen. Außerdem könnte die hohe Anzahl an PCR Zyklen während der drei PCR Phasen dazu beitragen, dass Nebenprodukte überamplifiziert werden oder eine Verschiebung der Antikörper-Diversität zugunsten der öfters vorhandenen oder im Primerbereich weniger mutierten (bzw. besser hybridisierenden) Antikörpersequenzen stattfindet. Dies lässt sich vermeiden, indem auf die erste PCR Phase (PCR1) verzichtet wird und für das Einführen der MIDs sowie der Bindestellen der Sequenzierprimer nur eine PCR mit folgenden Primern durchgeführt wird: Fw-Fusion-Primer aus Primer A und dem genspezifischen Fw-Primer sowie Rv-Fusion-Primer aus Primer B, dem MID und dem Ig-Isotyp-spezifischen Rv-Primer (Bindesequenz in CH1 nahe an VH) für die zweite PCR Phase. Da die Rv-Primer für PCR1 keine Kreuzreaktivität aufweisen [49], wäre beim Verzicht auf die erste PCR Phase ein Kreuzreaktivitätstest der Ig-Isotyp-spezifischen Rv-Primer notwendig. Zusätzlich könnte im Protokoll die Gelaufreinigung entfallen, da dafür viele DNA Moleküle notwendig sind und somit eine hohe Anzahl an PCR Zyklen erforderlich ist. Da es nicht möglich ist, die maximale Anzahl von Molekülen ($10^{11}$ [57]) nicht zu überschreiten und gleichzeitig sichtbare DNA Banden auf einem Agarosegel für eine Gelaufreinigung zu erhalten, würde sich folgende Methode anbieten: Zunächst wird die für die Einhaltung des Sequenzierkriteriums erforderliche Anzahl an PCR Zyklen mit der OpenOffice.org Calc-Tabelle im Voraus berechnet und die PCR wird durchgeführt. Der Ansatz wird in zwei geteilt, für die erste Hälfte wird die notwendige Anzahl an PCR Zyklen für den Gelelektrophorese-Nachweis kalkuliert. Nach stattgefundener PCR und erfolgreichem Beweisen von DNA Fragmenten mit der zu erwartenden Größe kann ein Teil der zweiten Ansatz-Hälfte als Template für die darauffolgende (eventuell nötige, je nach Protokoll) PCR Phase oder für die Sequenzierung eingesetzt werden. Zwar würden auch Primer übertragen werden, allerdings können Daten nur aus DNA Molekülen mit beiden Sequenzierprimerbindestellen gewonnen werden. Eventuell könnte mit einem Unterschuss an

Primern gearbeitet werden, um diesem Einfluss entgegenzuwirken. Die vorgeschlagenen Änderungen im Protokoll würden die Nebenproduktkonzentration reduzieren und gleichzeitig die Antikörper-Diversität in den Proben erhalten und somit zu Sequenzierdaten von besserer Qualität führen.

Parallel zu der Sequenzierungsvorbereitung wurden bioinformatische Untersuchungen durchgeführt und es wurde festgestellt, dass die minimal erforderliche Read-Länge 400 bp beträgt: sonst könnten die V Gene bei der VDJ Analyse nicht korrekt identifiziert werden. Im Zeitraum der Diplomarbeit konnte jedoch nicht auf das Roche Titanium Kit für die Amplicon Sequenzierung zurückgegriffen werden, welches eine Read-Länge von 400 bp ermöglicht [31]. Da jedoch mit den Standard-Reagenzien nur eine Read-Länge von 200-300 bp erreicht werden kann, können die Antikörpergene mit dem Titanium Kit für die Shotgun Sequenzierung untersucht werden: bei dem Design der Fusion-Primer können die Bindesequenzen der Amplicon Sequenzierprimer (Primer A und B) mit den Bindesequenzen der Primer des lieferbaren Titanium-Kits für die Shotgun-Sequenzierung ersetzt werden. Dabei findet eine quasi 100-% „Ligation“ statt, was gegenüber einer Ligation der amplifizierten Antikörpergene während des üblichen Shotgun-Verfahrens mehrere Vorteile aufweist: Einerseits sind bei der Herstellung der sstDNA Bibliothek keine Verluste von PCR-Fragmenten durch doppelte Ligation der gleichen Sequenzierprimerbindestelle zu erwarten. Andererseits kann nicht sichergestellt werden, ob alle Sequenzen mit MID erfolgreich ligiert werden. Dieses Problem kann jedoch durch Antikörpergenamplifizierung mit Fusion-Primern umgangen werden. Außerdem ist zu beachten, dass die Sequenzierprimerbindestellen zufällig zu den DNA-Fragmenten ligiert werden und somit die Sequenzierrichtung nicht im Voraus vorgegeben werden kann. Als Folge könnte bei kürzeren Reads die IgD- oder IgG-Sequenz am 3'-Ende eventuell nicht vollständig ermittelt werden. Dies könnte nur durch Einsatz mehrerer MIDs vermieden werden: statt zwei MIDs (zur Kennzeichnung des Gesundheitszustandes) wären vier MIDs erforderlich (für Gesundheitszustand sowie Ig Isotyp). Bei den momentan erhältlichen 12 MIDs wäre es demnach möglich, maximal sechs Proben in einem Lauf parallel zu sequenzieren.
Die ausgewerteten Next Generation Daten beruhen auf Ergebnissen einer Shotgun-Sequenzierung nach Austausch der Primerbindestellen. Somit wurde mit den durchgeführten Experimenten die Basis für den erfolgreichen Testlauf mit dem *Genome Sequencer FLX Instrument* geschaffen.

Bioinformatische Untersuchungen

Um eventuelle Probleme bei der Auswertung von Next Generation Sequenzierdaten, wie eine kürzere Readlänge oder möglicherweise fehlerhafte Ermittlung der Homopolymer-Nukleotidanzahl vorherzusagen, sollten diese Parameter in bioinformatischen Vorarbeiten untersucht werden. Diese sind sowohl für die Anwendungsprogrammierung hilfreich sowie liefern wertvolle Informationen, wie z.B. über die Primer *Coverage*. Für die Tests wurden zwei Sets von Daten eingesetzt: die VBASE2 V Gene sowie die scFv-Testsequenzen. Das erste Set setzt sich aus allen VBASE2 Genen zusammen, jedes V Gen ist jedoch nur einmal enthalten. Das zweite Set besteht aus *Rearrangements* von V(D)J Genen, die im Vergleich zu den Keimbahngenen mutiert sein können; jedoch sind in diesem Set nicht alle V Gene vorhanden und teilweise weisen die *Rearrangements* mehrmals das gleiche V Gen auf. Dementsprechend sind die Ergebnisse der beiden Datensets nicht immer miteinander vergleichbar. Außerdem sollte darauf hingewiesen werden, dass nur 8 IGKV scFv-Testsequenzen vorhanden waren. Dies erschwert statistisch relevante Aussagen.

Als Erstes wurde überprüft, wie genau die Identifizierung von V Genen in gekürzten Sequenzen ist, da die Sequenzierung der Antikörpergene in 3'-5' Richtung erfolgt und somit der Anfang der V Gene bei kürzeren Reads eventuell nicht erreicht wird. Für den Test wurden die Sequenzen in beiden Datensets in Schritten von 10 bp bis zum Erreichen einer maximalen Kürzungslänge von 100 bp gekürzt und anschließend mit „*DNAPLOT Query*“ ausgewertet. Dabei werden von den ungekürzten VBASE2 Genen bis zu maximal 10 % (hauptsächlich Pseudogene und Orphans ohne zugeordnete Genfamilie) nicht korrekt identifiziert. Dies liegt an dem Arbeitskonzept des DNAPLOT Programms: die zu identifizierenden Sequenzen werden zunächst gegen die so genannten Master-Sequenzen aligniert und erst dann gegen die V Gene. Die Master-Sequenzen sind so gewählt, dass ein Alignieren der funktionellen Gene gewährleistet wird und somit haben sie Vorrang vor Pseudogenen und Orphans. Die ungekürzten IGHV und IGLV Testsequenzen enthalten ihrerseits Sequenzen (ebenfalls bis 10 %), bei denen mehrere V Gene mit den gleichen Scores gefunden worden sind. Dies liegt daran, dass die scFv Sequenzen mutiert sind, und impliziert, dass alle mit maximalem Score identifizierten Keimbahngene mit der gleichen Wahrscheinlichkeit Teil des *Rearrangements* sein können. Somit kann nicht unterschieden werden, welches der Gene das ursprüngliche war und dementsprechend sind keine weitere Vergleiche durchführbar.

Weiterhin nimmt der Prozentsatz der Sequenzen mit korrekt identifizierten V Genen erwartungsgemäß mit zunehmender Kürzungslänge ab. Es werden Bereiche gekürzt, die *Mismatches* für die Unterscheidung zwischen den Genen beinhalten und so können nur noch 60 % der V Gene bzw. ca. 55 % der scFv-Testsequenzen bei 100 bp Kürzung eindeutig zugeordnet werden. Diese Verringerung ist am höchsten bei IGKV, gefolgt von IGHV und IGLV bei den VBASE2 Genen. Dies lässt sich durch die Verteilung der Gene in Familien erklären. Falls nur Genfamilien mit mehr als einem VBASE2 Klasse 1 Gen berücksichtigt werden, sind es bei IGKV lediglich 3 Familien, verglichen mit 5 IGHV bzw. 8 IGLV Genfamilien. Somit weisen die Klasse 1 IGKV Gene mehr Homologien auf und es ist wahrscheinlicher, dass bei einer Kürzung zwischen zwei (oder mehreren) Genen der gleichen Genfamilie nicht unterschieden werden kann. Demnach sind Sequenzen, bei denen mehrere Gene mit dem gleichen Score gefunden worden sind, nicht so problematisch wie Sequenzen, bei denen keins oder ein anderes Gen bzw. andere Gene identifiziert werden: im ersten Fall kann meistens die korrekte Genfamilie zugewiesen werden, während im zweiten Fall eine Kürzung zur falschen Identifizierung bzw. falschen Auswertung führt. Daher ist es für eine korrekte V Gen Identifizierung ratsam, keine gekürzte Sequenzen zu analysieren – oder zum Erreichen einer 80 % Genauigkeit eine maximale Kürzung von ca. 20 bp zuzulassen. Aus diesem Grund wurden die optimale sowie die minimal erforderliche Read-Länge mit Hilfe der scFv-Testsequenzlängen ermittelt. Für VH (IGH) Antikörpersequenzen wäre eine Read-Länge von 452 bp optimal; als minimale Read-Länge wären 404 bp zu empfehlen, um 75 % der Sequenzen in vollständiger Länge zu ermitteln und aufgrund der 20 bp Kürzung davon ca. 80 % korrekt zu identifizieren (insgesamt ca. 60 % Genauigkeit). Da das Titanium Kit eine Read-Länge von 400 bp erlaubt, wurden für die weiteren Analysen 400 bp als minimale Sequenzlänge eingesetzt.

Ein anderer Schwerpunkt der bioinformatischen Untersuchungen ist mit der eventuell nicht korrekten Ermittlung der Homopolymer-Nukleotidanzahl verbunden. Um mögliche Probleme bei der Auswertung im Voraus zu erkennen und zu lösen, wurden die beiden Datensets (V Gene und Testsequenzen) bezüglich der Länge und der Anzahl der Nukleotid-Homopolymere untersucht. Die Ergebnisse zeigen, dass bei beiden Datensets annähernd 100 % aller Sequenzen Homopolymere mit einer Länge von mindestens 4 bp enthalten. Wie zu erwarten, sinkt der Anteil der Homopolymer-enthaltenden Sequenzen mit Zunahme der Homopolymer-Länge, wobei dieser Anteil bei den scFv-Testsequenzen höher ist. Wichtig ist, dass bei den VBASE2 Genen IGKV einen deutlich geringeren Anteil an Sequenzen mit Homopolymeren ab 5 bp aufweist, während IGHV ca. 7,5- bis 9,1-facher Anteil an Sequenzen

mit Homopolymeren ab 6 bp als IGLV bzw. IGKV beinhaltet. Dementsprechend würden *Rearrangements* mit IGKV weniger Probleme bei der Auswertung verursachen, während bei IGHV die Wahrscheinlichkeit dafür größer ist. Wie vermutet, steigt die Anzahl an Homopolymeren in den Sequenzen mit kleiner werdender Homopolymerlänge: bei den V Genen sind bis zu 7, bei den scFv-Testsequenzen bis zu 12 Homopolymere pro Sequenz zu finden. Der erhöhte Anteil an scFv-Testsequenzen mit Homopolymeren sowie die größere Anzahl an Homopolymeren in diesen Sequenzen sind dadurch zu erklären, dass sie nicht nur aus V, sondern auch aus (D)J Gensegmenten bestehen. Insgesamt ist zu erwarten, dass je größer die Homopolymer-Anzahl in einer Sequenz ist und je länger die Homopolymere sind, desto höher wird die Wahrscheinlichkeit für eine Leserasterverschiebung. Zwar wurden nicht so viele Sequenzen mit Homopolymeren ab 7 bp gefunden (maximal 12,5 % bei den IGKV scFv-Testsequenzen) und laut Margulies et al. [28] wird die Ermittlung der Homopolymerlänge erst ab 8 bp inkorrekt, trotzdem würde erst die Auswertung von Next Generation Sequenzierdaten Aufschluss über die möglichen Leserasterverschiebungen geben können.

Als Nächstes wurde der Mutationsanteil in den scFv-Testsequenzen ermittelt, um daraus einen programmiertechnischen Parameter (*Cut Off*, Prozentsatz der Identität zwischen der getesteten Sequenz und dem Keimbahngen) zu bestimmen. Falls die Identität einer Sequenz mit dem Keimbahngen kleiner als der berechnete *Cut Off* Wert ist, wird diese Sequenz als zu stark mutiert für die weiterfolgenden Analysen betrachtet und nicht ausgewertet. Aufgrund der statistischen Auswertung wurden drei *Cut Off* Werte ermittelt: 97,5 % für IGHV, 94,6 % für IGKV bzw. IGLV sowie 97,0 % für die J Gensegmente.

Da die Anzahl der Mutationen in FR1 bei der Mutationsauswertung der scFv-Testsequenzen nach FRs und CDRs überraschend groß war, wurde zusätzlich eine Analyse zur Ermittlung der durch Primer verursachten Mutationen durchgeführt. Die Ergebnisse zeigen, dass bei einer Beibehaltung der Mutationsrate in FR1 über 70 % (IGHV bzw. IGLV) sowie ca. 42 % (IGKV) der Mutationen auf die 23 bp lange Primerbinderegion zurückzuführen sind. Falls die Primer keine Mutationen in den scFv-Testsequenzen hervorrufen würden, wäre der Mutationsanteil in den V Genen um ca. 1-1,2 % (bei IGHV bzw. bei IGLV) sowie um ca. 0,4 % (bei IGKV) geringer. Diese Analysen verdeutlichen die Bedeutsamkeit des Primerdesigns, besonders da die Unterschiede zwischen den VBASE2 V Keimbahngenen teilweise im Primerbereich liegen. Dafür spricht der relativ hohe Anteil an Sequenzen mit mehreren Treffer mit gleichen Scores (zwischen 4,4 und 7,7 % bei der 20 bp Kürzung der V Gene sowie bis zu 10 % bei den ungekürzten, aber mutierten scFv-Testsequenzen). Zusammenfassend: Je weniger Mutationen die Primer verursachen, umso korrekter wird die Genidentifizierung.

Da bei dieser Analyse die scFv-Testsequenzen (hergestellt mit Primern, die an der TU Braunschweig benutzt werden) ausgewertet worden sind, stellt sich die Frage, wie viele Mutationen die weniger degenerierten für das Projekt vorgesehenen MPIMG-Primer verursachen würden. Außerdem wäre es wichtig zu wissen, ob alle V Gene durch die MPIMG-Primer amplifiziert werden. Daher wurde die *Coverage* der 576 VBASE2 V Gene durch diese Primer ermittelt, insbesondere weil das Primerdesign auf bisher veröffentlichen Primern basiert [66]. Eine *Coverage* von insgesamt 89,1 % ergibt sich bei minimaler Länge der überlappten Region von 18 bp vom 3' Ende des Primers und bei maximal 4 zugelassenen *Mismatches*. Aufgrund der Existenz von Genen, die zu keiner Genfamilie zugeordnet sind, wurde die Auswertung familienbezogen durchgeführt: 79,7 % der V Gene gehören einer Genfamilie, davon werden 93,9 % mit dem Primerset amplifiziert (bzw. 100 % der zu untersuchenden Familie VH4); von den restlichen V Genen werden 70,1 % abgedeckt. Diese *Coverage* ist sehr hoch, jedoch spiegeln die ermittelten Werte die Funktionalität der Gene nicht wieder. Daher wurden die V Gene zusätzlich nach Funktionalität aufgrund der VBASE2 Klassifikation [8] aufgeteilt und die Analyse wurde wiederholt, um noch präziser die *Coverage* ermitteln zu können. Somit wird ein Schwerpunkt auf die Amplifizierung von V Genen gelegt, die tatsächlich in *Rearrangements* vorkommen. Unter Vernachlässigung der Pseudogene bzw. der Orphans ergibt sich eine *Coverage* aller funktionellen (Klasse 1) sowie möglicherweise funktionellen V Gene (Klasse 2 und 3) von 94,1 % (bzw. 99,3 % der Klasse 1 V Gene). Anscheinend wurden die Primer [66] für die Amplifizierung von familienzugeordneten, funktionellen (VBASE2 Klasse 1) V Genen designed. Ziel dieses Projektes ist jedoch der Vergleich der Nutzung aller Antikörpergene. Aus diesem Grund werden vier neue Primer vorgeschlagen. Diese ermöglichen die experimentelle Untersuchung, ob 11 bisher nur auf genomischer Ebene bewiesene V Gene tatsächlich in der V(D)J Rekombination teilnehmen. Somit werden nur 7 von den 321 funktionellen bzw. möglicherweise funktionellen Genen nicht amplifiziert: deren *Coverage* erhöht sich auf 97,8 % (bzw. auf 100 % der Klasse 1 V Gene). Daher erfolgt mit dem erweiterten Primer Set eine nahezu 100%ige Abdeckung der funktionellen bzw. der möglicherweise funktionellen Antikörpergene.

Zusätzlich wurde durch diese Primerauswertung gezeigt, dass der veröffentlichte VH4 Primer [66] einen Fehler enthält: das degenerierte Nukleotid (S) ist an der falschen Stelle. Jedoch ist es in diesem Fall zu empfehlen, statt der Korrektur der Primersequenz C anstelle von S in der Sequenz einzubauen. Sonst würde einer der drei Unterschiede (*Mismatches*) des VH4 Primers zu VH4(DP63) nicht mehr existieren und folglich wäre die Differenz in der Bindungsspezifität der beiden Oligonukleotide reduziert. Weiterhin wurde festgestellt, dass

drei der Primer nicht alle Familien amplifizieren, die in der Veröffentlichung von Sblattero und Bradbury [66] genannt wurden. Außerdem wäre zu empfehlen, den Primer VL15910 nicht einzusetzen, da durch ihn die *Coverage* der Genfamilien VL1, VL5 und VL9 dupliziert wird. Somit könnten falsche Schlussfolgerungen über die Genverteilung gezogen werden.
Die Gründe für die festgestellten Unterschiede liegen einerseits an dem damaligen Wissensstand: für das Design der Primer wurde die VBASE Datenbank [34] benutzt, sie enthält jedoch nur 25,7 % der V Gene in der VBASE2 [8]. Andererseits wurden anderen Kriterien für das Primerdesign verwendet: mindestens 16 bp Homologie und keine maximale Anzahl an *Mismatches* zwischen den Primern und den V Genen (verglichen mit 18 bp Homologie und Höchstgrenze von vier *Mismatches*). Es lässt sich zusammenfassen, dass durch die Primerauswertung bestehende Fehler korrigiert werden können, um somit eine repräsentative Analyse über die Antikörperverwendung durchführen zu können.

Auswertung der Next Generation Sequenzierdaten

Ziel der Analyse der Next Generation Sequenzierdaten ist ein tieferes Verständnis der Unterschiede zwischen Gesunden und Autoimmunpatienten bezüglich der Vorgänge, die zu Antikörperdiversität führen, wie z.B. die bei der V(D)J Rekombination entstehende Genverteilung der Antikörper-Isotypen IgG und IgD. Weiterhin stand die Untersuchung der Genfamilie VH4 und speziell des V Gens VH4(DP63) im Vordergrund.
Für die Auswertung der Sequenzen wurde eine neue VBASE2 Anwendung („*Statistic Analysis*“) programmiert. Sie gewährleistet den Vergleich der Gennutzung bei zwei zu untersuchenden Gruppen, die je nach Fragestellung zusammengestellt werden können. Jedoch werden keine VDJ Kombinationen, sondern die V, D und J Gene unabhängig voneinander gezählt und ausgegeben. Hingegen bietet die entwickelte relationale Datenbank nextIGbase die Möglichkeit, viele zusätzliche Analysen durchzuführen, wie z.B. die Ermittlung der VDJ Kombinationen, die Bestimmung der Anzahl an Mutationen in den Sequenzen, die Ermittlung der P und N Nukleotidanzahl, die Suche nach Sequenzen mit identischen CDRs und die Aufteilung nach *Uniques* (Sequenzen, die identisch sind und mehrmals in der zu untersuchenden Gruppe vorkommen). Das Webinterface bietet außerdem die Option, bei einer Veröffentlichung weltweit auf die Datenbank zuzugreifen.

Da die Next Generation Sequenzierdaten mittels Pyrosequenzierung gewonnen wurden, mussten zunächst Untersuchungen über die zu erwartenden Probleme, wie z.B. eventuell fehlerhafte Ermittlung der Homopolymer-Nukleotidanzahl sowie kürzere Reads, durchgeführt werden. Es wurde festgestellt, dass entgegen publizierter Daten [28] das *Genome Sequencer FLX Instrument* auch bei kurzen Homopolymersequenzen (3 bp lang) Schwierigkeiten hat, die korrekte Nukleotidanzahl zu ermitteln. Bei einer falschen Anzahl an Homopolymer-Nukleotiden entstehen *Frameshifts* bei der Auswertung mit „*DNAPLOT Query*“, da das DNAPLOT Programm keine Lücken (Gaps) einführen kann. Diese *Frameshifts* können mit Perl-Skripten nicht behoben werden und eine manuelle Korrektur der Sequenzen muss aufgrund der großen Anzahl an Sequenzen ausgeschlossen werden. Daher wurde beabsichtigt, die *Frameshifts* durch Vergleich mit der CDR3 Region von korrekt alignierten Sequenzen zu identifizieren, um sie anschließend entfernen zu können. Dieser Ansatz konnte jedoch nicht realisiert werden, da sich die restlichen Regionen auch bei identischer CDR3 durch die Anzahl und die Position der Punktmutationen im V bzw. im J Gen unterscheiden können. Somit konnten von insgesamt 25601 Sequenzen, die länger als die ermittelte minimale Read-Länge sind, nur 10055 korrekt alignierten Sequenzen (ohne Pseudogen-*Rearrangements*) analysiert werden. In der Zukunft könnte eine Möglichkeit zur Entfernung von *Frameshifts* erprobt werden: Bei einer 100%igen Übereinstimmung von korrekt alignierten Sequenzen mit dem korrekt alignierten Teil der problematischen Sequenzen (Teilsequenzen) ist anzunehmen, dass die jeweiligen Sequenzen bei Nicht-Auftreten von *Frameshift(s)* identisch sind. Eine andere Möglichkeit wäre der Einsatz von BLAST [67] für die V Gen Identifizierung, jedoch wäre nur eine Bestimmung der Anzahl an Mutationen im ganzen V Gen und keine in den FR und CDR Regionen realisierbar. Außerdem würden die mutierten Nukleotide am 3'-Ende des V Gens berücksichtigt werden, während „*DNAPLOT Query*“ die V-Gen-Sequenz nur bis zum Ende der Homologie mit dem Keimbahngen berücksichtigt.

Als Nächstes wurde die Spezifität der Primerbindung an ihre Zielsequenzen ermittelt; die Primer amplifizieren fast ausschließlich die Sequenzen, für die sie designed wurden. Dies ermöglicht semiquantitative Analysen mittels Agarose-Gelektrophorese [68]. Trotzdem wurden nur die eigentlichen Zielsequenzen bei der Auswertung der Next Generation Sequenzierdaten berücksichtigt, um eventuelle Verfälschung der Ergebnisse zu vermeiden.

Weiterhin wurde überprüft, ob die ermittelten *Cut Off* Werte (97,5 % bzw. 97 % für V bzw. J Gensegmente) für die Analyse der Sequenzierdaten einsetzbar sind: einerseits sollen Sequenzen mit internen *Frameshift*-Bereichen ausgeschlossen werden, andererseits sollen genügend Sequenzen ausgewertet werden. Ein *Cut Off* Wert für das D Segment kann zwar auch

eingesetzt werden, allerdings könnte aufgrund der kurzen Länge der D Gene nur eine einzige Mutation zu ca. 80%iger Identität mit dem Keimbahngen führen. Deswegen wurde auf den Einsatz eines D *Cut Off* Wertes verzichtet.

Die ermittelten V bzw. J *Cut Off* Werte (97,5 % bzw. 97 %) erwiesen sich als zu restriktiv: nur 99 IgG Sequenzen könnten demnach analysiert werden. Auf Basis von so wenigen Sequenzen wäre kein Vergleich der Gennutzung zwischen IgG und IgD möglich und dementsprechend wurde ein *Cut Off* Wert von 90 % für V und J Gene eingesetzt, wobei keine internen *Frameshifts* gefunden wurden. Somit entsprechen 5127 Sequenzen den gewählten *Cut Off* Kriterien. Zu empfehlen wäre, bei zukünftigen Analysen den *Cut Off* Wert schrittweise zu senken und zu erhöhen, um somit den optimalen Wert für die Analysen herauszufinden.

Außerdem muss ausgeschlossen werden, dass aufgrund einer Überamplifizierung Sequenzen bei der Auswertung mehrmals und überproportional berücksichtigt werden. Dementsprechend wurde die Verteilung von gleichen Nukleotidsequenzen pro *Unique* ermittelt. Dabei wurde die von „*DNAPLOT Query*“ ermittelte VDJ Sequenz als maßgebend betrachtet, da identische VDJ Sequenzen in mehreren vom *Genome Sequencer FLX Instrument* ermittelten Gesamtsequenzen gefunden wurden. Der Unterschied kann demnach nur auf Mutationen in der IgG bzw. IgD CH1 Region zurückzuführen sein, die vermutlich auf Problemen bei der Detektion der Homopolymerlänge beruhen.

Da 68,3 % der 2852 *Uniques* nur ein einziges Mal vorkommen und ca. 92,7 % bis zu maximal dreimal vorhanden sind, wurde schlussgefolgert, dass keine Überamplifizierung während der Sequenzierungsvorbereitung stattfand und somit multiples Vorhandensein der B-Zellen der Grund für die mehrmals vorkommenden Sequenzen sein muss. Daher wurden alle Sequenzen ausgewertet, die dem 90 % *Cut Off* Wert entsprachen, und nicht nur die *Uniques.*

Ein weiterer Nachteil der Next Generation Sequenzierung neben der möglicherweise inkorrekten Homopolymerlänge-Detektion ist die im Vergleich zum Standard (Sanger-Sequenzierung) kürzere Read-Länge [26, 27]. Durch fehlende Abdeckung des V-Gen-Anfangs aufgrund der 3'-5'-Sequenzierrichtung könnte das ursprüngliche Gen möglicherweise nicht korrekt identifiziert werden. Um solche Probleme zu vermeiden, wurde die minimal erforderliche Read-Länge berechnet (400 bp) und als Auswahlkriterium für die Sequenzen für die Gennutzung-Analyse eingesetzt. Dennoch muss sichergestellt werden, dass trotz dieser Sicherheitsmaßnahme keine Gene mit multipler V-Gen-Zuordnung ausgewertet werden. Es wurde festgestellt, dass nur 0,67 % der 5127 Sequenzen eine multiple V-Gen-Zuordnung aufgrund von kürzeren Read-Längen aufweisen, während 74,4 % der Sequenzen mit multipler V-Gen-Zuordnung aufgrund von Mutationen im V Gen auftreten. Dabei werden bei über 90 %

der Sequenzen mit nicht abgedecktem V-Gen-Anfang die V Gene korrekt identifiziert. Aufgrund dieser Ergebnisse ist zu empfehlen, für zukünftige Analysen die minimale Read-Länge zu senken. Somit wird die Anzahl der Sequenzen für die Auswertung der Gennutzung erhöht. Zwar wird auch die Anzahl der Sequenzen mit multipler V-Gen-Zuordnung größer, jedoch ist es mit Hilfe von Datenbank-Abfragen möglich, sie auszuschließen, und dennoch die korrekt identifizierten Gene zu berücksichtigen.
Zusammenfassend: Durch die Voruntersuchungen wurde der Einfluss von Fehlern bei der Ermittlung der Homopolymerlänge minimiert und es wurde sichergestellt, dass keine Überamplifizierung der Sequenzen stattfand. Zusätzlich wurde gezeigt, dass bei der gewählten minimalen Read-Länge kaum gekürzte Sequenzen vorkommen (0,66 % aller 5127 Sequenzen bei 90 % *Cut Off*), die eine Identifizierung der V Gene erschweren.

Dementsprechend wurden alle Voraussetzungen erfüllt, die einen Vergleich der Nutzung von Antikörpergenen hinsichtlich der Antikörperentstehung sowie -reifung bei Gesunden und bei Autoimmunpatienten ermöglichen. Daher wurden drei Schwerpunkte für die Datenauswertung festgelegt: (i) die Häufigkeit der Rekombination zwischen bestimmten V, D und J Gensegmenten, (ii) die Länge der hinzugefügten nicht-kodierten N Nukleotide zwischen den Gensegmenten und (iii) die Anzahl an Mutationen in den Sequenzen aufgrund der somatischen Hypermutation. Dabei wurden Unterschiede zwischen den Antikörper-Isotypen IgG und IgD bei der VH4 Familie und speziell bei dem Gen VH4(DP63) untersucht, wobei komplette *Rearrangements* mit identifizierten V und J Genen und mit korrekt aligniertem V-Gen-Anfang bei einem *Cut Off* Wert von 90 % eingesetzt wurden.
Zu erwarten sind einerseits Differenzen in der Gennutzung zwischen IgG und IgD aufgrund der verschiedenen Antigenbindungsspezifität. Andererseits sind bei IgG Unterschiede zwischen Gesunden und Autoimmunpatienten möglich, da bei RA spezifische Autoantikörper wie z.B. Rheumafaktoren [11] und ACPA [12, 13, 14] auftreten, während dies bei Gesunden nicht der Fall ist. Außerdem könnten die IgD Sequenzen bei beiden Gesundheitszuständen voneinander abweichen, da IgM anti-IgG Teil der Rheumafaktoren sind (IgM und IgD werden durch alternatives m-RNA-Spleißen gebildet [2, 3, 4]). Weiterhin sind Differenzen bei der Gennutzung von VH4(DP63) zu vermuten [69, 70].

Zunächst wurde die V(D)J Rekombination bei Sequenzen mit V, D und J Einzel-Genzuordnung analysiert. Identifiziert wurden 15 V Gene bzw. V Genallele, dementsprechend ist es nicht immer möglich, die untersuchten Gruppen miteinander zu vergleichen: In der vom Primer VH4(DP63)-Fw amplifizierten Gruppe ist nur ein einziges Gen enthalten, während bei VH4-

Fw dies für insgesamt 14 Genallele zutrifft. In der VBASE2 Datenbank sind drei VH4(DP63) Allele enthalten, wobei aufgrund des benutzten Primers nur zwei identifiziert werden können: die Differenz zwischen humIGHV201 und humIGHV033 ist eine einzige Mutation im Primerbereich. Es wurde außerdem festgestellt, dass acht bisher nur auf genomischer Ebene identifizierte V Gene bzw. Genallele (VBASE2 Klasse 2) tatsächlich funktionell sind. Das Vorhandensein von unterschiedlichen Allelen bei den untersuchten Sequenzen wird durch den Einsatz von cDNA Pools (keine einzelnen Patientenproben) erklärt. Zusätzlich wurden D Segmente durch Inversion rekombiniert.

Insgesamt wurden 568 VDJ Kombinationen gefunden. Davon beinhalten 103 das zu untersuchende Gen VH4(DP63), wobei bedeutende Unterschiede zwischen Gesunden und RA-Patienten bei dem IgG Isotyp festgestellt wurden. Jeweils zwei verschiedene VH4(DP63)-DJ Kombinationen sind bei 84,1 % (RA) bzw. 65,5 % (ND) der untersuchten Sequenzen zu finden. Bis zu maximal 33 % davon beruhen auf *Uniques*, die mehr als 20 Mal gefunden wurden. Dementsprechend werden diese VDJ Kombinationen tatsächlich bevorzugt und nicht aufgrund von einzelnen, überproportional amplifizierten Sequenzen gefunden. Bei den Genen der VH4 Familie wurde eine einzige VDJ Kombination gefunden, die bei ca. 20 % der RA IgG Sequenzen vorkommt, und zwar mit dem Gen humIGHV168 (IGHV4-4*07). Da die untersuchte Gruppe 14 V Genallele enthält, könnte zusätzlich eine Analyse der VDJ Kombinationen nur bei diesem Gen durchgeführt werden, um so den tatsächlichen Prozentsatz dieser VDJ Kombination beim humIGHV168 herauszufinden. Desweiteren wurden bei Analyse der VDJ-Kombinationen keine Sequenzen mit multipler V, D oder J Genzuordnung berücksichtigt, jedoch könnten bei solchen Sequenzen zumindest die Genfamilien ermittelt werden. Daher wäre eine Analyse der VDJ-Genfamilienkombination empfehlenswert.

Aufgrund der Ergebnisse lässt sich schlussfolgern, dass bedeutende Unterschiede zwischen Gesunden und Autoimmunpatienten bei dem IgG Isotyp bezüglich der VH4 Gennutzung bestehen. Im Gegensatz dazu konnte kein Beweis für solche Differenzen bei dem IgD Isotyp gefunden werden. Der Grund dafür könnte sein, dass die erhöhte Spezifität der Bindung von IgG zu Antigenen eine größere Anzahl von produzierenden B-Zellen und somit eine erhöhte Herstellung von löslichen Antikörper gegen oft erkannte Antigene zur Folge hat, während IgD hauptsächlich auf der Oberfläche von naiven, nicht mittels Antigenbindung aktivierten B-Lymphozyten membrangebunden vorliegt.

Zweiter Schwerpunkt der Auswertung der Next Generation Sequenzierdaten ist die Länge der CDR3 Regionen bzw. die damit verbundene Länge der D Gensegmente sowie die N

Nukleotidanzahl hinsichtlich des Hinzufügens von palindromischen (P) und nicht-kodierten (N) Nukleotiden zwischen den VDJ Gensegmenten.

Bei der Untersuchung der CDR3 Länge wurden gravierende Unterschiede zwischen Gesunden und RA-Patienten nur bei IgG VH4(DP63) nachgewiesen. Dies korreliert mit den Ergebnissen der Analyse der VDJ-Kombinationen und hängt wahrscheinlich von der Gennutzung ab.

Bei der Längen-Spannweite wurde eine durchschnittliche Differenz zwischen IgD und IgG von sieben Aminosäuren festgestellt. Dies könnte wie erwähnt auf die Funktion der Ig Isotypen zurückzuführen sein. Außerdem existieren bei IgD CDR3 Regionen mit einer Länge von deutlich über 25 Aminosäuren. Auch bei der Vorauswertung sind viele RA IgD VDJ *Rearrangements* des Gens VH4(DP63) aufgrund ihrer überdurchschnittlich langen CDR3 Regionen aufgefallen, jedoch wiesen die meisten davon *Frameshifts* auf. Es wäre zu empfehlen, nach Entfernung der *Frameshifts* die Sequenzen mit langen CDR3 erneut auszuwerten und auf Unterschiede zwischen Gesunden und Autoimmunpatienten zu untersuchen. Eine andere Variante wäre, trotz *Frameshifts* im V Gen die ab CDR3 korrekt alignierten Sequenzen zu analysieren.

Lange CDR3 Regionen können nur auf längeren D Gensegmenten oder auf längeren N Nukleotidsequenzen beruhen, da die P Nukleotidanzahl im Vergleich zur N Nukleotidanzahl gering ist. Dementsprechend wurden die Längen dieser Regionen ermittelt und mittels Chi-Quadrat-Vierfeldertests eine statistisch signifikante Differenz ($P < 0{,}0001$) zwischen IgD und IgD Sequenzen mit längeren D Gensegmenten (ab 20bp) nachgewiesen. Somit werden bei IgG und IgD unterschiedlich lange Teile der D Gensegmente rekombiniert. Weiterhin wurde ermittelt, dass 71,1 % der ND IgG VH4(DP63) Sequenzen über 20 bp N Nukleotide enthalten, wobei die Differenz zu den anderen untersuchten Gruppen ca. 59-69 % beträgt und mit einer Wahrscheinlichkeit $P < 0{,}0001$ statistisch signifikant ist. Somit sind Unterschiede zwischen Gesunden und Autoimmunpatienten bezüglich dieses Gens bei IgG festgestellt worden. Aufgrund der hohen N Nukleotidanzahl wurde manuell überprüft, ob D-D Fusionen vorliegen. Die für die Analyse eingesetzte Anwendung „*DNAPLOT Query*“ ermittelt dies nicht automatisch, sondern übernimmt nur das D Segment bzw. die D Segmente mit dem höchsten Score für die Export-Tabelle. Trotz widersprüchlicher Angaben in verschiedenen Publikationen [43, 44, 71-75] wurden D-D Fusionen sowie durch Inversion rekombinierte D Segmente in der Tat identifiziert. Daher ist zu empfehlen, einerseits die Sequenzen mit langen N Nukleotidregionen auf D-D Fusionen zu überprüfen und andererseits zu untersuchen, wie oft bei diesen Fusionen D Segmente durch Inversion eingesetzt werden.

Der dritte Schwerpunkt für die Auswertung der Next Generation Sequenzierdaten ist die Bestimmung der Anzahl an Mutationen, die während der somatischen Hypermutation auftreten. Bei der Interpretation der Ergebnisse ist zu beachten, dass ein *Cut Off* Wert von 90 % eingesetzt wurde: somit wird eine obere Grenze für die Mutationsanzahl gesetzt. Bei zukünftigen Analysen kann der *Cut Off* Wert jedoch beliebig angepasst werden.
Die analysierten RA Sequenzen enthalten mehr Mutationen als jene der Gesunden. Außerdem werden weniger Mutationen bei den *Rearrangements* des VH4(DP63) Gens im Vergleich zu den *Rearrangements* der restlichen Gene der VH4 Familie gezählt. Bei den VH4(DP63)-DJ *Rearrangements* wurden sehr geringe Unterschiede zwischen ND und RA bezüglich der Anzahl der Mutationen festgestellt, jedoch wurden nachweislich verschiedene D und J Gene mit VH4(DP63) rekombiniert. Es wurde jedoch bestätigt, dass der Ig Isotyp am meisten Einfluss auf die somatische Hypermutation hat: der Anteil an IgD Sequenzen mit maximal 9 Mutationen beträgt 50,6 – 93,1%, während die Werte bei IgG deutlich niedriger sind ( 2,5 – 7,9 %). Die Differenz ist statistisch signifikant ($P < 0,0001$, Chi-Quadrat-Vierfeldertest). Auch bei den unmutierten Sequenzen wurde ein statistisch signifikanter Unterschied nachgewiesen ($P < 0,0001$, Chi-Quadrat-Vierfeldertest). Dies erklärt, wieso so wenig IgG Sequenzen eine Identität mit den Keimbahngenen über den ermittelten *Cut Off* Werten (97,5 % bzw. 97 % für V bzw. J Gensegmente) aufwiesen. Die höhere Mutationsanzahl bei IgG ist wahrscheinlich eine Folge der erhöhten Spezifität der Ig Bindung zu Antigenen und wird während der somatischen Hypermutation vor dem Class Switch verursacht [4, 76].

Abschließend wurde untersucht, wie sich die drei erwähnten Faktoren (VDJ Kombinationen, N Nukleode, somatische Hypermutation) insgesamt auf die Diversität der Antikörper auswirken, indem die Kombinationen von den drei für die Antigenbindung zuständigen CDR Regionen bestimmt wurden.
Von den untersuchten 5127 Nukleotidsequenzen sind 55,6 % anhand ihrer VDJ Sequenz, 54,0 % anhand ihrer Aminosäuresequenz und 43,6 % anhand ihrer CDR Kombination *Uniques*. Daher bleibt die Diversität der Antikörpergenrekombinationen fast vollständig auf Aminosäureebene erhalten. Erstaunlicherweise wurde jedoch festgestellt, dass keine einzige der insgesamt 2235 CDR Kombinationen bei den anderen untersuchten Gruppen zu finden ist und somit bedeutende Unterschiede zwischen Gesunden und Autoimmunpatienten bezüglich Ig Isotyp sowie Nutzung der Gene der VH4 Familie besteht.

Fazit der hier dargestellten Arbeit ist, dass sowohl der Ig Isotyp als auch der Gesundheitszustand ein Einfluss auf die Antikörperentstehung im Hinblick auf die Gennutzung der VH4 Familie sowie auf die Antikörperreifung und dementsprechend auf die Diversität haben. Die festgestellten Unterschiede sind ein guter Startpunkt für weitere bioinformatische sowie biologische Untersuchungen. Zum Beispiel könnte die Auftretenshäufigkeit von Aminosäuren mit bestimmten chemischen Eigenschaften in den CDRs analysiert werden. Möglich wäre außerdem die Herstellung der am häufigsten identifizierten Sequenzen auf Proteinebene. Diese könnten z.B. mittels Phage Display [6] oder Protein Array Technologie [77, 78] auf spezifische Bindung an humane Proteine untersucht werden, oder mittels ELISA (Enzyme-Linked Immunosorbent Assay) [ 79, 80] auf Bindung bereits bekannter Autoantigene wie. z.B. Rheumafaktor [11] oder citrullinierte Peptide [12, 13, 14] untersucht werden. Vorstellbar wäre es auch, dass anhand dieser Untersuchungen weitere spezifische Antikörper gefunden werden könnten, die für die therapeutische Anwendung bei RA (auch in rekombinanter Form) geeignet wären [81].

# 6. Zusammenfassung

Antikörper sind Glykoproteine der adaptiven Immunantwort. Ihre Funktion ist die Abwehr von Fremdkörpern, z.B. Krankheitserregern, jedoch führen Störungen des Immunsystems zu Autoimmunerkrankungen. Die Next Generation Sequenziertechnologie ist geeignet, einen schnellen Vergleich zwischen Gesunden und an Rheumatoider Arthritis (RA) leidenden Patienten auf Antikörpergenebene zu ermöglichen. Davor müssen jedoch die experimentellen sowie bioinformatischen Methoden für den Einsatz der Technologie etabliert werden.

Im experimentellen Teil dieser Arbeit wurde mittels quantitativer *Real-Time* Polymerase-Kettenreaktion (qRT-PCR) und Polymerase-Kettenreaktion (PCR) die Basis für den erfolgreichen Testlauf mittels *Genome Sequencer FLX Instrument* geschaffen. Im bioinformatischen Teil wurde mit Hilfe von Perl-Skripts und einer speziell für die Next Generation Sequenzierdaten programmierten PostgreSQL-Datenbank eine schnelle Analyse von großen Mengen an Next Generation Sequenzierdaten ermöglicht.

Somit konnten signifikante Differenzen in der Expression von IgG und IgD mittels qRT-PCR festgestellt werden. Außerdem wurden anhand der Next Generation Sequenzierdaten Unterschiede bezüglich der Antikörperentstehung, -reifung sowie -diversität zwischen Gesunden und RA-Patienten im Hinblick auf die Gennutzung von VH4(DP63) bei IgG und IgD festgestellt. Weitere Untersuchungen sind auf biologischer Ebene notwendig, um die bioinformatisch ermittelten Unterschiede in der Gennutzung zu erforschen und möglicherweise für die therapeutische Anwendung bei RA geeignete Antikörper zu identifizieren.

# 7. Danksagung

Durch diese Diplomarbeit lernte ich sehr viel – nicht nur fachlich, sondern auch persönlich. Es ist gleichzeitig ein Fluch und ein Segen, sich selbst zu sein. Damit ständig das Gute zum Vorschein kommt, bedarf es jedoch Unterstützung – für die ich mich ganz herzlich bedanken möchte:

Bei Prof. Dr. Stefan Dübel und Prof. Dr. Hans Lehrach für die Möglichkeit, dieses so interessante Thema erforschen zu dürfen.

Bei meinen Betreuern Dr. Michael Hust, Dr. Zoltán Konthur und Dr. Thomas Schirrmann: für die Möglichkeit, viele Gele laufen lassen zu können; für die (hoffentlich erfolgreichen) Versuche mir beizubringen, wie eine wissenschaftliche Arbeit sprachlich ohne viel Poesie zu gestalten ist; für die hilfreichen Vorschläge und die intensiven Diskussionen; für die Freiheit, meiner Neugierde keine Grenzen setzen zu müssen. Allen einen herzlichen Dank für die hervorragende Betreuung!

Bei Prof. Dr. Werner Müller für seine Erlaubnis, VBASE2, DNAPLOT sowie die Perl-Skripte bei der Arbeit nutzen zu dürfen.

Bei Theam Soon Lim für die wertvolle Hilfe und Ermunterungen bei den PCR Experimenten.

Bei Dr. Volker Sievert für die bioinformatische Hilfe.

Bei Florian Rubelt für die Sequenzierdaten.

Bei Chenna Reddy Galiveti für die Unterstützung bei den qRT-PCR Experimenten.

Bei Angela Filbry, Yanica Grachenova, Sandra Halecker und Anna Velkova - dafür, dass ihr da für mich seid, für eure Unterstützung und für euren Glauben an mich.

Bei meinem Verlobten Sebastian Buhlmann - für deine Geduld, für deine Hilfe, für deine Liebe.

Bei meiner Familie – für alles!

# 8. Literaturverzeichnis


1. **Cornish-Bowden A:** Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* 1985, 13(9):3021–3030.

2. **Janeway CA, Travers P, Walport M, Shlomchik M:** Immunologie, 5. Auflage. **Spektrum Akademischer Verlag** 2002.

3. **Murphy KM, Travers P, Mark Walport M, Kenneth M, Janeway C:** Janeway's Immunobiology, 7th Revised edition. *Taylor & Francis* 2008.

4. **Holländer GA:** Immunologie: Grundlagen für Klinik und Praxis. *Elsevier* 2005.

5. **Breitling F, Dübel S:** Rekombinante Antikörper. *Spektrum Akademischer Verlag* 1997.

6. **Hust M., Toleikis L, Dübel S:** Antibody phage display. In: **Dübel S** (Ed.): Handbook of therapeutic antibodies. *Willey-VCH* 2007, 45-68.

7. **Vladutiu AO:** Immunoglobulin D: Properties, Measurement, and Clinical Relevance. *Clin Diagn Lab Immunol.* 2000, 7(2):131–140.

8. **Retter I, Althaus HH, Münch R, Müller W:** VBASE2, an integrative V gene database. *Nucleic Acids Res.* 2005, 33(Database issue):D671-4.

9. **Klein U, Goossens T, Fischer M, Kanzler H, Braeuninger A, Rajewsky K, Kuppers R:** Somatic hypermutation in normal and transformed human B cells. *Immunol Rev.* 1998 Apr;162:261-80.

10. **Schatz DG, Oettinger MA, Schlissel MS:** V(D) J RECOMBINATION: Molecular Biology and Regulation Annu. *Rev. Immunol1*. 992. 10:359~83.

11. **Roessner A:** Allgemeine Pathologie und Grundlagen der Speziellen Pathologie, 11. Auflage. *Urban & Fischer Verlag* 2008.

12. **Firestein SG:** Evolving concepts of rheumatoid arthritis. *Nature* 2003, **423**, 356-361.

13. **Pratt AG, Isaacs JD, Mattey DL:** Current concepts in the pathogenesis of early rheumatoid arthritis. *Best Pract Res Clin Rheumatol.* 2009, 23(1):37-48.

14. **van Gaalen FA, Linn-Rasker SP, van Venrooij WJ, de Jong BA, Breedveld FC, Verweij CL, Toes RE, Huizinga TW:** Autoantibodies to cyclic citrullinated peptides predict progression to rheumatoid *arthritis in patients with undifferentiated arthritis: a prospective cohort study. Arthritis Rheum* 2004, 50:709-715.

15. **Tan EM, Kunkel HG:** Characteristics of a soluble nuclear antigen precipitating with sera of patients with systemic lupus erythematosus. *J. Immunol.* 1966, 96: 464-471.

16. **Akizuki M, Powers R, Holman HR:** A soluble acidic protein of the cell nucleus which reacts with serum from patients with systemic lupus erythermatosus and Sjögren's syndrome. *J Clin Invest.* 1977, 59(2):264-272.

17. **Munoz LE, Gaipl US, Herrmann M:** Predictive value of anti-dsDNA autoantibodies: importance of the assay. *Autoimmun Rev.* 2008, 7(8):594-597.

18. **Witte T:** IgM antibodies against dsDNA in SLE. *Clin Rev Allergy Immunol.* 2008, 34(3):345-347.

19. **Cepeda EJ, Reveille JD:** Autoantibodies in systemic sclerosis and fibrosing syndromes: clinical indications and relevance. *Curr Opin Rheumatol.* 2004, 16(6):723-732.

20. **Grassegger A, Pohla-Gubo G, Frauscher M, Hintner H:** Autoantibodies in systemic sclerosis (scleroderma): clues for clinical evaluation, prognosis and pathogenesis. *Wien Med Wochenschr.* 2008, 158(1-2):19-28.

21. **Fertig N, Domsic RT, Rodriguez-Reyna T, Kuwana M, Lucas M, Medsger TA Jr, Feghali-Bostwick CA:** Anti-U11/U12 RNP antibodies in systemic sclerosis: a new serologic marker associated with pulmonary fibrosis. *Arthritis Rheum.* 2009, 61(7):958-965.

22. **Sanger F, Coulson, AR:** A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 1975, 94:441–448.

23. **Shendure J, Mitra RD, Varma C, Church, GM:** Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* 2004, 5:335–344.

24. Herstellerinformation: Applied Biosystems® SOLiD™ 3 Plus System, *Applied Biosystems* 2009.

25. **Morozova O, Marra MA:** Applications of next-generation sequencing technologies in functional genomics. *Genomics* 2008, 92(5):255-264.

26. **Shendure J, Ji H:** Next-generation DNA sequencing. *Nat Biotechnol.* 2008, 26(10):1135-45.

27. **Mardis ER:** Next-Generation DNA Sequencing Methods. *Annu. Rev. Genomics Hum. Genet.* 2008. 9:387–402

28. **Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM:** Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, 437:376-380.

29. **Droege M, Hill B:** The Genome Sequencer FLX$^{TM}$ System—Longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology* 2008*,* 136:3–10.

30. **Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR:** High-throughput sequencing of the zebrafish antibody repertoire. *Science* 2009, 324(5928):807-810.

31. **Herstellerinformation:** http://www.454.com/

32. **Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N:** 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* 2006, 7:275.

33. **Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM:** Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 2007, 8(7):R143.

34. **VBASE:** http://vbase.mrc-cpe.cam.ac.uk/

35. **Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc MP:** IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.* 2006, 34(Database issue):D781-784.

36. **Giudicelli V, Chaume D, Lefranc MP:** IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 2005, 33(Database issue):D256-261.

37. **IgBLAST:** http://www.ncbi.nlm.nih.gov/igblast/

38. **Mollova S, Retter I, Müller W:** Visualising the immune repertoire. *BMC Systems Biology* 2007, 1(Suppl 1):P30.

39. **Mollova S:** Development of a web tool for analysis of antibody genes. *Student research project, unpublished* 2007.

40. **Mollova S, Retter I, Hust M, Dübel S, Müller W:** Analysis of single chain antibody sequences using the VBASE2 Fab Analysis tool. In Preparation 2009.

41. **Brochet X, Lefranc MP, Giudicelli V:** IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res. 2008,36(Web Server issue):W503-508.

42. **Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP:** IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. *Bioinformatics* 2004, 20:379–385.

43. **Souto-Carneiro MM, Longo NS, Russ DE, Sun HW, Lipsky PE:** Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J Immunol.* 2004, 172(11):6790-802.

44. **Ohm-Laursen L, Nielsen M, Larsen SR, Barington T:** No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology* 2006, 119(2):265-77.

45. **Gaeta BA, Malming HR, Jackson KJ, Bain ME, Wilson P, Collins AM:** iHMMune-align: Hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* 2007, 23:1580-1587.

46. **Lefranc MP, Pommie C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G:** IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol.* 2003, 27(1):55-77.

47. **Hust M, Toleikis L, Dübel S:** Antibody phage display. Handbook of therapeutic antibodies, Ed. Dübel, S., *Wiley-VCH* 2007, 45-68.

48. **Kirsch MI, Hülseweh B, Nacke C, Rülker T, Schirrmann T, Marschall HJ, Hust M, Dübel S:** Development of human antibody fragments using antibody phage display for the detection and diagnosis of Venezuelan equine encephalitis virus (VEEV). *BMC-Biotechnology* 2008, 8:66.

49. **Lim TS, Mollova S, Dübel S, Lehrach H, Konthur Z:** V–gene amplification revisited – An optimised procedure for human antibody isotype and idiotype amplification. *New Biotechnol.* in revision 2009.

50. AMPD1 – Bulletin, *Sigma®,* 2005.

51. SuperScript II Reverse Transcriptase Manual, *Invitrogen™*, 2003.

52. **Heid CA, Stevens J, Livak KJ, Williams PM:** Real time quantitative PCR. *Genome Res.* 1996, 6(10):986-94.

53. ABI PRISM 7700 Sequence Detection System. *User Bulletin #2, Applied Biosystems* 2001.

54. **Yuan JS, Wang D, Stewart Jr. CN:** Statistical methods for efficiency adjusted real-time PCR quantification *Biotechnol. J.* 2008, 3:112–123.

55. **Vandesompele J et al:** Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, *Genome Biology* 2002, 3(7):research0034.1–0034.11.

56. **Jarvie T:** Amplicon Sequencing. *Application Note No. 5, Roche Applied Science* 2007

57. Guide To Amplicon Sequencing, *Roche*, 2006.

58. **Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H:** Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology* 1986, 51(1):263-73.

59. **Frey B, Suppmann B:** Demonstration of the Expand(tm) PCR System's Greater Fidelity and Higher Yields with a *lac*I-based PCR Fidelity Assay. *Biochemica* 1995, 2: 8-9.

60. **Reddy MS, Vaze MB, Madhusudan K, Muniyappa K:** Binding of SSB and RecA protein to DNA-containing stem loop structures: SSB ensures the polarity of RecA polymerization on single-stranded DNA. Biochemistry 2000, 39(46):14250-62.

61. **Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS:** 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.* 1991, 19:4008.

62. **Chang M, Lee HJ:** Gradient polymerase chain reaction performance using regular thermal cycle machine.*Anal Biochem*. 2005, 340(1):174-7.

63. Kurzprotokoll Cycle Pure Kit, PEQLAB Biotechnogie GmbH.

64. **Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. E.:** Molekulare Zellbiologie, 4. Auflage. *Spektrum Akad.Verlag* 2001.

65. **Cochrane G, Akhtar R, Aldebert P, Althorpe N, Baldwin A, Bates K, Bhattacharyya S, Bonfield J, Bower L, Browne P, Castro M, Cox T, Demiralp F, Eberhardt R, Faruque N, Hoad G, Jang M, Kulikova T, Labarga A, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Plaister S, Robinson S, Sobhany S, Vaughan R, Wu D, Zhu W, Apweiler R, Hubbard T, Birney E:** Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 2008, 36(Database issue):D5-12.

66. **Sblattero D, Bradbury A:** A definitive set of oligonucleotide primers for amplifying human V regions. *Immunotechnology* 1998, 3:271-278.

67. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ:** Basic Local Alignment Search Tool. *J. Mol. Biol.* 1990, 215:403–410.

68. **Lim TS, Konthur Z:** persönliche Kommunikation.

69. **Williams DG, Moyes SP, Mageed RA:** Rheumatoid factor isotype switch and somatic mutation variants within rheumatoid arthritis synovium. *Immunology.* 1999, 98(1):123–136.

70. **Hara Y, Nakamura N, Kuze T, HashimotoY, Sasaki Y, Shirakawa A, Furuta M, Yago K, Kato K, Abe M:** Immunoglobulin Heavy Chain Gene Analysis of Ocular Adnexal Extranodal Marginal Zone B-Cell Lymphoma. *Investigative Ophthalmology and Visual Science* 2001, 42:2450-2457.

71. **Brezinschek HP, Brezinschek RI, Lipsky PE:** Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *J. Immunol.* 1995, 155:190.

72. **Collins AM., Ikutani M, Puiu D, Buck GA, Nadkarni A, Gaeta B:** Partitioning of rearranged Ig genes by mutation analysis demonstrates D-D fusion and V gene replacement in theexpressed human repertoire. *J. Immunol.* 2004, 172:340.

73. **Meek KD, Hasemann CA, Capra JD:** Novel rearrangements at the immunoglobulin D locus: inversions and fusions add to IgH somatic diversity. *J. Exp. Med.* 1989, 170:39.

74. **Corbett SJ, Tomlinson IM, Sonnhammer EL, Buck D, Winter G:** Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J. Mol. Biol.* 1997, 270:587.

75. **Kompfner E, Oliveira P, Montalbano A, Feeney AJ:** Unusual germline DSP2 gene accounts for all apparent V-D-D-J rearrangements in newborn, but not adult, MRL mice. *J. Immunol.* 2001, 167:6933.

76. **Liu YJ, Malisan F, de Bouteiller O, Guret C, Lebecque S, Banchereau J, Mills FC, Max EE, Martinez-Valdez H:** Within germinal centers, isotype switching of immunoglobulin genes occurs after the onset of somatic mutation. *Immunity* 1996, 4(3):241-50???.

77. **Eickhoff H, Konthur Z, Lueking A, Lehrach H, Walter G, Nordhoff E, Nyarsik L, Büssow K:** Protein array technology: the tool to bridge genomics and proteomics. *Adv Biochem Eng Biotechnol.* 2002, 77:103-112.

78. **Büssow K, Konthur Z, Lueking A, Lehrach H, Walter G:** Protein array technology. Potential use in medical diagnostics. *Am J Pharmacogenomics* 2001, 1(1):37-43.

79. **Faith A, Pontesilli O, Unger A, Panayi GS, Johns P:** ELISA assays for IgM and IgG rheumatoid factors. *J Immunol Methods* 1982, 55:169–177.

80. **Engvall E, Perlman P:** Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry* 1971, 8:871-874.

81. **Dübel S (Ed.):** Handbook of therapeutic antibodies, *Wiley-VCH* 2007.

# 9. Anhang (Tabellen und Abbildungen)

**Tabelle A1:** MID Sequenzen für Amplicon Sequenzierung durch das *Genome Sequencer FLX Instrument*.

| Name | 5'-3' Sequenz | Name | 5'-3' Sequenz | Name | 5'-3' Sequenz |
|---|---|---|---|---|---|
| MID1 | ACGAGTGCGT | MID5 | ATCAGACACG | MID9 | TAGTATCAGC |
| MID2 | ACGCTCGACA | MID6 | ATATCGCGAG | MID10 | TCTCTATGCG |
| MID3 | AGACGCACTC | MID7 | CGTGTCTCTA | MID11 | TGATACGTCT |
| MID4 | AGCACTGTAG | MID8 | CTCGCGTGTC | MID12 | TACTGAGCTA |

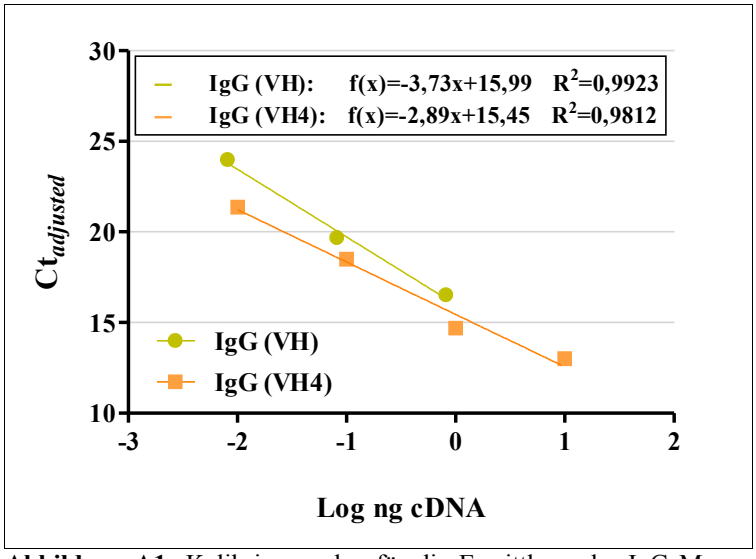



**Abbildung A1:** Kalibriergeraden für die Ermittlung der IgG Menge in den cDNA-Proben.

**Tabelle A2:** Berechnung der erforderlichen Anzahl von PCR Zyklen mit der programmierten OpenOffice.org Calc Tabelle. Dargestellt ist die Kalkulation bei steigender Anzahl von PCR Zyklen und die davon abhängige Detektierbarkeit mittels Gelelektrophorese bzw. Überschreitung der maximalen Anzahl DNA-Moleküle (mit Schrift in Fett hervorgehoben). Für eine Vereinfachung der Darstellung wurde mit nur einer einzigen PCR statt mit der geplanten drei gerechnet (Annahme: IgG Template hat eine Gesamtlänge von ca. 1450 bp).

| Eingabe | | | | | Ausgabe | | | |
|---|---|---|---|---|---|---|---|---|
| Anzahl der PCR Zyklen | Template [ng] | Länge der Template [bp] | Länge des PCR Produkts [bp] | PCR Ansatz-volumen [µL] | PCR Produkt & Template [Moleküle] | PCR End-konzentration [ng/µL] | Anzahl Moleküle (< $10^{11}$) | Auf dem Gel (bei Auftragung von 5 µL) |
| **7** | 1,0 | 1450 | 465 | 50 | 7,92E+10 | 0,83 | **ok** | **undetektierbar** |
| **8** | 1,0 | 1450 | 465 | 50 | 1,59E+11 | 1,66 | **zu viele** | **undetektierbar** |
| **9** | 1,0 | 1450 | 465 | 50 | 3,19E+11 | 3,30 | **zu viele** | **detektierbar** |

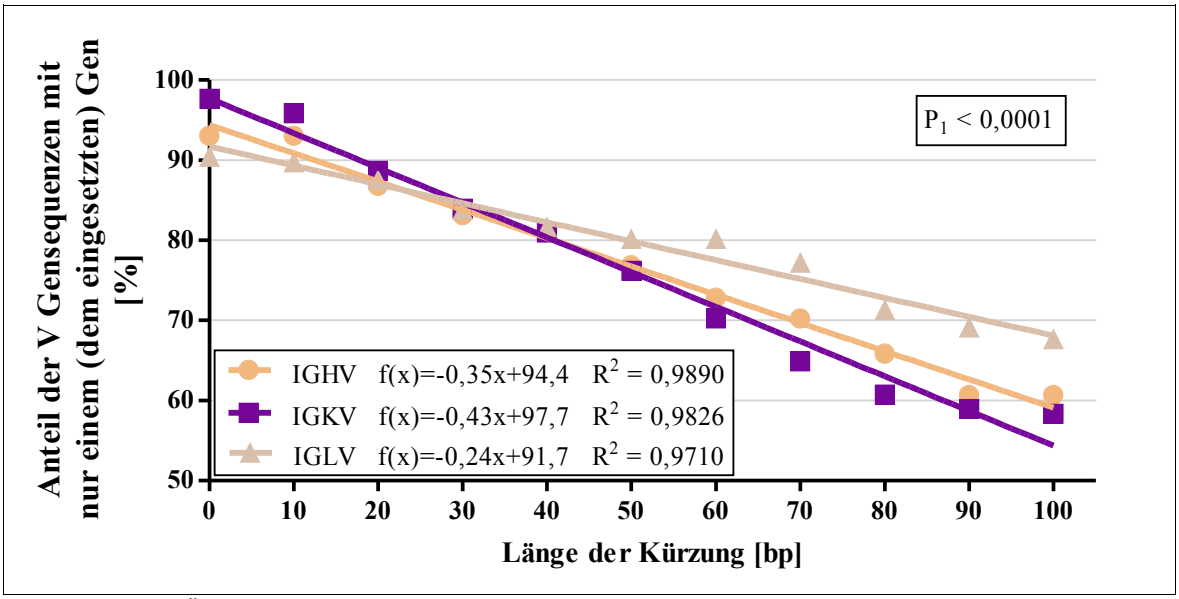



**Abbildung A2:** Änderung vom Anteil der V Keimbahngensequenzen mit nur einem (dem eingesetzten) Gen bei zunehmender Länge der Kürzung. Dargestellt sind auch die Geradengleichungen sowie die Regressionskoeffizienten (unten links) und die Wahrscheinlichkeit, dass die Steigungen identisch sind (oben rechts).

```
> humIGHV201 291 bp
CAGGTGCAGCTACAGCAGTGGGGCGCAGGACTGTTGAAGCCTTCGGAGACCCTGTCCCTCACCTGC
GCTGTCTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGG
CTGGAGTGGATTGGGGAAATCAATCATAGTGGAAGCACCAACTACAACCCGTCCCTCAAGAGTCGA
GTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGCTCTGTGACCGCCGCG
GACACGGCTGTGTATTACTGTGCGAGA
```

**Abbildung A3:** Sequenz des humIGHV201 mit dargestellten Nukleotid-Homopolymeren. Längen der Homopolymere: gelbgrün für 4 bp und hellgrün für 5 bp.

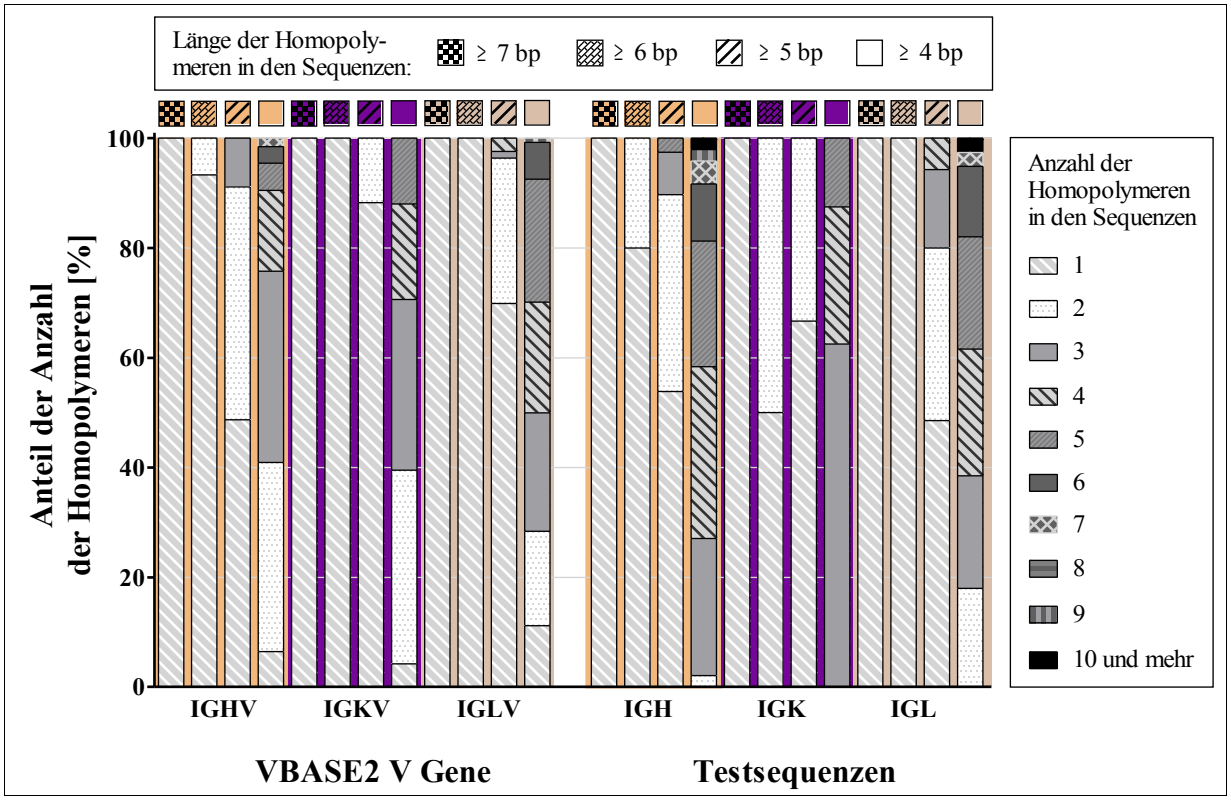



**Abbildung A4:** Verteilung der Anzahl der Homopolymerregionen, aufgeteilt nach Homopolymerlänge.

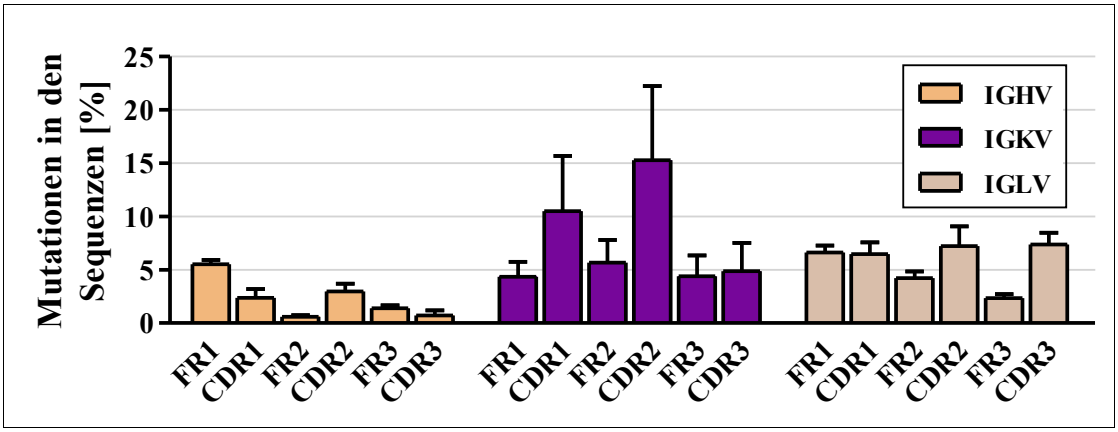



**Abbildung A5:** Histogramm mit den Mutationsanteilen in den V Genen der scFv-Testsequenzen.

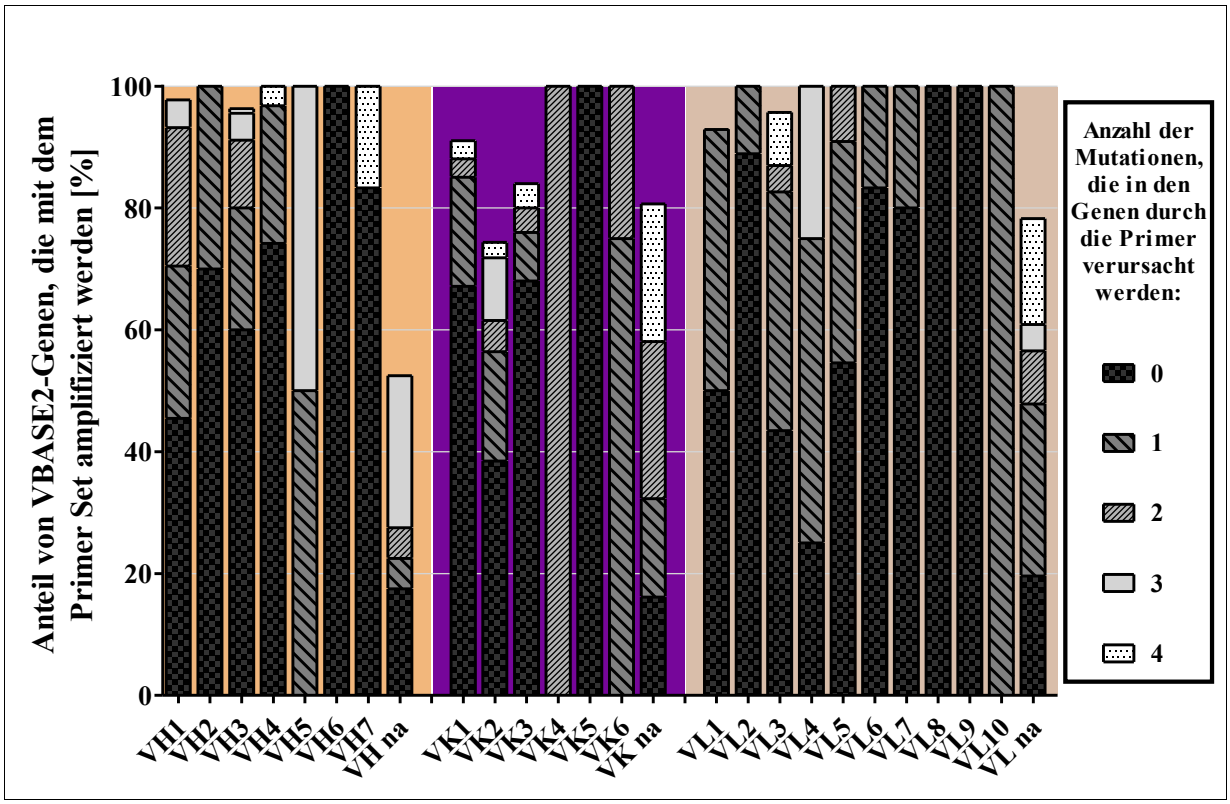



**Abbildung A6:** Veranschaulichung der *Coverage* der VBASE2-Gene mit dem am MPIMG eingesetzten Primer Set.

**Tabelle A3:** Zusammenfassung der Analyse des am MPIMG eingesetzten Primer Sets: Anzahl und Anteil der amplifizierten V Gene. In Kursiv sind die Anteile, bezogen auf der Anzahl aller VBASE2 V Gene.

| | | Gene [Anzahl *(Anteil)*] | Amplifizierte Gene [Anzahl (Anteil)] | |
|---|---|---|---|---|
| | | | Ursprungliches Primer Set | Erweitertes Primer Set |
| | **V Gene in der VBASE2** | 576 | 513 (89,1 %) | 526 (91,3 %) |
| **Familien-zuordnung** | **Gene, die einer Genfamilie zugeordnet sind** | 459 *(79,7 %)* | 431 (93,9 %) | 441 (96,0 %) |
| | **Gene, die keiner Genfamilie zugeordnet sind** | 117 *(20,3 %)* | 82 (70,1 %) | 85 (72,6 %) |
| **Funktiona-lität** | **Funktionelle und möglicherweise funktionelle Gene** | 321 *(55,7 %)* | 302 (94,1 %) | 314 (97,8 %*) |
| | **Pseudogene und Orphans** | 255 *(44,3 %)* | 211 (82,7 %) | 212 (83,1 %) |

* 100 % der funktionellen V Gene (VBASE2 Klasse 1)

**Tabelle A4:** Anzahl der Zielsequenzen sowie der unterschiedlichen Zielsequenzen (*Uniques*) in der erstellten *next*IGbase Datenbank.

| Gesundheits-zustand | Ig Isotyp | Primer | Zielsequenzen | Unterschiedliche Zielsequenzen (Uniques) | |
|---|---|---|---|---|---|
| | | | | 454 | VBASE2 |
| ND | IgD | VH4(DP63)-Fw | 767 | 565 | 550 |
| | | VH4-Fw | 729 | 543 | 533 |
| | IgG | VH4(DP63)-Fw | 1370 | 928 | 843 |
| | | VH4-Fw | 1169 | 780 | 740 |
| RA | IgD | VH4(DP63)-Fw | 1099 | 754 | 737 |
| | | VH4-Fw | 895 | 683 | 664 |
| | IgG | VH4(DP63)-Fw | 2107 | 1238 | 1122 |
| | | VH4-Fw | 1419 | 966 | 881 |

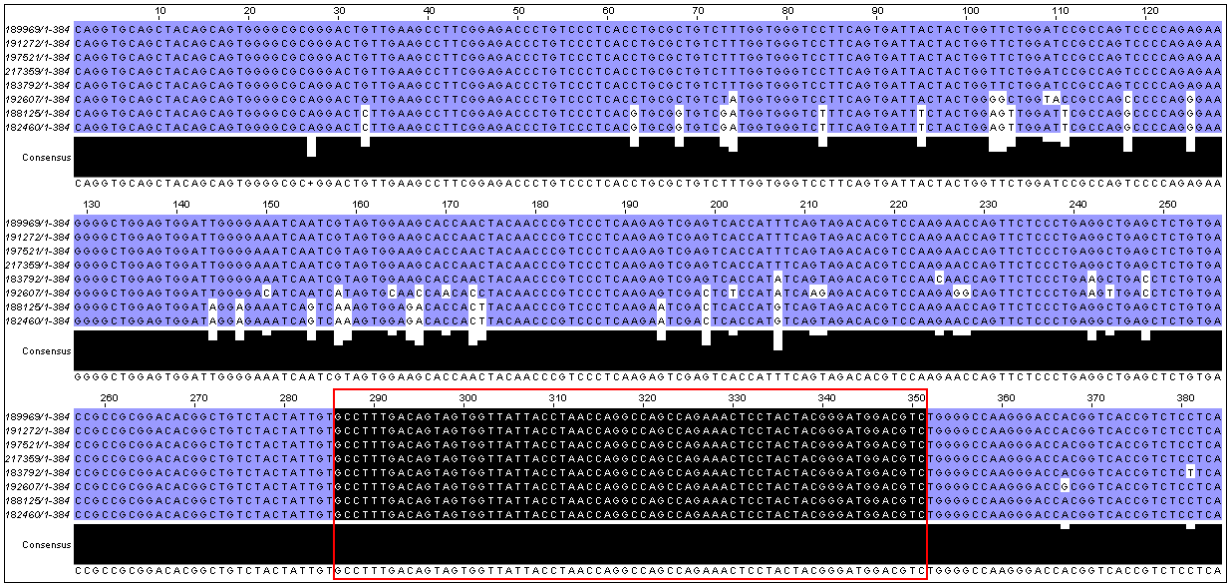



**Abbildung A7:** Screenshot des mit ClustalW2 erstellten Alignments von acht VDJ Sequenzen mit identischer CDR3. Das Farbschema BLOSUM62 Score wurde gewählt. Die oberen vier (189969, 191272, 197521 und 217359) sowie die unteren zwei (182460 und 188125) Nukleotidsequenzen sind jeweils identisch. Der rot umrandete Bereich mit den markierten Nukleotiden entspricht der CDR3 Region.

**Alignment for D segment (218698)**

```
218698                 score     GGCACACTCGCTGGCTGGCCGAACCGCAACCGCTACTGGTTCGAC
X97051 IGHD6-19*01      50.2     ..GTATAG.AG.......TAC
J00233 IGHD2-8*02       50.0     _____________________A.G.TATTG.......GGTGTATGCTATACC
X13972 IGHD1-14*01      49.6     ___________________..TAT.....G....A.
```

**Abbildung A8:** D-D Fusion von drei D Gensegmenten am Beispiel der Sequenz 218698: Screenshot des mit „*DNAPLOT Query*“ erstellten Alignments des D Segments.

## Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Diplomarbeit selbständig verfasst und die genutzten Quellen und Hilfsmittel vollständig angegeben zu haben.

Braunschweig, den 17.11.2009

_____

Svetlana Mollova