Statistical analysis of sequence populations in virology and immunology

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für

Biologie

an der

Universität Duisburg-Essen

vorgelegt von

Bettina Budeus

aus

Hagen

March 26, 2016

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden am Institut für Bioinformatik der Universität Duisburg-Essen oder an einer anderen gleichwertigen Einrichtung durchgeführt.

1.Gutachter: Prof. Dr. Daniel Hoffmann

2.Gutachter: Prof. Dr. med. Stefan Ross

3.Gutachter: Prof. Dr. Matthias Gunzer

Vorsitzender des Prüfungsausschusses: Prof. Dr. Markus Kaiser

Tag der mündlichen Prüfung: 07.Oktober 2016

Danksagung

Zunächst einmal möchte ich mich bei Prof. Dr. Daniel Hoffmann bedanken, der mir überhaupt ermöglich hat diese Arbeit in seiner Arbeitsgruppe schreiben zu können und vor allem ein Thema hatte, dass eigentlich viele Themen war. Außerdem gilt mein Dank Prof. Dr. Dominik Heider, ohne den ich wohl gar nicht von der Möglichkeit diese Arbeit zu schreiben erfahren hätte und der mir in vielen wissenschaftlichen Themen immer zur Seite stand.

Diese Arbeit hätte ohne die zahlreichen Kooperationspartner nicht gefertigt werden können. Dabei sind vor allem Dr. Maren Lipskoch, Dr. Marc Seifert und Dr. Stefanie Schweigele de Reynoso zu nennen, welche die ersten Daten und zahlreiche Ideen für zwei der Projekte lieferten. Fehlen dürfen in dieser Liste aber auch folgende Personen nicht, die alle auf die eine oder andere Art und Weise bei der Erstellung dieser Arbeit geholfen haben: Dr. Christoph Wilms, Dr. Reda Rawi und Dr. Anja Lange aus der Arbeitsgruppe Bioinformatik in Essen, Prof. Dr. Ralf Küppers und Martina Przekopowitz von der Tumorforschung in Essen, Dr. Anke Kraft von der Medizinische Hochschule Hannover, Dr. Andreas Walker und Prof. Dr. Jörg Timm aus der Virologie in Düsseldorf, Dr. Helenie Kefalakes, Dr. Jens Verheyen und Prof. Dr. Mengji Lu aus der Virologie in Essen und Dr. Rongge Yang und Dr. Yan Wang aus der Virologie in Wuhan in China.

Außerdem danken möchte ich meiner Biologielehrerin Silke Hinz, die uns Schülern in der Oberstufe so viel von dem was sie an der Uni gelernt hatte beigebracht hat, das das erste Semester im Biologiestudium relativ langweilig war.

Summary

In this thesis I have examined various topics regarding the relationship between viruses and the human immune system. I expanded and refined a tool (which can now be found as R-package `SeqFeatR` on C-RAN) for the analysis of sequence data and features of this sequences like HLA type or tropism (see chapter 4) and checked with this tool if there are differences between some multiple correction approaches for sequence data, and how Bayesian inference could be used in this context (see chapter 5). It could be shown that Bayesian inference is superior to the frequentistic methods for this kind of problem, because multiple correction approaches ignore the fact that different positions in a sequence alignment may be connected in the protein product of this sequence and are therefor not independent.

Furthermore, I have examined sequences from HCV with a form of bootstrap algorithm to find sequence areas which can be used in unknown transmission cases in court. Two areas were found, one in the hypervariable region and the other at the end of the non-structural protein NS5B (see chapter 9).

Proteasomal cleavage of alien amino acid sequences inside human cells leads to a presentation of fragments of these sequences on the surface of the cell as epitopes. To present such a fragment, not only must it bind to the MHC, but also needs to be in the correct length to be presented. Therefore viral evolution should favor those viruses, which cannot be cut into presentable epitopes. With epitope data from IEDB and predicted viral sequences which bind the MHC, I searched for amino acids inside the flanking regions around the epitope that may indicate a possible escape mutation against the proteasomal cleavage processes. Fourteen such amino acids and positions were found (see chapter 7).

I created a model of HBV reverse transcriptase to check if mutations in certain positions could influence binding with the nucleotide analogue reverse transcriptase inhibitor Tenofovir. Mutations which were inside the binding pocket for Tenofovir showed, in an experimental design by the group of Mengji Lu, a decreased affinity towards the drug (see chapter 10).

Together with Ralf Küppers group I examined NGS from different types of B cells to search for almost identical sequences between those. We found similar to identical sequences from two, three and even four kinds of cells in the blood samples of both donors (see chapter 6).

Zusammenfassung

In dieser Dissertation bearbeitete ich verschiedene Themen aus dem Bereich der humanpatho-genen Viren und des menschlichen Immunsystems. Zu diesem Zweck entwarf ich ein Programm (welches auf dem R-Archiv C-RAN unter dem Namen SeqFeatR zu finden ist) mit dem sich der Zusammenhang zwischen Sequenzdaten und spezifischen Eigenschaften, wie etwa HLA Typ oder Tropismus, analysieren läßt (s.h Kapitel 4). Mit diesem Programm untersuchte ich ob ein Unterschied zwischen den Verfahren zur Korrektur von Alphafehler-Kumulierung bei Sequenzdaten besteht und in welchem Maße die Verfahren der Bayesschen Statistik besser für diese geeignet sind (s.h. Kapitel 5). Dabei stellte sich heraus, dass letztere für diese Klasse von Problemen eher verwendet werden sollten, da Alphafehler-Kumulierungskorrekturen möglichen Abhängigkeite zwischen verschiedenen Sequen-zpositionen, welche sich unter Umständen erst im fertigen Protein offenbaren, ignorieren.

Weiterhin untersuchte ich HCV Sequenzen mittels einer Variante des Bootstrap-Algorithmus um jene Sequenz-Bereiche zu finden, die im Falle von ungeklärten Übertragungswegen zur Identifizierung dieser genutzt werden können. Dabei stellten sich zwei Bereiche als besonders geeignet heraus: Die hypervariable Region sowie ein Bereich am Ende des Nicht-Struktur Protein NS5B (s.h. Kapitel 9).

Die Spaltung von fremden Aminosäuresequenzen innerhalb von menschlichen Zellen durch das Proteasom kann zu einer Präsentation dieser Fragmente auf der Zelloberfläche als Epitope führen. Um solche Fragmente präsentieren zu können, müssen diese nicht nur an das spezifische MHC Molekül binden, sondern auch eine optimale Länge besitzen. Daher sollte der evolutionäre Prozess solche Viren fördern, deren Sequenzen sich nicht in entsprechende Stücke zerteilen lassen. Durch eine Kombination von Epitopdaten aus der IEDB und vorhergesagten viralen Sequenzen, welche sicher an MHC Moleküle binden, untersuchte ich, ob innerhalb der flankierenden Regionen um das jeweilige Epitop Sequenzpositionen existieren, welche auf eine Mutation hinweisen, die den Schnittmechanismus der Zelle verhindert. Ich fand vierzehn Aminosäuren und Positionen, die einen solchen Zusammenhang besitzen können (s.h. Kapitel 7).

Um heraus zu finden ob es in der reversen Transkriptase von HBV Positionen gibt, welche die Bindung mit dem nukleotidischen Reverse-Transkriptase-Inhibitor Tenofovir beeinflussen, erstellte ich ein Modell dieses Enzyms. Mutationen, die innerhalb der Bindetasche für Tenofovir lagen, führten in einer Versuchsreihe von der Gruppe von

Mengji Lu zu einer verringerten Affinität zwischen Enzym und Medikament (s.h. Kapitel 10).

Zusammen mit der Gruppe von Ralf Küppers untersuchte ich Hoch-Durchsatz-Sequenzdaten von verschiedenen Arten von B Zellen um ähnliche Sequenzen zu finden. Wir fanden ähnliche und sogar identische Sequenzen zwischen zwei, drei und sogar allen vier Arten von Zellen jeweils innerhalb der Blutproben jedes der beiden Spender (s.h Kapitel 6).

Table of contents

Glossary

IC_{50}	Half maximal inhibitory concentration. 67, 102, 108, 110
a priori	Bayesian inference: the probability before seeing the data. 65, 66
antigen	Any substance which provokes an adaptive immune response. 11–15
B cell	Cells with a B cell receptor on the cell surface. In mammals B cells are formed in the **b**one marrow. 3, 12, 13, 16, 81, 82, 102
CD	Cluster of differentiation. Protocol used for the identification of cell surface molecules providing targets for immunophenotyping of cells. 8, 12
epitope	The part of an antigen that is recognized by the immune system. iv, 13, 16
FASTA	File format for sequences with one header line for every sequence. 129, A-20
poly-N	Repetitious occurrence of the same nucleotide (length > 2). 36
T cell	Cells with a T cell receptor on the cell surface. Mature in the **t**hymus (although some also mature in the tonsils). 12, 13, 15, 16, 102, 104, 110, 113, 116

Abbreviations

bp	Base pairs. 36, 130, 131
kb	Kilo base pair, equal to 1,000 nucleotides. 100
nM	Nanomolar. 102, 108, 110, 111
AIDS	Acquired immune deficiency syndrome. 8
ANN	Artificial neural network. 67
ANOVA	Analysis of variance. 74
AUC	Area under the curve. 73, 75

BCR B cell receptor. 81
BH Benjamini–Hochberg procedure. 71, 74, 75
BY Benjamini–Hochberg–Yekutieli procedure. 71, 74, 75

C-RAN The Comprehensive R Archive Network. iv, vi, 38, 61
CCR5 Chemokine receptor type 5/ CD 195. 8
CTL Cytotoxic T cell. 13, 101, 114
CXCR4 Chemokine receptor type 4/ CD 184. 8

DNA Deoxyribonucleic acid. 6–9, 12, 36, 103, 106, 151, A-20

ER Endoplasmic reticulum. 15, 101, 115
ERAAP ER aminopeptidase associated with antigen processing (mice). 101, 115
ERAP ER aminopeptidase associated with antigen processing (human). 101, 115

FDR False discovery rate. 64, 65, 71, 74, 75, 110, A-28
FN False negative. 62
FP False positive. 62
FWER Family-wise error rate. 64

GC Germinal center. 81, 82
GWAS Genome-wide association studies. 66

HBV Hepatitis B virus. iv, 7–10, 15, 16, 123, 139, 140, 151
HCV Hepatitis C virus. iv, 3, 7–10, 15, 16, 66, 67, 99–104, 106–116, 126–132, 151
HIV Human immunodeficiency virus. 7–10, 15, 16, 63, 101, 115, 116, 122, 127, 128, 151
HLA Human leukocyte antigen. iv, vi, 13, 66, 67, 71, 72, 74, 75, 102, 107, 110, 112, 114–116, 151, A-27
HVR Hypervariable region. 130, 132

IEDB Immune Epitope Database and Analysis Resource. iv, vi, 67, 74, 75, 102, 104, 107, 110, 113
IFN Interferon. 11, 14–16, 114
IRAP Insulin responsive aminopeptidase. 101

ISG	Interferon-stimulated genes. 11
MCMC	Markov chain Monte Carlo. 75
MHC	Major histocompatibility complex. iv, vi, 13–16, 99, 101, 102, 107, 110, 113–115, 150, A-26
MLE	Maximum likelihood estimate. 66
NGS	Next generation sequencing. v, 36
NK	Natural killer cells. 11
NS	Non-structural protein. iv, vi, 10, 100, 103, 106, 114, 130–132
PCR	Polymerase chain reaction. 130, 132, 151
RNA	Ribonucleic acid. 6–8, 10, 11, 15, 100, 103, 106, 107, 127
ROC	Receiver operating characteristic. 73–75
SMRT	Single molecule real time sequencing. 151
SNP	Single Nucleotide Polymorphism. 66
TAP	Transporter associated with antigen. 13, 14, 101
TDF	Tenofovir disoproxil. 139, 140
TGS	Third generation sequencing. 151
TMV	Tobacco mosaic virus. 6
TN	True negative. 62
TP	True positive. 62

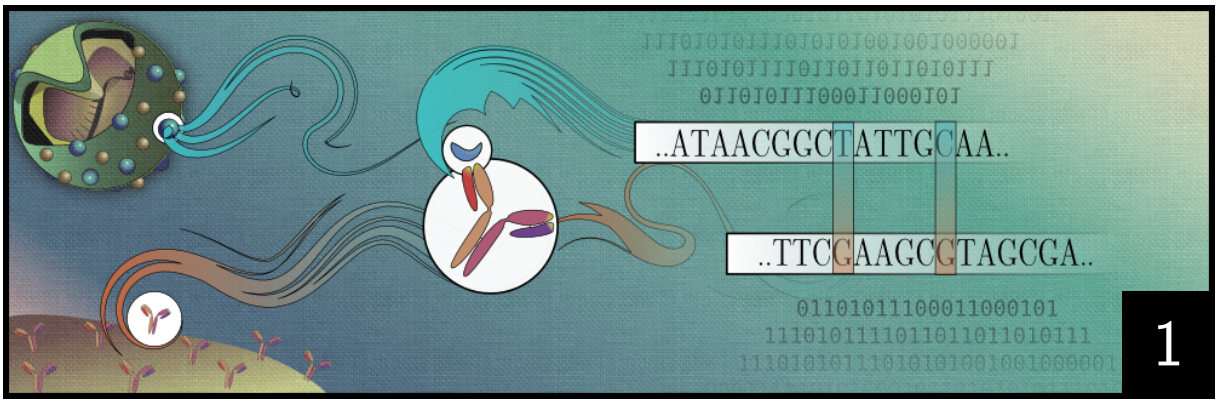
Reference genomes

H77	Hepatitis C virus strain H77 pCV-H77C polyprotein gene - used as reference genome. 102–104, 107, 110, 114, 128, 130

PART

I

INTRODUCTION AND GLOBALLY USED METHODS



Motivation

You have to learn the rules of the game.

ALBERT EINSTEIN

Statistical sequence analysis is a common task in modern biology and health care. As L.L. Larison Cudmore pointed out in 1977, "All living things need their instruction manual (even non-living things like viruses) and that is all they need, carried in one very small suitcase."[1]. Sequences are the foundation of life itself. Insights can be gained through knowledge of the secrets of these sequences. Even if the given sequences and questions asked are diverse, the computer scientist senses their similarity. On the sequence level, a question about the diversification of B cells is roughly the same as the analysis of transmission chains in Hepatitis C Virus (HCV), although the obvious object of interest, B cells and HCV, are quite different.

In this work, several sequence based problems were tackled with different bioinformatic methods, such as sequence alignment, phylogenetics, or homology modeling. A summary of the different organisms and methods used can be found in table Table 1.1.

Sequences are essential for many different kinds of analyses, which range from the interaction between pathogens and their hosts up to understanding relationships between organisms. The focus in this thesis was placed on the former: severe human pathogenic viruses and the human immune system.

Table 1.1: Different methods used in this work according to treated topics.

	statistics		**sequence analysis**			
topic	significance test[1]	correction of multiple testing errors	sequence alignment	phylo-genetics	epitope prediction	homology modeling[2]
B cell	✓		✓	✓		
HBV			✓			✓
HCV	✓	✓	✓		✓	
HIV	✓	✓	✓	✓	✓	

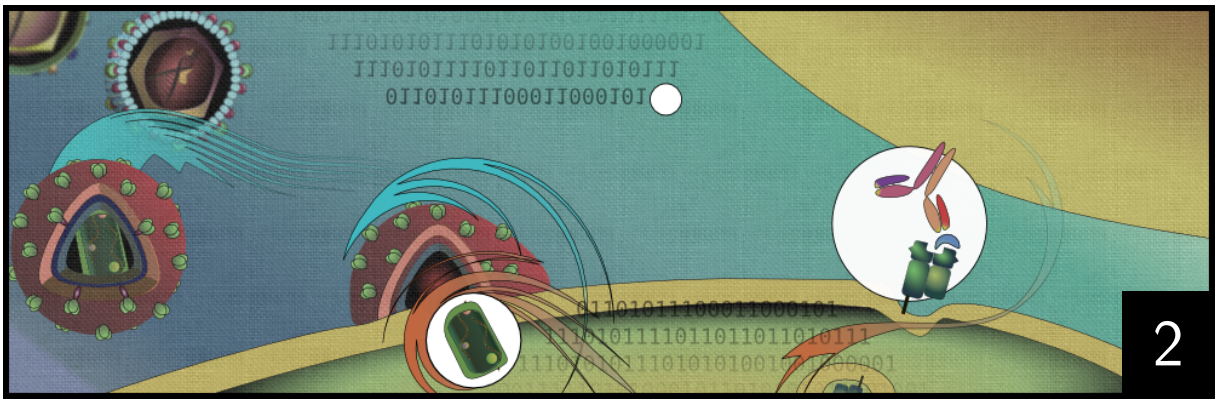
[1] Fisher's exact test, t-test, wilcoxon test, etc.
[2] homology modeling combines sequence analysis with structure analysis

1.1 Explanatory note

Although all projects in this thesis are based on sequences, each of them has a unique combination of basic and advanced methods. Because of this, only the most basic methods and backgrounds are described ahead of the projects in this introduction, and each project has additional, further information about the exact methods for that project. This enables the advanced reader to skip the introduction and basic methods section, if he or she already knows what sequence alignments, phylogenetics or homology modeling are, and read only the desired chapter of interest. Inexperienced readers should read the introduction first to understand the more complex context of the following chapters.

References

[1] L. Cudmore. *The Center of Life: A Natural History of the Cell*. Quadrangle Paperbacks. Quadrangle/New York Times Book Company, 1978. ISBN: 9780812962932. URL: https://books.google.de/books?id=IrMuAAAAIAAJ (cit. on p. 3).



Viruses and immune system - an overview

Be careful about reading health books. You may die of a misprint.

MARK TWAIN

2.1 Viruses

A virus is a small agent which replicates only in living cells of other organisms. The first found virus was the tobacco mosaic virus (TMV) in 1892 discovered by Dmitri Ivanovsky, who found out that filtrated fluids from infected plants could infect healthy ones [82]. Viruses range in size between 20 nm and 300 nm, some have a total length up to 1400 nm [23] and are (mostly) classified by the type of genetic information molecule they encapsulate. DNA viruses use the DNA-dependent DNA polymerase from the host cell and can be either single-stranded or double-stranded. RNA viruses may utilize RNA polymerase instead, if they have negative-sense RNA [93, 99]. Positive-sense RNA viruses can be immediately translated by the host cell [54, 91]. A special form of RNA viruses are Retroviruses, which use their own reverse transcriptase enzymes to generate DNA that is included into the host genome [131, 143]. Besides the genetic material all viruses consist of at least one protein to form a kind of hull as protection called a capsid [28]. Some viruses also have an envelope containing lipids from the host cell membrane [67, 127]. Adjacent to these proteins are often proteins which manage the cell entry into the host cell, and inside the hull proteins are often found that stabilize the genetic material. All in all viruses can have many different shapes [114].

RNA viruses have a much higher mutation rate than DNA viruses, because of the differences in proofreading between DNA polymerases and RNA polymerases. RNA polymerases usually lack such proofreading functionality and many mutations are built into the copies of the viral sequences [29, 30, 96]. DNA viruses mutate more slowly, whereas the reverse transcribing viruses, which use reverse transcription mechanisms without proofreading to insert their internal single strand RNA or double strand DNA molecule into the hosts cell, have a high mutation rate [32, 110]. This high mutation rate results in a higher evasion rate against the immune system and - in the case of pathogens - drugs and vaccines [34, 36, 61]. A consequence of the lack of proofreading are viruses which differ only in a few nucleotides from one another [55, 126]. This difference can have a high impact on the virulence of the particular viral particle, either in a good or in a catastrophic way [25] and combined with the selection pressure from the hosts immune system, a viral quasispecies is created, which 'is a well-defined distribution of mutants that is generated by a mutation-selection process' (see [94]). New techniques for sequencing can generate an overview of such quasispecies for the first time [7, 16, 20], to take a deeper insight into viral evolution and immune evasion mechanics.

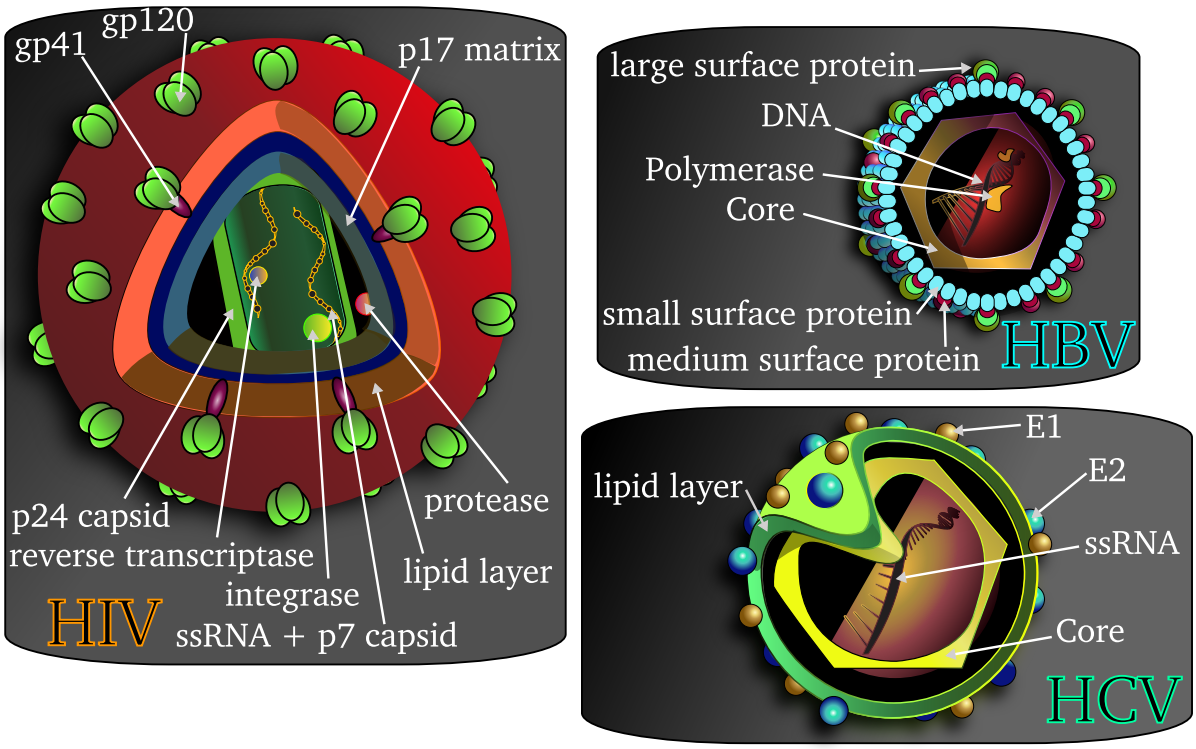


Figure 2.1: Model structure of HIV, HBV and HCV

Mya Breitbart and Forest Rohwer suggest that there are 10^{31} viruses on earth [14]. As of writing researchers know about 5,000 viruses in detail [28] (1062 RNA viruses in the RNA virus database [8]). The 'viralzone' currently (August, 2015) contains 129 human pathogenic viruses [58], but since better detection methods are advancing this number is increasing. Of these 129 viruses only a few are high risk and possibly deadly for the infected person and some are still an increasing danger in some countries. Apart from high risk and deadly viruses such as the Ebola and Pox viruses, the biggest threat posed is that of viruses with a long incubation period, even if they are only transmitted via contaminated blood [12, 24, 86]. Examples of such viruses are HIV, HBV and HCV, which are often called 'the big three', because they are widespread (and still spreading), often persistent, and there is no known vaccine, except for HBV [59].

2.1.1 HIV

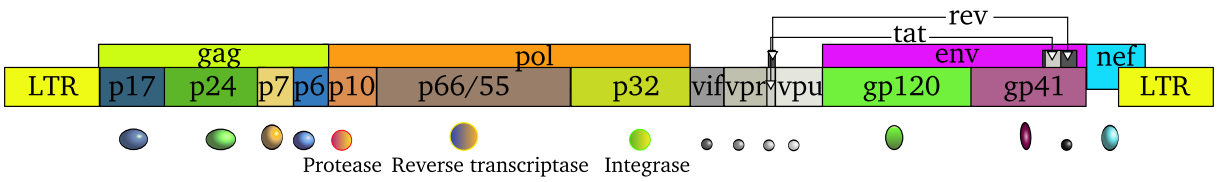


Figure 2.2: Human Immunodeficiency Virus I genome and proteins

The human immunodeficiency virus (HIV), which causes the acquired immunodeficiency syndrome (AIDS), was first clinically observed in 1981 in the United States and first described in 1983 [120]. Today about 34 million people worldwide are infected with HIV-1 [139], which is one of two forms of the virus. HIV-2 is mostly found in West Africa and not necessarily fatal. HIV-1 is a roughly spherical retrovirus with a diameter of 120 nm [33, 41] and is composed of two copies of positive single-stranded RNA. This RNA encodes normally nine genes (*gag, pol, env, tat, rev, nef, vif, vpr, vpu*), which encode 19 proteins (see Figure 2.2). Three of the genes, *gag, pol*, and *env*, contain information enabling the creation of the structural proteins for new virus particles [38] (see Figure 2.1). Especially *env*, more precisely the two cleavage products *gp41* and *gp120*, are needed to interact with the host cell and establish a fusion of both virus particle and cell [142]. HIV-1 can infect a variety of immune cells. It connects with it's virion envelope glycoproteins (*gp120*), the $CD4^+$ receptor [18, 145], and, depending on the tropism of the virus particle, one of two co-receptors: CCR5 or CXCR4 [9, 45, 102]. Due to the fact that HIV-1 integrates a reverse transcribed copy of itself in the host DNA, a cure for patients once infected

is difficult and not yet discovered. There exist drugs that inhibit the replication and spreading of the virus in different ways, but since the process of reverse transcription is error-prone the resulting random mutations may cause drug resistance [10, 110]. It is therefore vital to analyze and identify important regions in the viral genome to better understand the interaction of virus and host cell.

2.1.2 HBV

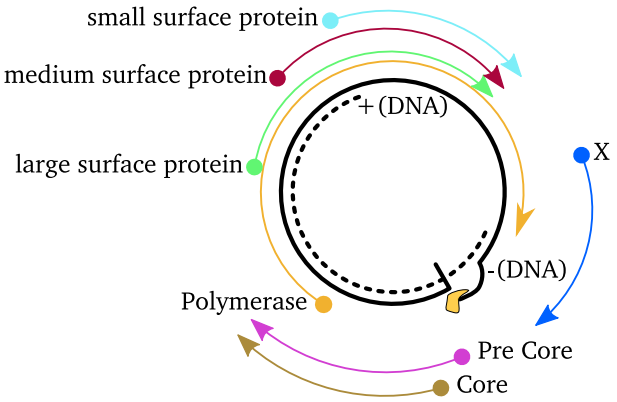


Figure 2.3: Hepatitis B Virus genome

In contrast to HIV, a vaccine against the hepatitis B virus (HBV) exists. HBV has been known since 1885 [81], but there are still many new infections, often in combination with HIV [70, 101] and HCV [50], even in the Western world were the vaccine is easily accessible, due to drug use or unprotected sexual contact [4, 75]. Available drugs can only diminish the infection and so far cannot cure it if chronic [49, 98], which is the case for 350 million people worldwide [76]. HBV is 30 nm - 42 nm in diameter and one of the smallest enveloped animal viruses known [26]. The envelope consists of three different surface proteins (see Figure 2.1). Although HBV has a DNA molecule inside the capsid, it also uses a reverse transcriptase enzyme like a retrovirus and is therefore error-prone [5]. The DNA molecule encodes only four genes (C, X, P, and S), is not fully double-stranded, and one end of the longer strand is linked with a viral polymerase [57, 72, 128] (see 2.3). In replication, the viral genome is extended to fully double-stranded and inserted into the cell nucleus [111]. Then it is transcribed by host enzymes and besides translation of the viral proteins is reverse transcribed back into the viral DNA [129]. Because of this special replication, the severe symptoms in chronically infected patients, and the still relatively high prevalence in certain areas, it is a important field of research.

2.1.3 HCV

Hepatitis C is a common disease in humans with around 170 million people worldwide infected with the virus, which was found relatively late in the 1970s and isolated in 1989 [22]. Like HIV and HBV, the main transmission mechanisms of HCV is via contaminated blood, drug use or sexual contact [3, 147]. HCV mainly infects the hepatocytes of the liver and produces lots of new virions. The virus may also replicate in peripheral blood mononuclear cells [148]. Up until now there has been no vaccination against HCV, but several antiviral drugs inhibit the replication of the virus [79, 100]. HCV therapy is difficult, because of the high error rate of the virus' RNA-dependent RNA polymerase, which increases the possibility of escape mutations [108] and generates a highly diverse quasispecies inside the patient [84]. Another treatment used for patients with liver damage through HCV is liver transplantation [141]. HCV consists of a positive sense single-stranded RNA genome with a single open reading frame [21, 62], surrounded by the capsid and a lipid envelope from the host cell [90] (see Figure 2.1). The translated product of the genome is then processed further to create 10 active viral proteins (core protein, E1 and E2 and p7; non-structural proteins include NS2, NS3, NS4A, NS4B, NS5A, and NS5B) [46]. This cleavage mechanism is mediated mainly by two viral proteases - NS2 and NS3-4A [35, 53, 103] and inhibition of one or both results in a greatly reduced viral replication [71] (see 2.4). Currently HCV is classified into seven genotypes with subgenotypes depending on sequence similarity. Genotype 1 is found world wide and occurs in a majority of the patients in the Western world [89]. Like HIV, HCV is still an important topic of research and particularly next-generation sequencing promises a deeper insight into the variability of the viral sequences, escape or compensatory mutations.

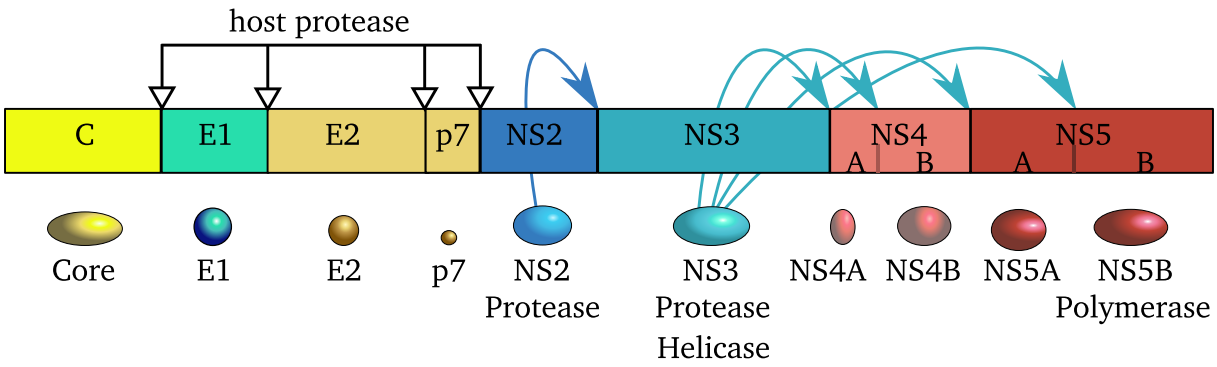


Figure 2.4: Hepatitis C virus genome and proteins

2.2 Immune system

The immune system is a combination of biological processes and structures to prevent or hinder infections by other organisms or viruses. It can be classified by time into innate immune response and adaptive immune response, or by involved structure into humoral immunity and cell-mediated immunity [60, 68].

2.2.1 Innate immune response

The innate immune response is also called non-specific immune system [60]. It is the first defense mechanism against invasion by pathogens and consists of many different structures and functions. The main components are: cytokine production to recruit immune cells, identification and removal of foreign substances, activation of the complement cascade to clear pathogens or mark them for destruction by other cells, acting as a physical barrier, and activation of the adaptive immune system [60]. Involved leukocytic cells are natural killer cells (NK cells), mast cells, eosinophils and basophils, and the phagocytic cells. Mast cells release certain second messengers which cause inflammation and recruiting of other leukocytic cells. Eosinophils and basophils kill or inhibit the growth of pathogens with toxins and respiratory burst [149] whereas all phagocytic cells engulf particles and pathogens [2, 60]. Phagocytic cells are further sub classified as macrophages, neutrophils, and dendritic cells [2, 60]. In contrast to other types of innate immune cells, NK cells destroy infected host cells using a similar mechanism as the cells from the adaptive immune response. NK cells recognize certain proteins with bound self antigens on the surface and compromise the cell, if a certain self pattern is not shown [60]. Antiviral host defense is mediated by type I interferons (IFN). Especially the RNA of a virus is recognized by certain proteins inside the cell, which then activate certain pathways that lead to the production of IFN. IFN induces the expression of hundreds of interferon-stimulated genes (ISG) which products inhibit or degrade viral proteins and viral RNA [1, 6, 113].

2.2.2 Adaptive immune response

In contrast to the innate immune response, the adaptive immune response occurs later after the infection has taken place and is highly specific [60]. This high specificity is a result of hypermutation in a DNA region called antigen receptor gene segments, where the genes are then also rearranged, which is called V(D)J recombination or rearrangement [31, 56, 144] (see Figure 2.5). This recombination is irreversible and all clones of these cells

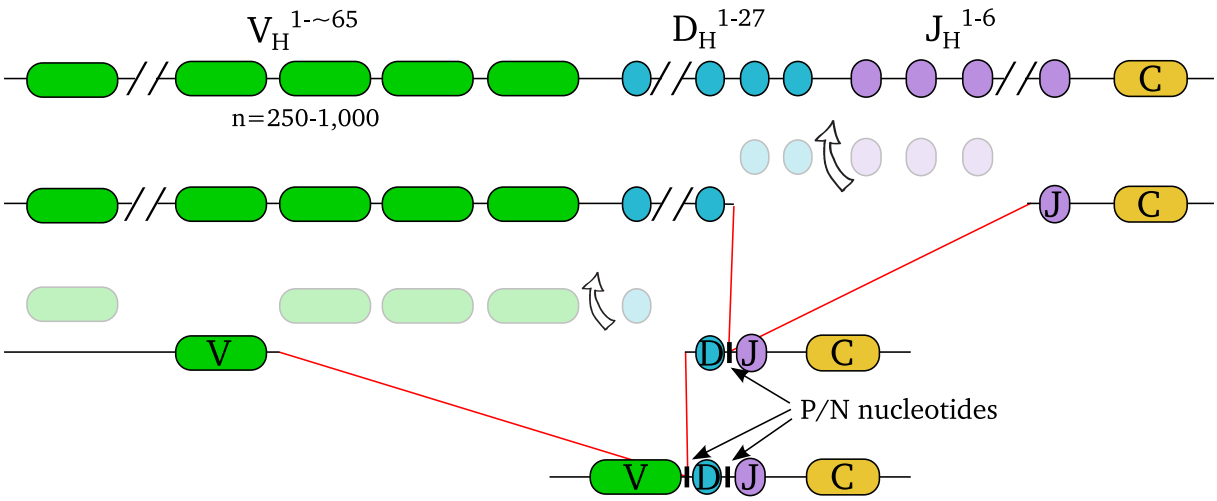


Figure 2.5: Scheme of the VDJ recombination. The V(D)J recombination occurs in developing lymphocytes in early stages of T cell and B cell maturation. Shown here is the recombination of immunoglobulin heavy chains. In this process, variable (V), diversity (D), and joining (J) gene segments are selected and combined randomly. Between V and D, and D and J random nucleotides (palindromic (P) and non-templated (N)) can be inserted and also removed later. This leads to a high diversity in the final rearrangement [42, 80]. The numbers above the first line indicate how many gene segments are available according to [60]. n is the number of possible genes.

will recognize the same antigen [69], which is important for the immunological memory [2]. The major functions of the adaptive immune system are: the recognition of non-self antigens, and the development of a immunological memory in the form of signature antibodies or T cell receptor [60]. The leukocytic cells of the adaptive immune response exist in three different stages: naive cells, which have not encountered their cognate antigen; effector cells, which have encountered their antigen and take part in the acute immune response; and memory cells, which are long-living cells beyond the acute infection [2]. Apart from the T cells, which are one kind of cell-mediated immunity, B cells are also involved in the adaptive immune response [60] as humoral immunity. T cells, which are called "T" because of their maturation region in the thymus, are subdivided according to their function and some proteins on the cells surface like CD4 or CD8 and are either

activated through other T cells (called T helper cells) or binding of the cognate antigen MHC complex of another host cell [60, 74]. B cells are involved in the creation of antibodies in blood plasma and lymph and they can be distinguished from other lymphocytes by the presence of a certain receptor (simply called B cell receptor), which is composed of a membrane-bound antibody [2, 125]. Like T cells, B cells are subdivided according to their function and the type of antibodies they secrete [60]. The critical difference between T cells and B cells is the recognition of their cognate antigen. Whereas B cells recognize native, soluble antigens, T cells can only recognize a bound form with the MHC molecule from another cell [60].

2.2.3 MHC and epitopes

Higher vertebrates have a complex system to eliminate infected cells from the body. The major genes involved in these pathways are called *MHC* (Major histocompatibility complex) which produce certain proteins. These gene products are labeled HLA (Human leukocyte antigen) in humans. In other vertebrates they are called by the same name as the gene. The *MHC* genes can be divided into three categories: MHC-class I, MHC-class II and MHC-class III. MHC-class I and II mediate antigen presentation, MHC-class III genes are involved in other (not necessarily immune) functions and proteins such as cytokines and heat shock proteins [92].

The MHC-I antigen procession and presentation pathway consists of the following steps: in the cytosol, proteins are degraded by the proteasome, some of them at the end of their useful lifetime, around 40% of them directly after synthesis [119]. Most of the peptide fragments generated by the proteasome are further degraded by other cytosolic proteases into single amino acids used for the synthesis of new proteins. Some of the non-degraded peptides are transported into the endoplasmic reticulum (ER) by the membrane spanning transporter TAP. There, the peptides can be degraded by aminopeptidases again [116, 121, 146] or exported back into the cytosol, unless they are able to bind to an empty MHC-I molecule. Once a peptide binds, the MHC-I - peptide complex is transported to the cell surface, where it is presented to cytotoxic T-lymphocytes (CTL cells) as an epitope (see 2.6 - MHC-I pathway). The main difference between the MHC-class I pathway and the MHC-class II pathway is the origin of antigenic proteins, the enzymes responsible for peptide generation, and the location of loading the peptide on the MHC molecule. MHC-class II only present proteins from the endosomes or lysosomes, degraded by proteases inside those (see 2.6 - MHC-II pathway). To summarize, MHC-class I presents antigens

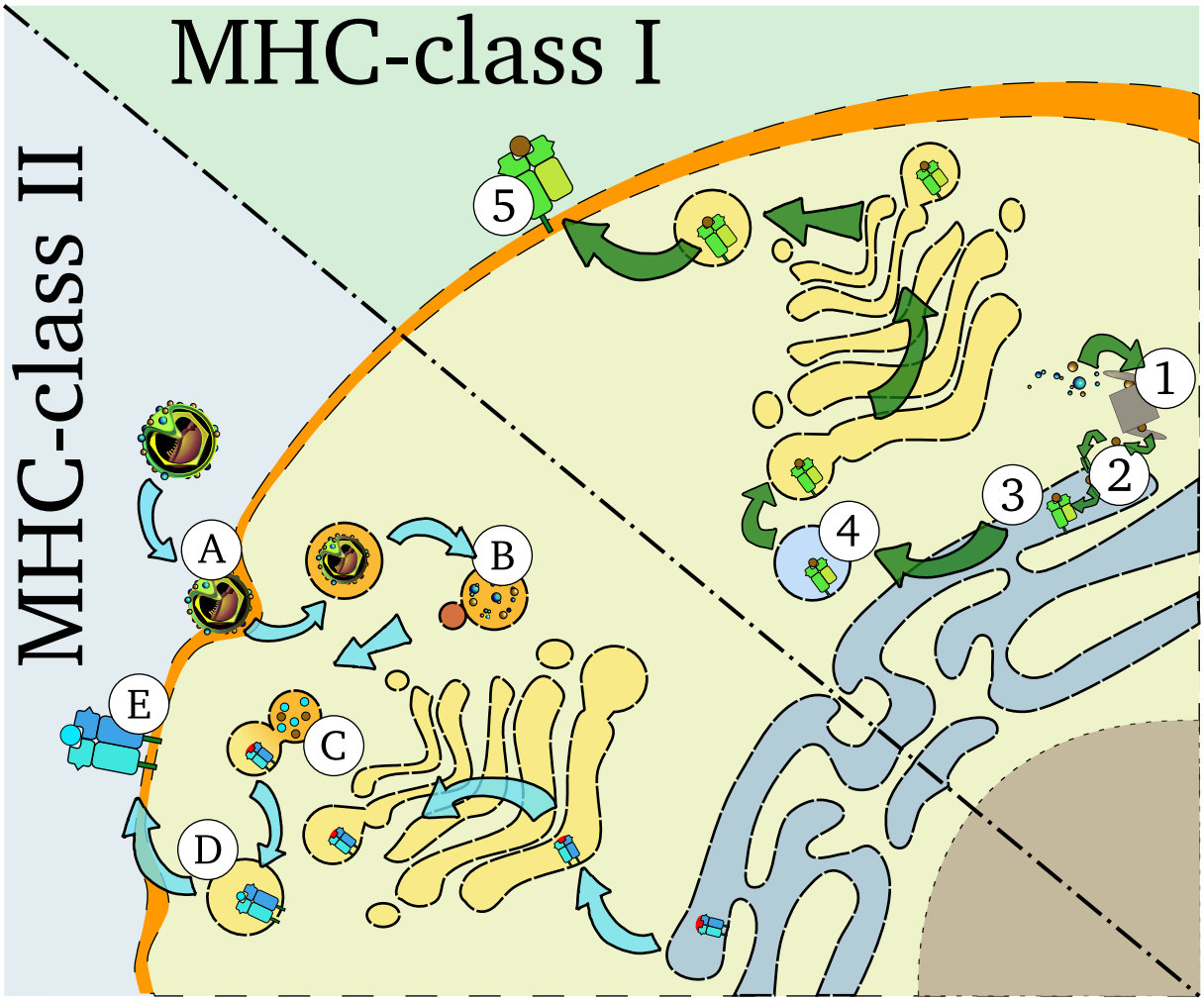


Figure 2.6: Overview of the MHC pathway system: MHC-I pathway: 1. The proteasome cuts viral (and other) particles in shorter peptides; 2. Those peptides are transported inside the ER lumen via TAP; 3. In the ER further modifications are made and the peptide is then loaded onto the MHC-class I molecule; 4. This molecule plus peptide is then transported via Golgi apparatus to the cell surface; 5. The peptide is presented on the surface. **MHC-II pathway:** A. Viral particles are ingested by endosomal compartments; B. Inside those particles and with the aid of lysosomes, proteolytic degradation produces peptides out of the viral proteins; C. The phagolysosome fuse together with Golgi vesicle including inactivated MHC-class II molecules; D. The MHC-class II molecules binds a suitable peptide; E. The peptide is presented on the surface.

which come from proteins produced inside the cell, whereas MHC-class II presents antigens from outside the cells.

The proteasome, a multi-subunit protein complex, is responsible for the majority of protein degradation in the cytosol [47, 136] and is an important part of the MHC-class I pathway [112]. All proteasomes isolated from eukaryotes contained the 20S proteasome, which has three catalytic cores β–1, β–2 and β–5 [48]. It is known that the subunits of 20S change under influence of IFN–γ, and it is assumed that these immunoproteasomes

increase the antigen procession capability of the cell [51]. The proteasome produces peptide fragments with a length of between three and 18 amino acids [17]. The subunits of 20S have different cut preferences, which can be identified through fluorogene activating peptide substrates (FAP): the β–1 subunit cuts after acidic amino acids, β–2 hydrolyzes the peptide bond after basic amino acids and β–5 shows a chymotrypsin-like activity with cut preferences after hydrophobic or aromatic amino acids [97]. This enzyme specificity is much lower with physiological substrates [66]. The proteasome splits inactive protein sequences mostly after basic and hydrophobic side chains [13], but the actual preferences also depend on the surrounding amino acids [95].

The common size of an antigen for a MHC-class I molecule is between 8–10 amino acids [15, 39, 88, 137]. Since only a small fraction of the peptides generated by the proteasome are of the correct size, most of them must be processed further either by peptidases in the cytosol or in the ER [66]. In the ER antigens can bind directly onto the MHC-class I molecule, after correct trimming by ER-aminopeptidases and removing of the precursor [117]. Those parts of the antigen which are recognized by the immune system are called epitopes.

2.2.4 The big three and the immune system

There are big differences between HIV-1, HBV and HCV in the early immune responses. HBV seems to avoid the induction of strong innate immune responses in the beginning of an infection [140], whereas HCV induces a strong response, but is able to avoid the consequences by the interaction of viral proteins with IFNs [11, 37, 130]. Antiviral effects of type I IFNs against HBV haven been shown in transgenic mice, although the strong immune response is missing [85]. IFN induces mechanisms which inhibit the formation of viral capsids and degenerate HBV-RNA and seems to be proteasome dependent [85, 109]. HIV-1 mucosal infections arose from a single virus as is shown in studies in human [64, 115] and rhesus macaques [65] even if they are infected with a huge viral quasispecies. Innate immune response may even support viral replication due to the recruitment of additional susceptible T cells to the site [77]. A down regulation of HBV replication via IFN-γ occurs in the adaptive immune response [52, 135]. In contrast to this, HCV specific T cells appear much later in the infection [132, 133] and IFN-γ inhibits protein synthesis and RNA replication of subgenomic and genomic HCV replicons [40]. HIV-1 has a special status in the adaptive immune response, since it infects particularly those cells, which

then spread the virus through the body [43, 87]. HIV-1 can even activate innate cells, B cells and T cells [27, 44, 124] but does not activate IFN-γ [83].

Immunological memory and protective immunity is also very different between the big three. Recovery from HBV results in lifelong immunity [63], and also lifelong viral particles in the blood [106], whereas HCV-specific antibodies may be lost after 10–20 years after clearance [105]. Studies show that reinfections of HCV result in lower HCV titers and no evidence of liver damage [73, 123]. It is also shown that escape mutations in HBV are mainly mediated by drugs and not through the immune system. Variants of HBV typically remain in low abundance in acute hepatitis [138] and even in chronic hepatitis T cell escape mutants are rare [107]. In contrast to this, HCV shows escape mutations mediated by epitope processing[134], MHC binding[19], T cell receptor stimulation [19] and antibodies[122], which is feasible through a large number of quasispecies when adaptive immune response occurs [104]. HIV-1 is difficult to grab for immunological memory and protective immunity. It almost always leads to a chronic infection [118], however there are some individuals who remain sero-negative after definite exposure to HIV-1, which appears to be unconnected with the host genotype [78]. This is therefore a constant field of research.

References

[1] S. Akira, S. Uematsu, and O. Takeuchi. "Pathogen Recognition and Innate Immunity". In: *Cell* 124.4 (2006), pp. 783–801. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/j.cell.2006.02.015 (cit. on p. 11).

[2] B. Alberts et al. *Molecular Biology of the Cell*. 4th. New York: Garland Science, 2002 (cit. on pp. 11–13).

[3] M. J. Alter. "Epidemiology of hepatitis C". In: *Hepatology* 26.S3 (1997), 62S–65S. ISSN: 1527-3350. DOI: 10.1002/hep.510260711. URL: http://dx.doi.org/10.1002/hep.510260711 (cit. on p. 10).

[4] M. J. Alter. "Epidemiology and Prevention of Hepatitis B". In: *Seminars in Liver Disease* 23.1 (2003), pp. 039–046. DOI: 10.1055/s-2003-37583 (cit. on p. 9).

[5] A. Bartholomeusz and S. Locarnini. "Hepatitis B virus mutations associated with antiviral therapy". In: *Journal of Medical Virology* 78.S1 (2006), S52–S55. ISSN: 1096-9071. DOI: 10.1002/jmv.20608 (cit. on p. 9).

[6] A. Baum, R. Sachidanandam, and A. García-Sastre. "Preference of RIG-I for short viral RNA molecules in infected cells revealed by next-generation sequencing". In: *Proceedings of the National Academy of Sciences* 107.37 (2010), pp. 16303–16308. DOI: 10.1073/pnas.1005077107 (cit. on p. 11).

[7] N. Beerenwinkel and O. Zagordi. "Ultra-deep sequencing for the analysis of viral populations". In: *Current Opinion in Virology* 1.5 (2011), pp. 413–418. ISSN: 1879-6257. DOI: http://dx.doi.org/10.1016/j.coviro.2011.07.008 (cit. on p. 7).

[8] R. Belshaw et al. "The RNA Virus Database". In: *Nucleic Acids Research* 37.suppl 1 (2009), pp. D431–D435. DOI: 10.1093/nar/gkn729 (cit. on p. 8).

[9] E. Berger et al. "A new classification for HIV-1". In: *Nature* 391 (1998), p. 240 (cit. on p. 8).

[10] B. Berkhout, A. T. Das, and N. Beerens. "HIV-1 RNA Editing, Hypermutation, and Error-Prone Reverse Transcription". In: *Science* 292.5514 (2001), p. 7. DOI: 10.1126/science.292.5514.7a (cit. on p. 9).

[11] C. B. Bigger, K. M. Brasky, and R. E. Lanford. "DNA Microarray Analysis of Chimpanzee Liver during Acute Resolving Hepatitis C Virus Infection". In: *Journal of Virology* 75.15 (2001), pp. 7059–7066. DOI: 10.1128/JVI.75.15.7059-7066.2001 (cit. on p. 15).

[12] F. Bihl et al. "Transfusion-transmitted infections." In: *J Transl Med* 5 (2007), p. 25. DOI: 10.1186/1479-5876-5-25 (cit. on p. 8).

[13] B. Boes et al. "Interferon gamma stimulation modulates the proteolytic activity and cleavage site preference of 20S mouse proteasomes." In: *The Journal of Experimental Medicine* 179.3 (1994), pp. 901–909. DOI: 10.1084/jem.179.3.901. eprint: http://jem.rupress.org/content/179/3/901.full.pdf+html. URL: http://jem.rupress.org/content/179/3/901.abstract (cit. on p. 15).

[14] M. Breitbart and F. Rohwer. "Here a virus, there a virus, everywhere the same virus?" In: *Trends in Microbiology* 13.6 (2005), pp. 278–284. ISSN: 0966-842X. DOI: http://dx.doi.org/10.1016/j.tim.2005.04.003 (cit. on p. 8).

[15] S. R. Burrows, J. Rossjohn, and J. McCluskey. "Have we cut ourselves too short in mapping {CTL} epitopes?" In: *Trends in Immunology* 27.1 (2006), pp. 11–16. ISSN: 1471-4906. DOI: http://dx.doi.org/10.1016/j.it.2005.11.001. URL: http://www.sciencedirect.com/science/article/pii/S1471490605002826 (cit. on p. 15).

[16] M. R. Capobianchi, E. Giombini, and G. Rozera. "Next-generation sequencing technology in clinical virology". In: *Clinical Microbiology and Infection* 19.1 (2013), pp. 15–22. ISSN: 1469-0691. DOI: 10.1111/1469-0691.12056. URL: http://dx.doi.org/10.1111/1469-0691.12056 (cit. on p. 7).

[17] P. Cascio et al. "26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide". In: *The EMBO Journal* 20.10 (2001), pp. 2357–2366. DOI: 10.1093/emboj/20.10.2357. URL: http://emboj.embopress.org/content/20/10/2357 (cit. on p. 15).

[18] D. Chan and P. Kim. "HIV entry and its inhibition". In: *Cell* 93 (5) (1998), pp. 681–4 (cit. on p. 8).

[19] K. M. Chang et al. "Immunological significance of cytotoxic T lymphocyte epitope variants in patients chronically infected by the hepatitis C virus." In: *The Journal of Clinical Investigation* 100.9 (Nov. 1997), pp. 2376–2385. DOI: 10.1172/JCI119778 (cit. on p. 16).

[20] S. Chevaliez, C. Rodriguez, and J. Pawlotsky. "New Virologic Tools for Management of Chronic Hepatitis B and C". In: *Gastroenterology* 142.6 (2012). Viral Hepatitis: A Changing Field, 1303–1313.e1. ISSN: 0016-5085. DOI: http://dx.doi.org/10.1053/j.gastro.2012.02.027 (cit. on p. 7).

[21] Q. L. Choo et al. "Genetic organization and diversity of the hepatitis C virus." In: *Proceedings of the National Academy of Sciences* 88.6 (1991), pp. 2451–2455. DOI: 10.1073/pnas.88.6.2451 (cit. on p. 10).

[22] Q. Choo et al. "Isolation of a cDNA derived from a blood-borne non-A, non-B hepatitis genome". In: *Science* 244 (1989), pp. 359–362 (cit. on p. 10).

[23] L. H. Collier, P. Kellam, and J. S. Oxford. *Human virology.* 4th ed. Oxford, England: Oxford University Press, c2011., 2011 (cit. on p. 6).

[24] N. Crofts et al. "Blood-borne virus infections among Australian injecting drug users: Implications for spread of HIV". English. In: *European Journal of Epidemiology* 10.6 (1994), pp. 687–694. ISSN: 0393-2990. DOI: 10.1007/BF01719282 (cit. on p. 8).

[25] S. Crotty, C. E. Cameron, and R. Andino. "RNA virus error catastrophe: Direct molecular test by using ribavirin". In: *Proceedings of the National Academy of Sciences* 98.12 (2001), pp. 6895–6900. DOI: 10.1073/pnas.111085598 (cit. on p. 7).

[26] D. Dane, C. Cameron, and M. Briggs. "Virus-like particles in serum of patients with australia-antigen-associated hepatitis". In: *The Lancet* 295.7649 (1970). Originally published as Volume 1, Issue 7649, pp. 695–698. ISSN: 0140-6736. DOI: http://dx.doi.org/10.1016/S0140-6736(70)90926-8 (cit. on p. 9).

[27] S. G. Deeks et al. "Immune activation set point during early HIV infection predicts subsequent CD4+ T-cell changes independent of viral load". In: *Blood* 104.4 (2004), pp. 942–947. DOI: 10.1182/blood-2003-09-3333 (cit. on p. 16).

[28] N. J. Dimmock, A. J. Easton, and K. N. Leppard. *Introduction to Modern Virology*. 6th edition. Oxford , UK: Blackwell Publishing, 2007 (cit. on pp. 6, 8).

[29] E. Domingo and J. J. Holland. "RNA virus mutations and fitness for survival". In: *Annual Review of Microbiology* 51.1 (1997). PMID: 9343347, pp. 151–178. DOI: 10.1146/annurev.micro.51.1.151 (cit. on p. 7).

[30] J. W. Drake and J. J. Holland. "Mutation rates among RNA viruses". In: *Proceedings of the National Academy of Sciences* 96.24 (1999), pp. 13910–13913. DOI: 10.1073/pnas.96.24.13910 (cit. on p. 7).

[31] W. J. Dreyer and J. C. Bennett. "The molecular basis of antibody formation: a paradox." eng. In: *Proc Natl Acad Sci U S A* 54.3 (Sept. 1965), pp. 864–869 (cit. on p. 12).

[32] S. Duffy, L. A. Shackelton, and E. C. Holmes. "Rates of evolutionary change in viruses: patterns and determinants". In: *Nat Rev Genet* 9.4 (Apr. 2008), pp. 267–276. ISSN: 1471-0056 (cit. on p. 7).

[33] L. S. Ehrlich et al. "HIV-1 Capsid Protein Forms Spherical (Immature-Like) and Tubular (Mature-Like) Particles in Vitro: Structure Switching by pH-induced Conformational Changes". In: *Biophysical Journal* 81.1 (2001), pp. 586–594. ISSN: 0006-3495. DOI: http://dx.doi.org/10.1016/S0006-3495(01)75725-6 (cit. on p. 8).

[34] S. F. Elena and R. Sanjuán. "Adaptive Value of High Mutation Rates of RNA Viruses: Separating Causes from Consequences". In: *Journal of Virology* 79 (2005), pp. 11555–11558. DOI: 10.1128/JVI.79.18.11555-11558.2005 (cit. on p. 7).

[35] C. Failla, L. Tomei, and R. De Francesco. "Both NS3 and NS4A are required for proteolytic processing of hepatitis C virus nonstructural proteins." In: *Journal of Virology* 68.6 (1994), pp. 3753–3760 (cit. on p. 10).

[36] B. B. Finlay and G. McFadden. "Anti-Immunology: Evasion of the Host Immune System by Bacterial and Viral Pathogens". In: *Cell* 124.4 (2006), pp. 767–782. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/j.cell.2006.01.034 (cit. on p. 7).

[37] E. Foy et al. "Regulation of Interferon Regulatory Factor-3 by the Hepatitis C Virus Serine Protease". In: *Science* 300.5622 (2003), pp. 1145–1148. DOI: 10.1126/science.1082604 (cit. on p. 15).

[38] A. D. Frankel and J. A. T. Young. "HIV-1: Fifteen Proteins and an RNA". In: *Annual Review of Biochemistry* 67.1 (1998). PMID: 9759480, pp. 1–25. DOI: 10.1146/annurev.biochem.67.1.1 (cit. on p. 8).

[39] D. Fremont et al. “Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb”. In: *Science* 257.5072 (1992). cited By (since 1996)476, pp. 919–927 (cit. on p. 15).

[40] M. Frese et al. “Interferon-γ inhibits replication of subgenomic and genomic hepatitis C virus RNAs”. In: *Hepatology* 35.3 (2002), pp. 694–703. ISSN: 1527-3350. DOI: 10.1053/jhep.2002.31770. URL: http://dx.doi.org/10.1053/jhep.2002.31770 (cit. on p. 15).

[41] B. Ganser-Pornillos, M. Yeager, and O. Pornillos. “Assembly and Architecture of HIV”. English. In: *Viral Molecular Machines*. Ed. by M. G. Rossmann and V. B. Rao. Vol. 726. Advances in Experimental Medicine and Biology. Springer US, 2012, pp. 441–465. ISBN: 978-1-4614-0979-3. DOI: 10.1007/978-1-4614-0980-9_20 (cit. on p. 8).

[42] G. H. Gauss and M. R. Lieber. “Mechanistic constraints on diversity in human V(D)J recombination.” eng. In: *Mol Cell Biol* 16.1 (Jan. 1996), pp. 258–269 (cit. on p. 12).

[43] T. B. Geijtenbeek et al. “DC-SIGN, a Dendritic Cell–Specific HIV-1-Binding Protein that Enhances trans-Infection of T Cells”. In: *Cell* 100.5 (2000), pp. 587–597. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/S0092-8674(00)80694-7. URL: http://www.sciencedirect.com/science/article/pii/S0092867400806947 (cit. on p. 16).

[44] J. V. Giorgi et al. “Predictive Value of Immunologic and Virologic Markers After Long or Short Duration of HIV-1 Infection.” In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 29.4 (2002). ISSN: 1525-4135 (cit. on p. 16).

[45] M. Goodenow and R. Collman. “HIV-1 coreceptor preference is distinct from target cell tropism: a dual-parameter nomenclature to define viral phenotypes”. In: *J Leukoc Biol* 80 (2006), pp. 965–972 (cit. on p. 8).

[46] A. Grakoui et al. “Expression and identification of hepatitis C virus polyprotein cleavage products.” In: *Journal of Virology* 67.3 (1993), pp. 1385–1395 (cit. on p. 10).

[47] M. Groettrup et al. “Peptide antigen production by the proteasome: complexity provides efficiency”. In: *Immunology Today* 17.9 (1996), pp. 429–435. ISSN: 0167-5699. DOI: http://dx.doi.org/10.1016/0167-5699(96)10051-7 (cit. on p. 14).

[48] M. Groll et al. “Structure of 20S proteasome from yeast at 2.4A resolution”. In: *Nature* 386.6624 (Apr. 1997), pp. 463–471. URL: http://dx.doi.org/10.1038/386463a0 (cit. on p. 14).

[49] L. G. Guidotti and F. V. Chisari. “To kill or to cure: options in host defense against viral infection”. In: *Current Opinion in Immunology* 8.4 (1996), pp. 478–483. ISSN: 0952-7915. DOI: http://dx.doi.org/10.1016/S0952-7915(96)80034-3 (cit. on p. 9).

[50] R. S. Harania et al. “HIV, hepatitis B and hepatitis C coinfection in Kenya”. In: *AIDS* 22.10 (2008). ISSN: 0269-9370 (cit. on p. 9).

[51] S. Heink et al. “IFN-γ-induced immune adaptation of the proteasome system is an accelerated and transient response”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.26 (2005), pp. 9241–9246. DOI: 10.1073/pnas.0501711102. URL: http://www.pnas.org/content/102/26/9241.abstract (cit. on p. 15).

[52] T. Heise, L. G. Guidotti, and F. V. Chisari. "Characterization of Nuclear RNases That Cleave Hepatitis B Virus RNA near the La Protein Binding Site". In: *Journal of Virology* 75.15 (2001), pp. 6874–6883. DOI: 10.1128/JVI.75.15.6874-6883.2001 (cit. on p. 15).

[53] M. Hijikata et al. "Proteolytic processing and membrane association of putative nonstructural proteins of hepatitis C virus". In: *Proceedings of the National Academy of Sciences* 90.22 (1993), pp. 10773–10777 (cit. on p. 10).

[54] J. J. Holland. "Evolving virus plagues." In: *Proc Natl Acad Sci U S A* 93.2 (Jan. 1996), pp. 545–546 (cit. on p. 6).

[55] J. Holland, J. De La Torre, and D. Steinhauer. "RNA Virus Populations as Quasispecies". English. In: *Genetic Diversity of RNA Viruses*. Ed. by J. Holland. Vol. 176. Current Topics in Microbiology and Immunology. Springer Berlin Heidelberg, 1992, pp. 1–20. ISBN: 978-3-642-77013-5. DOI: 10.1007/978-3-642-77011-1_1 (cit. on p. 7).

[56] N. Hozumi and S. Tonegawa. "Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions." eng. In: *Proc Natl Acad Sci U S A* 73.10 (Oct. 1976), pp. 3628–3632 (cit. on p. 12).

[57] J. F. Hruska et al. "Structure of hepatitis B Dane particle DNA before and after the Dane particle DNA polymerase reaction." In: *Journal of Virology* 21.2 (1977), pp. 666–672 (cit. on p. 9).

[58] C. Hulo et al. "ViralZone: a knowledge resource to understand virus diversity". In: *Nucleic Acids Research* 39.suppl 1 (2011), pp. D576–D582. DOI: 10.1093/nar/gkq901 (cit. on p. 8).

[59] W. Irving. "Viral Transmission: Infection Acquired by the Blood-Borne Route". In: *Principles and Practice of Clinical Virology*. Ed. by A. J. Zuckerman et al. John Wiley & Sons, Ltd, 2009. Chap. 2, pp. 29–41. ISBN: 9780470741405. DOI: 10.1002/9780470741405.ch2 (cit. on p. 8).

[60] C. J. Janeway et al. "Immunobiology: The Immune System in Health and Disease". In: 5th. New York: Garland Science, 2001. Chap. Principles of innate and adaptive immunity. (Cit. on pp. 11–13).

[61] B. A. Jude et al. "Subversion of the innate immune system by a retrovirus." In: *Nat Immunol* 4.6 (June 2003), pp. 573–578. DOI: 10.1038/ni926. URL: http://dx.doi.org/10.1038/ni926 (cit. on p. 7).

[62] O. Kalinina, H. Norder, and L. O. Magnius. "Full-length open reading frame of a recombinant hepatitis C virus strain from St Petersburg: proposed mechanism for its formation". In: *Journal of General Virology* 85.7 (2004), pp. 1853–1857. DOI: 10.1099/vir.0.79984-0 (cit. on p. 10).

[63] T. Kawatani et al. "Incidence of hepatitis virus infection and severe liver dysfunction in patients receiving chemotherapy for hematologic malignancies". In: *European Journal of Haematology* 67.1 (2001), pp. 45–50. ISSN: 1600-0609. DOI: 10.1034/j.1600-0609.2001.067001045.x. URL: http://dx.doi.org/10.1034/j.1600-0609.2001.067001045.x (cit. on p. 16).

[64] B. F. Keele et al. "Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection". In: *Proceedings of the National Academy of Sciences* 105.21 (2008), pp. 7552–7557. DOI: 10.1073/pnas.0802203105 (cit. on p. 15).

[65] B. F. Keele et al. “Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1”. In: *The Journal of Experimental Medicine* 206.5 (2009), pp. 1117–1134. DOI: 10.1084/jem.20082831 (cit. on p. 15).

[66] A. F. Kisselev et al. “The Sizes of Peptides Generated from Protein by Mammalian 26 and 20 S Proteasomes: Implications for understanding the degenerative mechanism and the antigen presentation”. In: *Journal of Biological Chemistry* 274.6 (1999), pp. 3363–3371. DOI: 10.1074/jbc.274.6.3363. URL: http://www.jbc.org/content/274/6/3363.abstract (cit. on p. 15).

[67] H.-D. Klenk et al. “Viral Glycoproteins as Determinants of Pathogenicity”. English. In: *Molecular Basis of Viral and Microbial Pathogenesis*. Ed. by R. Rott and W. Goebel. Vol. 38. Colloquium der Gesellschaft für Biologische Chemie 9.–11. April 1987 in Mosbach/Baden. Springer Berlin Heidelberg, 1988, pp. 25–38. ISBN: 978-3-642-73216-4. DOI: 10.1007/978-3-642-73214-0_3 (cit. on p. 6).

[68] G. R. Klimpel. “Immune Defenses”. In: ed. by S. Baron. University of Texas Medical Branch at Galveston, 1996. Chap. 50 (cit. on p. 11).

[69] N. R. Klinman and J. L. Press. “The B Cell Specificity Repertoire: Its Relationship to Definable Subpopulations”. In: *Immunological Reviews* 24.1 (1975), pp. 41–83. ISSN: 1600-065X. DOI: 10.1111/j.1600-065X.1975.tb00165.x (cit. on p. 12).

[70] A. P. Kourtis et al. “HIV–HBV Coinfection — A Global Challenge”. In: *New England Journal of Medicine* 366.19 (2012). PMID: 22571198, pp. 1749–1752. DOI: 10.1056/NEJMp1201796 (cit. on p. 9).

[71] D. Lamarre et al. “An NS3 protease inhibitor with antiviral effects in humans infected with hepatitis C virus”. In: *Nature* 426.6963 (Nov. 2003), pp. 186–189. ISSN: 0028-0836 (cit. on p. 10).

[72] T. A. Landers, H. B. Greenberg, and W. S. Robinson. “Structure of Hepatitis B Dane Particle DNA and Nature of the Endogenous DNA Polymerase Reaction”. In: *Journal of Virology* 23.2 (1977), pp. 368–376 (cit. on p. 9).

[73] R. E. Lanford et al. “Cross-Genotype Immunity to Hepatitis C Virus”. In: *Journal of Virology* 78.3 (2004), pp. 1575–1581. DOI: 10.1128/JVI.78.3.1575-1581.2004 (cit. on p. 16).

[74] A. Lanzavecchia. “Antigen-specific interaction between T and B cells”. In: *Nature* 314.6011 (Apr. 1985), pp. 537–539. URL: http://dx.doi.org/10.1038/314537a0 (cit. on p. 13).

[75] D. Lavanchy. “Worldwide epidemiology of {HBV} infection, disease burden, and vaccine prevention”. In: *Journal of Clinical Virology* 34, Supplement 1 (2005). Emerging Variants on Hepatitis B Virus From Research to the Clinic, S1–S3. ISSN: 1386-6532. DOI: http://dx.doi.org/10.1016/S1386-6532(05)00384-7 (cit. on p. 9).

[76] W. M. Lee. “Hepatitis B Virus Infection”. In: *New England Journal of Medicine* 337.24 (1997). PMID: 9392700, pp. 1733–1745 (cit. on p. 9).

[77] Q. Li et al. “Glycerol monolaurate prevents mucosal SIV transmission”. In: *Nature* 458.7241 (Apr. 2009), pp. 1034–1038. DOI: 10.1038/nature07831. URL: http://dx.doi.org/10.1038/nature07831 (cit. on p. 15).

[78] J. R. Lingappa et al. “Genomewide Association Study for Determinants of HIV-1 Acquisition and Viral Set Point in HIV-1 Serodiscordant Couples with Quantified Virus Exposure”. In: *PLoS ONE* 6.12 (Dec. 2011), e28632. DOI: 10.1371/journal.pone.0028632. URL: http://dx.doi.org/10.1371/journal.pone.0028632 (cit. on p. 16).

[79] V. Lohmann et al. “Replication of Subgenomic Hepatitis C Virus RNAs in a Hepatoma Cell Line”. In: *Science* 285.5424 (1999), pp. 110–113. DOI: 10.1126/science.285.5424.110 (cit. on p. 10).

[80] H. Lu, K. Schwarz, and M. R. Lieber. “Extent to which hairpin opening by the Artemis:DNA-PKcs complex can contribute to junctional diversity in V(D)J recombination”. In: *Nucleic Acids Research* 35.20 (Nov. 2007), pp. 6917–6923. DOI: 10.1093/nar/gkm823. URL: http://dx.doi.org/10.1093/nar/gkm823 (cit. on p. 12).

[81] A. Lurman. “Eine icterus epidemic”. In: *Berl Klin Woschenschr* 22.20 (1885), p. 23 (cit. on p. 9).

[82] A. Lustig and A. J. Levine. “One hundred years of virology.” In: *J Virol* 66.8 (Aug. 1992), pp. 4629–4631 (cit. on p. 6).

[83] N. Manel and D. R. Littman. “Hiding in Plain Sight: How HIV Evades Innate Immune Responses”. In: *Cell* 147.2 (2011), pp. 271–274. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/j.cell.2011.09.010. URL: http://www.sciencedirect.com/science/article/pii/S0092867411010683 (cit. on p. 16).

[84] M. Martell et al. “Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution.” In: *Journal of Virology* 66.5 (1992), pp. 3225–3229 (cit. on p. 10).

[85] H. McClary et al. “Relative Sensitivity of Hepatitis B Virus and Other Hepatotropic Viruses to the Antiviral Effects of Cytokines”. In: *Journal of Virology* 74.5 (2000), pp. 2255–2264. DOI: 10.1128/JVI.74.5.2255-2264.2000 (cit. on p. 15).

[86] C. de Mendoza et al. “Emerging viral infections - a potential threat for blood supply in the 21st century.” In: *AIDS Rev* 14.4 (2012), pp. 279–289 (cit. on p. 8).

[87] S. Moir et al. “B Cells of HIV-1 Infected Patients Bind Virions through Cd21–Complement Interactions and Transmit Infectious Virus to Activated T Cells”. In: *The Journal of Experimental Medicine* 192.5 (2000), pp. 637–646. DOI: 10.1084/jem.192.5.637 (cit. on p. 16).

[88] F. Momburg et al. “Peptide size selection by the major histocompatibility complex-encoded peptide transporter”. In: *Journal of Experimental Medicine* 179.5 (1994). cited By (since 1996)140, pp. 1613–1623 (cit. on p. 15).

[89] M. U. Mondelli and E. Silini. “Clinical significance of hepatitis C virus genotypes”. In: *Journal of Hepatology* 31 (1995), pp. 65–70 (cit. on p. 10).

[90] D. Moradpour, F. Penin, and C. M. Rice. “Replication of hepatitis C virus”. In: *Nat Rev Micro* 5.6 (June 2007), pp. 453–463. ISSN: 1740-1526 (cit. on p. 10).

[91] F. Murphy et al. “Introduction to the Universal System of Virus Taxonomy”. English. In: *Virus Taxonomy*. Ed. by F. Murphy et al. Vol. 10. Archives of Virology Supplement 10. Springer Vienna, 1995, pp. 1–13. ISBN: 978-3-211-82594-5. DOI: 10.1007/978-3-7091-6607-9_1 (cit. on p. 6).

[92] K. Murphy, P. Travers, and M. Walport. *Janeway's Immunobiology*. Taylor & Francis Ltd., 2008 (cit. on p. 13).

[93] G. Neumann, M. A. Whitt, and Y. Kawaoka. "A decade after the generation of a negative-sense RNA virus from cloned cDNA – what have we learned?" In: *Journal of General Virology* 83.11 (2002), pp. 2635–2662 (cit. on p. 6).

[94] M. A. Nowak. "What is a quasispecies?" In: *Trends in Ecology & Evolution* 7.4 (1992), pp. 118–121. ISSN: 0169-5347. DOI: http://dx.doi.org/10.1016/0169-5347(92)90145-2 (cit. on p. 7).

[95] A. K. Nussbaum et al. "Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1". In: *Proceedings of the National Academy of Sciences* 95.21 (1998), pp. 12504–12509. DOI: 10.1073/pnas.95.21.12504. URL: http://www.pnas.org/content/95/21/12504.abstract (cit. on p. 15).

[96] N. Ogata et al. "Nucleotide sequence and mutation rate of the H strain of hepatitis C virus." In: *Proceedings of the National Academy of Sciences* 88.8 (1991), pp. 3392–3396. DOI: 10.1073/pnas.88.8.3392 (cit. on p. 7).

[97] M. Orlowski and S. Wilk. "Catalytic Activities of the 20 S Proteasome, a Multicatalytic Proteinase Complex". In: *Archives of Biochemistry and Biophysics* 383.1 (2000), pp. 1–16. ISSN: 0003-9861. DOI: http://dx.doi.org/10.1006/abbi.2000.2036. URL: http://www.sciencedirect.com/science/article/pii/S0003986100920368 (cit. on p. 15).

[98] M. K. Osborn and A. S. F. Lok. "Antiviral options for the treatment of chronic hepatitis B". In: *Journal of Antimicrobial Chemotherapy* 57.6 (2006), pp. 1030–1034. DOI: 10.1093/jac/dkl123 (cit. on p. 9).

[99] P. Palese et al. "Negative-strand RNA viruses: genetic engineering and applications". In: *Proceedings of the National Academy of Sciences* 93 (21) (1996), pp. 11354–11358 (cit. on p. 6).

[100] J.-M. Pawlotsky. "Treatment failure and resistance with direct-acting antiviral drugs against hepatitis C virus". In: *Hepatology* 53.5 (2011), pp. 1742–1751. ISSN: 1527-3350. DOI: 10.1002/hep.24262 (cit. on p. 10).

[101] R. Ranjbar et al. "HIV/HBV Co-Infections: Epidemiology, Natural History, and Treatment: A Review Article". In: *Iran Red Crescent Med J* 13.12 (2011), pp. 855–862. ISSN: 2074-1812 (cit. on p. 9).

[102] N. Ray and R. Doms. "HIV-1 coreceptors and their inhibitors". In: *Curr Top Microbiol Immunol* 303 (2006), pp. 97–120 (cit. on p. 8).

[103] K. E. Reed, A. Grakoui, and C. M. Rice. "Hepatitis C virus-encoded NS2-3 protease: cleavage-site mutagenesis and requirements for bimolecular cleavage." In: *Journal of Virology* 69.7 (1995), pp. 4127–36 (cit. on p. 10).

[104] B. Rehermann and M. Nascimbeni. "Immunology of hepatitis B virus and hepatitis C virus infection". In: *Nature Reviews Immunology* 5.3 (Mar. 2005), pp. 215–229. DOI: 10.1038/nri1573. URL: http://dx.doi.org/10.1038/nri1573 (cit. on p. 16).

[105] B. Rehermann et al. “Cellular immune responses persist and humoral responses decrease two decades after recovery from a single-source outbreak of hepatitis C”. In: *Nature Medicine* 6.5 (May 2000), pp. 578–582. DOI: 10.1038/75063. URL: http://dx.doi.org/10.1038/75063 (cit. on p. 16).

[106] B. Rehermann et al. “The hepatitis B virus persists for decades after patients’ recovery from acute viral hepatitis despite active maintenance of a cytotoxic T-lymphocyte response”. In: *Nat Med* 2.10 (Oct. 1996), pp. 1104–1108. DOI: 10.1038/nm1096-1104. URL: http://dx.doi.org/10.1038/nm1096-1104 (cit. on p. 16).

[107] B. Rehermann et al. “Hepatitis B virus (HBV) sequence variation of cytotoxic T lymphocyte epitopes is not common in patients with chronic HBV infection.” In: *The Journal of Clinical Investigation* 96.3 (Sept. 1995), pp. 1527–1534. DOI: 10.1172/JCI118191. URL: http://www.jci.org/articles/view/118191 (cit. on p. 16).

[108] K. Rispeter et al. “Hepatitis C Virus Variability: Sequence Analysis of an Isolate after 10 Years of Chronic Infection”. English. In: *Virus Genes* 21.3 (2000), pp. 179–188. ISSN: 0920-8569. DOI: 10.1023/A:1008135413215 (cit. on p. 10).

[109] M. D. Robek, S. F. Wieland, and F. V. Chisari. “Inhibition of Hepatitis B Virus Replication by Interferon Requires Proteasome Activity†”. In: *Journal of Virology* 76.7 (2002), pp. 3570–3574. DOI: 10.1128/JVI.76.7.3570-3574.2002 (cit. on p. 15).

[110] J. Roberts, K. Bebenek, and T. Kunkel. “The accuracy of reverse transcriptase from HIV-1”. In: *Science* 242.4882 (1988), pp. 1171–1173. DOI: 10.1126/science.2460925 (cit. on pp. 7, 9).

[111] W. S. Robinson, D. A. Clayton, and R. L. Greenman. “DNA of a Human Hepatitis B Virus Candidate”. In: *Journal of Virology* 14.2 (1974), pp. 384–391 (cit. on p. 9).

[112] K. L. Rock et al. “Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on {MHC} class I molecules”. In: *Cell* 78.5 (1994), pp. 761–771. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/S0092-8674(94)90462-6. URL: http://www.sciencedirect.com/science/article/pii/S0092867494904626 (cit. on p. 14).

[113] J. R. Rodriguez-Madoz et al. “Inhibition of the Type I Interferon Response in Human Dendritic Cells by Dengue Virus Infection Requires a Catalytically Active NS2B3 Complex”. In: *Journal of Virology* 84.19 (2010), pp. 9760–9774. DOI: 10.1128/JVI.01051-10 (cit. on p. 11).

[114] W. H. Roos, R. Bruinsma, and G. J. L. Wuite. “Physical virology”. In: *Nat Phys* 6.10 (Oct. 2010), pp. 733–743. ISSN: 1745-2473 (cit. on p. 6).

[115] J. F. Salazar-Gonzalez et al. “Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection”. In: *The Journal of Experimental Medicine* 206.6 (2009), pp. 1273–1289. DOI: 10.1084/jem.20090378 (cit. on p. 15).

[116] T. Saric et al. “An IFN-γ–induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I–presented peptides”. In: *Nature Immunology* 12 (2002), pp. 1169–1176. DOI: 10.1038/ni859. URL: http://www.nature.com/ni/journal/v3/n12/full/ni859.html (cit. on p. 13).

[117] L. Saveanu et al. “Complexity, contradictions, and conundrums: studying post-proteasomal proteolysis in HLA class I antigen presentation”. In: *Immunological Reviews* 207.1 (2005), pp. 42–59. ISSN: 1600-065X. DOI: 10.1111/j.0105-2896.2005.00313.x. URL: http://dx.doi.org/10.1111/j.0105-2896.2005.00313.x (cit. on p. 15).

[118] M. Schechter et al. “HIV-1 and the aetiology of AIDS”. In: *The Lancet* 341.8846 (1993). Originally published as Volume 1, Issue 8846, pp. 658–659. DOI: http://dx.doi.org/10.1016/0140-6736(93)90421-C (cit. on p. 16).

[119] U. Schubert et al. “Rapid degradation of a large fraction of newly synthesized proteins by proteasomes”. In: *Nature* 404.6779 (2000), pp. 770–774. URL: http://dx.doi.org/10.1038/35008096 (cit. on p. 13).

[120] R. M. Selik, H. W. Haverkos, and J. W. Curran. “Acquired immune deficiency syndrome (AIDS) trends in the United States, 1978–1982”. In: *The American Journal of Medicine* 76.3 (1984), pp. 493–500. ISSN: 0002-9343. DOI: http://dx.doi.org/10.1016/0002-9343(84)90669-7 (cit. on p. 8).

[121] T. Serwold et al. “ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum”. In: *Nature* 419.6906 (Oct. 2002), pp. 480–483. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature01074 (cit. on p. 13).

[122] Y. K. Shimizu et al. “Neutralizing antibodies against hepatitis C virus and the emergence of neutralization escape mutant viruses.” In: *Journal of Virology* 68.3 (1994), pp. 1494–1500 (cit. on p. 16).

[123] N. H. Shoukry et al. “Memory CD8+ T Cells Are Required for Protection from Persistent Hepatitis C Virus Infection”. In: *The Journal of Experimental Medicine* 197.12 (2003), pp. 1645–1655. DOI: 10.1084/jem.20030239 (cit. on p. 16).

[124] G. Silvestri et al. “Nonpathogenic {SIV} Infection of Sooty Mangabeys Is Characterized by Limited Bystander Immunopathology Despite Chronic High-Level Viremia”. In: *Immunity* 18.3 (2003), pp. 441–452. ISSN: 1074-7613. DOI: http://dx.doi.org/10.1016/S1074-7613(03)00060-8. URL: http://www.sciencedirect.com/science/article/pii/S1074761303000608 (cit. on p. 16).

[125] G. W. Siskind and B. Benacerraf. “Cell Selection by Antigen in the Immune Response”. In: *Advances in Immunology*. Ed. by F. Dixon and H. G. Kunkel. Vol. 10. Advances in Immunology. Academic Press, 1969, pp. 1–50. DOI: http://dx.doi.org/10.1016/S0065-2776(08)60414-9 (cit. on p. 13).

[126] D. B. Smith et al. “Virus ’quasispecies’: making a mountain out of a molehill?” In: *Journal of General Virology* 78.7 (1997), pp. 1511–9. eprint: http://vir.sgmjournals.org/content/78/7/1511.full.pdf+html. URL: http://vir.sgmjournals.org/content/78/7/1511.short (cit. on p. 7).

[127] E. B. Stephens and R. W. Compans. “Assembly of Animal Viruses at Cellular Membranes”. In: *Annual Review of Microbiology* 42.1 (1988), pp. 489–516. DOI: 10.1146/annurev.mi.42.100188.002421 (cit. on p. 6).

[128] J. Summers, A. O'Connell, and I. Millman. "Genome of hepatitis B virus: restriction enzyme cleavage and structure of DNA extracted from Dane particles". In: *Proceedings of the National Academy of Sciences* 72.11 (1975), pp. 4597–4601 (cit. on p. 9).

[129] J. Summers and W. S. Mason. "Replication of the genome of a hepatitis B-like virus by reverse transcription of an {RNA} intermediate". In: *Cell* 29.2 (1982), pp. 403–415. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/0092-8674(82)90157-X (cit. on p. 9).

[130] D. R. Taylor et al. "Inhibition of the Interferon- Inducible Protein Kinase PKR by HCV E2 Protein". In: *Science* 285.5424 (1999), pp. 107–110. DOI: 10.1126/science.285.5424.107 (cit. on p. 15).

[131] H. M. Temin. "Reverse transcription in the eukaryotic genome: retroviruses, para"-retroviruses, retrotransposons, and retrotranscripts." In: *Molecular Biology and Evolution* 2.6 (1985), pp. 455–468 (cit. on p. 6).

[132] R. Thimme et al. "Determinants of Viral Clearance and Persistence during Acute Hepatitis C Virus Infection". In: *The Journal of Experimental Medicine* 194.10 (2001), pp. 1395–1406. DOI: 10.1084/jem.194.10.1395 (cit. on p. 15).

[133] R. Thimme et al. "Viral and immunological determinants of hepatitis C virus clearance, persistence, and disease". In: *Proceedings of the National Academy of Sciences* 99.24 (2002), pp. 15661–15668. DOI: 10.1073/pnas.202608299 (cit. on p. 15).

[134] J. Timm et al. "CD8 Epitope Escape and Reversion in Acute HCV Infection". In: *The Journal of Experimental Medicine* 200.12 (2004), pp. 1593–1604. DOI: 10.1084/jem.20041006 (cit. on p. 16).

[135] L. V. Tsui et al. "Posttranscriptional clearance of hepatitis B virus RNA by cytotoxic T lymphocyte-activated hepatocytes". In: *Proceedings of the National Academy of Sciences* 92 (1995), pp. 12398–12402 (cit. on p. 15).

[136] D. Voges, P. Zwickl, and W. Baumeister. "THE 26S PROTEASOME: A Molecular Machine Designed for Controlled Proteolysis". In: *Annual Review of Biochemistry* 68.1 (1999), pp. 1015–1068. DOI: 10.1146/annurev.biochem.68.1.1015 (cit. on p. 14).

[137] T. Wenzel et al. "Existence of a molecular ruler in proteasomes suggested by analysis of degradation products". In: *{FEBS} Letters* 349.2 (1994), pp. 205–209. ISSN: 0014-5793. DOI: http://dx.doi.org/10.1016/0014-5793(94)00665-2. URL: http://www.sciencedirect.com/science/article/pii/0014579394006652 (cit. on p. 15).

[138] S. A. Whalley et al. "Evolution of hepatitis B virus during primary infection in humans: Transient generation of cytotoxic T-cell mutants". In: *Gastroenterology* 127.4 (2004), pp. 1131–1138. DOI: http://dx.doi.org/10.1053/j.gastro.2004.07.004 (cit. on p. 16).

[139] WHO. *Global HIV/AIDS Response: Epidemic update and health sector progress towards Universal Access*. Progress report. WHO, 2011 (cit. on p. 8).

[140] S. Wieland et al. "Genomic analysis of the host response to hepatitis B virus infection". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.17 (2004), pp. 6669–6674. DOI: 10.1073/pnas.0401771101 (cit. on p. 15).

[141] R. H. Wiesner, M. Sorrell, and F. Villamil. "Report of the First International Liver Transplantation Society Expert Panel Consensus Conference on Liver Transplantation and Hepatitis C". In: *Liver Transplantation* 9.11 (2003), S1–S9. ISSN: 1527-6473. DOI: 10.1053/jlts.2003.50268 (cit. on p. 10).

[142] C. B. Wilen, J. C. Tilton, and R. W. Doms. "HIV: Cell Binding and Entry". In: *Cold Spring Harb. Perspect Med.* 2 (2012) (cit. on p. 8).

[143] F. Wong-Staal and R. C. Gallo. "Human T-lymphotropic retroviruses". In: *Nature* 317.6036 (Oct. 1985), pp. 395–403 (cit. on p. 6).

[144] E. J. Wood. "Cellular and molecular immunology". In: ed. by A. A. K. and A. H. Lichtman. 5th. Vol. 32. 1. John Wiley & Sons Inc., 2004, pp. 65–66. DOI: 10.1002/bmb.2004.494032019997. URL: http://dx.doi.org/10.1002/bmb.2004.494032019997 (cit. on p. 12).

[145] R. Wyatt and J. Sodroski. "The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens". In: *Science* 280 (5371) (1998), pp. 1884–8 (cit. on p. 8).

[146] I. A. York et al. "The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues". In: *Nat Immunol* 3.12 (Dec. 2002), pp. 1177–1184. ISSN: 1529-2908. URL: http://dx.doi.org/10.1038/ni860 (cit. on p. 13).

[147] S. Zeuzem et al. "Risk factors for the transmission of hepatitis C". In: *Journal of hepatology* 24.2 Suppl (1996), pp. 3–10. ISSN: 0168-8278 (cit. on p. 10).

[148] A. L. Zignego et al. "Infection of peripheral mononuclear blood cells by hepatitis C virus". In: *Journal of Hepatology* 15.3 (1992), pp. 382–386. ISSN: 0168-8278. DOI: http://dx.doi.org/10.1016/0168-8278(92)90073-X (cit. on p. 10).

[149] T. G. Zreik and D. L. Olive. "Pathophysiology. The biologic principles of disease." eng. In: *Obstet Gynecol Clin North Am* 24.2 (June 1997), pp. 259–268 (cit. on p. 11).

3

Methods, tools and techniques used globally

> Any sufficiently advanced technology is indistinguishable from magic.
>
> *ARTHUR C. CLARKE*

A lot of tools and techniques are available for sequence analysis, starting with the "production" of such sequences up to the tools for analysis of the sequences. For most problems more than one solution exists, which is both good and bad. On the one hand it is good, because one can use whatever tool and technique is most suitable to the task, but on the other hand due to this redundancy a result from one tool may not be usable with another. One example for such an issue are file formats for phylogenetic trees, which are often not readable by another designated tool [7, 53].

In this work many methods, tools, and techniques from this huge sequence-tool-space were used. The most important ones are sequence alignments, phylogenetics, statistics in the case of methods used inside the computer, (next-generation) sequencing as basic tool to generate sequences, which were then analyzed, and the programming environment R, which was used to analyze most of the results presented here. Other tools including SQLite - a database programming language, LaTeX and even office tools like open office were used where appropriate.

3.1 Sequence alignments

The analysis of sequences is often preceded by an alignment. This alignment is the result of a method of identifying regions of similarity in DNA, RNA or protein sequences that may be a consequence of functional, structural, or evolutionary relationships between them. Often sequence alignments are displayed as a matrix, with different sequences in rows and sequence positions in columns (see Figure 3.1). In general, one can distinguish alignments either by the number of sequences (pairwise alignment vs. multiple sequences alignment) or by the method of alignment (global vs. local alignment, plus hybrid versions).

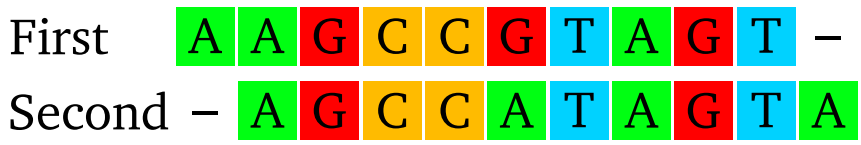


Figure 3.1: Example of a global sequence alignment with nucleotides from DNA. Colored by type of nucleotide. Horizontal lines are gaps, which indicate insertions or deletions inside or outside the other sequence(es) and are abbreviates indel.

Global alignments attempt to align every residue in every sequence and are useful for sequences of higher similarity and approximately of equal length.

Local alignments align regions of high similarity and should be used in the case of highly diverse sequences with certain motifs inside.

Pairwise alignments, alignments between exactly two sequences, are often solved with either the Needleman–Wunsch algorithm or the Smith-Waterman algorithm, both of which are examples of dynamic programming, [34, 49]. The Needleman–Wunsch algorithm is a global alignment algorithm and uses a simple technique which can be performed manually on a piece of paper if the two sequences are short enough (see example in Figure 3.2). In our example a simple scoring function is used, but often for protein alignments complex scoring matrices are more appropriate like PAM or BLOSUM [8, 16], which incorporate information about the structure of the amino acids to determine the value of a mismatch. Smith and Waterman transformed the Needleman–Wunsch algorithm to create an algorithm to align sequences locally. The basic idea and method is the same, but with two exceptions: all negative values are set to zero and backtracking starts at the highest scoring cell up to the first cell with a zero inside.

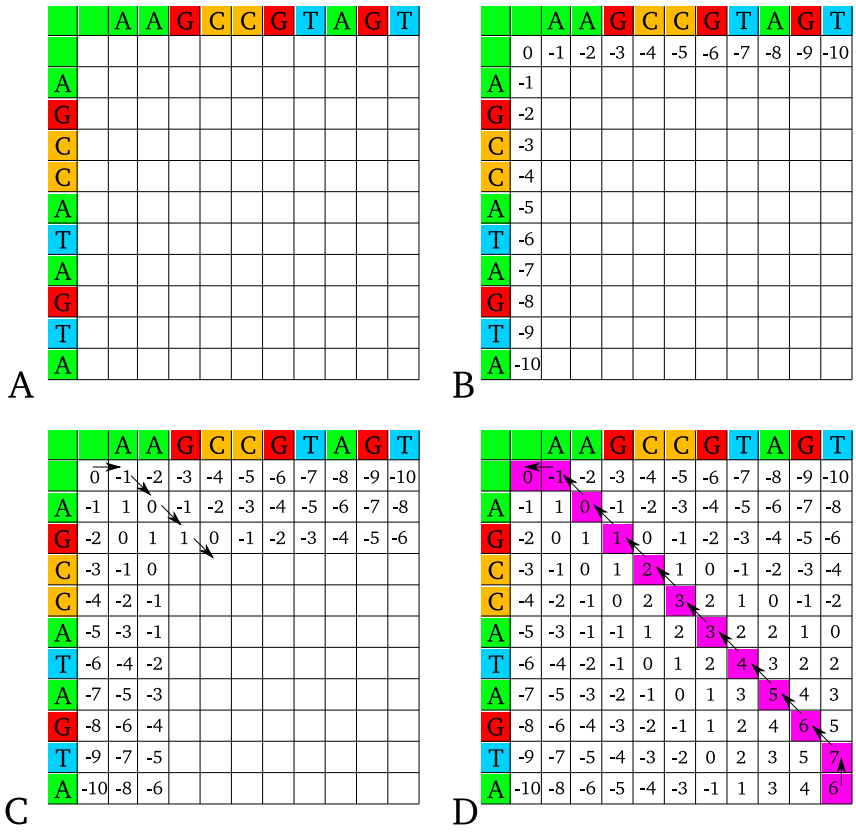


Figure 3.2: Needleman–Wunsch algorithm example. The task for the algorithm example presented here is to align the sequences AAGCCGTAGT and AGCCATAGTA. **A** First a matrix is created with row numbers and column numbers according to the length of the sequences plus one for the possibility of a gap. The scoring function used here was: match = 1, mismatch = -1 and gap = -1. After defining the scoring function the cells can be filled. The top left cell is always zero. Every other cell is the best possible score (i.e. highest) from existing scores to the left, top or top-left (diagonal). As a formula: $M_{i,j} = max(M_{i-1,j-1} + S_{i,j}, M_{i,j-1} + W, M_{i-1,j} + W)$, with i, j index of the matrix, S scoring function (match $S(i,j) = 1$, mismatch $S(i,j) = -1$), W gap penalty. Scores from the top or left represent an insertion or deletion, scores from the diagonal represent a match or mismatch. **B** The first row and column is easily filled as there are no top or top-left in the case of the first row or left and top-left in the case of the first column. **C** While filling in the matrix according to the formula above, the cell or cells from which the value was taken has to be marked. Usually an arrow is used. **D** When the whole table is filled like described here, the score in the bottom right cell represents the alignment score for the best alignment. From there the arrows can be traced back to get the alignment. A diagonal arrow is a letter above a letter, a top or left arrow is a gap. This results in the alignment as pictured in Figure 3.1 with an alignment score of 8.

Multiple sequences alignments use a type of phylogenetic approach instead, in which the sequences with the highest similarity are aligned first and the rest of the sequences successively until all sequences are incorporated into the solution. Most commonly used tools are Clustal, in different variants, or T-Coffee [17, 36]. Other methods for multiple sequences alignment often include either more information, as is the case with RNA/amino acid sequences alignments based on their secondary/tertiary structure, or use general optimization algorithms such as Hidden Markov models, which are less susceptible to

noise created by conservative or semi conservative substitutions [20].

3.2 Phylogenetics

Phylogenetic analyses generate information about the evolutionary relationships between sequences from molecular sequencing data, although they may be difficult to interpret, as is shown in Velasco [56]. With phylogenetics a researcher is able to hypothesize the evolutionary history of taxonomic groups and visualize it as a relationship tree, a well-known example of which is the tree of life [30]. Before phylogenetics were available, scientists used distance matrix-based methods based on overall similarity in observable traits, such as morphology.

Several scientific methods exist to do phylogenetics. The most common ones are distance-matrix methods, and maximum parsimony, maximum likelihood, and Bayesian inference. The latter use mathematical models describing the evolution of characters and can use sequence data, be it nucleotides or amino acids, although distance-matrix, maximum parsimony and Bayesian inference can be used with any type of suitable data. They are either parametric (maximum likelihood, Bayesian inference) or non-parametric (distance-matrix, maximum parsimony). The resulting trees of all methods have a certain terminology, as shown in Figure 3.3. Sometimes trees are visualized not as a tree, but unrooted or circular to better display certain features e.g. known differentiators such as the origin of the used sequence.

Distance-matrix methods use sequence distances from a multiple sequence alignment. The distances are often defined as the fraction of mismatches at aligned positions. Gaps may be ignored or counted as mismatches [13, 33]. Often used variants of the distance-matrix method are neighbor-joining for unrooted trees and UPGMA (Unweighted Pair Group Method with Arithmetic Mean) for rooted trees [44, 51]. Both use general data clustering analyses. Neighbor-joining does not assume a constant rate of evolution, whereas UPGMA does.

Maximum parsimony tries to find the phylogenetic tree that suppose the least evolutionary change to explain observed data [11]. This method is not statistically consistent, it is not guaranteed to produce the correct tree with high probability [12], and is therefore not used in this thesis.

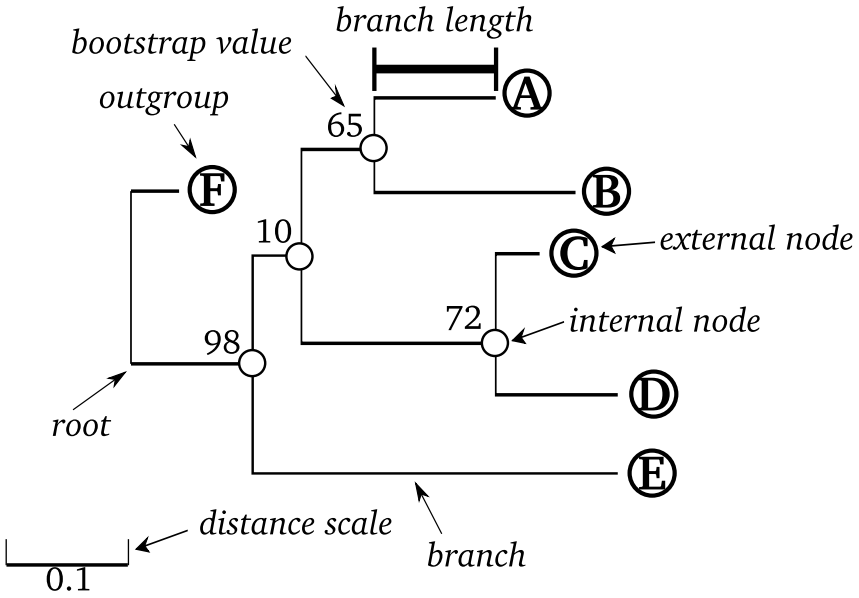


Figure 3.3: Terminology of phylogenetic trees.

Instead maximum likelihood is a parametric statistical method and it employs an explicit model of evolution. The quality of the tree highly depends on the chosen model, which has to be reasonable for the given data. Like maximum parsimony, each tree will get a score and the tree with the best score is the most likely candidate for the evolution of the given sequences. Maximum likelihood estimates the likelihood of trees. A likelihood is relative to the probability of a certain outcome out of all possible outcomes, e.g the number six on a fair sided dice has a likelihood of about 16,6% or 1/6 to fall. If one does not know the possible number of outcomes, the number of sides of the dice, but observes that the number six falls roughly 1/6 of all rolls, one can estimate that the number six is likely to fall one time out of six. The probability of a tree cannot be calculated, since no one knows the number of trees, but the probability of the data given a tree can be computed if one assumes a model. There exist several well-established models, e.g. Jukes-Cantor (the oldest model and often inappropriate) or GTR (the most general neutral, independent, finite-sites, time-reversible model), which all assume a certain (different) kind of evolutionary change [19, 55]. In contrast to maximum parsimony, the branch length in maximum likelihood has a certain meaning: the branch length is interpreted as being proportional to the average probability of change on that branch [13]. The drawback of this method is the model itself. If this model is incorrect, it will produce a biased result [13].

Bayesian inference uses the same models of evolutionary change as maximum likelihood,

but is different in both theory and application. It uses Bayes' theorem (see chapter 5), and does not produce a single or set of equally optimal trees. Bayesian phylogenetic analysis calculates the likelihoods of trees in a Markov chain Monte Carlo (MCMC) simulation and produces a credible sample of trees. One drawback is, of course, the same as with other Bayesian methods: the need to explicitly set out a set of prior probabilities, to infer from.

The standard protocol of a phylogenetic analyses includes DNA/Amino Acid sequence assembly, multiple sequence alignment, model-test and phylogeny reconstruction, an example of such a protocol can be found in Nature Protocols [4].

Several obstacles exist in getting a reasonable tree out of the given data. Some characters are more likely to evolve convergently than others, e.g. in 2013 researchers found convergent amino acid substitutions in genes implicated in hearing and vision in echo-locating bats and dolphins [37]. Maximum likelihood and Bayesian inference can take some account for this so-called homoplasy, but the weights further down the tree for those sites have to be inferred from the data. Horizontal gene transfer is an issue for all three methods. If genes are transferred not from parents to offspring as is the main assumption in phylogenetics, but from for example one bacterium to another, this transformed bacterium will be placed much farther away than it should be. A possible solution is the assumption that the largest set of genes that have been inherited together were transmitted vertically. Hybrids create the same difficulties as horizontal gene transfer and are not uncommon among plants [58]. It is important to choose a good region in the genome to create the tree, a region which is similar in highly related species and differs in less related ones, and of course missing data will also result in a less accurate tree.

Bootstrapping can be used to gain knowledge about the certainty of a particular inner node in the result tree. With bootstrapping, usually one column (of the sequence alignment) is left out of the analysis and the frequency whereby this exact node reappears in the result tree is then counted. This is done usually 100 to 1000 times out of performance reasons, and the bootstrap number at each node informs the reader how often this node occurred (see Figure 3.3).

3.3 Homology modeling

Homology modeling is used to generate a tertiary structure out of a protein sequence and an experimental homologous protein structure. It basically uses the method of pairwise/multiple sequence alignments to align the target sequence to one or more reference sequence(s). Since evolutionarily related proteins tends to show similar sequences and homologous proteins have a highly similar protein structure, homology modeling can create for example the tertiary structure of a protein without the usage of complex NMR or X-ray experiments. The quality of the model depends on the quality of the sequence alignment and on the quality of the template structure. Indels inside the alignment may reduce the quality by creating gaps inside the structure due to poor resolution. Regions of the target without a matching structure have to be modeled in different ways, for example loop modeling. Although homology modeling is prone to errors and should not be used for purposes that require precise atomic-resolution data such as drug design and protein–protein interaction predictions, it can be used to obtain qualitative conclusions regarding the biochemistry of the query sequence, especially relating to residues.

The normal approach in homology modeling is: template selection, sequence alignment, model construction, and model assessment [31]. The most important step is the search for a compatible template structure. This is usually done by a BLAST search of the target sequence, an alignment assisted database search through thousands of sequences [1]. The best result depends on several factors. The coverage of the aligned region, the part of the target which can be aligned to the reference, might be the important one. Other factors include the sequence similarity of both sequences or of their functions, the similarity of the predicted target, and the template secondary structure. The model itself is then generated from the given structure and the sequence alignment with fragment assembly, segment matching or satisfaction of spatial restraints [23, 45, 57]. The last one, a satisfaction of spatial restraints, is an often used variant and a frequently used implementation is MODELLER [14]. In the last step, the created structure of the target is assessed. With either statistical potentials or physics-based energy calculations an energy value is calculated, which can be used to get an indication of the accuracy of the model.

3.4 (Next-generation) sequencing

For about 25 years, biological sequences could only be examined with the so-called Sanger Sequencing method, named after it's creator Frederick Sanger [46]. Sanger Sequencing uses the normal cell mechanisms to copy DNA so that from a template sequence one nucleotide after another is added to the result sequence. With this method, sequences of 750 *bp* can be generated [47], but the actual number depends on the used technology. Sanger sequencing uses one template and is therefor very cost-intensive for a set of many sequences.

In 1996 a new method for sequencing was published by Ronaghi et al. [42], the so-called pyrosequencing, which utilizes the sequencing by synthesis principle. Pyrosequencing uses chemiluminescent enzymes and detects if a type of nucleotide solution added to the template sequences emits light or not. The biggest issue with pyrosequencing are the so-called poly-N regions, because the intensity difference from 2 or more nucleotide additions to the sequence is very small and often not measurable [39, 41, 43].

Several other methods have been published since then, which are now summarized in the term 'Next-generation sequencing' or short NGS. Besides the pyrosequencing, which is called 454 after the most used sequencing machine invented by 454 Life Science and distributed by Roche [21], a second method is widely utilized, the Illumina technology, which is also the name of the machine and the company distributing the machines [38, 48]. Illumina uses a technique invented in the mid to late 90's by the company Solexa, in which nucleotides labeled with different fluorescence marker are added by a DNA polymerase sequentially to a primer depending on the template DNA. After every added nucleotide a laser excites the bound nucleotide and a picture is taken to identify which nucleotide was bound. Since every used nucleotide serves as a terminator for the polymerization, poly-N occurring in the template are rarely a problem for this method [2, 35, 50]. Other than that mentioned there are less frequently used systems which also produce sequences on a large scale, such as single molecule real time sequencing from Pacific Bioscience, which uses phospho-linked nucleotides and detects in real time the addition of a single nucleotide to a single strand of DNA [22]. The answer to the question "which method is better" depends on what the researcher wishes to analyze. Longer sequences are better with the 454 technology, but have more insertion and deletion errors, many short sequences with fewer errors are possible with the Illumina technology [28, 32, 40].

3.5 Fileformats for sequences

3.5.1 FASTA format

Prior to the development of next-generation sequences, the most frequently used file format for sequence data was the so-called FASTA format. The original idea was developed by David J. Lipman and William R. Pearson in 1985 as part of the sequence alignment tool with the same name [27]. Later NCBI adopted this format for their tools and databases, and very quickly it became a standard in the field of bioinformatics. A sequence in FASTA format is represented by two items: the description line and the actual sequence. The description line starts with a single greater-than ('>') symbol. The following (usually 80) letters may contain any letter besides the greater-than (a few tools only allow certain characters). In the original format lines with a semicolon could follow this line, which were ignored by the tool. The sequences after the description line are in one-letter code and may contain gaps, asterisks and IUB/IUPAC letters [26]. The length of a FASTA-file is by definition not restricted.

Example of FASTA format:

```
>B.KR.1993.Donor_P_93KPS2_7012.HM210885
FFREDLAFPQ-GKAREFSSEQTRANSPTRR-ELQVWGRDNNSLSEAGADR
```

3.5.2 FASTQ format

The FASTQ format was originally developed to combine sequences in FASTA format with quality data by Sanger, but is now used for next-generation sequencing data as well [6]. Compared to the FASTA format with two lines, the FASTQ format contains four lines per sequence. The first two lines correspond to the FASTA format: line 1 is the description line starting with an '@' character instead of the greater-than ('>') symbol; the second line is the raw sequence in one-letter code. The third line starts with a plus '+' symbol and can contain the same description as the first line. The fourth one is the most complex line, containing the mentioned qualities encoded into one letter per sequence position and therefore with the same length as the second line. This encoding uses ASCII symbols and depends on the machine used to generate the data [24, 25]. The Sanger format of FASTQ can encode quality scores from 0 up to 93, whereas the first Illumina sequences

encoded quality scores between -5 to 62 [6]. The actual encoding should be known to decode correctly.

The quality score Q itself is an integer mapping of a probability that the corresponding base call is correct. The now often used Phred Quality score, which originated in the computer program Phred base-calling, is calculated by the following formula [9, 10]:

$$Q = -10 \log_{10} P \tag{3.1}$$

Thus a quality score of 30 tells the user that this base call is wrong in one of a thousand base calls. A good quality score depends on the sequences and sequencer and the task which should be solved, but usually base calls lower than 10 are considered as risky and MacManes recommends certain precautions for quality scores lower than 5 in RNA-seq [29].

Example of FASTQ format:

```
@M03645:23:000000000-ACB9C:1:1101:13351:1037 2:N:0:1
NGGGTTCCCTGCCCCATATGGGCTATCTTTCGCACAGTAATAACGCGTGTCCCAGGCTGN
+
#8ACCGGGGG#=BFGGF#=C##=####::C########:#::FFGGGGGGGGGGGGGGG#
```

3.6 R - The programming environment used in this thesis

R is a programming environment which was originally intended to be used for statistical problems [18]. Its language is based on S [5] and the functional programming language Scheme [54]. One or more packages exist for nearly all larger tasks, which anyone can use freely together with the programming language to solve their own problems. Everyone is free to submit their code to the central package storage called C-RAN. There is even a big package resource purely for packages from the bioscience called BioConductor [15]. Due to this system and the simple and easy to learn language, R has grown from a mere statistical tool to a fully grown programming language and is, besides C++, C#, Java, Pearl and Python, especially widely used in the field of biosciences [52].

One drawback of R is the speed of the execution of scripts [3]. For long scripts or larger tasks with loops, the calculation speed drops very quickly with increasing intricacy. R

was originally invented for small statistical tasks and is not optimized for large amounts of sequences. Keeping this in mind, R with its thousands of packages is the perfect programming language to use in sequence analysis problems.

References

[1] S. F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 1990), pp. 403–410. DOI: 10.1016/s0022-2836(05)80360-2. URL: http://dx.doi.org/10.1016/S0022-2836(05)80360-2 (cit. on p. 35).

[2] W. J. Ansorge. “Next-generation DNA sequencing techniques”. In: *New Biotechnology* 25.4 (2009), pp. 195–203. ISSN: 1871-6784. DOI: http://dx.doi.org/10.1016/j.nbt.2008.12.009 (cit. on p. 36).

[3] S. B. Aruoba and J. Fernández-Villaverde. *A Comparison of Programming Languages in Economics*. Working Paper 20263. National Bureau of Economic Research, June 2014. DOI: 10.3386/w20263. URL: http://www.nber.org/papers/w20263 (cit. on p. 38).

[4] F. Bast. “Sequence similarity search, Multiple Sequence Alignment, Model Selection, Distance Matrix and Phylogeny Reconstruction”. In: *Protocol Exchange* (July 2013). DOI: 10.1038/protex.2013.065. URL: http://dx.doi.org/10.1038/protex.2013.065 (cit. on p. 34).

[5] J. M. Chambers. “A statistical data language”. In: *Statistical computation. Proceedings of a Conference on Statistical Computation held at the Univ. of Wisconsin Madison, Wisconsin April 28-30, 1969.* (1969) (cit. on p. 38).

[6] P. J. A. Cock et al. “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants”. In: *Nucleic Acids Research* 38.6 (Dec. 2009), pp. 1767–1771. DOI: 10.1093/nar/gkp1137. URL: http://dx.doi.org/10.1093/nar/gkp1137 (cit. on pp. 37, 38).

[7] K. Cranston et al. “Best Practices for Data Sharing in Phylogenetic Research.” In: *PLOS Currents Tree of Life*. Edition 1 (June 2014). DOI: doi:10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645. (cit. on p. 29).

[8] M. O. Dayhoff and R. M. Schwartz. “Chapter 22: A model of evolutionary change in proteins”. In: *in Atlas of Protein Sequence and Structure*. 1978 (cit. on p. 30).

[9] B. Ewing and P. Green. “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities”. In: *Genome Research* 8.3 (1998), pp. 186–194. eprint: http://genome.cshlp.org/content/8/3/186.full.pdf+html. URL: http://genome.cshlp.org/content/8/3/186.abstract (cit. on p. 38).

[10] B. Ewing et al. “Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment”. In: *Genome Research* 8.3 (Mar. 1998), pp. 175–185. DOI: 10.1101/gr.8.3.175. URL: http://dx.doi.org/10.1101/gr.8.3.175 (cit. on p. 38).

[11] J. S. Farris. “Methods for Computing Wagner Trees”. English. In: *Systematic Zoology* 19.1 (1970), pp. 83–92. ISSN: 00397989. URL: http://www.jstor.org/stable/2412028 (cit. on p. 32).

[12] J. Felsenstein. “Cases in which Parsimony or Compatibility Methods Will be Positively Misleading”. English. In: *Systematic Zoology* 27.4 (1978), pp. 401–410. ISSN: 00397989. URL: http://www.jstor.org/stable/2412923 (cit. on p. 32).

[13] J. Felsenstein. *Inferring phylogenies*. Sinauer associates Sunderland, 2004 (cit. on pp. 32, 33).

[14] A. Fiser and A. Sali. "Modeller: generation and refinement of homology-based protein structure models." eng. In: *Methods Enzymol* 374 (2003), pp. 461–491. DOI: 10.1016/S0076-6879(03)74020-8. URL: http://dx.doi.org/10.1016/S0076-6879(03)74020-8 (cit. on p. 35).

[15] R. Gentleman et al. "Bioconductor: open software development for computational biology and bioinformatics". In: *Genome Biology* 5.10 (2004), R80. ISSN: 1465-6906. DOI: 10.1186/gb-2004-5-10-r80 (cit. on p. 38).

[16] S. Henikoff and J. G. Henikoff. "Amino acid substitution matrices from protein blocks." eng. In: *Proc Natl Acad Sci U S A* 89.22 (Nov. 1992), pp. 10915–10919 (cit. on p. 30).

[17] D. G. Higgins and P. M. Sharp. "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." eng. In: *Gene* 73.1 (Dec. 1988), pp. 237–244 (cit. on p. 31).

[18] R. Ihaka and R. Gentleman. "R: A Language for Data Analysis and Graphics". In: *Journal of Computational and Graphical Statistics* 5.3 (1996), pp. 299–314. DOI: 10.1080/10618600.1996.10474713 (cit. on p. 38).

[19] T. H. Jukes and C. R. Cantor. *Evolution of Protein Molecules*. Ed. by H. N. Munro. Academy Press, 1969 (cit. on p. 33).

[20] K. Karplus, C. Barrett, and R. Hughey. "Hidden Markov models for detecting remote protein homologies." eng. In: *Bioinformatics* 14.10 (1998), pp. 846–856 (cit. on p. 32).

[21] M. Kircher and J. Kelso. "High-throughput DNA sequencing – concepts and limitations". In: *BioEssays* 32.6 (2010), pp. 524–536. ISSN: 1521-1878. DOI: 10.1002/bies.200900181 (cit. on p. 36).

[22] M. J. Levene et al. "Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations". In: *Science* 299.5607 (2003), pp. 682–686 (cit. on p. 36).

[23] M. Levitt. "Accurate modeling of protein conformation by automatic segment matching." eng. In: *J Mol Biol* 226.2 (July 1992), pp. 507–533 (cit. on p. 35).

[24] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324. eprint: http://bioinformatics.oxfordjournals.org/content/25/14/1754.full.pdf+html. URL: http://bioinformatics.oxfordjournals.org/content/25/14/1754.abstract (cit. on p. 37).

[25] H. Li, J. Ruan, and R. Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". In: *Genome Research* 18.11 (2008), pp. 1851–1858. DOI: 10.1101/gr.078212.108. eprint: http://genome.cshlp.org/content/18/11/1851.full.pdf+html (cit. on p. 37).

[26] C. Liébecq. *Biochemical Nomenclature: And Related Documents : A Compendium 1992 (International Union of Biochemistry and Molecular Biology)*. Ed. by I.-I. J. C. on Biochemical Nomenclature and N. C. of IUBMB. 2nd ed. Portland Pr, 1992. ISBN: 1855780054 (cit. on p. 37).

[27] D. Lipman and W. Pearson. "Rapid and sensitive protein similarity searches". In: *Science* 227 (1985), pp. 1435–1441 (cit. on p. 37).

[28] N. J. Loman et al. “Performance comparison of benchtop high-throughput sequencing platforms”. In: *Nat Biotech* 30.5 (2012), pp. 434–439. ISSN: 1087-0156. URL: http://dx.doi.org/10.1038/nbt.2198 (cit. on p. 36).

[29] M. D. MacManes. “On the optimal trimming of high-throughput mRNA sequence data”. In: *Frontiers in Genetics* 5.13 (2014). ISSN: 1664-8021. DOI: 10.3389/fgene.2014.00013. URL: http://www.frontiersin.org/bioinformatics_and_computational_biology/10.3389/fgene.2014.00013/abstract (cit. on p. 38).

[30] D. R. Maddison, K.-S. Schulz, and W. P. Maddison. “The Tree of Life Web Project.” In: *Linnaeus Tercentenary: Progress in Invertebrate Taxonomy.* Vol. 1668. 1-766. Z.-Q. Zhang, W.A. Shear, 2007, pp. 19–40 (cit. on p. 32).

[31] M. A. Martí-Renom et al. “Comparative protein structure modeling of genes and genomes.” eng. In: *Annu Rev Biophys Biomol Struct* 29 (2000), pp. 291–325. DOI: 10.1146/annurev.biophys.29.1.291. URL: http://dx.doi.org/10.1146/annurev.biophys.29.1.291 (cit. on p. 35).

[32] M. L. Metzker. “Sequencing technologies - the next generation”. In: *Nat Rev Genet* 11.1 (Jan. 2010), pp. 31–46. ISSN: 1471-0056. URL: http://dx.doi.org/10.1038/nrg2626 (cit. on p. 36).

[33] D. W. Mount. *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor Laboratory Press, 2004. ISBN: 9780879696870. URL: https://books.google.de/books?id=M8pqAAAAMAAJ (cit. on p. 32).

[34] S. B. Needleman and C. D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” eng. In: *J Mol Biol* 48.3 (Mar. 1970), pp. 443–453 (cit. on p. 30).

[35] C. B. Nielsen et al. “Visualizing genomes: techniques and challenges”. In: *Nat Meth* 7.3s (Feb. 2010), S5–S15. DOI: 10.1038/nmeth.1422 (cit. on p. 36).

[36] C. Notredame, D. G. Higgins, and J. Heringa. “T-Coffee: A novel method for fast and accurate multiple sequence alignment.” eng. In: *J Mol Biol* 302.1 (Sept. 2000), pp. 205–217. DOI: 10.1006/jmbi.2000.4042. URL: http://dx.doi.org/10.1006/jmbi.2000.4042 (cit. on p. 31).

[37] J. Parker et al. “Genome-wide signatures of convergent evolution in echolocating mammals”. In: *Nature* 502.7470 (Sept. 2013), pp. 228–231. DOI: 10.1038/nature12511. URL: http://dx.doi.org/10.1038/nature12511 (cit. on p. 34).

[38] E. Pettersson, J. Lundeberg, and A. Ahmadian. “Generations of sequencing technologies”. In: *Genomics* 93.2 (2009), pp. 105–111. DOI: http://dx.doi.org/10.1016/j.ygeno.2008.10.003 (cit. on p. 36).

[39] M. Pop and S. L. Salzberg. “Bioinformatics challenges of new sequencing technology”. In: *Trends in Genetics* 24.3 (2008), pp. 142–149. ISSN: 0168-9525. DOI: http://dx.doi.org/10.1016/j.tig.2007.12.006 (cit. on p. 36).

[40] M. Quail et al. “A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers”. In: *BMC Genomics* 13.1 (2012), p. 341. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-341. URL: http://www.biomedcentral.com/1471-2164/13/341 (cit. on p. 36).

[41] M. Ronaghi, M. Uhlén, and P. Nyrén. "A Sequencing Method Based on Real-Time Pyrophosphate". In: *Science* 281.5375 (1998), pp. 363–365. DOI: 10.1126/science.281.5375.363 (cit. on p. 36).

[42] M. Ronaghi et al. "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release". In: *Analytical Biochemistry* 242.1 (1996), pp. 84–89. ISSN: 0003-2697. DOI: http://dx.doi.org/10.1006/abio.1996.0432 (cit. on p. 36).

[43] J. M. Rothberg and J. H. Leamon. "The development and impact of 454 sequencing". In: *Nat Biotech* 26.10 (Oct. 2008), pp. 1117–1124. ISSN: 1087-0156 (cit. on p. 36).

[44] N. Saitou and M. Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular Biology and Evolution* 4.4 (1987), pp. 406–425 (cit. on p. 32).

[45] A. Sali and T. L. Blundell. "Comparative protein modelling by satisfaction of spatial restraints." eng. In: *J Mol Biol* 234.3 (Dec. 1993), pp. 779–815. DOI: 10.1006/jmbi.1993.1626. URL: http://dx.doi.org/10.1006/jmbi.1993.1626 (cit. on p. 35).

[46] F. Sanger and A. Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of Molecular Biology* 94.3 (1975), pp. 441–448. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1016/0022-2836(75)90213-2 (cit. on p. 36).

[47] S. C. Schuster. "Next-generation sequencing transforms today's biology". In: *Nat Meth* 5.1 (Jan. 2008), pp. 16–18. ISSN: 1548-7091 (cit. on p. 36).

[48] J. Shendure and H. Ji. "Next-generation DNA sequencing". In: *Nat Biotech* 26.10 (Oct. 2008), pp. 1135–1145. ISSN: 1087-0156 (cit. on p. 36).

[49] T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences." eng. In: *J Mol Biol* 147.1 (Mar. 1981), pp. 195–197 (cit. on p. 30).

[50] B. Sobrino, M. Brión, and A. Carracedo. "SNPs in forensic genetics: a review on SNP typing methodologies". In: *Forensic Science International* 154.2–3 (2005), pp. 181–194. ISSN: 0379-0738. DOI: http://dx.doi.org/10.1016/j.forsciint.2004.10.020 (cit. on p. 36).

[51] R. R. Sokal and C. D. Michener. "A statistical method for evaluating systematic relationships". In: *University of Kansas Scientific Bulletin* 28 (1958), pp. 1409–1438 (cit. on p. 32).

[52] J. E. Stajich and H. Lapp. "Open source tools and toolkits for bioinformatics: significance, and where are we?" In: *Briefings in Bioinformatics* 7.3 (2006), pp. 287–296. DOI: 10.1093/bib/bbl026 (cit. on p. 38).

[53] A. Stoltzfus et al. "Publishing re-usable phylogenetic trees, in theory and practice". In: *iEvoBio 2011*. 2011. DOI: 10.1038/npre.2011.6048.1 (cit. on p. 29).

[54] G. J. Sussman and G. L. S. Jr. "Scheme: An interpreter for extended lambda calculus". In: *MEMO 349, MIT AI LAB*. 1975 (cit. on p. 38).

[55] S. Tavaré. "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". In: *American Mathematical Society: Lectures on Mathematics in the Life Sciences*. Vol. 17. Amer Mathematical Society, 1986, pp. 57–86. ISBN: 0821811673. URL: http://www.worldcat.org/isbn/0821811673 (cit. on p. 33).

[56] J. D. Velasco. "Phylogeny as population history". In: *Philosophy and Theory in Biology* 5.20150505 (2013). DOI: 10.3998/ptb.6959004.0005.002. URL: http://dx.doi.org/10.3998/ptb.6959004.0005.002 (cit. on p. 32).

[57] B. Wallner and A. Elofsson. "All are not equal: A benchmark of different homology modeling programs". In: *Protein Science* 14.5 (May 2005), pp. 1315–1327. DOI: 10.1110/ps.041253405. URL: http://dx.doi.org/10.1110/ps.041253405 (cit. on p. 35).

[58] J. F. Wendel and J. Doyle. "Molecular Systematics of Plants II". In: ed. by D. E. Soltis, P. S. Soltis, and J. J. Doyle. Boston: Springer US, 1998. Chap. DNA Sequencing, pp. 265–296. DOI: 10.1007/978-1-4615-5419-6. URL: http://dx.doi.org/10.1007/978-1-4615-5419-6 (cit. on p. 34).

PART

II

SEQFEATR AND STATISTICAL CONSIDERATIONS

The **SeqFeatR** web server for the discovery of viral CTL epitopes *

Budeus B.[1], Michalski M.[1] , Timm J.[2] , Hoffmann D.[1]

[1]Research Group Bioinformatics, Faculty of Biology, Center for Medical Biotechnology, University of Duisburg-Essen, Essen, Germany,
[2]Institute of Virology, University Hospital Dusseldorf, Dusseldorf, Germany

*as Feature

Use SEQFEATR for ...

... discovering statistically significant associations of the feature with the presence or absence of certain amino acids/nucleotides at certain sequence positions (for example HBV epitopes[1]).

https://seqfeatr.zmb.uni-due.de/

seqfeatr@uni-due.de

Link to webpage

What you see on the webpage

Input

Imagine the following alignment* of amino acid sequences in FASTA format, taken from 14 patients that either have a certain HLA type or do not have that feature.

count positions

```
>P01_HLA_A01_00_B01_02
LPDIQGNENMGYQPSWIFCGMETNGSQCLEEMFHCCWINC
>P02_HLA_A01_00_B01_02
MPDWNQKWGNDHLASINLD-WLKTIQQPGIEKHLRFYENW
>P03_HLA_A01_02_B01_02
VPDASGKHGIIGMDVTSSMERRHGMVQLPWPAMVWGRPHW
>P04_HLA_A01_00_B01_02
MPDVRGVGCARRDCLIVHRFCMPFNNQVYCKVWIVYWTYK
>P05_HLA_A01_00_B01_02
QPDTPKITRKEATAIHKCGIHWQTNCQKLSTVHPFHHQVD
>P06_HLA_A01_02_B01_02
SWDDFSDFTMVHQWYAQGTLGPYKAMQLKMIFQGVSIMEV
>P07_HLA_A01_02_B01_02
IPDEPCYCCVKNKILTVEIGVHHAKSQVRRNIDNIRRKTE
>P08_HLA_A04_03_B04_03
HPSTICPYIWKMYFTW-MGQKLVIQKVNGRTPPHCDECNQ
>P09_HLA_A04_03_B04_03
SNFTTTKLRDQHNLYP-AGLQEIEHKVDHQILGIYGQIWY
>P10_HLA_A04_03_B04_03
ETSTALRTQDQTFMLALRANYMVMLKVLDCISVKLFICWR
>P11_HLA_A04_03_B04_00
DSSTMDAECSTLQRFIWWHAHYAWIRVAKKPYCLDCPYAV
>P12_HLA_A04_03_B04_03
KKSTLGIARGIQRSHGWYWRQTHCVMVLTPSQHKMGEKSW
>P13_HLA_A04_03_B04_00
IQSTELCGCLINWPPMQWIVFAHMDDVNDSQTNTCDMRSQ
>P14_HLA_A04_03_B04_03
GRSTNARTMGGQDCAYMTHTLTKHIWVILAFDPIMIVHKP
```

*This bad alignment was chosen for one reason only: it demonstrates that it can be difficult to spot relationships between features and sequence positions.

Method

SeqFeatR will automatically discover in the sequences above a significant association between HLA*B01 and amino acid D at third sequence position:

- 7 sequences with HLA*B01 and D3
- 0 sequences with not HLA*B01 and with D3
- 0 sequences with HLA*B01 and not D3
- 7 sequences with not HLA*B01 and not D3

HLA-B*01

D3	+	-
+	7	0
-	0	7

Fisher's exact test yields a p-value < 0.001 and we have an OR of infinity. Thus we have a significant and strongly positive association of HLA*B01 and D3.

Type of sequences

Nucleotides / Amino acids

Input files

Sequence alignment to analyse (FASTA):* Durchsuchen Example_AA.fasta

Download example input

Known epitopes (CSV): Durchsuchen No file selected

Reference sequence (FASTA): Durchsuchen No file selected

Known binding motifs (CSV): Durchsuchen No file selected

Features

One feature / HLA types

Position of feature HLA-A:* 10 - 11 13 - 14

Position of feature HLA-B:* 17 - 18 20 - 21

Graphical output

Height of horizontal line: 0.01

Hight of star level: 0.5

Window size: 9

Further options

Minimal number of members: 1

P-value correction: bonferroni

Check for phylogenetic bias? yes / no

Matrix:* BLOSUM62

Results are stored online for 72 hours

Output

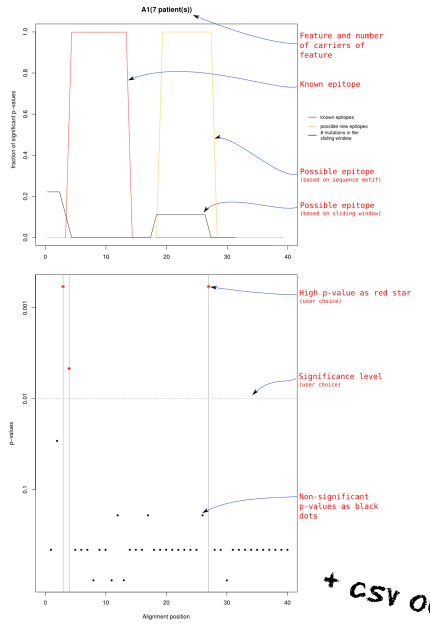


Figure A: Manhattan plot (p-values along sequence).

The so-called Manhattan plot, is a convenient means to discover significant associations of sequence alignment positions with features. SeqFeatR produces Manhattan plots consisting of two separate plots (*Figure A*):

The top half of the plot focuses on complete epitopes or putative epitopes comprising "windows" of several sequence positions (e.g. windows of 9 positions), while the bottom half gives a more detailed picture of the same data at the level of single sequence positions. The x-axis for both plots is the same, namely the positions in the input sequence alignment.

SeqFeatR can mark a significance level α (here: α = 0.01) with a horizontal line. Associations with $-\log_{10}$ p-values above that line (i.e. p-values < α) are shown with a special symbol (here: red stars) and considered significant.
To ease the visual localization of the highly significant positions, they are additionally marked with vertical lines that hit the sequence axis at the corresponding positions. (The resolution of the x-axis is usually to coarse to show single positions, but sufficient to localize significantly associated positions in the fully resolved csv-file which is given as second output file.)

If the features in your SeqFeatR input have been HLA types, and if you then see several sequence positions in close proximity showing up with high $-\log_{10}$ p values in the Manhattan plot, you may have found a HLA epitope.

+ csv output

Additional Input

The red curve (*optional*) allows the user to mark known epitopes, e.g. published in the literature or in a database. Here an example of what you could put into such a csv file:
4;12;A1
2;9;A3
This example marks two known epitopes with two lines of the form EpitopeStart; EpitopeStop; HLAtype. The bump is only shown if the HLA type in the csv file matches the HLA type in the sequence alignment.

With the yellow curve (*optional*) the user can mark alignment regions that are conforming with given sequence patterns:

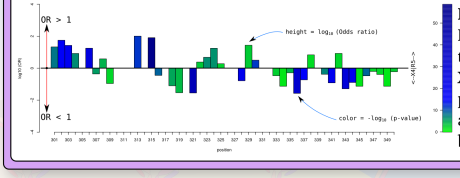
Genotype;Motif;Reference
A*01;x[PV]xxxx[DENQ]EN;SYFPEITHI

The header line (Genotype; Motif; Reference) describes the structure of the following lines.
The motif shown here covers nine amino acid positions. Here the nine amino acid positions are shown as indices:

$$x_1\ \underbrace{[P\ V\ ]}_{2}\ x_3\ x_4\ x_5\ x_6\ \underbrace{[DENQ]}_{7}\ E_8\ N_9$$

The letter x stands for: "this could be any amino acid". The two square brackets at positions 2 and 7 show which amino acids are allowed at these two positions.

Further possibilities - inside R-package



Example odds ratio plot. Here we analyzed amino acid sequences of HIV-1 gp120 protein variants, and we had as feature the co-receptor tropism of HIV-1, which can be R5 or not R5 (the latter is often called X4). The odds ratio (OR) plot shows for each sequence alignment position the association strength $\log_{10}$(OR) as bar height and the p-value as bar color. The plot demonstrates that high values of $\log_{10}$OR (long bars) and high statistical significance (blue color) are not the same.

[1] **Adaptation of the hepatitis B virus core protein to CD8+ T cell selection pressure**
Helenie Kefalakes, Bettina Budeus, Andreas Walker, Christoph Jochum, Gudrun Hilgard, Andreas Heinold, Falko Heinemann, Guido Gerken, Daniel Hoffmann, Joerg Timm

Funding: This work was supported by Deutsche Forschungsgemeinschaft [TRR60/B1 to J.T. and D.H.].

SeqFeatR for the discovery of feature-sequence associations

Goals are dreams with deadlines.

DIANA SCHARF

Abstract

Specific selection pressures often lead to specifically mutated genomes. The open source software SeqFeatR has been developed to identify associations between mutation patterns in biological sequences and specific selection pressures ("features"). For instance, SeqFeatR has been used to discover in viral protein sequences new T cell epitopes for hosts of given HLA types. SeqFeatR supports frequentist and Bayesian methods for the discovery of statistical sequence-feature associations. Moreover, it offers novel ways to visualize results of the statistical analyses and to relate them to further properties. In this article we demonstrate various functions of SeqFeatR with real data. The most frequently used set of functions is also provided by a web server.

SeqFeatR is implemented as R package and freely available from the R archive CRAN (http://cran.r-project.org/web/packages/SeqFeatR/index.html). The package includes a tutorial vignette. The software is distributed under the GNU General Public License (version 3 or later). The web server URL is https://seqfeatr.zmb.uni-due.de.

This chapter is based on the following publication:

Bettina Budeus, Jörg Timm, and Daniel Hoffmann (2015). **SeqFeatR for the discovery of feature-sequence associations.**

http://www.plosone.org/article/related/info%3Adoi%2F10.1371%2Fjournal.pone.0146409

PLOS | ONE

RESEARCH ARTICLE

SeqFeatR for the Discovery of Feature-Sequence Associations

Bettina Budeus[1], **Jörg Timm**[2], **Daniel Hoffmann**[1]*

1 Research Group Bioinformatics, Faculty of Biology, University of Duisburg-Essen, Essen, NRW, Germany, **2** Institute for Virology, University Hospital Düsseldorf, Düsseldorf, NRW, Germany

* daniel.hoffmann@uni-due.de

Abstract

Specific selection pressures often lead to specifically mutated genomes. The open source software SeqFeatR has been developed to identify associations between mutation patterns in biological sequences and specific selection pressures ("features"). For instance, SeqFeatR has been used to discover in viral protein sequences new T cell epitopes for hosts of given HLA types. SeqFeatR supports frequentist and Bayesian methods for the discovery of statistical sequence-feature associations. Moreover, it offers novel ways to visualize results of the statistical analyses and to relate them to further properties. In this article we demonstrate various functions of SeqFeatR with real data. The most frequently used set of functions is also provided by a web server. SeqFeatR is implemented as R package and freely available from the R archive CRAN (http://cran.r-project.org/web/packages/SeqFeatR/index.html). The package includes a tutorial vignette. The software is distributed under the GNU General Public License (version 3 or later). The web server URL is https://seqfeatr.zmb.uni-due.de.

Citation: Budeus B, Timm J, Hoffmann D (2016) SeqFeatR for the Discovery of Feature-Sequence Associations. PLoS ONE 11(1): e0146409. doi:10.1371/journal.pone.0146409

Editor: I. King Jordan, Georgia Institute of Technology, UNITED STATES

Received: August 19, 2015

Accepted: December 15, 2015

Published: January 5, 2016

Data Availability Statement: HBV sequences are available from GenBank (accession numbers KP856971-KP857118).

Funding: Funding for this work was provided by Deutsche Forschungsgemeinschaft (http://www.dfg.de), grant TRR 60 / B1 to JT and DH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

There is a widening gap between the surge of information rich sequence data, and the human resources available for analysis. This is a problem that severely hampers progress in biomedicine and other life sciences [1, 2]. Ideally, experimental or clinical researchers who are most familiar with and interested in their data should be enabled to analyze their data by themselves. While software for statistics and graphics, such as R [3] (http://www.R-project.org/), are freely available and well-suited for such analyses, the steep slope of the learning curve is often discouraging experimental and clinical researchers, who are fully occupied with managing experiments or clinical duties. A general and relevant field where this disparity has been expressed to the authors by clinical researchers, especially immunologists and virologists, is the association of features of clinical interest with sequences. A concrete example is the association of patients' HLA (Human Leukocyte Antigen) types with substitutions in a viral protein sequenced from these patients, as a way of identifying T-cell epitopes and immune escape mutations [4]. There are powerful computational tools for the identification of associations between sequences and features, for instance in the domain of genome wide association studies [5], or next-generation

sequencing exome or genome comparisons [6, 7], but these tools are optimized for specific application scenarios and not for ease of use in experimental or clinical laboratory settings.

We have developed the R-package SeqFeatR to allow experimental and clinical researchers easier access to the statistical and graphical capabilities of R for feature-sequence association studies. R was chosen since it is a powerful, free, open source suite that is available for all commonly used computing platforms.

SeqFeatR has been successfully introduced in several virological labs, and sequence-feature associations identified with SeqFeatR have been experimentally confirmed, as in the case of novel $CD8^+$ T-cell epitopes in HCV [8], or compensatory substitutions outside such epitopes [9].

These published examples have used the feature "HLA type" and amino acid sequences. However, SeqFeatR is completely agnostic about the type of feature used, as long as it can be labeled unequivocally, and it also processes nucleotide sequences. Both will be demonstrated in section "Examples beyond HLA-sequence association".

Core functionality of SeqFeatR

Given a set of related nucleotide or amino acid sequences, such as variants of a gene from several patients with certain phenotypes, SeqFeatR discovers in those sequences positions that are statistically associated with a "feature", for instance with one of the patient phenotypes. An example of a feature of great clinical importance is the HLA type of a patient. In a patient infected with highly variable virus, such as HIV, HCV, or HBV, the HLA system of that patient selects viral variants with immune escape mutations. Thus we can expect that mutations in viral genome sequences are associated with the patient feature "HLA type". SeqFeatR detects such associations, in other words: it finds among all alignment positions those that have a statistically significant association with a given feature. Analogous to the association of single alignment positions vs. features, SeqFeatR allows for the screening of associations of position pairs or tuples with sequence features, though at higher computational cost.

Technically, SeqFeatR reads FASTA formatted multiple sequence alignments, with each sequence labeled in its header line with the name of the feature, for instance the HLA type of the patient from whom the respective viral sequence has been extracted. The alignment should contain sequences that are positive for the feature of interest, and sequences that are negative. SeqFeatR steps through all alignment columns and applies frequentist or Bayesian methods to detect associations with the feature.

SeqFeatR itself does not implement alignment functionality, since there are many excellent programs for multiple sequence alignments that can be used to turn sequence sets into multiple sequence alignments, for instance MAFFT [10], T-Coffee [11], or Clustal omega [12].

Frequentist approach

In the frequentist approach used in SeqFeatR, Fisher's exact tests [13] are applied to contingency tables for all letters of the relevant alphabet (amino acid or nucleic acids) at all alignment positions vs. sequence features. This most frequently requested type of analysis is fast and also provided by the SeqFeatR web server. Logarithmically scaled p values are plotted along the alignment with single position resolution (Manhattan plots), or averaged over epitope sized windows. These association analyses are potentially affected by high numbers of false positives due to multiple testing. Therefore, SeqFeatR offers methods for multiple testing corrections, from the very conservative Bonferroni correction to the more relaxed control of False Discovery Rates (FDRs) [14].

Beyond Manhattan plots, SeqFeatR provides some novel visualization tools for advanced exploratory analyses, for instance an odds-ratio plot that simultaneously shows, along a

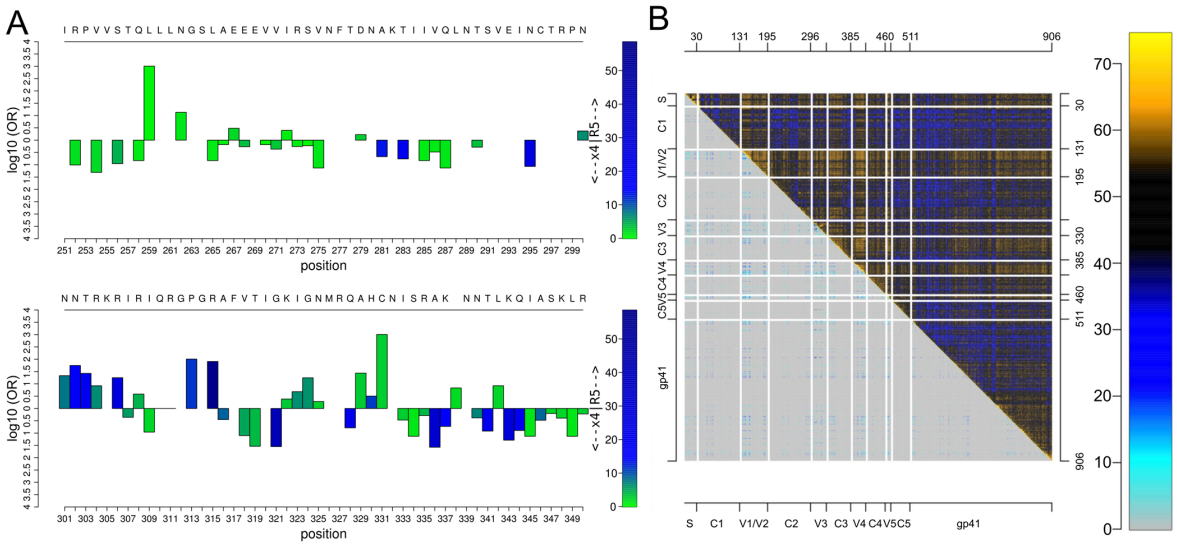


Fig 1. Odds-ratio plot and Tartan plot for visualization of statistical associations. A Odds-ratio plot, based on an alignment of region of HIV-1 gp120 around the V3 loop (C296-C331). Here, the feature is the predicted co-receptor tropism of HIV-1 [17] (R5 vs. X4 tropic). Bar heights and colors indicate logarithms of odds ratios and negative logarithms of *p* values, respectively. A reference sequence and sequence positions can be added in the top and bottom rows for orientation. **B** Tartan plot for the synopsis of two alignment pair association measures, here: –log *p* from association test between alignment position pairs (upper right triangle) vs. Direct Information between these pairs (lower left triangle). Association strengths are color coded (color legend on the right). For orientation, axes can be annotated and sequence substructures can be indicated by lines.

doi:10.1371/journal.pone.0146409.g001

sequence, odds-ratios and *p* values as two aspects of association strength (Fig 1A). Another new visualization tool is the "Tartan plot" for a synopsis of two arbitrary scalar measures of sequence position pair association, e.g. (in Fig 1B) – log *p* from statistical association testing of amino acids at each pair (*i*, *j*) of alignment positions vs. the Direct Information between *i*, *j* [15, 16]. The synoptic plotting quickly reveals structure in such data, such as in Fig 1B the strong association of V1/V2 loops of HIV-1 gp120 protein with the other variable loops and parts of gp41, both in terms of *p* values from amino acid pair-association tests, and the more refined Direct Information.

Bayesian approaches

While the frequentist approach works well in many cases, it has also drawn criticism, for instance because *p* values are often abused or misinterpreted [18]. Another problem with the frequentist approach occurs in situations where the same test is applied to multiple hypotheses, such as testing for associations with phenotype features for all positions in a multiple sequence alignment. As mentioned before, it is customary to "correct" the *p* values, e.g. by the very conservative Bonferroni correction or other more liberal alternatives, to avoid an increasing number of false positive tests results. As a more consistent alternative to deal with these problems, SeqFeatR offers also Bayesian inference methods [19], namely Bayes factors (BFs) and hierarchical models (though these are posing other problems, such as the necessity to specify priors). In the following we describe the implemented BF approach. For the hierarchical models we only mention that the SeqFeatR R-package has an interface to the Gibbs sampling engine JAGS [20]; a detailed account of hierarchical models for sequence feature association analyses will be given in a separate publication.

The BF for two hypotheses H_0 and H_1, given sequence and feature data D, is the ratio of posterior odds and the corresponding prior odds: $BF = (p(H_1|D)/p(H_0|D))/(\pi_1/\pi_0)$. In other words, the BF equals the posterior odds ratio if the prior probabilities π_0, π_1 are equal and thus the prior odds ratio is 1. In our case H_1 is the hypothesis that a feature is associated with an amino acid or nucleotide at an alignment position, and H_0 is the hypothesis that there is no such association. The higher the BF, the more likely H_1 (association) and the less likely H_0 (no association). If the prior probability of association π_1 is known, the ratio of posterior probabilities of association over non-association can be computed as $BF \cdot \pi_1/(1 - \pi_1)$.

Here we use a BF for the hypothesis H_1 that feature and amino acid at an alignment position are *close* to independence vs. H_0 that they are independent. A model H_1 close to independence will often be more relevant than a "uniform model" that, for instance, assumes a uniform distribution of contingency table cell probabilities. Albert *et al.* have derived a BF expression for the ratio of a close-to-independence model over an independent model based on Dirichlet distributed elements of contingency tables [21, 22]:

$$BF_K(\{y_{rc}\}) = \frac{\int \frac{\mathrm{Dir}(\{K\eta_r\eta_c + y_{rc}\})}{\mathrm{Dir}(\{K\eta_r\eta_c\})} d\{\eta_r\} d\{\eta_c\}}{\mathrm{Dir}(\{y_r + 1\})\mathrm{Dir}(\{y_c + 1\})}, \quad (1)$$

where y_{rc} are the observed contingency table counts with row index r and column index c; $\mathrm{Dir}(\{\alpha_i\}) = 1/B(\{\alpha_i\})\prod_i p_i^{\alpha_i - 1}$ is the Dirichlet distribution of probabilities p_i (here: probabilities of contingency table elements) with normalizing multinomial Beta function B and concentration parameters α_i; y_r, y_c are the row and column sums of the observed contingency table; K is a precision hyperparameter; η_r, η_c are hyperparameters corresponding to probabilities of row r and column c of tables with row-column independence. Curly brackets indicate that we have sets of two or more parameters. For instance, in the case of a 2 × 2 contingency table (amino acid present or absent at an alignment position versus feature present or absent), the Dirichlet distributions in the integrand depend on four parameters (two columns, two rows) and the integration therefore runs over four parameters. The prior belief in the independence is expressed by K: the higher this hyperparameter, the more dominant the independence structure imposed by η_r η_c will be in comparison to the observed counts y_{rc} in the numerator of Eq (1), and for $K \to \infty$ complete independence is achieved. The BF is computed numerically as an average by importance sampling of Eq (1) using η_r, η_c values that are randomly drawn from a Dirichlet distribution with concentration parameters evaluated from the entries y_{rc} of the observed contingency table. The procedure is detailed in Ref. [23]. $BF_K(\{y_{rc}\})$ is reported by SeqFeatR.

While SeqFeatR allows for setting an explicit K value, it may not be easy to specify an appropriate value of K that is applicable to all alignment positions. In such cases, a new empirical Bayes variant of this BF is convenient. In this variant, an individual value of K is estimated from each contingency table itself. To derive this value, we first acknowledge that the sum S of absolute values of differences between the actually observed counts in the contingency table and the counts expected under independence is a measure of how confident we are that columns and rows are *dependent*:

$$S = \sum_{rc} \left| y_{rc} - \frac{\sum_k y_{rk} \sum_k y_{kc}}{N} \right|, \quad (2)$$

with total table count $N = \sum_{rc} y_{rc}$. Clearly, for perfectly independent rows and columns, the value of S reaches its minimum of zero. The maximum of $S = N$ is attained for strong dependence of rows and columns, for instance for a 2 × 2 table with $y_{11} = y_{22} = N/2 = n$ and $y_{12} = y_{21} = 0$. To

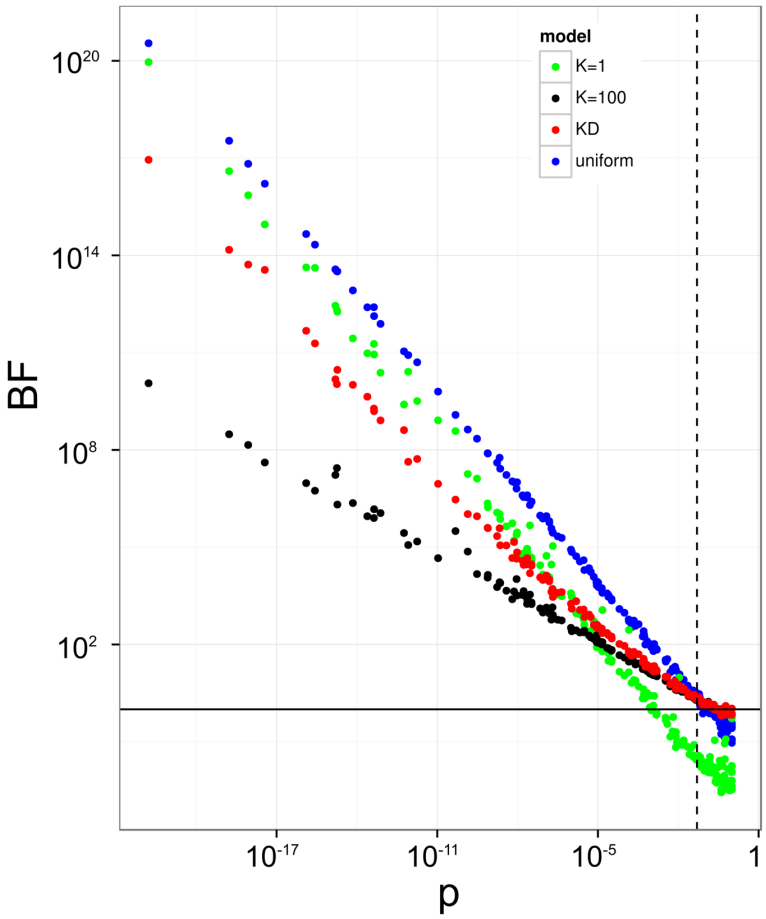


Fig 2. Comparison of statistical indicators of association. 200 random contingency tables with total count $N = 100$, a typical order of magnitude for analyses of sequence-feature association in practice, are analyzed by Fisher's exact test, yielding p values for the rejection of independence (horizontal axis, not corrected for multiple testing), and by four different BF models, namely $K = 1$, $K = 100$, K_D, and uniform model, with corresponding BFs on vertical axis. Solid horizontal black line at $BF = 1$ and dashed vertical line at $p = 0.05$ for orientation.

doi:10.1371/journal.pone.0146409.g002

recast S into a measure of prior belief in *independence* we use as precision hyperparameter in Eq (1) instead of K the difference K_D:

$$K_D = N - S. \tag{3}$$

A simple interpretation of K_D is that if all N counts in the contingency table support independence, we have $S = 0$ and therefore $K_D = N$ (maximum prior belief in independence), while if all counts support association, we have $S = N$ and therefore $K_D = 0$ (minimum prior belief in independence). SeqFeatR also offers the option of using K_D.

Fig 2 shows that for contingency tables for which independence cannot be rejected as indicated by $p \approx 1$ from Fisher's exact test (lower right corner), K_D and K_{100} yield approximately

the same $BF \approx 1$, i.e. association and independence are given approximately equal weights. In this corner, the uniform model and even more so the model with low confidence in independence ($K = 1$) have BFs much closer to zero, both favoring independence over dependence. At the other end of the p value range, on the left side of the plot, the low p values lead to rejection of independence, and concordant with this, high BFs that favor association over independence. Here, the increase of BFs in the K_D model follows those of lower K models. Effectively, the K_D model suppresses noise by collapsing weak-association cases to $BF \approx 1$ (similar to high K models), while it readily supports stronger associations (similar to low K or uniform models).

Comparison of frequentist and Bayesian approaches for discovery of HLA escape substitutions

Recently, we have reported the discovery and experimental confirmation of several HLA escape substitutions in Hepatitis B Virus (HBV) from chronically infected patients [24] (sequences available from GenBank, accession numbers KP856971-KP857118). In that report, we had used SeqFeatR with the frequentist approach for the discovery. In Fig 3 we compare the latter approach (without correction for multiple testing) and Bayes factors with precision hyperparameters $K = 1$ and K_D. For this comparison, we have chosen two significant associations identified in Ref. [24], namely the strongest (alignment position 66 with HLA type A*01, corresponding to position 38 of HBV core protein reference) and the weakest (alignment position 96 with HLA type B*44, corresponding to position 67 of HBV core protein reference).

For all three analyses of the association with HLA A*01, alignment position 66 clearly sticks out with extremely small p value and high values of $BF_{K=1}$, and BF_{K_D} (top row of Fig 3). A frequentist would not seriously consider any other position as associated with this HLA, and most of the positions have $p \approx 1$. For $K = 1$ we have a wide spread around $BF = 1$, or $\log_{10} BF = 0$. Two BFs other than at position 66 lie slightly above $BF = 10$ (or $\log_{10} BF = 1$), a threshold often used to mark "substantial" evidence [25]. However, in contrast to the BF at position 66, these two BFs are not clearly separated from the bulk of the other BFs. Towards lower BFs, many values reach down to 10^{-2} or lower, indicating preference for independence over association at these alignment positions. For K_D we see the noise suppression mentioned earlier as the spread of the low BFs is constrained to a much smaller range than for $K = 1$.

For feature HLA B*44 we had only about 21 sequences (compared to 41 for HLA*A01), leading to a weaker association signal (bottom row of Fig 3). Still, the frequentist analysis shows position 96 with a p value that is clearly separated from the rest (panel D). However, a Bonferroni correction collapses all p values to 1, while the FDR correction collapses all to 1, except for position 96 with a corrected value of 0.16 (S1 Fig). The BFs with $K = 1$ do not favor association at any position (panel E). Conversely, for K_D position 96 has a clearly elevated BF (panel F). In summary, the frequentist approach with a strict correction for multiple comparisons, or the BF approach with $K = 1$ would both have led to a missing of the experimentally validated association at position 96, while the frequentist approach without correction, or BF_{K_D} both identify this association.

Detection of phylogenetic bias

Sequences analyzed with SeqFeatR can often be considered samples from different branches of the same phylogenetic tree, evolved from a common ancestor under selection pressure related to the "feature". A good example are again viral genome sequences evolved under selection pressure by the HLA systems (= features) of infected persons [8, 9, 24]. Under these circumstances, it is possible that SeqFeatR reports apparent sequence-feature associations that are due to a phylogenetic bias in the data. For instance, consider transmission of a virus from a mother

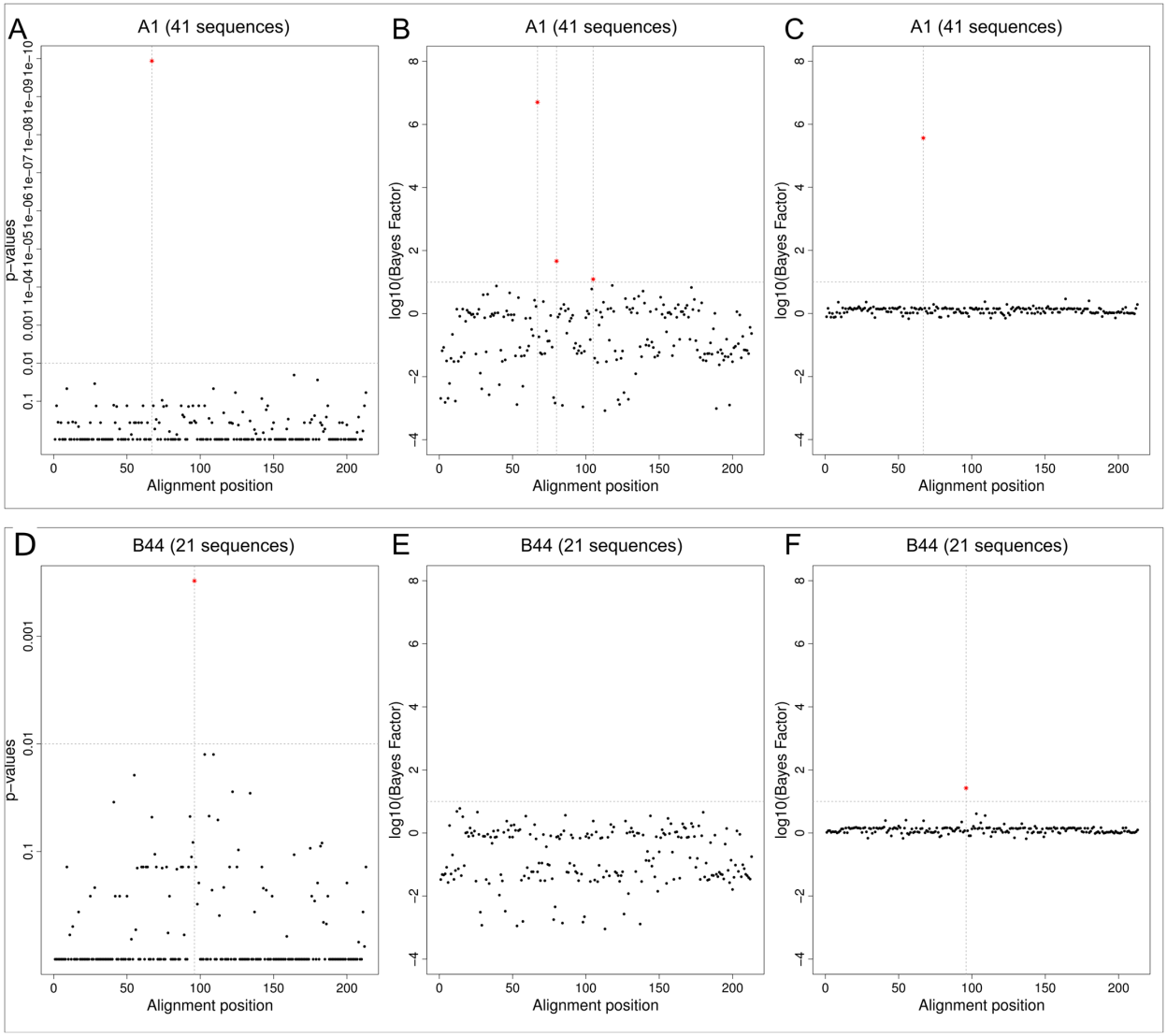


Fig 3. Comparison of frequentist approach and Bayes factors (BF). Discovery of association of alignment positions of HBV core proteins with patient HLA types, here: A*01 (top row) and B*44 (bottom row). Sequence numbers in panel titles are feature-carrying fractions of the total of 148 sequences included in the alignment. Association of sequences with feature HLA were analyzed by Fisher's exact test (panels A, D), BF with $K = 1$ (panels B, E), and BF with K_D (panels C, F). Alignment positions with association above certain thresholds (horizontal dashed lines) are marked by red stars and vertical dashed lines, namely $p < 0.01$ (A, D), or $BF > 10$ (B, C, E, F). The p values and BFs shown are the best for each alignment position (lowest p values, highest BFs).

doi:10.1371/journal.pone.0146409.g003

to several children, all having the same HLA type. In this case, not only HLA escape mutations of viral proteins are associated with this HLA type, but apparently also mutations specific to the founder virus of the mother that are transmitted to the children, but unrelated to the HLA type. A mutation of a viral protein that is really associated with HLA type should co-occur with the HLA in other parts of the phylogenetic tree (i.e. outside this mother-child transmission), while this repeated co-occurrence is less likely for mutations that, due to phylogenetic bias, are only apparently associated with HLA.

SeqFeatR computes a simple quantitative indicator B of the strength of the phylogenetic bias for a given feature as follows. We expect that a phylogenetic bias is likely, if evolutionary distances within the group of sequences that carry the feature are much smaller than typical evolutionary distances in the total set of analyzed sequences. Thus, we define B as

$$B = 1 - \frac{\langle d_{ij} \rangle_{feature}}{\langle d_{ij} \rangle_{all}}, \quad (4)$$

where d_{ij} is the Levenshtein distance between sequences i and j. The ratio gives the mean distance between sequences carrying the feature over the mean distance in the total sequence sample. B lies then between values that typically are close to zero or even become negative for low bias, and a maximum of 1 for the strongest bias. For instance, in Fig 4E, feature-carrying sequences are spread out over different parts of the phylogenetic tree of all sequences in the sample, and consistent with this $B = 0.05$ signals low bias. Conversely, in Fig 4B feature-carrying leaves are concentrated in a sub-tree, and $B = 0.26$ indicates higher bias.

If detection of specific substitutions is desired that are associated with the feature, and not due to phylogenetic bias, a high B suggests extension of the set of sequences, especially with evolutionarily less closely related sequences that carry the feature.

Examples beyond HLA-sequence association: HIV-1 co-receptor tropism and genetic species differences

In the above examples we have focused on the HLA type as feature and amino acid sequences. However, SeqFeatR is agnostic about the type of feature and sequence and therefore can be applied to other features and nucleotide sequences, too. To illustrate this we give in the following two examples.

HIV-1 co-receptor tropism. The Human Immunodeficiency Virus 1 (HIV-1) enters cells after contact with the cellular receptor CD4 and one of two co-receptors, either CCR5 or CXCR4 [26]. The choice of the co-receptor (or "co-receptor tropism") is encoded in the viral genome, specifically in the third variable loop (V3) of the viral glycoprotein 120 [27]. Since the co-receptor tropism has implications for prognosis [28] and therapy [29], its determination from V3 sequence has attracted a lot of interest. Here we demonstrate that SeqFeatR recovers V3 sequence patterns known to be associated with co-receptor tropism.

To simplify the alignment, we used only V3 sequences of 35 amino acids (S1 Alignment), the by far most frequent length, from a dataset published earlier [30]. This led to 84 V3 sequences of CXCR4-tropic virus and 928 V3 sequences of CCR5-tropic virus. We then applied SeqFeatR with co-receptor tropism as feature. The resulting Manhattan plot (S2 Fig) shows many positions with highly significant deviations between CXCR4- and CCR5-tropic virus. One of the patterns recognized early on as specific for CXCR4 is the occurrence of positively charged amino acids at positions 11 and 25, the so-called 11/25 rule [31]. In fact, in the SeqFeatR output both positions 11 and 25 have significant deviations between CXCR4- and CCR5-tropic virus with p-values less than 10^{-4}. Inspection of the alignment confirms that in

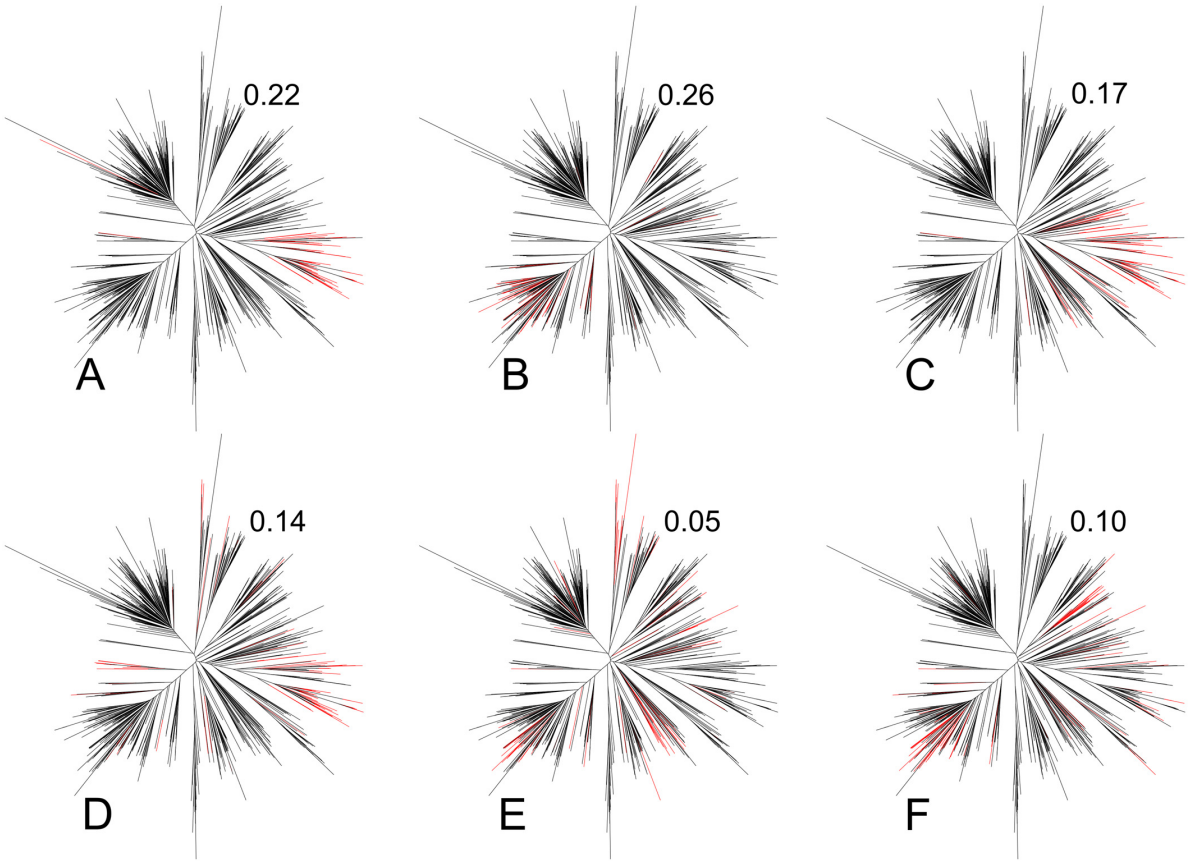


Fig 4. Phylogenetic distribution of feature-carrying sequences and phylogenetic bias indicator *B*. The distance-based phylogenetic tree in all six panels was computed for the same set of 788 East Asian HIV-1 gag protein sequences obtained from the HIV sequence database at http://www.hiv.lanl.gov. In each panel, those branches are colored red that correspond to sequences that carry an amino acid substitution apparently associated with a certain HLA type. The numbers to the upper right of each tree are the corresponding values of the bias indicator *B*, Eq (4).

doi:10.1371/journal.pone.0146409.g004

CXCR4-tropic virus both positions 11 and 25 are significantly enriched in positively charged amino acids Arginine and Lysine in comparison to CCR5-tropic virus.

Genetic species differences. SeqFeatR can be used to discover genetic differences between species or other taxonomic levels. For the following example we have retrieved from the SILVA database [32], version 123, RNA sequences of the small ribosomal subunit (SSU) of two closely related green algae, *Chlamydomonas applanata* (9 sequences) and *Chlamydomonas reinhardtii* (10 sequences). The input alignment is provided as S2 Alignment. Using these two species as features, we found with SeqFeatR 29 positions with highly significant differences (red stars in Manhattan plot S3 Fig). Nucleotide sequence differences such as these can be used to understand genetic bases of species differences or to design species specific PCR primers [33].

SeqFeatR addresses various needs and levels of expertise

SeqFeatR has three modes of use, addressing users with different levels of expertise and different needs: For users not versed in R programming and with sequence material and features

that can be transmitted over the Internet, we offer the SeqFeatR web server. For reproducibility and documentation, the web server generates a detailed report for the user. If the data must not leave the respective institution, inexperienced users may still use a simple Tcl/Tk-based graphical user interface (GUI) that can be started by the SeqFeatR_GUI() command from R. Experienced users can access the full range of SeqFeatR commands in R-scripts. Training material such as tutorial texts (https://cran.r-project.org/web/packages/SeqFeatR/vignettes/SeqFeatR_tutorial.pdf) and videos are provided for users at all levels.

Supporting Information

S1 Fig. Frequentist approach with correction for multiple testing. Association of alignment positions of HBV core protein with patient HLA types A*01 (**A**) and B*44 (**B**). Sequence numbers in panel titles are feature-carrying fractions of the total of 148 sequences included in the alignment. Association of sequences with feature HLA were analyzed with Fisher's exact test, and resulting *p* values were corrected for multiple testing with FDR option.
(TIFF)

S2 Fig. Association of V3 sequence positions with HIV-1 co-receptor tropism. Manhattan plot output of SeqFeatR showing sites in the V3 amino acid sequences S1 Alignment that are significantly associated with co-receptor tropism.
(PDF)

S3 Fig. Association of *Chlamydomonas* SSU nucleotide sequence position with species. Manhattan plot output of SeqFeatR showing sites in the SSU nucleotide sequence alignment S2 Alignment that are significantly associated with *Chlamydomonas* species, here: *Chlamydomonas reinhardtii* (RH) vs *Chlamydomonas applanata* (AP).
(PDF)

S1 Alignment. V3 amino acid sequences of CCR5- and CXCR4-tropic HIV-1. S2 Fig was produced by SeqFeatR with this input. All sequences (84 from CXCR4-tropic and from 928 CCR5-tropic virus) have the same length of 35 amino acids and have not been submitted to an extra alignment step. Note that the feature labels "X4" (for CXCR4-tropic) and "R5" (for CCR5-tropic) have been added at the end of the FASTA headers after a semicolon.
(FA)

S2 Alignment. Alignment of SSU nucleotide sequences from *Chlamydomonas*. Alignment of RNA sequences of small ribosomal subunit sequences: 9 from *Chlamydomonas applanata*, 10 from *Chlamydomonas reinhardtii*. S3 Fig was generated by SeqFeatR with this input. Note again that the last element of the FASTA header stands for the feature, here: RH for *reinhardtii* and AP for *applanata*.
(FA)

Acknowledgments

We thank Michael Michalski for technical assistance with the web server.

Author Contributions

Performed the experiments: BB DH. Analyzed the data: BB JT DH. Wrote the paper: BB JT DH. Selected and devised methods: DH. Developed software: BB. Provided data: JT.

References

1. Schmidt C. Cancer: Reshaping the cancer clinic. Nature. 2015 Nov; 527(7576):S10–S11. Available from: http://dx.doi.org/10.1038/527S10a PMID: 26536216
2. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genet Med. 2013; 15:802–809. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3906918/ doi: 10.1038/gim.2013.121 PMID: 24008998
3. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: https://www.R-project.org/
4. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science. 2002 May; 296(5572):1439–43. doi: 10.1126/science.1069660 PMID: 12029127
5. Rentería ME, Cortes A, Medland SE. Using PLINK for Genome-Wide Association Studies (GWAS) and data analysis. Methods Mol Biol. 2013; 1019:193–213. doi: 10.1007/978-1-62703-447-0_8 PMID: 23756892
6. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar; 22 (3):568–76. doi: 10.1101/gr.129684.111 PMID: 22300766
7. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics. 2013 Jun; 29(12):1498–503. doi: 10.1093/bioinformatics/btt183 PMID: 23620360
8. Ruhl M, Knuschke T, Schewior K, Glavinic L, Neumann-Haefelin C, Chang DI, et al. CD8(+) T-cell response promotes evolution of hepatitis C virus nonstructural proteins. Gastroenterology. 2011 Jun; 140(7):2064–73. doi: 10.1053/j.gastro.2011.02.060 PMID: 21376049
9. Ruhl M, Chhatwal P, Strathmann H, Kuntzen T, Bankwitz D, Skibbe K, et al. Escape from a dominant HLA-B*15-restricted CD8+ T cell response against hepatitis C virus requires compensatory mutations outside the epitope. J Virol. 2012 Jan; 86(2):991–1000. doi: 10.1128/JVI.05603-11 PMID: 22072759
10. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul; 30(14):3059–66. doi: 10.1093/nar/gkf436 PMID: 12136088
11. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302(1):205–17. doi: 10.1006/jmbi.2000.4042 PMID: 10964570
12. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol. 2014; 1079:105–116. Available from: http://dx.doi.org/10.1007/978-1-62703-646-7_6 PMID: 24170397
13. Fisher RA. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society. 1922; 85(1):87–94. Available from: http://www.jstor.org/stable/2340521 doi: 10.2307/2340521
14. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57(1):289–300. Available from: http://dx.doi.org/10.2307/2346101
15. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 2011 Dec; 108(49):E1293–301. doi: 10.1073/pnas.1111471108 PMID: 22106262
16. Wang Y, Rawi R, Wilms C, Heider D, Yang R, Hoffmann D. A Small Set of Succinct Signature Patterns Distinguishes Chinese and Non-Chinese HIV-1 Genomes. PLoS One. 2013; 8(3):e58804. doi: 10.1371/journal.pone.0058804 PMID: 23527028
17. Heider D, Dybowski JN, Wilms C, Hoffmann D. A simple structure-based model for the prediction of HIV-1 co-receptor tropism. BioData Min. 2014; 7:14. Available from: http://dx.doi.org/10.1186/1756-0381-7-14 PMID: 25120583
18. Nuzzo R. Scientific method: statistical errors. Nature. 2014 Feb; 506(7487):150–152. Available from: http://dx.doi.org/10.1038/506150a PMID: 24522584
19. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet. 2009 Oct; 10(10):681–90. doi: 10.1038/nrg2615 PMID: 19763151
20. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003); 2003.
21. Albert JH, Gupta AK. Mixtures of Dirichlet Distributions and Estimation in Contingency Tables. Ann Statist. 1982 12; 10(4):1261–1268. Available from: http://dx.doi.org/10.1214/aos/1176345991

22. Albert JH. A bayesian test for a two-way contingency table using independence priors. Canadian Journal of Statistics. 1990; 18(4):347–363. Available from: http://dx.doi.org/10.2307/3315841

23. Albert J. Bayesian Computation with R. Springer Verlag; 2009.

24. Kefalakes H, Budeus B, Walker A, Jochum C, Hilgard G, Heinold A, et al. Adaptation of the hepatitis B virus core protein to CD8(+) T-cell selection pressure. Hepatology. 2015 Feb;Available from: http://dx.doi.org/10.1002/hep.27771 PMID: 25720337

25. Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995; 90 (430):773–795. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

26. D'Souza MP, Harden VA. Chemokines and HIV-1 second receptors. Confluence of two fields generates optimism in AIDS research. Nat Med. 1996 Dec; 2(12):1293–300. doi: 10.1038/nm1296-1293 PMID: 8946819

27. Hwang SS, Boyle TJ, Lyerly HK, Cullen BR. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science. 1991 Jul; 253(5015):71–4. doi: 10.1126/science.1905842 PMID: 1905842

28. Koot M, Keet IP, Vos AH, de Goede RE, Roos MT, Coutinho RA, et al. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. Ann Intern Med. 1993 May; 118(9):681–8.

29. Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, et al. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. Antimicrob Agents Chemother. 2005 Nov; 49(11):4721–32. doi: 10.1128/AAC.49.11.4721-4732.2005 PMID: 16251317

30. Dybowski JN, Heider D, Hoffmann D. Prediction of co-receptor usage of HIV-1 from genotype. PLoS Comput Biol. 2010 Apr; 6(4):e1000743. doi: 10.1371/journal.pcbi.1000743 PMID: 20419152

31. Xiao L, Owen SM, Goldman I, Lal AA, deJong JJ, Goudsmit J, et al. CCR5 coreceptor usage of non-syncytium-inducing primary HIV-1 is independent of phylogenetically distinct global HIV-1 isolates: delineation of consensus motif in the V3 domain that predicts CCR-5 usage. Virology. 1998 Jan; 240 (1):83–92. Available from: http://dx.doi.org/10.1006/viro.1997.8924 PMID: 9448692

32. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013 Jan; 41 (Database issue):D590–D596. Available from: http://dx.doi.org/10.1093/nar/gks1219 PMID: 23193283

33. Epp LS, Stoof-Leichsenring KR, Trauth MH, Tiedemann R. Molecular profiling of diatom assemblages in tropical lake sediments using taxon-specific PCR and Denaturing High-Performance Liquid Chromatography (PCR-DHPLC). Mol Ecol Resour. 2011 Sep; 11(5):842–853. Available from: http://dx.doi.org/10.1111/j.1755-0998.2011.03022.x PMID: 21592311

The multiple testing problem and SeqFeatR

To kill an error is as good a service as, and sometimes even better than, the establishing of a new truth or fact.

CHARLES DARWIN

Abstract

Hypotheses testing is very common in science, especially in medical science. There is a ongoing conflict between the camps of a classical hypothesis testing background and those of a Bayesian background. For classical hypothesis testing scientists the multiple testing problem and therefore a multiple testing correction seems natural. Bayesian Inference sees itself without such multiple testing problems. Sequence data, as investigated in this work, is problematic for multiple testing correction, as it is for independent tests, and nucleotides or amino acids are all but random and independent.

I investigated the best method to analyze the given data (sequence data and features) for my R-package `SeqFeatR`. Bayesian Inference solutions showed much better results for our test set than the multiple testing corrections available in C-RAN-package stats. Thus I included Bayes Factors and Hierarchical Bayes into `SeqFeatR`.

5.1 Introduction

5.1.1 Overview: Error rates and their corrections

Hypotheses testing is very common. For a given hypothesis the results of one or more experiments are evaluated to see if the null hypothesis (H_0) can be rejected (and the alternative hypothesis (H_1) can be accepted) or not. The decision value is often the p-value, with which the user can estimate if the null hypothesis should be rejected. The p-value is the probability of getting the results (or more extreme results), given that the null hypothesis is true. For a set of rejection areas R (rejection area = the set of values for the test statistic that leads to rejection of H_0) the p-value of an observed statistic S = s is:

$$p(s) = \min_{s \epsilon R} P(S \epsilon R | H_0 \, true) \tag{5.1}$$

In every statistical conclusion one can make two types of error: the first are Type-I errors, where the null hypothesis is rejected although it is true (= false positive), the second are Type-II errors, where the null hypothesis is not rejected although it is false (= false negative). The probability of making at least one false positive decision is α, and of not making this error 1-α (see Table 5.1). The value α usually equals the significance level of a test.

If there is not just one test, but m tests, the probability of not making an error is $(1-\alpha)^m$ and of making at least one error $1-(1-\alpha)^m$. The probability for one false positive in 100 tests may be close to one, depending on the actual value of α, which can be very critical

Table 5.1: Overview of Type-I and Type-II errors in hypothesis testing

	actual situation "Truth"	
Decision	H_0 true	H_0 false
Do not reject H_0	Correct decision (1- α) TN	Incorrect decision (Type II error β) FN
reject H_0	Incorrect decision (Type I error α) FP	Correct decision (1- β) TP

in certain areas like tests for serious illnesses. These tests can have a big impact on the life of a person, such as a male patient who lost his job and home because of the positive result in a HIV test in 1985 [39]. This would be bad enough if the test was correct, but if it was a false positive the situation would be even worse [8]. The false positive rate - the probability of having a false positive - can be calculated in different ways, which makes it difficult to choose the right one for any given situation. The first calculation method simply divides the number of false positives by the total number [9], the second divides the number of false positives by the sum of true positives and false positives [15] and the third divides the number of false positives by the sum of false positives and true negatives [37]:

$$1.FPR = \frac{FP}{Total} \tag{5.2}$$

$$2.FPR = \frac{FP}{TP + FP} = \mathbb{P}(H_0\ true|reject\ H_0) = 1 - \text{positive prediction value} \tag{5.3}$$

$$3.FPR = \frac{FP}{TN + FP} = \mathbb{P}(reject\ H_0|H_0\ true) = 1 - \text{specificity} \tag{5.4}$$

These three possible interpretations of a false positive rate result in big differences as Germanson noted in 1989 [21]. The second variant is often used today. Since a false positive is unwanted, there exist many different methods to control the false positive rate and to correct the values to minimize the chance for a false (positive) result:

```
V = # false positives
R = # rejected hypothesis
𝔼 = the expected value
```

$$PCER = \mathbb{E}(V)/m \text{ Per-comparison error rate} \tag{5.5}$$

$$PFER = \mathbb{E}(V) \text{ Per-family error rate} \tag{5.6}$$

$$FWER = P(V \geq 1) \text{ Family-wise error rate} \tag{5.7}$$

$$FDR = \mathbb{E}(\frac{V}{R}|R > 0)P(R > 0) \text{ False discovery rate} \tag{5.8}$$

$$pFDR = \mathbb{E}(\frac{V}{R}|R > 0) \text{ Positive false discovery} \tag{5.9}$$

The most well-known correction is the Bonferroni [50] correction, a method to control the

Family-wise error rate (FWER). The Bonferroni correction simply corrects the p-value by the number of tests m. It rejects any hypothesis with p-value $\leq \frac{\alpha}{m}$.

$$\tilde{p}_j = \min[mp_j, 1] \tag{5.10}$$

However the interpretation of findings depends on the number of tests and not the test itself. Likewise there is an increased Type-II error, where one cannot reject the general H_0 (= all single H_0 are true), even if there are effects. Perneger describes the Bonferroni method as "unnecessary at best and deleterious at worst" [32]. Another FWER method was invented by Holm [24], who in contrast to the single step approach by Bonferroni, used a sequential approach, where the p-values are sorted and ranked in ascending order and each value is multiplied not with m, but with a decreasing m:

$$\tilde{p}_j = min[(m - j + 1) \cdot p_j, 1] \tag{5.11}$$

This method is not as strict as Bonferronis but Pernegers objections are still applicable. FWER methods are appropriate if any false positive would be bad.

False discovery rate is a method which allows a certain number of false positives. It controls the proportion of false positives among the set of rejected hypothesis. Benjamini and Hochberg developed the FDR and formalized it in 1995 [4]. For an FDR level δ the p-values are sorted and ranked in ascending order and divided into those which are considered significant ($\leq$ rank j) and those which are not ($>$ rank j). For all significant p-values $p(j)$:

$$p(j) \leq \delta \frac{j}{m} \tag{5.12}$$

Storey and Tibshirani extended the FDR into the pFDR, which is the calculated FDR assuming that there is at least one positive hypothesis test and named a measure of statistical significance *q-value*, which numbers the expected proportion of false positives for calling a certain single result significant [44]. In contrast to the p-value, the q-value for a set of rejection areas R of an observed statistic S = s was defined as:

$$q(s) = \min_{s \epsilon R} \; pFDR(R) \; = \; \min_{s \epsilon R} \; P(H_0 \, true | S \epsilon R) \tag{5.13}$$

The q-value is the p-value with the conditional statement and event of interest reversed, so that it is more intuitive in certain questions.

5.1.2 The problem with sequence data

All of the above corrections are for multiple tests. "'Multiple"' refers to the same experiment, where the tests themselves are repeated but independently. If the tests are not independent, multiple comparison corrections give unwanted results [16, 35, 40]. The (position-wise) analysis of sequence data, be it nucleotides or amino acids, is not independent, since the change of one nucleotide or amino acid often results in the change of one or more nucleotides or amino acids at other positions.

Problem: *Can we even consider using multiple corrections for sequence data? And if we can use them, how should we include the connection of different positions?*

5.1.3 A possible answer - Bayesian Inference

Bayesian Inference is a statistical method derived from a special case of a statistical problem solved by Thomas Bayes and generalized by Pierre-Simon Laplace [42]. It uses "'Bayes' Rule"', the form of Bayes' theorem and therefore a method to analyze conditional probabilities with two types of evidence. One type of evidence is known *a priori* through knowledge, and the other evidence is the result of some form of test [43]. Bayes' rule is not limited to only one pair of prior and test results, but can be used with the posterior from one analysis and another test - which is then called "'Hierarchical Bayes"' or "'Bayesian Updating"'. Bayesian Inference for a parameter θ can be formalized:

$$\mathbb{P}(\theta|data) = \frac{\mathbb{P}(data|\theta) \cdot p(\theta)}{p(data)} \rightarrow posterior \propto likelihood \cdot prior \tag{5.14}$$

with $p(\theta)$ being the prior and the probability before seeing the data and $\mathbb{P}(data|\theta)$ the probability of the data given the parameter. The posterior $\mathbb{P}(\theta|data)$ represents the information combination of the viewed data and the prior. One of the bonuses of the Bayesian approach is that there is no need for multiple testing correction. Bayesian testing has a built-in penalty for this [19, 26]. In a Bayes Factor approach, a Bayesian Inference variant in which two models or hypotheses are compared, the evidence of association can be weighed against a prior probability and, unlike for example Bonferroni, without reference to the number of sequence positions tested. This may lead to an increased rate of false positive associations, but since the expected number of true positive associations will also increase, the FDR will remain roughly constant [41].

5.1.4 Bayes and amino acid/nucleotide sequence data

Bayes' Theorem cannot be applied directly to sequence data, since at first sight there is one vital piece of information missing: The *a priori* probability for an amino acid or nucleotide at a certain position being a certain letter. A possible way out of this problem is an analysis of as many sequences of the same kind as possible and the usage of those frequencies as a 'good guess', as was done by Garabed in 2008 [17]. Stephens recommended another approach in his review article on Bayes in GWAS settings: a Bayes Factor is combined with an *a priori* probability, so that the single nucleotide polymorphism (SNP) is associated with the phenotype in question [41]. These *a priori* probabilities may be the same value for the whole genome or may vary across SNPs and should be in the range of 10^{-4} to 10^{-6} [11]. Wakefield recommended instead - assuming a normal prior distribution with a mean of zero and variance W - a Bayes Factor calculated from the maximum likelihood estimate (MLE) of the log OR and its sampling variance, which results by specifying W as a function of an assumed effect size distribution, in a Bayes Factor, which is unaffected by sample size [38, 48].

A more important question is if Bayes is better than a direct multiple comparison correction on the results from a classical hypothesis testing approach for sequences based data. This was tested with the R-package `SeqFeatR`, HCV sequences and three datasets of epitopes from HCV.

5.2 Methods

5.2.1 Data

I used an amino acid dataset from the Anti-D cohort [13, 49] consisting of 81 HCV sequences from Core up to NS2 with known HLA types of the sequences, and reduced this dataset to all sequences, which were HLA-A*02 (29 sequences).

5.2.2 Sample epitopes for the sequences

I used three different approaches to get epitopes for the sequences:

1. Predicted epitopes from the IEDB prediction tool [27], which is a meta-predictor (uses more than one method): I predicted all epitopes for the 29 sequences of HLA-A*02:01 and HLA-A*02:06. An epitope with an ANN (Artificial Neural Networks) IC_{50} value of 500 or less was considered a positive match after the recommendations from IEDB and You [51]. All epitopes with higher values were considered a mismatch (414 positive epitopes).

2. Known epitopes from the Los Alamos database [53]: I took all epitopes for our selected region from the Los Alamos database (54 epitopes, none from the NS2 region).

3. Known epitopes from IEDB database [47]: I downloaded the known positive epitopes from IEDB for HCV and all HLA-A*02 alleles and compared the epitope sequence to H77 to get start and end points for the epitopes (32 epitopes).

5.2.3 Bayesian Inference in `SeqFeatR` - Bayes factors

A variant of the Bayesian method for evaluation is a Bayesian model comparison with the usage of the so-called Bayes Factor. The Bayes Factor is used to evaluate the probabilities of two different models and therefore a possibility to estimate which model is better. A detailed description of Bayes Factors in `SeqFeatR` is given in chapter 4.

The BF for two hypotheses H_0 and H_1, given data D, is the ratio of posterior odds and the corresponding prior odds: $\text{BF} = (p(H_1|D)/p(H_0|D)) \,/\, (\pi_1/\pi_0)$.

It is important to note that with this technique the better of two models is favored, which must not necessarily be the best model for the data. `SeqFeatR` has an added extra column in the csv output which provides an estimate of the simulation standard error of the computed value of the Bayes factor.

5.2.4 Bayesian Inference in SeqFeatR - Hierarchical - with example

A more complex variant of Bayesian Inference is "'Hierarchical Bayes"'. In Bayesian Inference a hierarchical model is characterized by priors which are in themselves probability distributions called hyperpriors. In a basic model, only one position and one amino acid is taken into consideration for the evaluation of a significant difference. In a more general approach all amino acids and all positions are sampled in one model. This type of model is more realistic in that amino acids and sequence positions are by no means independent. Instead, it is very common that one sequence position influences another one due to binding affinities inside the secondary or tertiary structure of the protein.

Alignment positions, e.g 1-20-21-35, can be simulated with a multinomial distribution, which is used to model n observations that fall into one of a finite number of mutually exclusive categories. The categories in this case are all alignment positions, and those positions are mutually exclusive. Likewise amino acids or nucleotides can be simulated with a multinomial distribution. A multinomial distribution is a more general form of the binomial distribution, in which more than two outcomes are possible.

$$\begin{aligned} n_i &= \text{the number of times outcome i occurs} \\ p_i &= \text{the probability of outcome i} \\ p &= \frac{n!}{(n_1!)(n_2!)\dots(n_k!)} p_1^{n1} p_2^{n2} \dots p_k^{nk} \end{aligned} \tag{5.15}$$

The likelihood function of the multinomial distribution can be described as:

$$f(x_1, \dots, x_k | \theta_1, \dots, \theta_k) = \left(\frac{n!}{x_1! \dots x_k!} \right) \theta_1^{x_1} \dots \theta_k^{x_k} \tag{5.16}$$

The probability that an observation falls into category i is given by Θ_i, whereas the vector $X = (X_1, \dots, X_k)$ counts how many of these observations fall into each category.

A conjugated priori distribution for the multinomial distribution is the so-called Dirichlet distribution. The Dirichlet distribution - a generalized form of the Beta distribution - is parametrized by a vector α of positive real numbers. The multinomial distribution

Table 5.2: Example for hierarchical Bayes with two alignment positions and three amino acids.

Feature	A	P	Y
1	3	5	0
2	0	3	5

may return for example the probabilities of an amino acid being an Alanine, Proline or Tyrosine, the Dirichlet distribution how probable a specific distribution of Alanines, Prolines and Tyrosines is. If those probabilities were uniformly distributed (each amino acid has the same probability to occur) the vector α would contain the same values for each possibility. The posterior probability distribution for a Dirichlet distribution is again Dirichlet distributed, but takes the vector X into account and has the parameters $\alpha_1 + X_1, \ldots, \alpha_k + X_k$ [10, 22].

Example

For purposes of illustration I classify the sequences' feature according to the number of amino acids and assume only one alignment position.

Observations within each alignment position are easily depicted as being multinomial (with k = 3).

$$(X_A, X_P, X_Y)|(\Theta_A, \Theta_P, \Theta_Y) \sim Multinomial(\Theta_A, \Theta_P, \Theta_Y) \tag{5.17}$$

Θ_i is the probability that an observation falls into category i, thus Θ_A is the probability that Alanine was observed. The vector (X_A, X_P, X_Y) is a vector of counts of the number of amino acids in each of the three categories.

If we assume that our alignment positions are independent, the prior distribution for the multinomial parameters are also independent and uniform (for example equal one):

$$(\Theta_A, \Theta_P, \Theta_Y) \sim Dirichlet(1, 1, 1) \tag{5.18}$$

Thus the posterior probability equals:

$$(\Theta_A, \Theta_P, \Theta_Y)|(X_A, X_P, X_Y) \sim Dirichlet(1 + X_A, 1 + X_P, 1 + X_Y) \tag{5.19}$$

For the example mentioned above, this leads to:

$$(\Theta_{f1_A}, \Theta_{f1_P}, \Theta_{f1_Y})|(X_{f1_A}, X_{f1_P}, X_{f1_Y}) \sim Dirichlet(3 + 1, 5 + 1, 0 + 1) \tag{5.20}$$

$$\mathbb{E}[(\Theta_{f1_A}, \Theta_{f1_P}, \Theta_{f1_Y})|(X_{f1_A}, X_{f1_P}, X_{f1_Y})] = (0.36, 0.55, 0.09)^{1} \tag{5.21}$$

$$(\Theta_{f2_A}, \Theta_{f2_P}, \Theta_{f2_Y})|(X_{f2_A}, X_{f2_P}, X_{f2_Y}) \sim Dirichlet(0 + 1, 3 + 1, 5 + 1) \tag{5.22}$$

$$\mathbb{E}[(\Theta_{f2_A}, \Theta_{f2_P}, \Theta_{f2_Y})|(X_{f2_A}, X_{f2_P}, X_{f2_Y})] = (0.09, 0.36, 0.55) \tag{5.23}$$

$$\tag{5.24}$$

The alignment positions in this example were treated as independent in which each position does not interact with another, but this is not generally the case. Usually the amino acid at one sequence position interacts at least with their direct neighbors, but may also interact with amino acids several positions up- or downstream. Therefore we must expect that the alignment positions are not independent. We cannot model the multinomial parameters with the uniform Dirichlet distribution, but can use a common hyperparameter that reflects a common central tendency. I use the Gamma distribution with shape and scale = 1 for those pseudo-counts (representing the number of observations that we have already seen), because it can be easily calculated and the parameters within the vector α are always positive. For the posterior distribution, we must add the counts for the "new" observations X_i.

$$(\Theta_A, \Theta_P, \Theta_Y) \sim Dirichlet(A_A, A_P, A_Y) \tag{5.25}$$

$$(\Theta_A, \Theta_P, \Theta_Y)|(X_A, X_P, X_Y) \sim Dirichlet(A_A + X_A, A_P + X_P, A_Y + X_Y) \tag{5.26}$$

With this slight change, the calculation becomes quite complex and should be solved with

[1] Example calculation of the expected value $\mathbb{E}$: $\mathbb{E}[(\Theta_{f1_A}, \Theta_{f1_P}, \Theta_{f1_Y})|(X_{f1_A}, X_{f1_P}, X_{f1_Y})] = (\frac{3+1}{3+1+5+1+0+1}, \frac{5+1}{3+1+5+1+0+1}, \frac{0+1}{3+1+5+1+0+1})$

special tools and programs, e.g via the R-packages R2jags[46], rjags[34], and Jags [33] - a program for the analysis of Bayesian Hierarchical models using Markov Chain Monte Carlo (MCMC) simulations. A basic interface to this tools is integrated into `SeqFeatR`.

If we calculate this depended version, the results are:

$$\mathbb{E}[(\Theta_{f1_A}, \Theta_{f1_P}, \Theta_{f1_Y})|(X_{f1_A}, X_{f1_P}, X_{f1_Y})] \approx (0.35, 0.57, 0.08) \tag{5.27}$$
$$\mathbb{E}[(\Theta_{f2_A}, \Theta_{f2_P}, \Theta_{f2_Y})|(X_{f2_A}, X_{f2_P}, X_{f2_Y})] \approx (0.08, 0.40, 0.51) \tag{5.28}$$
$$\tag{5.29}$$

These values are similar to the results from the independent approach but not identical, even for such a simple example.

However as mentioned above, sequence positions are linked together in some form and theoretically we can use the same technique to combine sequence position, feature and amino acid together.

5.2.5 SeqFeatR prediction

`SeqFeatR` was used with all available corrections for multiple testing (Bonferroni, Holm, Hochberg, Hommel, BH, BY, FDR [4, 5, 23–25]) and Bayesian Inference on a given set of sequences with HLA information and the results for HLA-A*02 were selected for further analysis. Two variants of Bayes Factor calculation were used. The standard calculation with a fixed Dirichlet Precision Parameter and `SeqFeatR`s Bayes Factors with a calculated Dirichlet Precision Parameter (see chapter 4).

5.2.6 Comparison of multiple corrections and Bayesian Inference

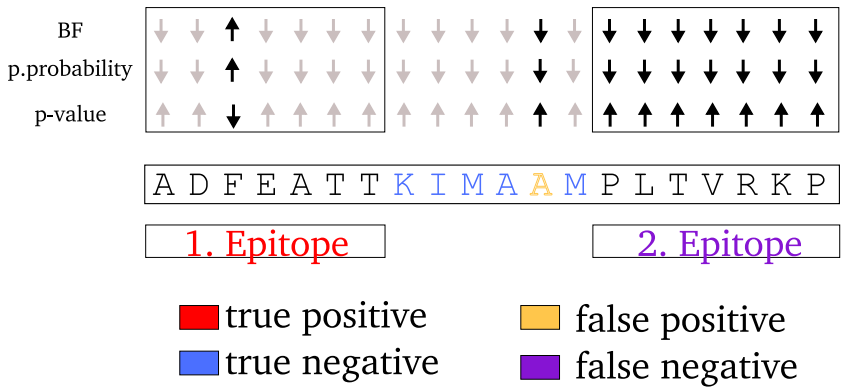


Figure 5.1: Example of true/false positives and true false negatives in the comparison between different multiple corrections and Bayesian Inference with epitope data from HCV sequences calculated by `SeqFeatR`. The first epitope is a true positive, the second epitope a false negative. `SeqFeatR` was used to search for hints of an possible epitope at each sequence position in an alignment of 81 amino acid HCV sequences with given HLA types.

To compare the results from the different methods, true and false negatives and true and false positives had to be defined. A true positive - an epitope found, where an epitope was in the original data from the database - was noted if the result had one or more lower p-values or higher posterior probabilities/Bayes Factors at the epitope position (usually eight to twelve positions plus two in both of the flanking regions, because there could also be mutations outside of but belonging to the epitope, to permit processing of the epitope (see chapter 7)). A false negative was noted, if there was no low p-value or high probability/Bayes Factors at a given epitope position. A true negative was noted for every amino acid outside all of the epitope sequence positions, which had a high p-value or low posterior probability/Bayes Factor, whereas a false positive was noted for every amino acid outside all epitope sequence positions, which had a low p-value or high posterior probability/Bayes Factor (see Figure 5.1).

5.3 Results

5.3.1 Results from SeqFeatR

The multiple comparison correction for SeqFeatR corrected all of the p-values to 1, regardless of the chosen method. Only for no correction ("none") and Bayesian Inference/Bayes Factors were the result values for some positions different.

5.3.2 ROC Curves and AUC values

The ROC for the three different kinds of epitope datasets showed Bayesian Inference as a superior model to compensate for multiple testing in every case. "'No correction"' had slightly to moderate inferior results, whereas all other multiple corrections were no better than pure guesses (see Figure 5.2). Of the three Bayesian methods, the Bayes Factor with a fixed Dirichlet Precision Parameter showed the highest AUC values, and thus seems to be the best choice for this dataset.

The area under the curve (AUC) values for the ROC confirmed those impressions (see Table 5.3).

5.4 Discussion

The results show that for sequence data and epitopes multiple correction procedures are not appropriate, since every tested method of correction eliminated all signals. The mathematical reason for this total elimination of sequence data signal as used by SeqFeatR is that there is a very high number of tests and thus the multiple corrections increase all p-values to one. SeqFeatR checks every position for every occurring amino acid, and tests which one has the lowest p-value. For a sequence set with only 5 amino acids and 3 different kinds of features, and on every position occurrence of all possible amino acids, SeqFeatR makes 300 tests. This increases dramatically with the sequence length and the number of features.

In the comparison of the different methods it could be shown that Bayesian Inference is almost always better than no multiple correction, and always better than all used multiple

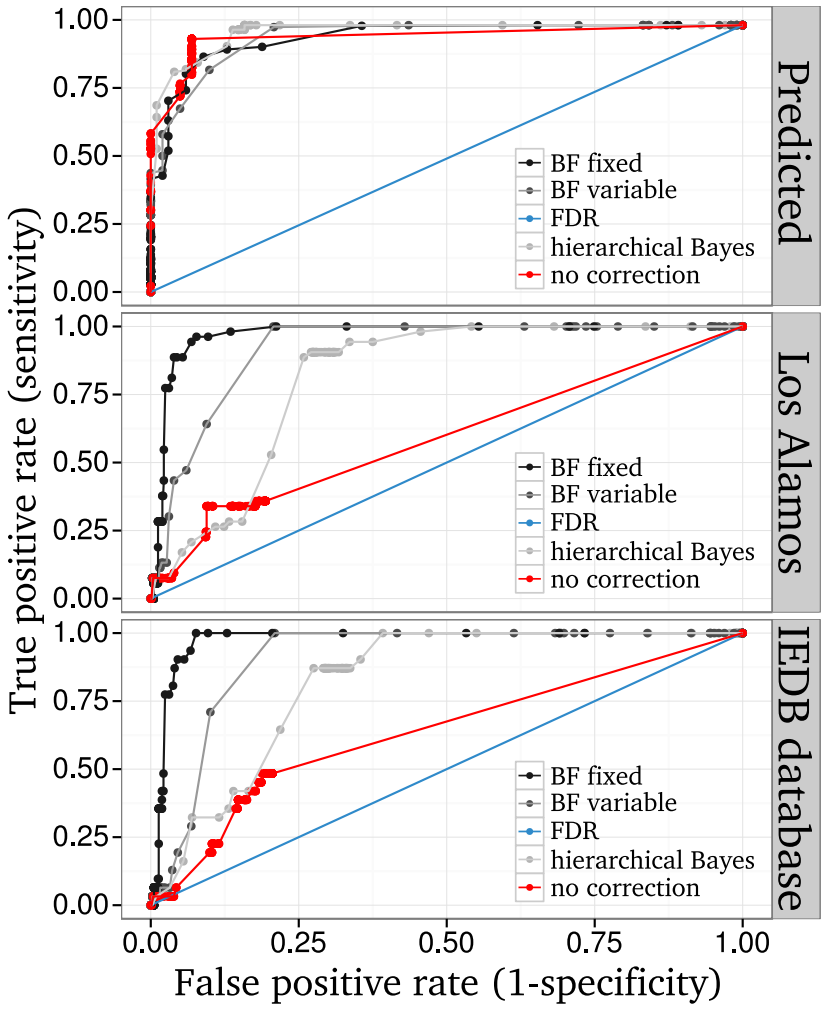


Figure 5.2: ROC for each of the three datasets. "Predicted" was taken from the prediction tool on the IEDB website, "Los Alamos" from the Los Alamos database and "IEDB" from the IEDB database. All datasets were compared with the results from SeqFeatRs prediction for a set of 81 amino acid HCV sequences with given HLA types. Shown are the ROC for HLA-A*02 and two versions of Bayes Factors (fixed or variable Dirichlet Precision Parameter), hierarchical Bayes, no multiple correction on p-values and FDR multiple correction. *FDR* is a representative for all of the results from multiple corrections (Bonferroni, Holm, Hochberg, Hommel, BH, BY, FDR).

corrections (see 5.2). Only in the case of predicted epitopes, "'no correction"' has similar lower false positive rate at a medium sensitivity and is therefore similar to Bayesian Inference. But one has to keep in mind that predicted epitopes means precisely that, predicted and not verified, and that this dataset has to be treated with caution. Also, I used a very simple hierarchical model to calculate the results, which is not uncommon among statisticians [1, 30], and there are several more complex methods to insert the prior probability. Those analyses could incorporate a Poisson prior and an ANOVA analysis as

Table 5.3: AUC values for the ROC. "Predicted" was taken from the prediction tool on the IEDB website, "Los Alamos" from the Los Alamos database and "IEDB" from the IEDB database. All datasets were compared with the results from SeqFeatRs prediction for a set of 81 amino acid HCV sequences with given HLA types. Shown are the ROC for HLA-A*02 and two versions of Bayes Factors (fixed or variable Dirichlet Precision Parameter), hierarchical Bayes, no multiple correction on p-values and FDR multiple correction. *FDR* is a representative for all of the results from multiple corrections (Bonferroni, Holm, Hochberg, Hommel, BH, BY, FDR).

Data set	method	AUC value
Predicted	Bayes Factor fixed	0.94
	Bayes Factor variable	0.94
	Hierarchical Bayes	**0.96**
	multiple correction	0.5
	no correction	0.94
Los Alamos	Bayes Factor fixed	**0.97**
	Bayes Factor variable	0.92
	Hierarchical Bayes	0.82
	multiple correction	0.5
	no correction	0.59
IEDB database	Bayes Factor fixed	**0.97**
	Bayes Factor variable	0.90
	Hierarchical Bayes	0.83
	multiple correction	0.5
	no correction	0.64

in the example used by Kruschke [28]. Another simple possibility would be to incorporate the R-package *conting*, which analyzes contingency tables with Bayesian Inference and uses Poisson priors and MCMC to evaluate the probability density [31].

It is not remarkable in itself that our Bayesian approach showed better results than the "'corrected"' ones, because many scientists have discovered similar findings in the last few years. Bayesian Inference seems to be better suited for their problems than a classical hypothesis testing method like a t-test or Fisher's exact Test [2, 18, 36]. But it is often difficult to estimate which is better. In 2001 Austin et al. studied the differences between a Bayesian and a classical hypothesis testing approach for profiling hospitals, and found that they differ, but could not tell which one is better because he did not know the correct values [3]. In 2007 a similar comparison was made by Storvik et al. for DNA Databases [45]. Efron and Roderick J Little suggested to use a combination of both methods [14, 29], which has been already done by some groups [6, 7, 12, 20], and was further implemented later on by other groups [29, 52]. In the future, such combinations of both sides could also be used for the problem given here.

References

[1] A. Agresti and D. Hitchcock. “Bayesian inference for categorical data analysis”. English. In: *Statistical Methods and Applications* 14.3 (2005), pp. 297–330. ISSN: 1618-2510. DOI: 10.1007/s10260-005-0121-y. URL: http://dx.doi.org/10.1007/s10260-005-0121-y (cit. on p. 74).

[2] P. M. E. Altham. “Exact Bayesian Analysis of a 2 ⋆ 2 Contingency Table, and Fisher’s "Exact" Significance Test”. English. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 31.2 (1969), pp. 261–269. ISSN: 00359246 (cit. on p. 75).

[3] P. C. Austin, C. D. Naylor, and J. V. Tu. “A comparison of a Bayesian vs. a frequentist method for profiling hospital performance”. In: *Journal of Evaluation in Clinical Practice* 7.1 (2001), pp. 35–45. ISSN: 1365-2753. DOI: 10.1046/j.1365-2753.2001.00261.x. URL: http://dx.doi.org/10.1046/j.1365-2753.2001.00261.x (cit. on p. 75).

[4] Y. Benjamini and Y. Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal Statistical Society. Series B. Methodological* 57.1 (1995), pp. 289–300. ISSN: 0035-9246 (cit. on pp. 64, 71).

[5] Y. Benjamini and D. Yekutieli. “The control of the false discovery rate in multiple testing under dependency”. In: *Ann. Statist.* 29 (4) (2001), pp. 1165–1188 (cit. on p. 71).

[6] J. O. Berger, L. D. Brown, and R. L. Wolpert. “A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing”. In: *The Annals of Statistics* 22.4 (1994), pp. 1787–1807. ISSN: 00905364 (cit. on p. 75).

[7] J. Berger, B. Boukai, and Y. Wang. “Properties of Unified Bayesian-Frequentist Tests”. English. In: *Advances in Statistical Decision Theory and Applications*. Ed. by S. Panchapakesan and N. Balakrishnan. Statistics for Industry and Technology. Birkhäuser Boston, 1997, pp. 207–223. ISBN: 978-1-4612-7495-7. DOI: 10.1007/978-1-4612-2308-5_14 (cit. on p. 75).

[8] R. Bhattacharya, S. Barton, and J. Catalan. “When good news is bad news: psychological impact of false positive diagnosis of HIV”. In: *AIDS Care* 20.5 (2008), pp. 560–564. DOI: 10.1080/09540120701867206 (cit. on p. 63).

[9] D. S. Burke et al. “Measurement of the False Positive Rate in a Screening Program for Human Immunodeficiency Virus Infections”. In: *New England Journal of Medicine* 319.15 (1988), pp. 961–964. DOI: 10.1056/NEJM198810133191501 (cit. on p. 63).

[10] R. J. Connor and J. E. Mosimann. “Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution”. In: *Journal of the American Statistical Association* 64.325 (1969), pp. 194–206. DOI: 10.1080/01621459.1969.10500963. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1969.10500963. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10500963 (cit. on p. 69).

[11] W. T. C. C. Consortium. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.” eng. In: *Nature* 447.7145 (June 2007), pp. 661–678. DOI: 10.1038/nature05911. URL: http://dx.doi.org/10.1038/nature05911 (cit. on p. 66).

[12] G. S. Daita and J. K. Ghosh. "On priors providing frequentist validity for Bayesian inference". In: *Biometrika* 82.1 (1995), pp. 37–45. DOI: 10.1093/biomet/82.1.37 (cit. on p. 75).

[13] S. Dittmann et al. "Long-term persistence of hepatitis C virus antibodies in a single source outbreak". In: *Journal of Hepatology* 13.3 (1990), pp. 323–327. URL: http://www.journal-of-hepatology.eu/article/0168-8278(91)90076-N/abstract (cit. on p. 66).

[14] B. Efron. "Bayesians, Frequentists, and Scientists". In: *Journal of the American Statistical Association* 100 (2005), pp. 1–5. DOI: 10.1198/016214505000000033 (cit. on p. 75).

[15] J. Fleiss. *Statistical Methods for Rates and Proportions*. 2nd ed. New York: Wiley, 1981 (cit. on p. 63).

[16] P. A. Games. "An Improved t Table for Simultaneous Control on g Contrasts". In: *Journal of the American Statistical Association* 72.359 (1977), pp. 531–534 (cit. on p. 65).

[17] R. B. Garabed. "Micro and macro approaches to the analytical epidemiology of foot-and-mouth disease". PhD thesis. University of California, Davis, 2008 (cit. on p. 66).

[18] A. Gelman. "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing*". In: *International Statistical Review* 71.2 (2003), pp. 369–382. DOI: 10.1111/j.1751-5823.2003.tb00203.x (cit. on p. 75).

[19] A. Gelman. "Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics". In: *Statistical Science* 24.2 (May 2009), pp. 176–178. DOI: 10.1214/09-STS284D (cit. on p. 65).

[20] C. Genovese and L. Wasserman. "Bayesian Frequentist Multiple Testing". In: (2002) (cit. on p. 75).

[21] T. Germanson. "Screening for HIV: Can we afford the confusion of the false positive rate?" In: *Journal of Clinical Epidemiology* 42.12 (1989), pp. 1235–1237. ISSN: 0895-4356. DOI: http://dx.doi.org/10.1016/0895-4356(89)90122-4 (cit. on p. 63).

[22] I. Good. *The estimation of probabilities: an essay on modern Bayesian methods*. Research monograph. M.I.T. Press, 1965. URL: https://books.google.de/books?id=wxLvAAAAMAAJ (cit. on p. 69).

[23] Y. Hochberg. "A sharper Bonferroni procedure for multiple tests of significance". In: *Biometrika* 75.4 (1988), pp. 800–802. DOI: 10.1093/biomet/75.4.800 (cit. on p. 71).

[24] S. Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70. URL: http://www.jstor.org/stable/4615733 (cit. on pp. 64, 71).

[25] G. Hommel. "A stagewise rejective multiple test procedure based on a modified Bonferroni test". In: *Biometrika* 75.2 (1988), pp. 383–386. DOI: 10.1093/biomet/75.2.383 (cit. on p. 71).

[26] W. H. Jefferys and J. O. Berger. "Ockham's Razor and Bayesian Analysis". In: *American Scientist* Vol. 80, No. 1 (1992), pp. 64–72 (cit. on p. 65).

[27] Y. Kim et al. "Immune epitope database analysis resource". In: *Nucleic Acids Research* 40.W1 (2012), W525–W530. DOI: 10.1093/nar/gks438 (cit. on p. 67).

[28] J. K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. 1st. Academic Press, 2010. ISBN: 0123814855, 9780123814852 (cit. on p. 75).

[29] R. J. Little. "Calibrated Bayes". In: *The American Statistician* 60 (2006), pp. 213–223. DOI: 10.1198/000313006X117837 (cit. on p. 75).

[30] M. R. Novick. "Baysian methods in psychological testing". In: *ETS Research Bulletin Series* 1969.1 (1969), pp. i–23. ISSN: 2333-8504. DOI: 10.1002/j.2333-8504.1969.tb00572.x. URL: http://dx.doi.org/10.1002/j.2333-8504.1969.tb00572.x (cit. on p. 74).

[31] A. M. Overstall and R. King. "conting: An R Package for Bayesian Analysis of Complete and Incomplete Contingency Tables". In: *Journal of Statistical Software* 58.7 (2014), pp. 1–27. URL: http://www.jstatsoft.org/v58/i07/ (cit. on p. 75).

[32] T. V. Perneger. "What's wrong with Bonferroni adjustments". In: *BMJ* 316.7139 (1998), pp. 1236–1238. DOI: 10.1136/bmj.316.7139.1236 (cit. on p. 64).

[33] M. Plummer. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. 2003 (cit. on p. 71).

[34] M. Plummer. *rjags: Bayesian graphical models using MCMC*. R package version 3-14. 2014. URL: http://CRAN.R-project.org/package=rjags (cit. on p. 71).

[35] K. J. Rothman. "No Adjustments Are Needed for Multiple Comparisons." In: *Epidemiology* 1.1 (1990) (cit. on p. 65).

[36] J. Rouder et al. "Bayesian t tests for accepting and rejecting the null hypothesis". English. In: *Psychonomic Bulletin & Review* 16.2 (2009), pp. 225–237. ISSN: 1069-9384. DOI: 10.3758/PBR.16.2.225 (cit. on p. 75).

[37] D. Sackett, R. Haynes, and P. Tugwell. *Clinical epidemiology*. Boston: Little, Brown and co, 1985 (cit. on p. 63).

[38] P. C. Sham and S. M. Purcell. "Statistical power and significance testing in large-scale genetic studies". In: *Nat Rev Genet* 15.5 (May 2014), pp. 335–346. ISSN: 1471-0056. URL: http://dx.doi.org/10.1038/nrg3706 (cit. on p. 66).

[39] R. Sherer. "Physician use of the hiv antibody test: The need for consent, counseling, confidentiality, and caution". In: *JAMA* 259.2 (1988), pp. 264–265. DOI: 10.1001/jama.1988.03720020066039 (cit. on p. 63).

[40] Z. Sidak. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". In: *Journal of the American Statistical Association* 62.318 (1967), pp. 626–633 (cit. on p. 65).

[41] M. Stephens and D. J. Balding. "Bayesian statistical methods for genetic association studies". In: *Nat Rev Genet* 10.10 (Oct. 2009), pp. 681–690. ISSN: 1471-0056. URL: http://dx.doi.org/10.1038/nrg2615 (cit. on pp. 65, 66).

[42] S. M. Stigler. "Thomas Bayes's Bayesian Inference". In: *Journal of the Royal Statistical Society. Series A (General)* 145.2 (1982), pp. 250–258 (cit. on p. 65).

[43] J. V. Stone. *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Sebtel Press, 2013. ISBN: 0956372848, 9780956372840 (cit. on p. 65).

[44] J. D. Storey and R. Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445. DOI: 10.1073/pnas.1530509100 (cit. on p. 64).

[45] G. Storvik and T. Egeland. "The DNA Database Search Controversy Revisited: Bridging the Bayesian–Frequentist Gap". In: *Biometrics* 63.3 (2007), pp. 922–925. ISSN: 1541-0420. DOI: 10.1111/j.1541-0420.2007.00751.x (cit. on p. 75).

[46] Y.-S. Su and M. Yajima. *R2jags: A Package for Running jags from R*. R package version 0.05-01. 2015. URL: http://CRAN.R-project.org/package=R2jags (cit. on p. 71).

[47] R. Vita et al. "The immune epitope database 2.0." In: *Nucleic Acids Res.* 38 (Database issue) (Jan. 2010), pp. D854–62 (cit. on p. 67).

[48] J. Wakefield. "Bayes factors for genome-wide association studies: comparison with P-values". In: *Genetic Epidemiology* 33.1 (2009), pp. 79–86. ISSN: 1098-2272. DOI: 10.1002/gepi.20359. URL: http://dx.doi.org/10.1002/gepi.20359 (cit. on p. 66).

[49] M. Wiese et al. "Low Frequency of Cirrhosis in a Hepatitis C (Genotype 1b) Single-Source Outbreak in Germany: A 20-Year Multicenter Study". In: *Hepatology* 32.1 (2000), pp. 91–96. ISSN: 0270-9139. DOI: http://dx.doi.org/10.1053/jhep.2000.8169. URL: http://www.sciencedirect.com/science/article/pii/S0270913900040635 (cit. on p. 66).

[50] K. J. Worsley. "An Improved Bonferroni Inequality and Applications". English. In: *Biometrika* 69.2 (1982), pp. 297–302. ISSN: 00063444. URL: http://www.jstor.org/stable/2335402 (cit. on p. 63).

[51] L. You et al. "Understanding Prediction Systems for HLA-Binding Peptides and T-Cell Epitope Identification". In: *Pattern Recognition in Bioinformatics*. Ed. by J. Rajapakse, B. Schmidt, and G. Volkert. Vol. 4774. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pp. 337–348. ISBN: 978-3-540-75285-1. DOI: 10.1007/978-3-540-75286-8_32 (cit. on p. 67).

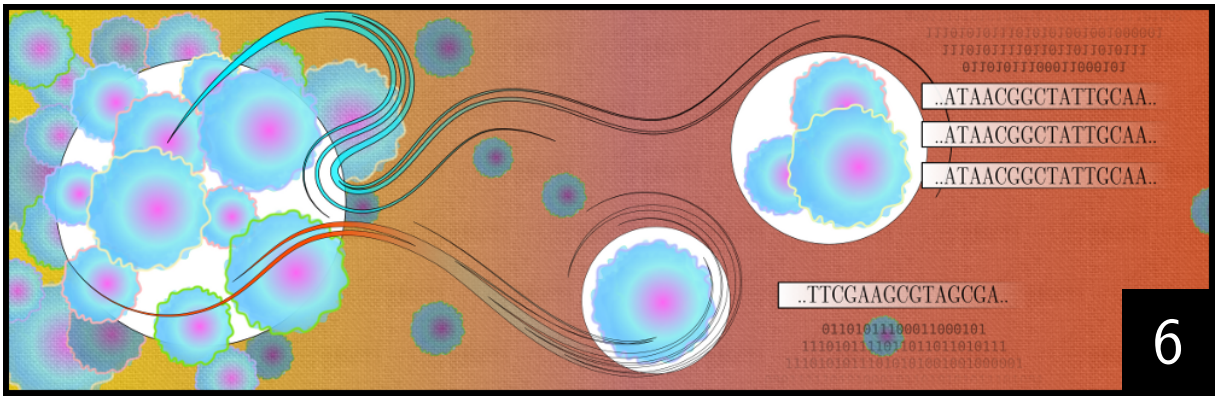
[52] A. Yuan. "Bayesian frequentist hybrid inference". In: *The Annals of Statistics* 37.5A (Oct. 2009), pp. 2458–2501. DOI: 10.1214/08-AOS649 (cit. on p. 75).

[53] K. Yusim et al. "Los Alamos Hepatitis C Immunology Database". In: *Applied Bioinformatics* 4.4 (2005), pp. 217–225. ISSN: 1175-5636. DOI: 10.2165/00822942-200504040-00002. URL: http://dx.doi.org/10.2165/00822942-200504040-00002 (cit. on p. 67).

PART

III

SEQUENCE ANALYSIS OF SANGER AND NGS SEQUENCES



Complexity of the human memory B cell compartment is determined by the versatility of clonal diversification in germinal centres

I would rather live in a world where my life is surrounded by mystery than live in a world so small that my mind could comprehend it.

HARRY EMERSON FOSDICK

Abstract

Our knowledge about the clonal composition and intraclonal diversity of the human memory B cell compartment as well as the relationship between memory B cell subsets is still limited, although these are central issues for our understanding of adaptive immunity. We performed a deep sequencing analysis of rearranged immunoglobulin (Ig) heavy chain genes from biological replicates, covering more than 100,000 memory B lymphoctes from two healthy adults. We reveal a highly similar BCR repertoire among the four main human IgM^+ and IgG^+ memory B cell subsets. Strikingly, in both donors 45% of sequences could be assigned to expanded clones, demonstrating that the human memory B cell compartment is characterized by many, often very large B cell clones. Twenty percent of the clones consisted of class switched and IgM^+ (IgD^+) members, a feature that correlated significantly with clone size. Hence, we provide strong evidence that the vast majority of Ig mutated B cells – including $IgM^+IgD^+CD27^+$ B cells – are post-germinal center (GC) memory B cells. Clone members showed high intraclonal sequence diversity and

high intraclonal versatility in Ig class and IgG subclass composition, with particular patterns of memory B cell clone generation in GC reactions. In conclusion, GC produce amazingly large, complex and diverse memory B cell clones, equipping the human immune system with a versatile and highly diverse compartment of IgM^+ (IgD^+) and class-switched memory B cells.

This chapter is based on the following publication: Bettina Budeus, Stefanie Schweigle, Martina Przekopowitz, Daniel Hoffmann, Marc Seifert, and Ralf Küppers (2015). **Complexity of the human memory B-cell compartment is determined by the versatility of clonal diversification in germinal centers.**

http://www.pnas.org/content/early/2015/08/28/1511270112.long

Complexity of the human memory B-cell compartment is determined by the versatility of clonal diversification in germinal centers

Bettina Budeus[a,1], Stefanie Schweigle de Reynoso[b,1], Martina Przekopowitz[b], Daniel Hoffmann[a], Marc Seifert[b,2], and Ralf Küppers[b,2,3]

[a]Bioinformatics, Faculty of Biology, University of Duisburg-Essen, 45117 Essen, Germany; and [b]Medical Faculty, Institute of Cell Biology (Cancer Research), 45122 Essen, Germany

Edited by Klaus Rajewsky, Max-Delbruck-Center for Molecular Medicine, Berlin, Germany, and approved August 12, 2015 (received for review June 9, 2015)

Our knowledge about the clonal composition and intraclonal diversity of the human memory B-cell compartment and the relationship between memory B-cell subsets is still limited, although these are central issues for our understanding of adaptive immunity. We performed a deep sequencing analysis of rearranged immunoglobulin (Ig) heavy chain genes from biological replicates, covering more than 100,000 memory B lymphocytes from two healthy adults. We reveal a highly similar B-cell receptor repertoire among the four main human IgM$^+$ and IgG$^+$ memory B-cell subsets. Strikingly, in both donors, 45% of sequences could be assigned to expanded clones, demonstrating that the human memory B-cell compartment is characterized by many, often very large, B-cell clones. Twenty percent of the clones consisted of class switched and IgM$^+$(IgD$^+$) members, a feature that correlated significantly with clone size. Hence, we provide strong evidence that the vast majority of Ig mutated B cells—including IgM$^+$IgD$^+$CD27$^+$ B cells—are post-germinal center (GC) memory B cells. Clone members showed high intraclonal sequence diversity and high intraclonal versatility in Ig class and IgG subclass composition, with particular patterns of memory B-cell clone generation in GC reactions. In conclusion, GC produce amazingly large, complex, and diverse memory B-cell clones, equipping the human immune system with a versatile and highly diverse compartment of IgM$^+$(IgD$^+$) and class-switched memory B cells.

IgV gene repertoire | human memory B cell subsets | IgM memory | clonal composition

The diversity of B lymphocytes is granted by the variability of their B-cell receptors (BCRs). This variability is generated in recombination processes during B-lymphocyte development in the bone marrow, where Ig variable (*V*), diversity (*D*), and joining (*J*) gene segments are combined to form antibody heavy and light chain *V* region genes (*D* segments only for the heavy chain). As a consequence, each naive B cell is equipped with a unique BCR (1). If B cells are activated by recognition of an antigen and T-cell help is provided, these B cells are driven into germinal center (GC) reactions where they undergo strong proliferation and further diversify their BCRs. The process of somatic hypermutation (SHM), which introduces point mutations and also some deletions and insertions into the V region genes at a very high rate, is activated in proliferating GC B cells (2, 3). Mutated GC B cells are then selected by interaction with follicular T helper and dendritic cells for improved affinity (4). GC B cells with unfavorable mutations undergo apoptosis. A large fraction of GC B cells performs class switch recombination to exchange the originally expressed IgM and IgD isotypes by IgG, IgA, or IgE (5). GC B cells undergo multiple rounds of proliferation, mutation, and selection, so that large GC B-cell clones are generated. Positively selected GC B cells finally differentiate into long-lived memory B cells or plasma cells (6).

The human memory B-cell compartment was originally thought to be mainly or exclusively composed of class-switched B cells, which typically account for about 25% of peripheral blood (PB) B cells (7). However, the detection of somatically mutated IgM$^+$ B cells pointed to the existence of non–class-switched memory B cells (8). Besides rare CD27$^+$ B cells with high IgM but low or absent IgD expression (IgM-only B cells; typically less than 5% of PB B cells) also IgM$^+$IgD$^+$CD27$^+$ B cells harbor mutated V genes, whereas IgM$^+$IgD$^+$CD27$^-$ B cells are mostly unmutated, naive B cells (9, 10). Hence, the two IgM$^+$CD27$^+$ populations were proposed to represent post-GC memory B-cell subsets (10). As both subsets together comprise about 25% of PB B cells and are detectable at similar frequencies in secondary lymphoid tissues (11), they represent a substantial fraction of the human B-cell pool. Moreover, as CD27 is also expressed on class-switched memory B cells, CD27 was proposed as a general memory B-cell marker (10, 12). Further studies refined this picture and revealed that about 10–20% of IgG$^+$ B cells are CD27 negative, so that presumably also CD27$^-$ memory B cells exist (13).

However, there are still major controversies and unresolved issues regarding the human memory B-cell compartment. First, the origin of the IgM$^+$IgD$^+$CD27$^+$ B-cell subset is debated, and it has been proposed that these cells are not post-GC B cells but either "effector B cells," derived from a particular developmental pathway with SHM as primary BCR diversification mechanism (14), or memory B cells generated in T-independent (TI) immune responses (15). Moreover, another study proposed the existence of a subset of IgM$^+$IgD$^+$CD27$^+$ B cells that represent human (GC independent) B1 B cells (16), although this is controversially

Significance

The complexity of the human memory B-lymphocyte compartment is a key component to depict and understand adaptive immunity. Despite numerous prior investigations, the generation of certain memory B-cell subsets, the dependency on T-cell help, and the composition, size, and diversity of clonal expansions are either poorly understood or debated. Here we provide an extensive and tightly controlled immunoglobulin heavy chain variable (*IGHV*) gene repertoire analysis of four main human memory B-cell subpopulations, revealing that an ordered diversification in germinal centers determines a highly versatile memory B-cell compartment in humans with surprisingly many very large B-cell clones.

Author contributions: M.S. and R.K. designed research; S.S.d.R. and M.P. performed research; B.B., M.P., D.H., and M.S. analyzed data; and B.B., M.S., and R.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank Sequence Read Archive (accession no. SRP062460).

[1]B.B. and S.S.d.R. contributed equally to this work.

[2]M.S. and R.K. contributed equally to this work.

[3]To whom correspondence should be addressed. Email: ralf.kueppers@uk-essen.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1511270112/-/DCSupplemental.

discussed (17). The existence of $CD27^+$ B-cell precursors in fetal liver (18) and of (infrequently and lowly) mutated $IgM^+IgD^+CD27^+$ B cells before birth and also in immunodeficient patients considered to lack GC indeed support a GC independent generation (whereas IgM-only B cells are missing in these instances, so that they are generally considered to represent post-GC memory B cells) (19, 20). The seemingly close relationship of PB $IgM^+IgD^+CD27^+$ B cells and splenic marginal zone B cells (21), which are considered to be key players for TI immune responses, has been taken as argument for an origin of these cells from TI immune responses (15). However, a prior focused IgV gene study showed that for large IgG^+ memory B-cell clones often also $IgM^+IgD^+CD27^+$ members can be found, arguing for a GC origin of at least a fraction of the latter cells (22). Second, the relationship between the various memory B-cell subsets is unclear. Are these subsets generated in common GC reactions that give rise to distinct types of memory B cells, or are they typically derived from independent immune responses or GC reactions? Third, how diverse is the pool of memory B cells generated from a GC B-cell clone in terms of intraclonal IgV gene diversity, and how large can memory B-cell clones be?

Next-generation sequencing (NGS) of IgV genes allows a comprehensive overview on the composition and diversity of the lymphocyte compartment (23–26). Several previous studies already analyzed human *IGHV* gene repertoire diversity. Although important findings were made, these studies did not include all PB memory B-cell subsets, e.g., CD27-negative class-switched B cells or IgM-only B cells, and/or were mostly based on small samples sizes and, thus, limited in estimating the complexity and clonal composition of the memory B-cell pool (27–29). Especially the clonal relationship between $IgM^+IgD^+CD27^+$ and post-GC memory B cells—although existing in principle (22)—has been claimed to be rare (27), thus supporting the view of a GC-independent generation of this large human B-cell subset. However, revelation of clonal overlap and expansion of the highly complex memory B-cell compartment requires the analysis of extensive cell numbers and independent testing procedures, thus far lacking in previous approaches. Thus, the clonal composition and the IgV mutation patterns of the four major human PB IgM^+ and IgG^+ memory B lymphocyte subsets, as well as their clonal relationship, are unclarified. In this study, we explore these issues comprehensively by NGS and provide molecular evidence for a reevaluation of our understanding of memory B-cell subsets.

Results

Human PB Memory B-Cell Subsets Display Nearly Identical *IGHV* Gene Repertoires. We sort-purified the four major human IgM^+ and IgG^+ PB B-cell subsets carrying mutated IgV genes, i.e., $IgM^+IgD^{low/-}CD27^+$ (IgM-only), $IgM^+IgD^+CD27^+$, IgG^+CD27^+, and IgG^+CD27^- B cells, according to their relative frequency in PB, to a total of 200,000 B lymphocytes, from two adult healthy donors (Fig. S1 *A–C* and Table S1). Plasmablasts were excluded as $CD27^{high}$ cells. For each donor and B-cell subset, two cell aliquots were separately processed to test for reproducibility of the analysis and to be able to identify clone members with identical mutation patterns within one B-cell subset. *IGHV* genes of the three largest families 1, 3, and 4 were amplified by RT-PCR. After stringent quality filtering (*Materials and Methods* and Table S1), we included 66,652 *IGHV* gene rearrangements from donor 1 and 40,658 from donor 2 in the analysis.

The average *IGHV* gene frequency and use reflects the typical repertoire of the three major *IGHV* gene families in human PB (30) and appeared highly similar among the subsets (Fig. 1). To quantify this, we considered *IGHV* genes used with substantial frequency (>5% of rearrangements in at least one population) and differential expression (at least twofold change) between two B-cell subsets. These criteria were applied to exclude random variation, because biological replicates did not show statistically significant variations above these thresholds (Fig. S1*C*). Only few *IGHV* gene frequencies (7 of 42 per subset for donor 1, 6–9 of 42 per subset for donor 2) differed significantly ($P < 0.05$; Fig. 1) between populations. Thus, only a minor fraction of *IGHV* genes (on average 17% in both donors) is differentially used between memory B-cell subpopulations. The high similarity of the *IGHV* gene rearrangement repertoires of the four memory B-cell subsets is underlined by highly similar length distributions of the complementarity determining region (CDR)III (Fig. S1*D*).

Memory B-Cell Subsets Differ in Their Median Mutation Frequencies, Showing a Consistent Pattern Among Donors. The median *IGHV* mutation frequency of memory B-cell populations was 2.7% and 3.0% (donors 1 and 2, respectively) for IgG^+CD27^-, 7.1% and 7.1% for IgG^+CD27^+, 3.4% and 3.7% for $IgM^+IgD^+CD27^+$, and 4.8% and 4.7% for IgM-only B cells (Fig. 1*B*). These frequencies are similar to previously published data (10, 13, 31), showing that our analysis reproduces (and extends) existing smaller datasets. Moreover, not only the median mutation frequencies of the distinct B-cell subsets, but also their distribution, is strikingly similar between the donors (Fig. S2*A*). All populations showed a homogenous distribution of values, indicating that these are homogenous populations (Fig. S2).

$IgG3^+$ Memory B Cells May Include a Fraction of T Cell-Dependent, but GC-Independent, Generated Lymphocytes. In line with earlier studies (13, 31), $IgG3^+$ B cells represented a severalfold larger fraction among IgG^+CD27^- than IgG^+CD27^+ B cells (Fig. 2*A*). The IgG^+CD27^- subset contained more than 50% (donor 1) and 30% (donor 2) $IgG3^+$ B cells and in both donors less than 10% $IgG2^+$ B cells, whereas the IgG^+CD27^+ B-cell subset showed an opposite distribution (Fig. 2*A*). The average fraction of IgG1-switched B cells was similar between both subsets. IgG4 transcripts were not detected in donor 1 and only at a very low frequency in donor 2. The median mutation frequencies (Fig. 2*B*) and the distribution of mutations (Fig. S2*B*) were similar in both donors; however, they differed between IgG subclasses. The mutation load was lower in IgG^+CD27^- than IgG^+CD27^+ B cells, as previously published (13, 29). An interesting observation was that virtually unmutated IgV gene sequences (defined as ≥99% *IGHV* germ-line homology) were preferentially found among $IgG3^+CD27^-$ B cells: 11% and 6% (donors 1 and 2, respectively) and 12% and 14% of IgG1- and IgG2-switched sequences from $CD27^-$ B cells, respectively, were unmutated, whereas 25% and 18% of IgG3-switched sequences lacked somatic mutations ($P < 0.001$ by Fisher's exact test; Fig. 2*C*). In contrast, less than 1% of IgG^+CD27^+ sequences were unmutated. To assess whether this may be related to a frequent GC independent generation of $IgG3^+$ B cells (13), we analyzed the B-cell CLL/lymphoma 6 (*BCL6*) major mutation cluster (MMC) in conventional class-switched ($IgG^+IgG3^-CD27^+$), $IgG3^+CD27^+$ and $IgG3^+CD27^-$ B cells of two healthy donors as an indicator of a GC experience (22). We determined a mutation frequency of 0.08–0.11% in $IgG^+IgG3^-CD27^+$ lymphocytes (Table S2). $IgG3^+CD27^-$ B cells harbored on average significantly less *BCL6* mutations (0.05%), but still a substantial fraction of the cells showed mutations. Furthermore, among unmutated *IGHV* gene sequences from $IgG3^+CD27^-$ B cells belonging to clones, 65% and 34% (donors 1 and 2, respectively) of these sequences could be assigned to clones with somatically mutated members. Together, this may indicate that a small fraction of $IgG3^+CD27^-$ memory B cells is generated before GC diversification (*Discussion*).

The Human Memory B-Cell Pool Is Composed of $IgM^+(IgD^+)$ and IgG^+ Post-GC Cells, Which Are Often Common Members of Large Clones. We detected a surprisingly high fraction of 44% and 46% clonally related *IGHV* genes (Fig. 3*A*) in donors 1 and 2, respectively.

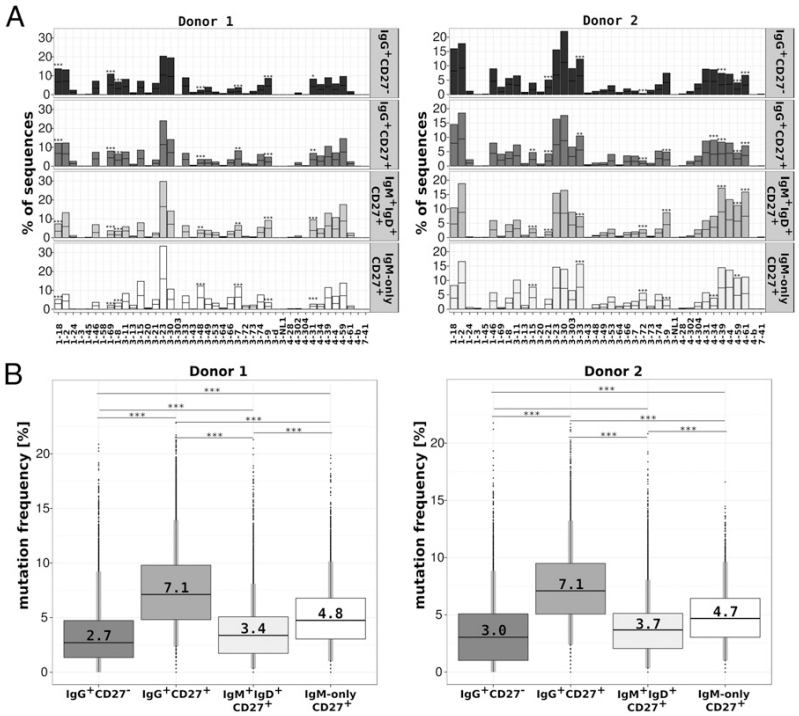


Fig. 1. BCR repertoire and mutation analysis of human PB memory B-cell subsets. (*A*) The relative use of individual *IGHV* gene segments of families 1, 3, 4, and 7 among memory B-cell subpopulations shows highly similar patterns. Statistically significant differences between individual subsets are marked (*P < 0.05, **P < 0.01, ***P < 0.001; Fisher's exact test). Test results were not corrected for multiple comparisons to estimate the maximum number of gene segments differing between B-cell subsets with the full specificity of Fisher's exact test. Only *IGHV* gene segments comprising at least 5% of total sequences in at least one condition and showing at least twofold difference in frequency between two B-cell subsets were considered. The separator in each column marks the amount of sequences contributed by each vial. (*B*) *IGHV* gene mutation frequencies (mutations/100 bp) of memory B-cell subsets are distinct. Median values (black bars) are given as numbers, and box plots represent 25 and 75 percentiles. ***P < 0.001; t test.

This observed clonality is not even saturated (Fig. 3*B*). A contamination with plasmablasts, containing up to 1,000-fold more Ig transcripts than memory lymphocytes, would lead to a significant overestimation of clonality. However, plasmablasts were largely excluded by cell sorting and a bias from cells with high Ig transcript level was eliminated by sequence collapsing (*Materials and Methods*). Moreover, the average number of identical sequences before collapsing did not differ significantly between clonally expanded (present in two replicates) and unexpanded sequences (present in one replicate), as determined by robust TOST, ε = 1 (https://cran.r-project.org/web/packages/equivalence/index.html), which would be the case if the observed clonality was simply derived from expanded plasmablasts. Similarly, an overestimation of clonality by PCR amplification is excluded by collapsing identical sequences as described in *Materials and Methods*.

An average of 33% (36% and 27% for donors 1 and 2, respectively) of the clones contained only IgG sequences, 47% (40% and 59%) contained only IgM sequences, and 20% (24% and 14%) of the clones were composed of IgM^+ and IgG^+ sequences (Fig. 3*A* and Table S3). The fraction of composite clones consisting of $IgM^+IgD^+CD27^+$ and IgM-only/IgG^+ B cells (considering that IgM-only B cells are accepted as post-GC B cells) increased with clone size and represented a large proportion (40.7% and 41.5%) of sequences assigned to clones with at least four members (Fig. 3*C*). Similar results were obtained, when $IgM^+IgD^+CD27^+$ and IgM-only were considered as one B-cell subset (Fig. S3*A*). Both parameters, clone size and composite clone structure, correlated significantly by power law with a smaller exponent for composite clones (−2.5 vs. −3.1). Thus, the more members a clone has, the higher the chance that it is composed of both IgG^+ (±IgM-only) and $IgM^+IgD^+CD27^+$ B cells. The frequency of sole IgG clones steadily decreased with rising clone size. However, there was still a considerable number of clones with only $IgM^+IgD^+CD27^+$ (Fig. 3*B*) or $IgM^+IgD^+CD27^+$ and IgM-only (Fig. S3*A*) members.

To clarify whether *IGHV* genes of $IgM^+IgD^+CD27^+$ B cells belonging to IgM^+/IgG^+ composite clones were distinct from sequences of unique $IgM^+IgD^+CD27^+$ B cells or clones with only $IgM^+IgD^+CD27^+$ B-cell members (potentially indicating a heterogeneity of the $IgM^+IgD^+CD27^+$ B-cell subset), their IgV gene rearrangement patterns were compared. However, *IGHV* gene use (Fig. S3*B*) and median CDRIII length (42 nucleotides) were practically identical. Although it seemed that the median mutation frequency of *IGHV* genes from unique $IgM^+IgD^+CD27^+$ B cells or clones composed of only these cells was mildly lower, this tendency was also detectable comparing unique and clonal IgM-only B-cell sequences (presumably reflecting that members of large clones have undergone on average more proliferation and hence SHM in the GC than members of small clones; Fig. S3*C*). We conclude that $IgM^+IgD^+CD27^+$ B cells without a detectable relationship to IgG^+ memory B cells and those being members of shared clones with IgG^+ memory B cells represent a homogenous population.

Taken together, our analysis shows a surprisingly high clonality among memory B-cell subpopulations. The fraction of IgM and IgG composite clones is substantial and, importantly, increases with clone size. Unique and clonal IgM^+ sequences are mostly identical in their BCR repertoire features.

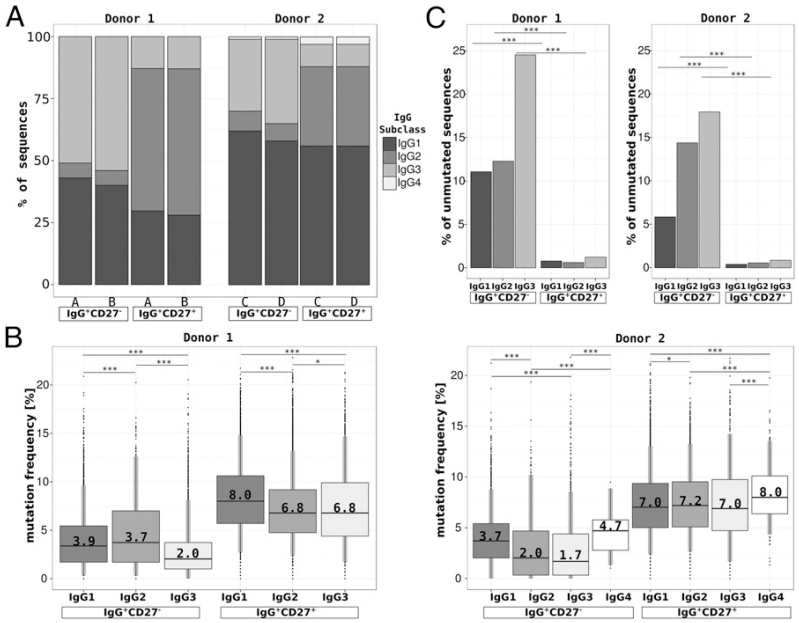


Fig. 2. IgG subclass use and mutation pattern of human PB memory B cells. (*A*) The IgG subclass composition of IgG^+CD27^+ and IgG^+CD27^- B-cell subsets revealed a significantly larger fraction of IgG3-switched B cells among the latter subpopulation ($P < 0.001$ by Fisher's exact test). (*B*) *IGHV* gene mutation frequencies (mutations/100 bp) of IgG memory B-cell subclasses. Median values (black bars) are given as numbers; box plots represent 25 and 75 percentiles. $*P < 0.05$, $***P < 0.001$; *t* test. (*C*) Ig-unmutated sequences were preferentially detectable among IgG^+CD27^- B cells, and among these, the $IgG3^+$ fraction showed the highest frequency of unmutated sequences.

$IgM^+IgD^+CD27^+$ and IgM-Only B Cells Represent a Homogenous Memory B-Cell Subset with Prolonged GC Participation of the Latter. The observation that IgM-only B cells show a mildly higher mutation frequency than $IgM^+IgD^+CD27^+$ B cells (Fig. 1*B*) may suggest a distinctness of both populations. However, a more detailed analysis revealed that their BCR repertoires are practically identical regarding *IGHV* gene use (Fig. S3*B*) and average CDRIII length (consistently 42 nucleotides). Minor variations in *IGHV* gene frequencies were similarly detectable in biological replicates, thus representing B-cell sampling variability. In line with this, IgM-only and $IgM^+IgD^+CD27^+$ B cells frequently belonged to common clones (Table S3). These results strongly indicate that IgM-only and $IgM^+IgD^+CD27^+$ B lymphocytes are one and the same population.

However, how can the different *IGHV* mutation frequencies be explained? Key to this question is the genealogical analysis of B-cell clones: from a total number of 9,312 (donor 1) and 5,301 (donor 2) clones, we selected 628 and 417 clones with at least seven members, respectively, for detailed analysis. Notably, in line with their higher mutation load, IgM-only sequences showed a significantly higher "mean distance to root" than clonally related $IgM^+IgD^+CD27^+$ sequences (0.72 vs. 0.68; $P < 0.001$, Fisher's exact test), i.e. IgM-only cells on average derived from more highly mutated members of a GC B cell clone than $IgM^+IgD^+CD27^+$ cells (Table S4 and Fig. S4).

Taken together, IgM-only and $IgM^+IgD^+CD27^+$ B cells derive from common GC reactions, and IgM-only memory B cells derive from GC B-cell clone members that acquired more mutations and down-regulated IgD and hence presumably typically resided longer in GCs.

Genealogic Analysis of Memory B-Cell Clones. B-cell clones often included members of distinct memory B-cell subsets. However, their distribution in clone dendrograms differed clearly. $IgM^+IgD^+CD27^+$ and/or IgM-only B cells in genealogic trees frequently showed a broadly diversified, "bushy" structure (Fig. 4 *A–C*) including many members with few mutations, presumably reflecting their generation in early phases of GC reactions. In contrast, IgG^+ clone members typically had more shared mutations (long branches) and mostly heavily mutated sequences (Fig. 4 *D–F*). Their single-rooted, narrow structures indicate generation in later GC-phases or on secondary GC passage. To further substantiate the distinctness of IgM and IgG memory B-cell clone patterns, we defined and measured typical genealogic tree parameters (32, 33) (Table S4 and Fig. S4). Whereas sole IgM or sole IgG clones can be clearly distinguished by dendrogram properties, composite clones show "composite tree structures" (Fig. 5). Thus, composite clones represent chimeras of sole IgM and IgG clones.

The interwoven pattern of IgM^+ and IgG^+ B cells in common clones substantiated that IgM memory B cells frequently derive from common GC reactions with class-switched memory B cells (Fig. 4 *G–L*, Fig. S5, and Table S5). A further interesting finding was that clones for which a substantial number of members were identified (1,045 clones with at least seven members) were rarely dominated by one cell type (Fig. 4 *B–H*), but usually a surprisingly heterogeneous clone composition in terms of B-cell subset was observed (Fig. 4 *J–L* and Table S5). The picture was different when considering IgG subclass use in clones with at least seven IgG members. For 79% of clones, only a single IgG subclass was identified. Dendrograms including more than one IgG subclass rarely showed subbranch-specific class switching events (Table S5 and Fig. S5).

Taken together, the genealogical analysis of memory B-cell clones strongly underlines the frequent generation of memory B-cell subsets in common GC reactions. Moreover, we revealed particular patterns of memory B-cell clone generation in GC reactions and relationships between distinct memory B-cell subsets.

Discussion

In the present work, we obtained detailed insight into the composition of the human memory B-cell pool in terms of *IGHV*

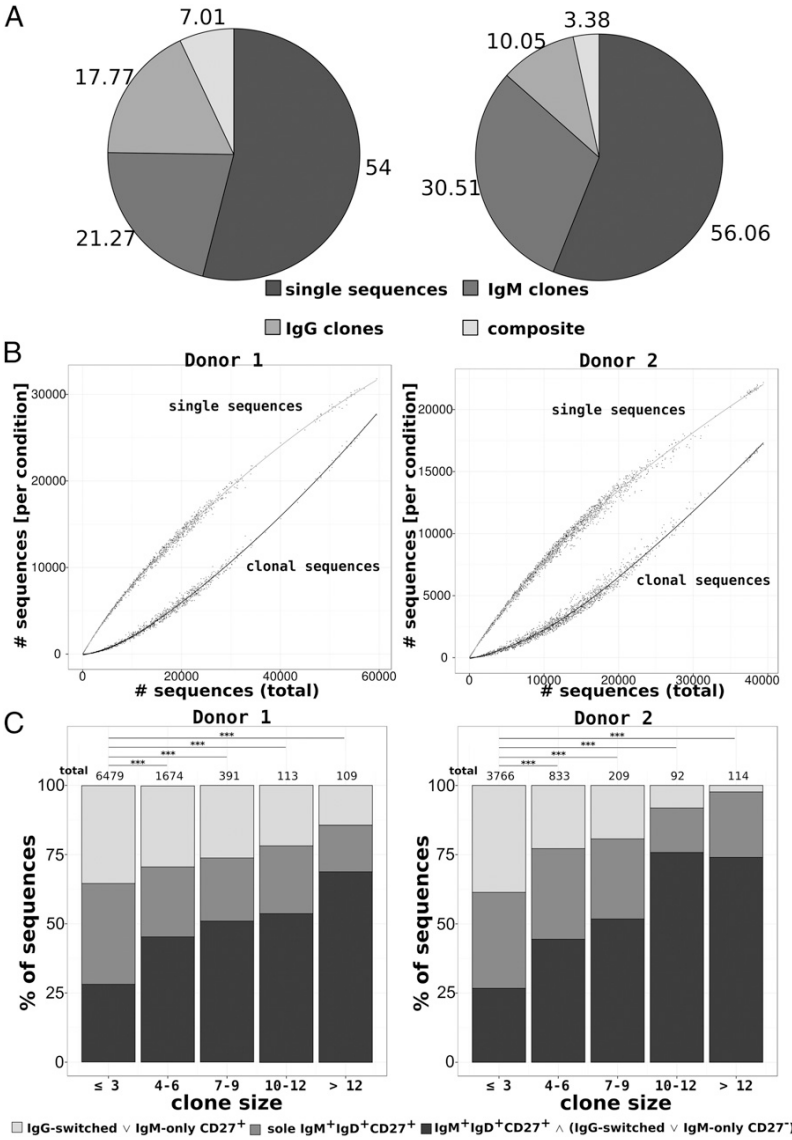


Fig. 3. Clonal composition of human memory B cells. (*A*) The relative fraction of single and clonal B-cell sequences per donor is given. Clonal sequences are split up into sole IgM ($IgM^+IgD^+CD27^+$ or IgM-only) clones, sole IgG clones and composite clones ($IgM^+IgD^+CD27^+$ and/or IgM-only and IgG^+ B-cell sequences). Numbers denote the fraction of sequences in each category. (*B*) To estimate the fraction of clones with increasing sample size, we determined clonality among randomly selected sequence samples per donor—sample sizes ranging from two to the maximum number of available sequences—by our CDRIII clustering approach. The regression curves (locally weighted scatter plot smoothing) revealed an unsaturated clonality of the memory B-cell pool in both donors. (*C*) Correlation of clone type fractions and clonal sizes. The larger a clone, the more likely it is of composite subtype ($IgM^+IgD^+CD27^+$ and IgM-only/IgG^+ B cells). This correlation is statistically highly significant (***$P < 0.001$; Fisher's exact test, composite vs. noncomposite) already for clone sizes of more than three members. In contrast, clones consisting only of IgG^+ B-cell sequences (with or without IgM-only B cells) are practically undetectable when sufficient numbers of B cells are analyzed. Bin sizes were chosen arbitrarily for clarity of the depiction, Fig. S3*D* shows a version of this figure lacking bins. In the legend, ∧ indicates "and," and ∨ indicates "or."

gene repertoire, clonal composition, complexity, and relationship between $IgM^+IgD^+CD27^+$, IgM-only, IgG^+CD27^+, and IgG^+CD27^- B cells, including IgG subclass information. Considering only sequences with at least twofold coverage, we reliably determined intraclonal diversity, as technical artifacts were largely eliminated. Additionally, replicate analyses significantly enhanced the reproducibility and stability of statistical evaluations and allowed for precise determination of clonal expansions. Evaluating 41,000 and 67,000 memory B cells of two donors, we obtained a representative overview of the memory *IGHV* gene repertoire and clonal composition. The four B-cell subsets analyzed were strikingly similar in *IGHV* gene use and CDRIII length, implying identical generation pathways and highly similar selection processes. However, there were significant differences in the *IGHV* gene mutation loads, with IgM-only B cells carrying on average more mutations than $IgM^+IgD^+CD27^+$ B cells and IgG^+CD27^+ memory B cells carrying more *IGHV* mutations than IgG^+CD27^- B cells. These observations validate and extend prior smaller studies (13, 27, 31). The distribution of mutation frequencies indicated that each B-cell subset was homogenous and not a mixture of two or more

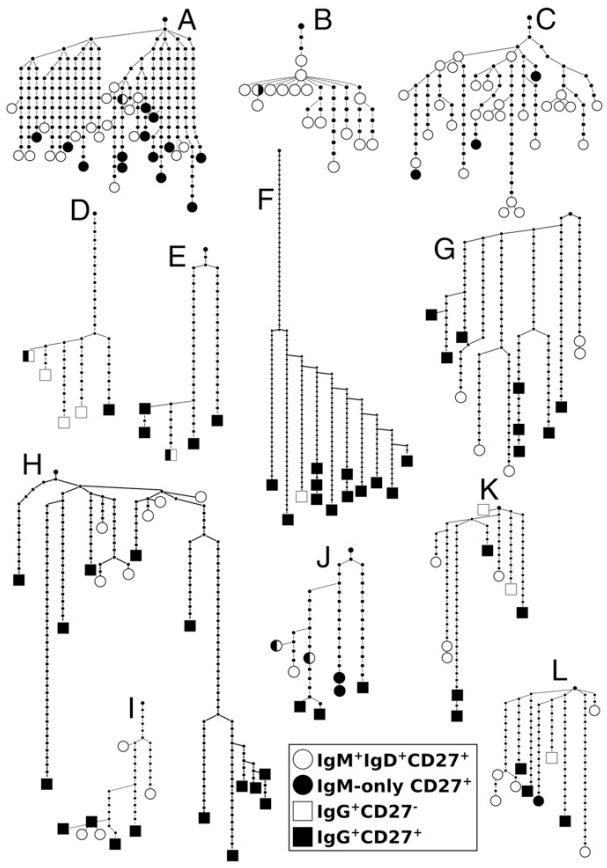


Fig. 4. Genealogic analysis of human memory B-cell clones. (*A–C*) Sole IgM clones usually show an early and broad diversification, leading to bushy dendrograms. (*D–F*) Sole IgG clones tend to have long roots and narrow shapes. (*G–L*) typical examples of composite clone dendrograms, where IgM and IgG memory B-cell subsets are intermingled. Specific subsets are indicated by a gray scale code (legend), germ-line *IGHV* (root), and sequential mutation events are marked as small black circles. Less than 1% of nodes or leaves represent more than one sequence; occasional contributions from independent subsets are given as circles with split shading.

major separate populations. The interindividually consistent mutation frequencies indicate that memory B-cell subsets are generated largely independent of individual immune histories (see discussion of clonal compositions).

A further notable observation was that IgG$^+$CD27$^-$ B cells include a considerable fraction of *IGHV*-unmutated sequences, particularly among IgG3$^+$CD27$^-$ B cells (25% and 18% of sequences unmutated in donors 1 and 2, respectively). An analysis of *BCL6* mutations, a molecular indicator of a GC experience as only GC B cells acquire *BCL6* mutations (22), revealed a potential generation of (a fraction of) IgG3$^+$CD27$^-$ B cells before GC differentiation, as recently described for some murine memory B cells (34). However, the involvement of T-cell support in the generation of these pre-GC memory B cells is supported by the considerable amount of unmutated IgG3$^+$CD27$^-$ B-cell sequences belonging to somatically diversified clones. Finally, as still 20% of IgG3$^+$CD27$^-$ vs. 30–40% of IgG$^+$CD27$^+$ memory B cells were *BCL6* mutated, this nevertheless indicates that the majority of IgG3$^+$CD27$^-$ B cells (harboring also a low *IGHV* mutation load) is GC derived.

A major finding of our analysis is the surprisingly high degree of clonal relation among memory B cells (45% of total sequences). Certainly, it is expected that, from a single GC clone, numerous memory B cells are generated. However, considering that only 41,000 and 67,000 cells of an estimated human adult PB memory B-cell pool of 2.6×10^8 cells were analyzed, this striking clonality was unexpected. Clearly, we still underestimate the extent of clonal relatedness, given the restricted numbers of memory B cells investigated, as is evident from the nonsaturated fraction of clonal sequences in both samples (Fig. 3*B*). Several clones with more than 50 members were detected, which may be projected to total sizes of more than 150,000 members in PB (and presumably more members in lymphoid tissues) (35). Importantly, the high clonality of the memory B-cell compartment does by no means entail that this compartment is restricted in its complexity, due to

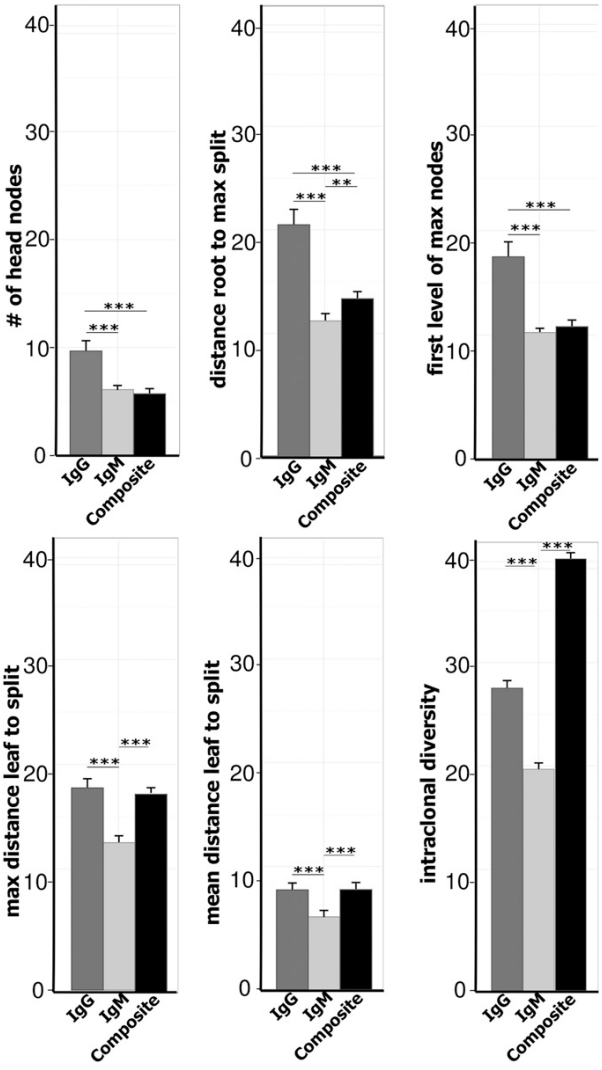


Fig. 5. Statistical comparison of genealogical trees of memory B-cell clone subtypes. Tree parameters are defined in Table S4. Whereas sole IgM and sole IgG clones are statistically significant different in their dendrogram patterns, composite clones show combinations of these patterns. The higher intraclonal diversity of mixed clones compared with sole IgM or IgG clones likely reflects that the combination of mutation patterns of IgM and of IgG members results in high values for intraclonal diversity (**$P < 0.01$, ***$P < 0.001$; paired *t* test, error bars show SEM).

intraclonal diversity and diversity in Ig isotype or IgG subclass composition (see below).

Because the origin of $IgM^+IgD^+CD27^+$ B cells is debated, we were particularly interested in the characterization of clone compositions. In both donors, 20% of clonal sequences belonged to composite IgG^+ and IgM^+ B-cell clones. At first glance, this implicates that most clones are either class-switched or non–class-switched, but importantly, the largest fraction of non-composite clones consists of only two or three members. Indeed, with increasing clone size, the fraction of composite clones significantly rises, representing more than 50% of clones with more than nine members (Fig. 3*B*), whereas the frequency of sole IgG clones steadily decreased, suggesting that for most if not all IgG^+ clones $IgM^+IgD^+CD27^+$ members can be identified if enough sequences are analyzed.

Moreover, the highly similar *IGHV* gene use and CDRIII length distribution of $IgM^+IgD^+CD27^+$ B cells in comparison with classical GC-derived IgG^+ and IgM-only memory B cells (with an accepted GC origin), is a further strong indication that also $IgM^+IgD^+CD27^+$ B cells are GC derived. This argument holds also true for unique $IgM^+IgD^+CD27^+$ B-cell sequences and (usually small) $IgM^+IgD^+CD27^+$ B-cell clones without IgG^+ members: the highly similar *IGHV* gene repertoires of $IgM^+IgD^+CD27^+$ B-cell clonally related to IgG^+ memory B cells and those without such a relationship detected argues that at least the majority of unique and sole IgM^+ clone sequences represent a homogenous population with those $IgM^+IgD^+CD27^+$ B cells with a clear GC origin. Perhaps, GC-independent $IgM^+IgD^+CD27^+$ B cells exist in young children (16, 18, 36), but become a minor B-cell population in adults. A frequent common GC origin of $IgM^+IgD^+CD27^+$ and IgG^+ memory B cells was already indicated from a previous small scale PCR study (22). Moreover, our recent global gene expression profiling study also indicated a close relationship of $IgM^+IgD^+CD27^+$ and IgG^+ B cells, sharing key features of post-GC memory B cells (37). The existence of human post-GC $IgM^+IgD^+CD27^+$ memory B cells is further indicated from a study of specific memory B cells (38). The main reason why prior NGS studies detected few (if any) clonally related IgM^+ and IgG^+ B cells is most likely that too few B cells were analyzed per donor; the chance to find composite clones simply increases with sample size (27–29). Further insight into the GC-dependent generation of IgM memory B cells is provided by genealogic tree analyses of more than 1,000 informative clone dendrograms. First, IgM and IgG clone members are often intermingled and thus derived from a single mutating GC B-cell clone, excluding that these mutated IgM^+ B cells were generated GC independently. Second, IgM members on average locate more close to the root, explaining their lower *IGHV* mutation frequency. Assuming that somatic mutations accumulate (probably not linearly but steadily) with additional cycles of proliferation and mutation of GC B cells, higher mutated B cells on average experienced an extended GC residence. Therefore, we propose shifts in the generation of distinct memory B-cell subsets in the course of the GC reaction. This idea is substantiated by distinct replication histories, showing on average less cell divisions for, e.g., $IgM^+IgD^+CD27^+$ than IgG^+CD27^+ B cells (31). Hence, the lower mutation load of IgM^+CD27^+ B cells does not indicate a separate (GC independent) origin of these cells, as has been proposed by others (21). Third, the relationship of IgM-only and $IgM^+IgD^+CD27^+$ B cells is resolved: previous studies discussed that these subsets represent developmentally or functionally different B-cell subtypes (20, 21, 39). However, their *IGHV* gene repertoire similarity is striking, as is the clonal overlap between both subsets. Finally, the significantly longer distance to root of IgM-only B cells in clone dendrograms may explain the reduced IgD expression and higher mutation load in IgM-only B cells by their generation from more advanced GC B cells. Notably, B cells down-regulate IgD on prolonged stimulation (40). The homogeneity of PB IgM-only and $IgM^+IgD^+CD27^+$ B cells is also supported by their identical gene expression patterns (37).

Similar to the relationship of $IgM^+IgD^+CD27^+$, IgM-only, and IgG^+ memory B cells, also IgG^+CD27^- and IgG^+CD27^+ B cells apparently are generated in an ordered pattern in common GC reactions, rather than in distinct immune responses, as CD27 is up-regulated in GC B cells and IgG^+CD27^+ lymphocytes show higher mutation loads, but both subsets show high clonal overlap in this and a previous study (28). Many examples showed that IgG^+CD27^- and IgG^+CD27^+ memory B cells were often derived from common GC B-cell clones.

Clones including multiple IgG^+ members are often dominated by a single IgG subclass. In these instances, class-switching might have occurred early in GC B-cell clone expansion and remained stable without consecutive switching events, or it occurred multiple times in the GC B-cell clone, but was repeatedly directed to the same IgG subclass. Indeed, targeting of specific Ig subclasses during class switching is well regulated (41). The few large IgM^+ B-cell clones lacking IgG^+ members (Fig. 4*A*) indicate that GC reactions without generation of IgG^+ memory B cells can also exist.

When analyzing clone dendrograms (Fig. 4 and Fig. S5), one must keep in mind that we sampled only a small fraction of the clone members, and some memory clones might be composed of members that underwent different numbers of GC reactions. Nevertheless, the analysis of typical, repeatedly identified dendrogram structures allows inferences on GC dynamics and selective pressures guiding memory B-cell development:

Our study indicates that the early and broadly diversified IgM^+ post-GC B cells represent a reservoir of flexible lymphocytes that facilitate immune adaptation to modified pathogens. In contrast, some sole IgG^+ and mixed clones (Fig. 4 *D* and *F* and Fig. S5 *B* and *F*) show a high number of shared mutations from the root to the first node, which in most or all instances might reflect a secondary (or higher order) GC passage of a memory B-cell that has acquired the shared mutations in (an) earlier GC passage(s). Indeed, murine and also human IgM memory B cells preferentially reenter GC reactions (37, 42), although also IgG memory B cells have this capacity (43). Some clones are characterized by many individual mutations per clone member (Fig. 4 *F* and Fig. S5 *A* and *I*). These high numbers of member-specific mutations may reflect that in these instances mostly only highly mutated GC B-cell clone members were selected into the memory B cell pool and that many members of subbranches did not survive the selection in the GC, with only particular combinations of mutations yielding a sufficient high affinity and fitness for selection into the memory B-cell pool. The detection of single or few clone members with many unique mutations (i.e., early branching) belonging to a distinct memory B-cell subset in clones otherwise dominated by (an) other memory B-cell subset(s) (Fig. 4*F* and Fig. S5*A*) indicates that the decision to undergo class-switching or to express CD27 can sometimes be made early in a member of a GC B-cell clone and be kept during multiple additional rounds of proliferation and mutation until the cells differentiate into memory B cells.

Finally, most clone members differ from each other by point mutations, meaning that throughout all phases of a GC reaction, an amazing variety of memory B cells is produced. This important immune strategy to produce a diverse antigen-specific memory B-cell compartment facilitates responses to variants of the original antigen.

Taken together, human PB memory B-cell subsets share a highly similar *IGHV* gene repertoire, and most if not all $IgM^+IgD^+CD27^+$ B cells in adults are post-GC memory B cells. Moreover, memory B cells show a surprisingly high clonality and often include very large clones, composed of most or even all combinations of memory subsets and IgG subclasses. Thus, IgM^+ and IgG^+ memory B cells often derive from common GC reactions. Their specific generation is dynamically and presumably chronologically regulated.

Materials and Methods

Cell Separation. Two healthy adult donors for PB (both male, 35 and 38 y of age, infection free for more than 6 mo) were recruited from the Medical School in Essen. The study protocol was approved by the Internal Review Board of the Medical School in Essen. PB mononuclear cells were isolated by Ficoll-Paque density centrifugation (Amersham) from 500 mL PB. $CD19^+$ B cells were enriched to >98% by magnetic cell separation using the MACS system (Miltenyi Biotech).

Cell Sorting. The B-cell suspension of each donor was split into two aliquots and stained with anti-CD27-APC, anti-IgD-PECy7, anti-CD23-PE and anti-IgM-FITC or anti-CD27-APC, anti-IgD-PE, and anti-IgG-FITC antibodies (all Becton Dickinson Biosciences) and sorted with a FACSAria cell sorter (Becton Dickinson Biosciences) as IgM^+IgD^+ memory ($IgM^+IgD^+CD27^+CD23^-$), IgM-only ($IgM^+IgD^-CD27^+CD23^{low/-}$), IgG^+CD27^+ memory ($IgD^-IgG^+CD27^+$), or IgG^+ $CD27^-$ memory ($IgD^-IgG^+CD27^-$) B cells according to the relative frequency of each population in two equal-sized replicates per population. Plasmablasts ($CD23^-CD27^{high}$) were excluded from the analysis. Purity was >99% for each population as determined by reanalysis on a FACSCanto flow cytometer (Becton Dickinson Biosciences) in combination with FACSDiva software.

Experimental Strategy. The four major human PB B-cell subsets carrying mutated IgV genes were sort-purified to a total of 200,000 B lymphocytes, according to their relative frequency in PB, from two adult healthy donors (Fig. S1 *A* and *B* and Table S1). To control for technical bias, each population was sorted in two equal-sized biological replicates (termed A and B for donor 1, and C and D for donor 2) and processed in parallel. The average mean fluorescence intensity (MFI) of IgD expression on sorted cells was 519 and 324 for IgM-only and 5,748 and 4,603 for $IgM^+IgD^+CD27^+$ B cells (donors 1 and 2, respectively), i.e., the two subsets of IgM^+CD27^+ B cells were clearly separated. RNA was extracted, and full-length *IGHV* gene rearrangements of the *IGHV1*, *3* and *4* families (the VH1 primer also amplifies *IGHV7* family gene segments), including the 5′ part of *IGHC*, were amplified. The *IGHC* primer for Cγ was designed to allow the determination of the Cγ subclass. PCR products were processed and sequenced on a Roche 454 Sequencer. To exclude artificial sequence variants, we aimed at 10-fold coverage of each rearrangement (retrieving on average 2×10^6 sequences per 200,000 cells per donor) so that after quality filtering (base calling, minimum length, and 454 error correction), we based data analysis on sequences that were detected at least twice (mean coverages: ninefold and fourfold for donors 1 and 2, respectively). Identical sequences in one cell aliquot were collapsed and counted once. With this strategy, we aimed at avoiding potential PCR-introduced biases of the repertoire (Table S1). This strategy also eliminates a potential bias due to contaminating plasmablasts, as identical transcripts from a plasmablast would be collapsed to a single sequence. To account for germ-line *IGHV* diversity, for every donor, the germ-line configuration of *IGHV1*, *3* and *4* alleles was determined from the two most frequently used *IGHV* alleles among sequences scored unmutated. As population-based PCR approaches can generate PCR hybrid artifacts, we determined the *IGHV* gene assignment of a random collection of 1,000 sequences for the 5′ and 3′ end of the *IGHV* region independently. Only for six sequences, the *IGHV* gene assignment was inconsistent, indicating a neglectable PCR hybrid artifact frequency in our sample.

RNA Isolation and *IGHV* RT-PCR. B cells were sorted into TRIzol lysis buffer (Sigma-Aldrich), and RNA was isolated with the RNeasy micro Kit (Qiagen) and reverse transcribed with primers specific for Cμ (5′-CCACGCTGCTCGTAT-3′) and Cγ (5′-TAGTCCTTGACCAGG-3′) for 1 h at 42 °C according to the Sensiscript protocol (Qiagen). *IGHV* gene PCR of *IGHV1* (including *IGHV7*), *3*, and *4* family rearrangements was carried out with leader peptide-specific primers 5′-CTCACCATGGACTGGACCTGGAG-3′ (VHL1), 5′-ACCATGGAGTTTGGGCTGAGCTG-3′ and 5′-ACCATGGAACTGGGGCTCCGCTG-3′ (VHL3.1 and 3.2, respectively), and 5′-AAGAACATGAAACACCTGTGGTTCTTC-3′ (VHL4) and 5′-GCTCGTATCCGACGGGGAATTCTCAC-3′ (Cμ) or 5′-GCAGCCCAGGGCSGCTGTGC-3′ (Cγ) specific primers at 60 °C annealing temperature for 35 cycles with the Phusion High-Fidelity DNA polymerase (Finnzymes; Thermo Scientific).

***Bcl6* Mutation Analysis.** The *Bcl6* major mutation cluster was amplified by seminested PCR strategy in 2.5 mM $MgCl_2$, 125 μM dNTPs, 0.125 μM each primer (5′-CGCTCTTGCCAAATGCTTTGGC-3′ and 5′-CTCTCGTTAGGAAGATCACGGC-3′), and 1.2 U High Fidelity DNA polymerase mix (Roche) in the first and 1.75 mM $MgCl_2$, 67 μM dNTPs, 0.125 μM each primer (5′-CGCTCTTGCCAAATGCTTTG-3′ and 5′-GACACGATACTTCATCTCATC-3′), and 1.2 U Fermentas Taq DNA polymerase in the second round of amplification. PCR products were purified with EZNA Cycle pure kit (VWR International), cloned with the pGEM-T Easy cloning kit (Promega), and sequenced from both strands with second-round amplification primers.

Generation of Amplicon Library. For unidirectional sequencing of *IGHV* gene rearrangements with the GS FLX Titanium emPCR Kit (Lib-L) (Roche), the appropriate adapter sequences with different barcodes defining the two donors, the B-cell populations, and the replicates were added by PCR to the amplified *IGHV* gene templates. Each library was gel-purified (QIAquick; Qiagen), and the appropriate amount of amplified DNA was pooled according to the relative size of each B-cell population in PB before sequencing with Roche 454 GS-FLX+ Titanium by LGC Genomics.

Determination of Clonality. Sequences were considered clonally related when using the same *IGHV* gene and sharing at least 90% CDRIII sequence identity, accounting for intraclonal diversity by SHM. A CDRIII length tolerance of 5% was included to consider insertions or deletions generated by SHM. Moreover, clonal sequences had to be present either in two different B-cell populations or replicates or had to have at least two nonshared substitutions in the *IGHV* segments, accounting for rare PCR-introduced nucleotide variants. These stringent parameters revealed a high number of clones, but also included presumably 5% false positives. The latter was estimated through manual evaluation of alignments. If, for example, N-nucleotides differed by several nucleotides or different *IGHD* or *IGHJ* segments were used, such sequences were treated as not expanded.

Bioinformatics. All statistical and bioinformatical evaluations were performed in R (www.R-project.org/) and based on the international ImMunoGeneTics information system (IMGT) database (www.imgt.org/). The determination of IgG subclass use was based on pairwise alignments of the amplified C regions with the germ line. Mutation frequencies and genealogical trees were calculated based on the number or relative position of nucleotide exchanges in the *IGHV* region of each sequence in comparison with the most similar allelic variant present in the respective donor (determined from unmutated sequences). Genealogic trees were calculated with IgTree (kindly provided by Ramit Mehr, Bar-Ilan Universität, Ramat-Gan, Israel) (44). Intraclonal diversity denotes the mean number of nonshared substitutions of all sequences belonging to a clone.

ACKNOWLEDGMENTS. We thank Julia Jesdinsky-Elsenbruch and Sarah Taudien for excellent technical assistance and Klaus Lennartz for valuable engineering support. This work was supported by the Deutsche Forschungsgemeinschaft through Grants Ku1315/8-1, GRK1431, SE1885/2-1, and TRR60/A2 and B1.

1. Alt FW, Blackwell TK, Yancopoulos GD (1987) Development of the primary antibody repertoire. *Science* 238(4830):1079–1087.
2. Goossens T, Klein U, Küppers R (1998) Frequent occurrence of deletions and duplications during somatic hypermutation: Implications for oncogene translocations and heavy chain disease. *Proc Natl Acad Sci USA* 95(5):2463–2468.
3. Küppers R, Zhao M, Hansmann ML, Rajewsky K (1993) Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. *EMBO J* 12(13):4955–4967.
4. MacLennan IC (1994) Germinal centers. *Annu Rev Immunol* 12:117–139.
5. Manis JP, Tian M, Alt FW (2002) Mechanism and control of class-switch recombination. *Trends Immunol* 23(1):31–39.
6. Rajewsky K (1996) Clonal selection and learning in the antibody system. *Nature* 381(6585):751–758.
7. Tarlinton D (2006) B-cell memory: Are subsets necessary? *Nat Rev Immunol* 6(10):785–790.
8. van Es JH, Meyling FH, Logtenberg T (1992) High frequency of somatically mutated IgM molecules in the human adult blood B cell repertoire. *Eur J Immunol* 22(10):2761–2764.
9. Klein U, Küppers R, Rajewsky K (1997) Evidence for a large compartment of IgM-expressing memory B cells in humans. *Blood* 89(4):1288–1298.
10. Klein U, Rajewsky K, Küppers R (1998) Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J Exp Med* 188(9):1679–1689.
11. Giesecke C, et al. (2014) Tissue distribution and dependence of responsiveness of human antigen-specific memory B cells. *J Immunol* 192(7):3091–3100.
12. Agematsu K, et al. (1997) B cell subpopulations separated by CD27 and crucial collaboration of CD27+ B cells and helper T cells in immunoglobulin production. *Eur J Immunol* 27(8):2073–2079.
13. Fecteau JF, Côté G, Néron S (2006) A new memory CD27-IgG+ B cell population in peripheral blood expressing VH genes with low frequency of somatic mutation. *J Immunol* 177(6):3728–3736.
14. Weller S, et al. (2004) Human blood IgM "memory" B cells are circulating splenic marginal zone B cells harboring a prediversified immunoglobulin repertoire. *Blood* 104(12):3647–3654.

15. Kruetzmann S, et al. (2003) Human immunoglobulin M memory B cells controlling Streptococcus pneumoniae infections are generated in the spleen. *J Exp Med* 197(7): 939–945.
16. Griffin DO, Holodick NE, Rothstein TL (2011) Human B1 cells in umbilical cord and adult peripheral blood express the novel phenotype CD20+ CD27+ CD43+ CD70-. *J Exp Med* 208(1):67–80.
17. Inui M, et al. (2015) Human CD43+ B cells are closely related not only to memory B cells phenotypically but also to plasmablasts developmentally in healthy individuals. *Int Immunol* 27(7):345–355.
18. McWilliams L, et al. (2013) The human fetal lymphocyte lineage: Identification by CD27 and LIN28B expression in B cell progenitors. *J Leukoc Biol* 94(5):991–1001.
19. Scheeren FA, et al. (2008) T cell-independent development and induction of somatic hypermutation in human IgM+ IgD+ CD27+ B cells. *J Exp Med* 205(9):2033–2042.
20. Weller S, et al. (2001) CD40-CD40L independent Ig gene hypermutation suggests a second B cell diversification pathway in humans. *Proc Natl Acad Sci USA* 98(3): 1166–1170.
21. Weill JC, Weller S, Reynaud CA (2009) Human marginal zone B cells. *Annu Rev Immunol* 27:267–285.
22. Seifert M, Küppers R (2009) Molecular footprints of a germinal center derivation of human IgM+(IgD+)CD27+ B cells and the dynamics of memory B cell generation. *J Exp Med* 206(12):2659–2669.
23. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135(3): 183–191.
24. Baum PD, Venturi V, Price DA (2012) Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. *Eur J Immunol* 42(11): 2834–2839.
25. Jackson KJ, Kidd MJ, Wang Y, Collins AM (2013) The shape of the lymphocyte receptor repertoire: Lessons from the B cell receptor. *Front Immunol* 4:263.
26. Michaeli M, Noga H, Tabibian-Keissar H, Barshack I, Mehr R (2012) Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front Immunol* 3:386.
27. Wu YC, et al. (2010) High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 116(7):1070–1078.
28. Wu YC, Kipling D, Dunn-Walters DK (2011) The relationship between CD27 negative and positive B cell populations in human peripheral blood. *Front Immunol* 2:81.
29. Jackson KJ, Wang Y, Collins AM (2014) Human immunoglobulin classes and subclasses show variability in VDJ gene mutation levels. *Immunol Cell Biol* 92(8):729–733.
30. Brezinschek HP, Brezinschek RI, Lipsky PE (1995) Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *J Immunol* 155(1):190–202.
31. Berkowska MA, et al. (2011) Human memory B cells originate from three distinct germinal center-dependent and -independent maturation pathways. *Blood* 118(8):2150–2158.
32. Horesh Y, Mehr R, Unger R (2006) Designing an A* algorithm for calculating edit distance between rooted-unordered trees. *J Comput Biol* 13(6):1165–1176.
33. Shahaf G, et al. (2008) Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: A large-scale simulation study. *J Theor Biol* 255(2):210–222.
34. Takemori T, Kaji T, Takahashi Y, Shimoda M, Rajewsky K (2014) Generation of memory B cells inside and outside germinal centers. *Eur J Immunol* 44(5):1258–1264.
35. Seifert M, et al. (2009) A model for the development of human IgD-only B cells: Genotypic analyses suggest their generation in superantigen driven immune responses. *Mol Immunol* 46(4):630–639.
36. Descatoire M, et al. (2014) Identification of a human splenic marginal zone B cell precursor with NOTCH2-dependent differentiation properties. *J Exp Med* 211(5): 987–1000.
37. Seifert M, et al. (2015) Functional capacities of human IgM memory B cells in early inflammatory responses and secondary germinal center reactions. *Proc Natl Acad Sci USA* 112(6):E546–E555.
38. Narváez CF, et al. (2012) Human rotavirus-specific IgM Memory B cells have differential cloning efficiencies and switch capacities and play a role in antiviral immunity in vivo. *J Virol* 86(19):10829–10840.
39. Werner-Favre C, et al. (2001) IgG subclass switch capacity is low in switched and in IgM-only, but high in IgD+IgM+, post-germinal center (CD27+) human B cells. *Eur J Immunol* 31(1):243–249.
40. Pascual V, et al. (1994) Analysis of somatic mutation in five B cell subsets of human tonsil. *J Exp Med* 180(1):329–339.
41. Xu Z, Zan H, Pone EJ, Mai T, Casali P (2012) Immunoglobulin class-switch DNA recombination: Induction, targeting and beyond. *Nat Rev Immunol* 12(7):517–531.
42. Dogan I, et al. (2009) Multiple layers of B cell memory with different effector functions. *Nat Immunol* 10(12):1292–1299.
43. McHeyzer-Williams LJ, Milpied PJ, Okitsu SL, McHeyzer-Williams MG (2015) Class-switched memory B cells remodel BCRs within secondary germinal centers. *Nat Immunol* 16(3):296–305.
44. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R (2008) IgTree: Creating Immunoglobulin variable region gene lineage trees. *J Immunol Methods* 338(1-2):67–74.

Supporting Information

Budeus et al. 10.1073/pnas.1511270112

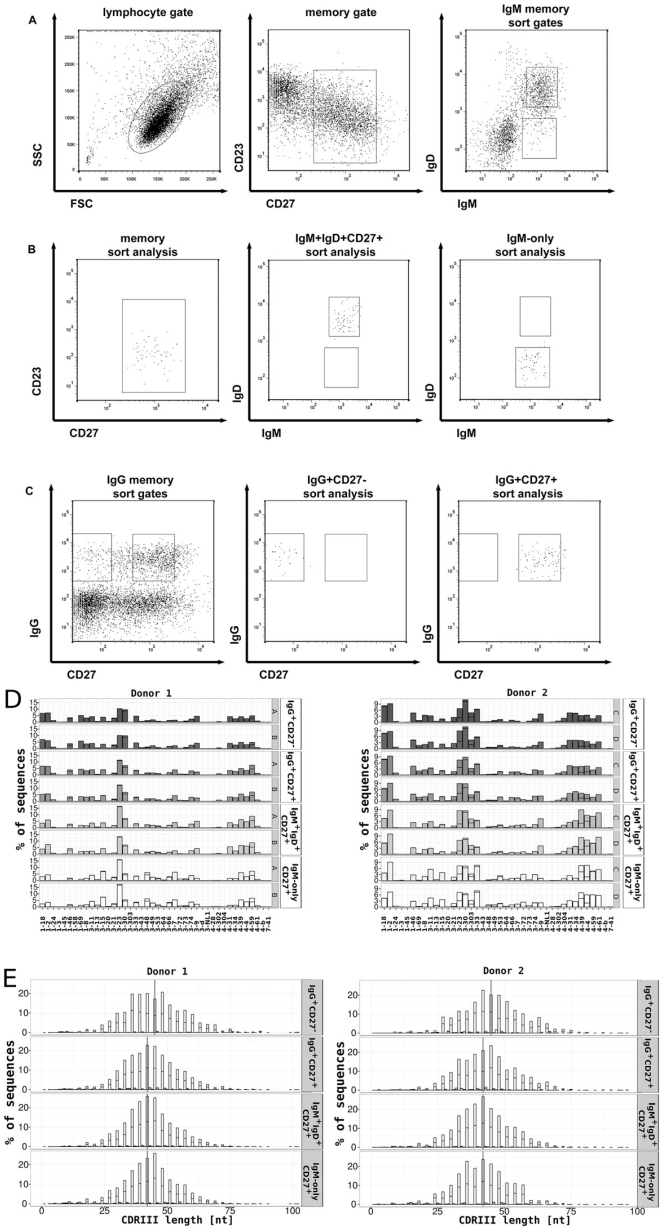


Fig. S1. Cell sorting strategy of memory B-cell subpopulations. (*A*) IgM$^+$IgD$^+$CD27$^+$ and IgM-only B cells are separated after enrichment of B cells by CD19 MACS according to their surface IgD level. CD27^{high} plasmablasts are excluded. (*B*) Postsort analysis of IgM$^+$IgD$^+$CD27$^+$ and IgM-only B cells from donor 1. (*C*) IgG$^+$CD27$^+$ and IgG$^+$ CD27$^-$ B cells are defined by surface IgG and CD27 expression. Pre- and postsort analysis of donor 1 is shown. (*D*) The relative use of individual *IGHV* gene segments of families 1, 3, 4, and 7 among memory B-cell subpopulations and biological replicates is highly similar. Minor variations between subpopulations are detectable at similar ranges in biological replicates. No statistically significant differences are detectable between any two conditions among single *IGHV* genes with >5% frequency and greater than twofold change. The separator in each column marks the amount of sequences contributed by each allelic variant of the respective IGHV gene. (*E*) CDRIII length spectratyping reveals highly similar distributions between memory B-cell subpopulations. The separator in each column marks the amount of sequences contributed by each vial.

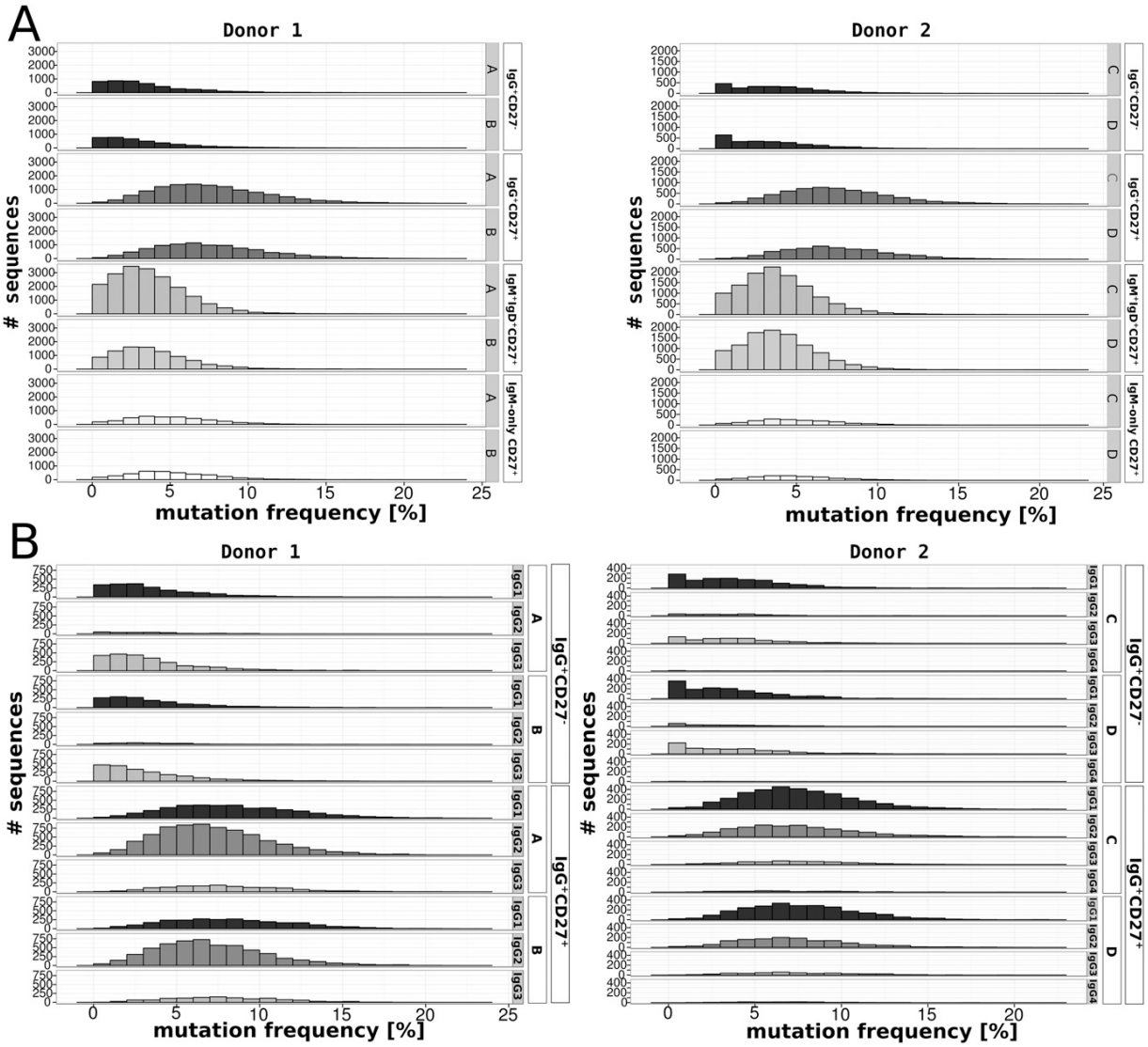


Fig. S2. Distribution of mutation frequencies in B-cell subpopulations and biological replicates. (*A*) In both donors, the distributions in mutation frequency (mutations/100 bp) are highly similar in replicate analyses and distinct for each memory B-cell population analyzed. (*B*) The distribution of mutation frequencies (mutations/100 bp) in IgG subclasses.

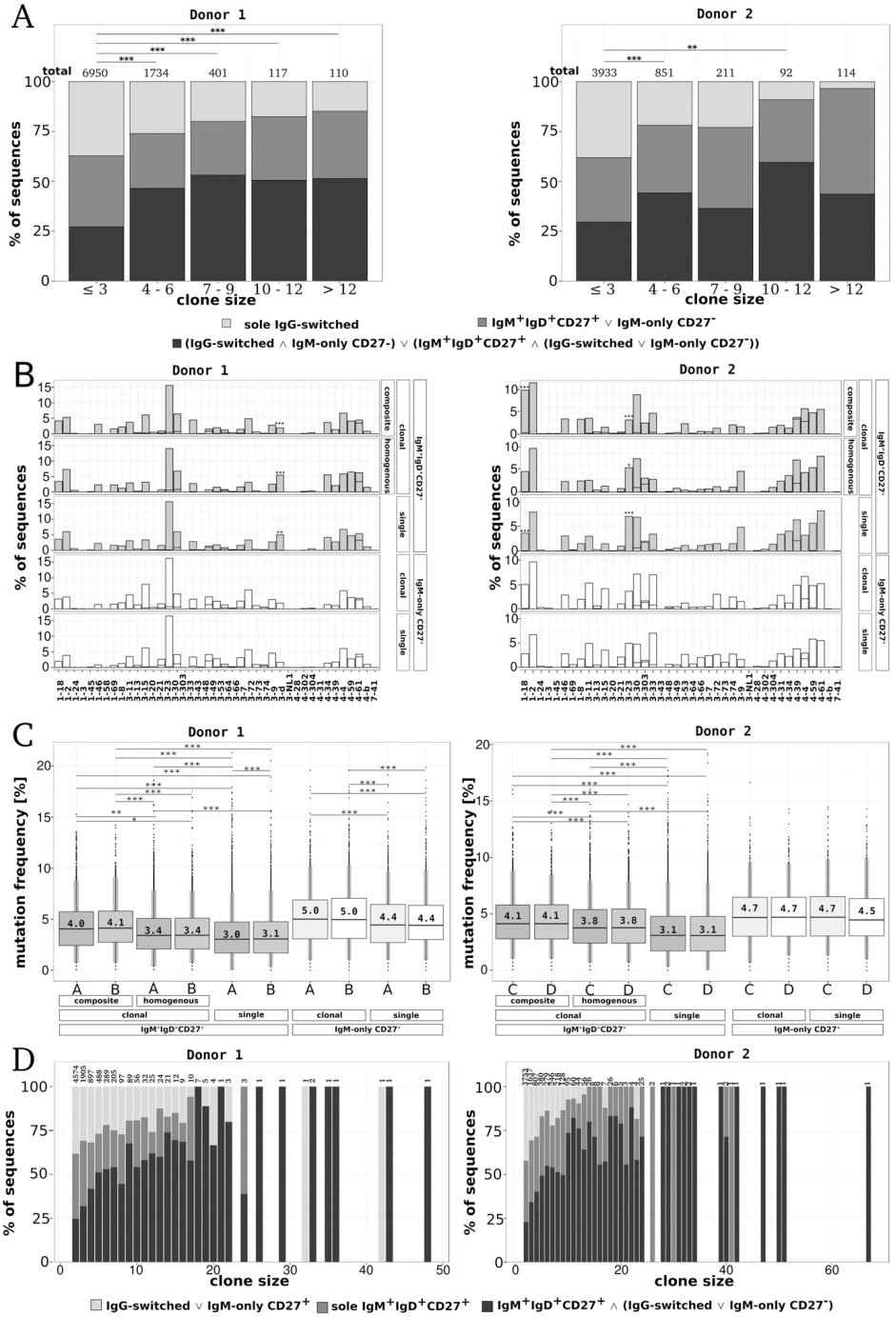


Fig. S3. Clonal composition of human memory B cells. (*A*) Correlation of clone type fractions and clonal sizes as in Fig. 3*B*, except that clones composed of $IgM^+IgD^+CD27^+$ and IgM-only sequences are included in the IgM clone fraction. ∧ indicates "and," and ∨ indicates "or." (*B*) *IGHV* gene use and (*C*) mutation frequencies (mutations/100 bp) of single or clonal sequences—no matter whether derived from composite or sole IgM (homogenous)—are very similar to each other. (*D*) Correlation of clone sizes and clone types as in Fig. 3*C*, nonstaggered depiction.

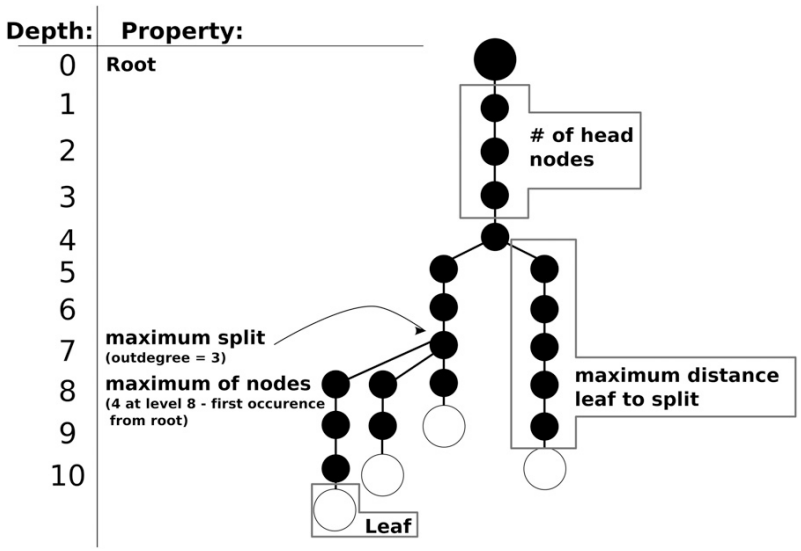


Fig. S4. Explanation of shape parameters used to describe genealogic dendrograms.

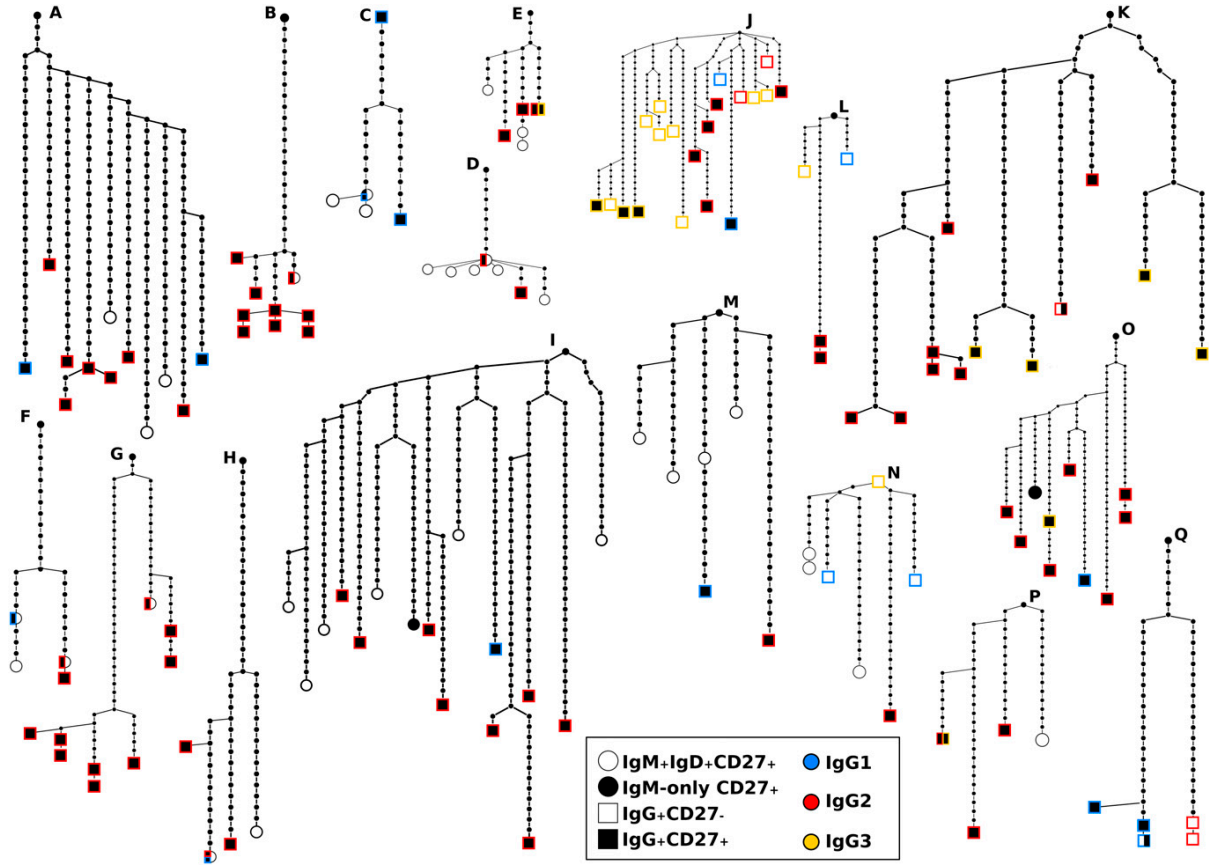


Fig. S5. Genealogic analysis of human memory B-cell clones. (*A–I*) Selected dendrograms of $IgM^+IgD^+CD27^+$, IgM-only, and IgG^+CD27^+ composite clones, where the B-cell subtypes are intermingled according to their *IGHV* gene mutation pattern. (*J–Q*) Selected dendrograms of B-cell clones including class switched B cells with more than one IgG subclass. Specific subsets are indicated by a gray scale code (legend); germ-line *IGHV* (root) and sequential mutation events are marked as small black circles. Less than 1% of nodes/leaves represent more than one sequence; occasional contributions from independent subsets or IgG subclasses are given as split-colored circles and additional colors, respectively.

Table S1. Input cell numbers and number of sequences retrieved by NGS

	Donor 1								
	IgG^+CD27^-		IgG^+CD27^+		$IgM^+IgD^+CD27^+$		IgM-only $CD27^+$		Total
Cell sample (replicate vial)	A	B	A	B	A	B	A	B	A+B
# input cells	8,000	8,000	45,000	45,000	50,000	50,000	5,500	5,500	217,000
# 454 reads retrieved	75,747	77,171	363,212	392,155	582,051	340,066	85,455	75,382	1,991,239
# reads> 300 bp length	73,841	75,112	320,145	365,358	577,016	331,317	83,696	74,287	1,900,772
# unique sequences (identical sequences collapsed)	28,677	26,567	79,599	78,315	148,274	81,009	22,525	23,078	488,044
# sequences with ≥ twofold coverage	7,710	6,949	21,155	20,271	30,285	15,807	5,366	5,076	112,619
# sequences after 454 error correction*	4,766	3,992	12,121	9,497	19,410	9,077	3,889	3,900	66,652
	Donor 2								
	C	D	C	D	C	D	C	D	C+D
# input cells	10,000	10,000	46,000	46,000	50,000	50,000	5,500	5,500	223,000
# 454 reads retrieved	79,630	93,677	360,117	519,422	547,384	409,085	67,107	52,783	2,129,205
# reads> 300 bp length	68,466	86,169	274,638	387,728	456,433	344,669	58,871	45,839	1,722,813
# unique sequences (collapsing of identical sequences)	39,313	49,170	103,627	115,420	235,542	183,475	29,655	24,712	780,914
# sequences with ≥ twofold coverage	4,549	6,038	13,534	16,068	21,179	16,206	2,750	2,126	82,450
# sequences after 454 error correction*	2,372	2,584	6,138	4,637	11,683	10,219	1,705	1,320	40,658

*Insertions/deletions arising from incorrect length determination of sequences of identical nucleotides, typically occurring by 454 base calling.

Table S2. BCL6 mutation analysis of human IgG3 class-switched memory B cells

Donor	Cell type	No. of total sequences	No. of mutated sequences	No. of sequence with *n* mutations						Fraction of mutated sequences (%)	Mutation frequency (%)	Fisher's exact test vs. $IgG^+IgG3^-CD27^+$
				1	2	3	4	5	6			
1	$IgG^+IgG3^-CD27^+$	21	10	8	1	1				47.6	0.08	
	$IgG3^+CD27^+$	30	15	7	4	1	2	1		50.0	0.14	
	$IgG3^+CD27^-$	47	9	7				1	1	21.3	0.05	
2	$IgG^+IgG3^-CD27^+$	20	7	5				1	1	35.0	0.11	
	$IgG3^+CD27^+$	35	10	4	4			2		28,6	0.17	
	$IgG3^+CD27^-$	31	6	3	2		1			19.4	0.05	
Total	$IgG^+IgG3^-CD27^+$	41	17	13	1	1		1	1	41.5	0.09	
	$IgG3^+CD27^+$	65	25	11	8	1	2	3		38.5	0.11	NS
	$IgG3^+CD27^-$	78	15	10	2		1	1	1	20.5	0.05	$P < 0.05$
	T cells*	18	0							0	0	

NS, not significant.
*As expected, the *BCL6* MMC of T cells, analyzed as background control, was unmutated.

Table S3. Number of sequences and clones retrieved from both donors

	Donor 1					Donor 2				
Population	IgG^+CD27^-	IgG^+CD27^+	$IgM^+IgD^+CD27^+$	$IgM^+onlyCD27^+$	Σ	IgG^+CD27^-	IgG^+CD27^+	$IgM^+IgD^+CD27^+$	$IgM^+onlyCD27^+$	Σ
# sequences	8,758	21,618	28,487	7,789	66,652	4,956	10,775	21,902	3,025	40,658
# single sequences (% total)	6,482 (74%)	9,861 (48%)	16,205 (57%)	3,416 (44%)	36,414 (54%)	4,024 (82%)	7,070 (66%)	10,450 (48%)	1,247 (41%)	22,791 (57%)
# clonal sequences (% total)	2,276 (26%)	11,307 (52%)	12,282 (43%)	4,373 (56%)	30,238 (46%)	932 (18%)	3,705 (34%)	11,452 (52%)	1,778 (59%)	17,867 (43%)
Sequences belonging to clones (# clones)										
IgG^+CD27^-	1,143 (458)	—	—	—		538 (234)	—	—	—	
IgG^+CD27^+	—	8,422 (2,890)	—	—		—	2,937 (1,059)	—	—	
$IgM^+IgD^+CD27^+$	—	—	7,821 (2,616)	—		—	—	8,161 (2,465)	—	
IgM^+only $CD27^+$	—	—	—	1,472 (546)		—	—	—	471 (187)	
IgG^+CD27^-/IgG^+CD27^+	780 (487)	1,046 (487)	—	—	1,826 (487)	288 (220)	322 (220)	—	—	610 (220)
$IgM^+IgD^+CD27^+/IgM^+$only $CD27^+$	—	—	2,766 (1,156)	2,115 (1,156)	4,881 (1,156)	—	—	2,602 (679)	1,172 (679)	3,774 (679)
$IgG^+CD27^-/IgM^+IgD^+CD27^+$	233 (183)	—	278 (183)	—	511 (183)	72 (66)	—	97 (66)	—	169 (66)
IgG^+CD27^-/IgM^+only $CD27^+$	19 (16)	—	—	21 (16)	40 (16)	7 (5)	—	—	7 (5)	14 (5)
$IgG^+CD27^+/IgM^+IgD^+CD27^+$	—	852 (475)	829 (475)	—	1,681 (475)	—	288 (184)	406 (184)	—	694 (184)
IgG^+CD27^+/IgM^+only $CD27^+$	—	500 (241)	—	385 (241)	885 (241)	—	67 (44)	—	56 (44)	123 (44)
$IgG^+CD27^-/IgG^+CD27^+/IgM^+IgD^+CD27^+$	71 (51)	135 (51)	78 (51)	—	284 (51)	17 (16)	25 (16)	34 (16)	—	76 (16)
$IgG^+CD27^-/IgG^+CD27^+/IgM^+$only $CD27^+$	4 (4)	6 (4)	—	4 (4)	14 (4)	1 (1)	4 (1)	—	1 (1)	6 (1)
$IgG^+CD27^-/IgM^+IgD^+CD27^+/IgM^+$only $CD27^+$	15 (11)	—	24 (11)	16 (11)	55 (11)	7 (5)	—	13 (5)	12 (5)	32 (5)
$IgG^+CD27^+/IgM^+IgD^+CD27^+/IgM^+$only $CD27^+$	—	333 (170)	456 (170)	350 (170)	1,139 (170)	—	59 (34)	131 (34)	56 (34)	246 (34)
$IgG^+CD27^-/IgG^+CD27^+/IgM^+IgD^+CD27^+/IgM^+$only $CD27^+$	11 (8)	13 (8)	30 (8)	10 (8)	64 (8)	2 (2)	3 (2)	8 (2)	3 (2)	15 (2)
Σ IgG/IgM (# clones) (% total)	353 (273)	1,839 (949)	1,695 (898)	786 (450)	4,673 (1,159)	106 (95)	446 (281)	689 (307)	135 (91)	1,376 (357)
Σ clonal sequences (# clones)	2,276 (1,218)	11,307 (4,326)	12,282 (4,670)	4,373 (2,152)	30,238 (9,312)	932 (549)	3,705 (1,560)	11,452 (3,451)	1,778 (957)	17,867 (4,901)
% clonal sequences	(16%)	(16%)	(14%)	(18%)	(15%)	(11%)	(12%)	(6%)	(8%)	(8%)

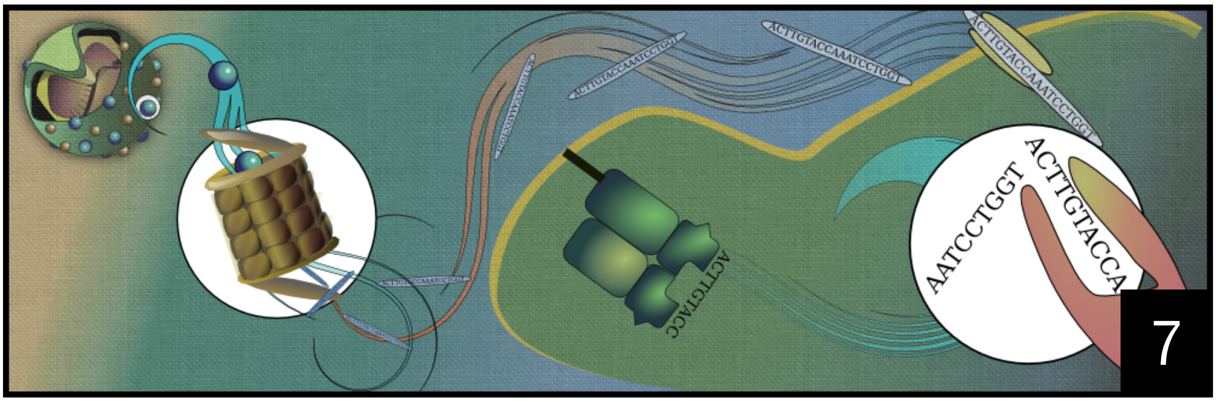
Table S4. Shape parameters of genealogic dendrograms

Parameter	Definition	Biological meaning
Intraclonal diversity	Average frequency of unique (i.e., not shared by all nodes/leaves) mutations within a tree	Unique mutations per B-cell clone member
# of head nodes	Number of mutations, shared by all nodes within a tree	Shared mutations per B-cell clone member
Distance root to maximum outdegree (depth of node with maximum outdegree)	Number of mutations to the node-level with maximum outdegree	Relative time point of clone member, from which a maximal number of distinct mutations occurred
#1 level of maximum nodes (maximal number of nodes with same depth)	Tree height with maximum number of nodes	Relative time point of maximum number of clone members considering clonal expansion
Mean distance to root	Average number of mutations per leaf	Average number of mutations per B-cell clone member
Maximum distance leaf to split	Maximum number of unique mutations per leaf (branch length)	Maximum number of unique mutations per sequence of a clone
Mean distance leaf to split	Average number of unique mutations per leaf (average branch length)	Average number of unique mutations per sequence of a clone

Table S5. Overview on sole B-cell subset/IgG subclass and composite clone numbers with at least seven clone members

Clone types	Count	% total
No. clones with more than six members* (total)	1,045	
$IgM^+IgD^+CD27^+$	297	28,42
IgM-only $CD27^+$	17	1,63
IgG^+CD27^+	155	14,83
IgG^+CD27^-	10	0,96
Composite	566	54,16
Composite with ≥90% subset dominance	86	8,23
No. clones with more than six IgG^+ members (total)	254	
IgG1	13	5,12
IgG2	132	51,97
IgG3	57	22,44
IgG4	0	0,00
Composite	52	20,47
Composite with ≥90% subclass dominance	10	3,94

*Clone sizes of more than six members were chosen arbitrarily, however, ensuring that a substantial clone sample number (>1,000) was available for genealogic analysis and sufficient leaves per dendrogram exist to allow differential structure analysis.



Proteasomal selection pressure on hepatitis C virus epitopes

I have not failed. I've just found 10,000 ways that won't work.

THOMAS A. EDISON

Abstract

Proteasomal cleavage of proteins in human cells leads to presentation of epitopes at the cell surface. Those presentations can be recognized by cells from the immune system, which then activate the immune response. Evasion of such recognizability is a constant drift for pathogen evolution, in which those pathogens are favored which have mutations that either disable the recognition itself, the binding to the presenting molecule, or the cleavage itself. Several escape mutations for either recognition or binding evasion are known for Hepatitis C virus (HCV). Escape mutations outside epitopes towards no cleavage by the proteasome may occur due to an inability to escape binding or recognition inside the epitope.

We analyzed the flanking regions of HCV epitopes from the Immune Epitope Database with MHC binding prediction, which were classified either as recognized or not, and searched for amino acids inside those regions, which showed a significant difference between both sets. We found fourteen differing amino acids inside the 15+15 amino acids long flanking region.

7.1 Introduction

7.1.1 Flaviviridae

Flaviviridae is a family of viruses which are mostly spread by arthropod vectors. The name originated from the Latin word flavus, which means yellow and labels a prominent virus of this family, the yellow fever virus [47]. Flaviviridae are divided into four genera: the Flavivirus, Hepacivirus, Pegivirus and Pestivirus. All but the last genera include viruses which can infect humans. The most prominent members of the Flaviviridae are, other than the mentioned yellow fever virus, Dengue Fever, West Nile Virus, and Tick-borne encephalitis virus of the Flavivirus, and Hepatitis C of the Hepacivirus [9]. All Flaviviridae viruses have linear, single-stranded RNA genomes of positive polarity with a length between 9.6 to 12.3 *kb*. The viral particles are enveloped, spherical, and about 40 nm - 60 nm in diameter [54]. From the RNA genome one polyprotein is synthesized, which is then processed through host and viral proteases. All polyproteins are organized similarly, although the members of the genera are only distantly related. In the N-terminal region are core and envelope proteins, in the C-terminal region are the non-structural ones [41]. A common feature is the location of serine protease and helicase activities in the NS3 region and an RNA-dependent RNA polymerase near the C-terminus of the polyprotein [40]. Most of the human pathogen Flaviviridae do not induce a chronic infection. Two exceptions are Hepatitis C and GBV-C, which can lead to chronic infections. In the case of HCV 80% of humans exposed to the virus develop a chronic infection [34]. GBV-C instead takes a chronic course in around 20%-30% [15, 25]. Up until now GBV-C does not seem to cause a human disease [5].

7.1.2 HCV

Hepatitis C is a disease that occurs frequently with around 170 million people worldwide infected with the HC-virus. Genomic studies from different genotypes suggest that this virus evolved around 1,100 to 1,350 years ago (95% credible region, 600 to >2,500 years ago)[39]. The different (sub)genotypes evolved 400 to 200 years ago from genotype 1b [42] and have an estimated rate of mutation of 1.8×10^{-4} (95% credible region $0.9 \times 10^{-}4$ to $2.9 \times 10^{-}4$)[39]. Compared to its long time in the human population it was discovered relatively late in 1970s and isolated in 1989 [7].

7.1.3 The proteasomal cleavage and ER procession

Proteins in human cells are constantly degraded by peptidases and processed to be presented by the MHC-class I system on the cell surface. Proteins in the cytosol are cleaved by the ubiquitin-proteasome system and peptides are generated [33]. Up until now, there are three proteasomes known: the constitutive proteasome, which can be found in every cell type, the immunoproteasome, which can be induced by interferon-γ, and the thymoproteasome in thymal cells [14]. The proteasome cleaves proteins into peptides with a length between 3-18 amino acids [4]. Sometimes new peptides are generated through a reversal of proteolysis and ligation of smaller fragments [53]. Those peptides are then transported via TAP into the lumen of the ER. TAP can transport peptides of a length of between 7 and 40 amino acids [37] and transfers the peptide in the ER to MHC-class I molecules. Those molecules are in complex with TAP/tapasin, ER60 and calreticulin [36]. In this so-called Peptide Loading Complex (PLC), peptides can bind directly but are cleaved by aminopeptidases if they are to long [21]. Currently there are three known aminopeptidases in this system in the ER: ERAP1, ERAP2, and IRAP, which is homologous to the first two [43] respectively ERAAP in mice [19]. Each of these aminopeptidases have different trimming preferences and specificity. The trimming of the antigenic precursor, which often removes only one additional amino acid, is largely affected by the N-terminal specificity of the aminopeptidase [4]. The preferred residue for ERAP1 is leucine, whereas for ERAP2 it is arginine [18, 50]. IRAP can cleave both substrates [31].

For HIV it is known that certain mutations in the flanking regions of an epitope lead to reduced presentation of the epitope [32], and it was also shown, that the cleavage of in vitro constructed flanking regions with known epitopes depends on the amino acid composition [49].

7.1.4 Main hypothesis

Our main hypothesis is that because of its long duration in the human population HCV is adapted to the human immune system not only with escape mutations inside the epitopes to evaded recognition through CTL, but also in the flanking regions of possible epitopes in a way, that the MHC-class I pathway cannot produce the antigen - MHC complex.

7.2 Methods

7.2.1 Data

A dataset from IEDB[52] was generated for HCV epitopes. All epitopes listed in the database are given either as T cell response, B cell response or MHC binding and sorted into positive ones (at least one positive test) and negative ones (no positive test so far). We retrieved the 194 epitopes for T cell response and HLA-A*02. For each epitope BLAST [1] was used to get the surrounding sequences out of H77, because the original sequence data was not available in most cases. Fifteen amino acid positions in front of the epitope and fifteen amino acids behind the epitope were taken into account for further analysis.

A prediction of all HLA-A*02 MHC binder, and thus potential epitopes, in the sequence of HCV reference genome H77 [3, 6, 23] was carried out with the prediction tools on the IEDB website [52]. The predictions were made using the IEDB analysis resource Consensus tool [22] which combines predictions from ANN respectively NetMHC (3.4) [27, 35], SMM [38] and Comblib [48]. The predicted epitopes were sorted by IC_{50} and epitopes from the IEDB database into a positive and a negative set: all predicted epitopes with an $IC_{50} < 5000\ nM$ and an entry in the database were selected as positive, all predicted epitopes with an $IC_{50} < 50\ nM$ and no entry in the database were selected as negative. For actual numbers see Table 7.1. The difference in IC_{50} refers to the fact that even a weak binder can be presented and recognized, but for the negative list we wanted to be sure that the entries are not simply in the list because of weak or no binding to the MHC molecule.

Table 7.1: Number of predicted HLA-A*02 MHC binder of HCV intersected with T cell epitopes. T cell epitopes were taken from IEDB database, MHC binder were predicted with tools from IEDB out of the HCV reference sequence H77.

type	number
T cell$^+$ + predicted $IC_{50} < 5000\ nM$	690
T cell$^+$ + predicted $IC_{50} < 500\ nM$	278
T cell$^-$ + predicted $IC_{50} < 50\ nM$	337

A second dataset from IEDB was generated from human HLA-A*02 positive epitopes. Additional refinement was necessary since these epitopes were not compared to predicted ones and some epitopes had a length larger than twelve or shorter than seven, which had

to be removed. Then all remaining epitopes were clustered to eliminate potential double entries in the data.

Further, we downloaded polymerase (NS5b) sequences, one from each, from different Flaviviridae (HCV, West Nile Fever, Tick-Born Encephalitis, Yellow Fever, and GBV-C), and human DNA-directed RNA polymerase I subunit RPA49 from Uniprot [8] to compare the amino acid frequencies inside those proteins and search for significant changes.

7.2.2 Analysis of RNA-dependent RNA polymerase from Flaviviridae

Every amino acid of the RNA-dependent RNA polymerase protein sequences from Flaviviridae was counted and then compared by a Fisher's exact test with all other counts of this amino acid.

7.2.3 Flanking regions

The flanking regions of an epitope are up- and downstream of the epitope. The size of the flanking regions for this analysis was 15 nucleotides. We call the flanking regions plus the epitope an extended epitope (see Figure 7.1).

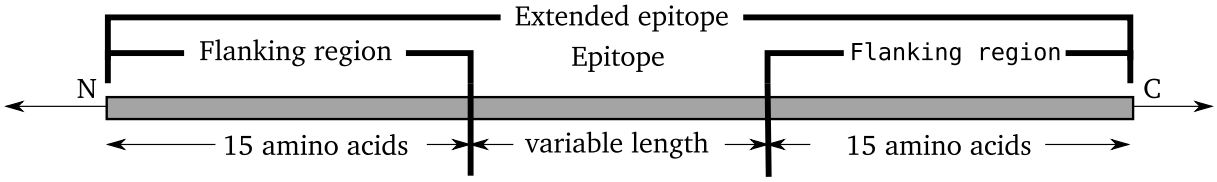


Figure 7.1: Scheme of an extended epitope. An extended epitope is the epitope plus 15 amino acids upstream and downstream of it.

All flanking regions were clustered before analysis, because some of them showed the same sequence despite a different epitope in the middle. This was because all flanking regions were taken from H77 and not the original sequence were the epitope was found. To eliminate this potential bias the amount of those flanking regions which showed the same sequence were reduced to one.

7.2.4 Analysis of significant differences between the HCV epitopes

For this analysis two different tests were carried out with all epitope datasets: a Fisher's exact test and a Mann-Whitney-U test (see Figure 7.2).

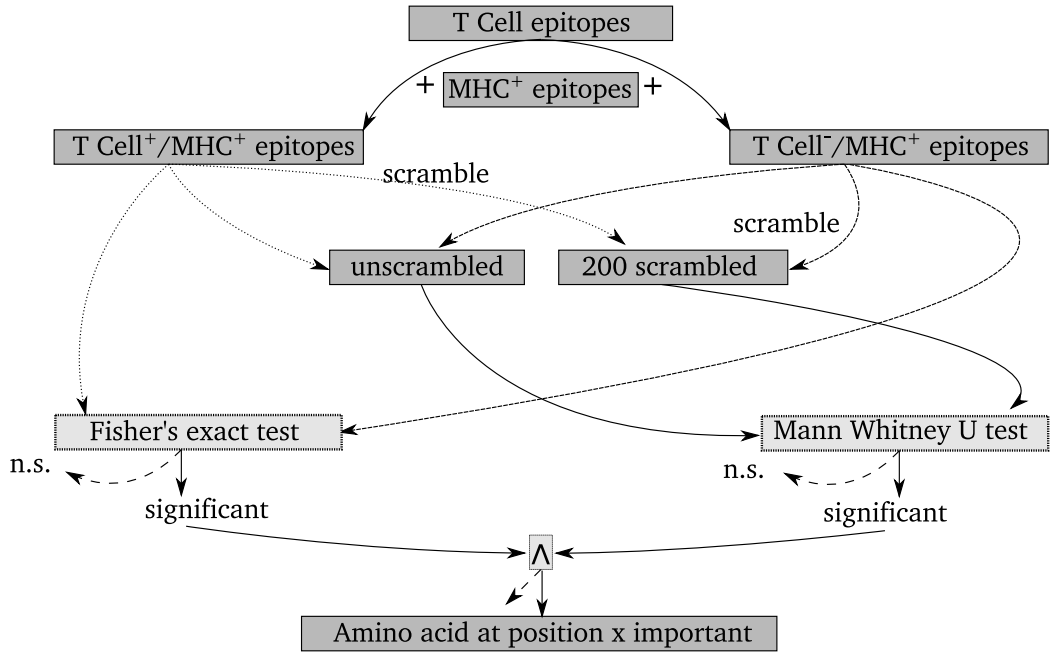


Figure 7.2: Flowchart for proteasomal selection pressure on hepatitis C virus epitopes. Dark gray boxes are input or output data, light gray boxes are statistical or mathematical functions. Measured T cell epitopes from IEDB database are combined with predicted epitopes out of H77. A position wise Fisher's exact test and a Mann-Whitney-U test of scrambled and unscrambled sequences are compared to extract those amino acids and positions, which are different between the sets of epitopes (HCV$^+$, HCV$^-$, and human$^+$)

Fisher's exact test for positive and negative epitopes

A Fisher's exact test was carried out for every flanking regions sequence position (i) and every amino acid (j) for both datasets to examine which position in the flanking regions of the epitopes differs between the positive (p) and the negative set (n).

$$p_{i,j,p} = \frac{\binom{|flanking_{i,j,p}|+|flanking_{i,j,n}|}{|flanking_{i,j,p}|}\binom{|flanking_{i,\neg j,p}|+|flanking_{i,\neg j,n}|}{|flanking_{i,\neg j,p}|}}{\binom{|flanking|}{|flanking_{i,j,p}|+|flanking_{i,\neg j,p}|}} \tag{7.1}$$

The p-values were than adjusted for multiple correction with the p.adjust function from the R-stats package with the method of "FDR".

Mann-Whitney-U test for original and scrambled flanking regions

Further tests for every position p and every amino acid a were carried out to check if the results from the Fisher's exact test were due to too few epitopes. Two hundred times a scrambled set(letters at position were inserted randomly) from every epitope was created to be used in the Fisher's exact test and compared with two hundred equally generated p-values created with the correct order of amino acids in a Mann-Whitney-U test [29]:

$$\begin{aligned} R_{p,a} &= \text{Rank values of p-values for correct amino acid order} \\ \overline{R_{p,a}} &= \text{Rank values of p-values for scrambled amino acid order} \\ n_{p,a_1} &= \#\text{ of p-values for correct amino acid order} = 200 \\ \overline{n_{p,a}} &= \#\text{ of p-values for scrambled amino acid order} = 200 \end{aligned}$$

$$U = min(n_{p,a_1}\overline{n_{p,a}} + \frac{n_{p,a_1}(n_{p,a_1}+1)}{2} - \sum R_{p,a_1},\ n_{p,a_1}\overline{n_{p,a}} + \frac{\overline{n_{p,a}}(\overline{n_{p,a}}+1)}{2} - \sum \overline{R_{p,a}}) \tag{7.2}$$

Since $n_{p,a_1}, \overline{n_{p,a}} > 25$ it was approximated by a normal distribution:

$$Z = \frac{U - \frac{n_{p,a_1}\overline{n_{p,a}}}{2}}{\sqrt{\frac{n_{p,a_1}\overline{n_{p,a}}(n_{p,a_1}+\overline{n_{p,a}}+1)}{12}}} \approx N(0;1) \tag{7.3}$$

7.3 Results

7.3.1 The amino acid frequencies of HCV NS5b are highly similar to GBV-C and human DNA-directed RNA polymerase

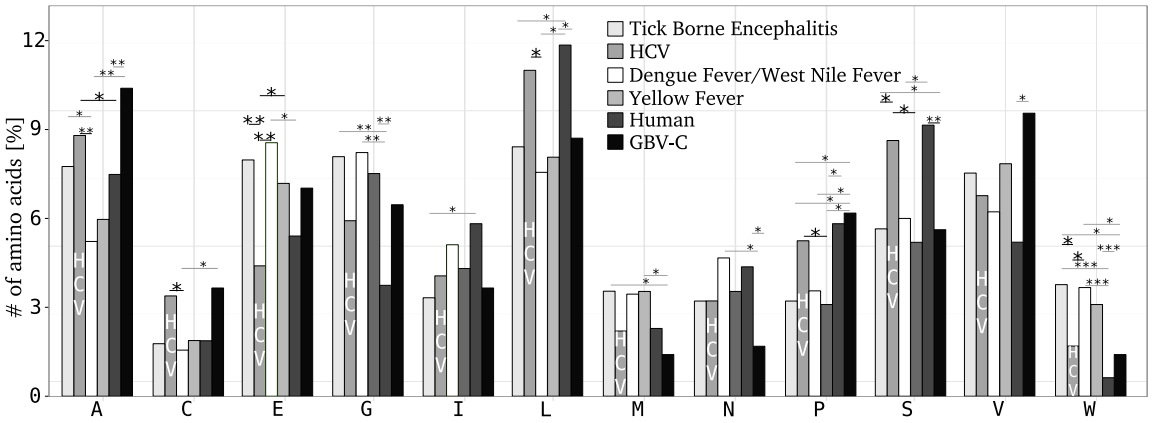


Figure 7.3: Percentage of amino acids in RNA Polymerase in HCV, Tick-Borne Encephalitis, Dengue Fever/West Nile Fever, Yellow Fever, GBV-C and human. Shown are those amino acids, which displayed a significant difference between at least two viruses, or virus and human. The sequences of Dengue Fever and West Nile Fever were completely identical, hence only one bar for both is shown. Fisher's exact test: * $<$ 0.05, ** $<$ 0.015, *** $<$ 0.001

The comparison of the amino acid frequencies of NS5b, the RNA-dependent RNA polymerase, of certain Flaviviridae and human DNA-directed RNA polymerase I showed significant differences in a Fisher's exact test. In twelve amino acids more than one pair of either viruses or virus and human showed differences (alanine, cystein, glutamic acid, glycine, isoleucine, leucine, methionine, asparagine, proline, serine, valine, tryptophan). In two amino acids at least one pairing showed disparity (isoleucine, valine). The amino acid frequencies of Dengue Fever and West Nile Fever had no difference at all and their sequences were completely identical. The percentage of more than half of the amino acids from HCV and either the human sequence or GBV-C showed certain similarities, being either higher or lower than the rest of the Flaviviridae (see Figure 7.3). The RNA polymerase of HCV and the RNA polymerase of GBV-C showed significantly increased amounts of alanine, cystein and proline, and significantly decreased amounts of tryptophan compared with the other Flaviviridae. The RNA polymerase of HCV and the RNA polymerase of humans showed significantly increased amounts of leucine, proline and serine and significantly decreased amounts of glutamic acid and tryptophan compared with

other Flaviviridae. Trends could be estimated in glycine (decreased amount in HCV and human) and methionine (decreased amount in HCV, GBV-C and human). In summary, there seems to be an interesting similarity between the composition of RNA polymerases of HCV, GBV-C and humans in contrast to the other Flaviviridae.

7.3.2 Epitopes in IEDB are not distributed equally over whole H77

We compared the positive and negative epitopes to the reference sequence H77 to get an overview of the distribution of epitopes in the sequence and searched if there were regions with a high density of epitopes and regions without any epitopes at all. The epitope distribution was not equal (see Figure 7.4) and regions existed with lots of epitopes and regions with only a few or none. The known secondary structure of the HCV proteins seem to have no connection with a high load of epitopes. Probably some regions are more often examined than others, due to better protocols or known interactions.

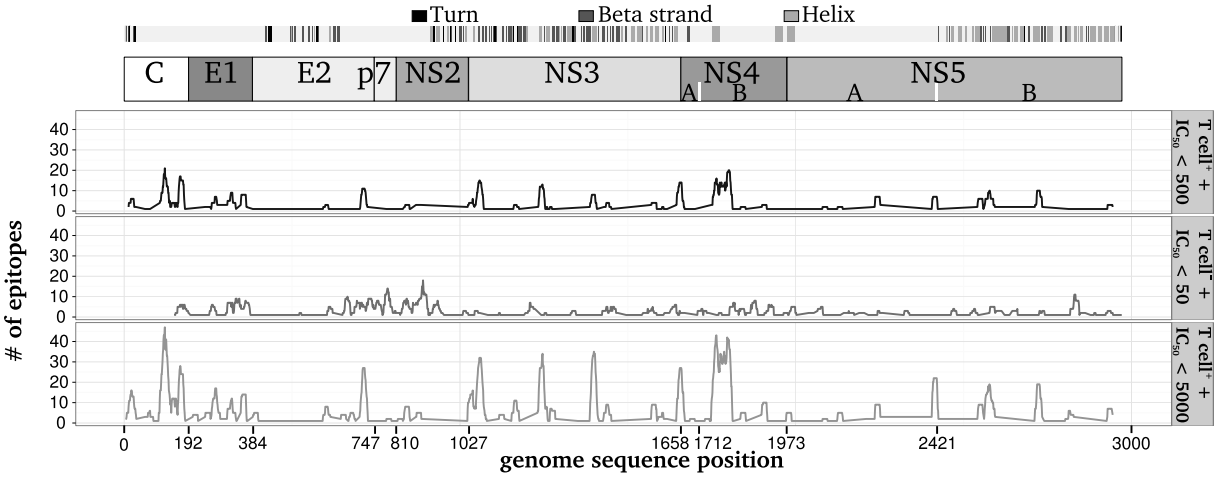


Figure 7.4: Overview of the genome sequence positions of the used HLA-A*02 epitopes. For every genome sequence position the number of epitopes covering this position was noted. The epitopes were taken from IEDB and intersected with a prediction of MHC binder of different IC limits with tools from IEDB. Breaks mark beginning and end of HCV proteins, which are noted above the plot, according to [28]. Secondary structure for H77 above the plot was obtained from Uniprot.

7.3.3 Amino acid amounts are different between HCV and human epitopes

The numbers for all amino acids in the extended epitope positions were collected and compared with each other to search for general differences in the amount of certain amino acids. Those differences would potentially show as significant in the Fisher's exact test, although they do arise only from general disparity. The comparison between both HCV datasets, alanine, glycine, and histidine showed a significant difference in the amount. In the comparison between the human and HCV negative datasets, alanine, aspartic acid, glutamic acid, glycine, lysine, asparagine, glutamine, serine, threonine, valine, and tryptophan showed a significant difference in the amounts. In the comparison between the human and HCV positive datasets, alanine, cystein, aspartic acid, glutamic acid, glycine, histidine, lysine, asparagine, glutamine, arginine, threonine, valine, and tryptophan showed a significant difference in the amounts (see Table 7.2 and Figure 7.5).

7.3.4 Fourteen amino acids are different between HCV^+ and HCV^-

The comparison between both HCV epitope datasets and the humane epitope dataset showed several differences (see Table 7.3). As expected, a second comparison with a dataset generated with $IC_{50} < 500nM$ showed fewer significant differences. The most frequent amino acids, which showed a significant difference between the human dataset and both HCV datasets, were glutamic acid and alanine. Both amino acids also showed a highly significant difference in overall frequency (see Figure 7.5), so it is likely that those amino acids have a general difference between human and HCV genomes.

The position wise significant p-values under consideration of the results from the Mann-Whitney$U$ Test leave 16 amino acids and positions in the first set (comparison of HCV positive and negative epitopes), eleven amino acids and positions in the second set (human epitopes and negative HCV epitopes), and 37 amino acids and positions in the third set (human epitopes and positive HCV epitopes) from all possible combinations. Four amino acids showed significantly increased frequencies in the flanking regions of positive HCV epitopes in the comparison of both HCV datasets, which indicates that those amino acids may be responsible for epitope processing in HCV, respectively may not hinder the processing. Twelve amino acids showed a significantly increased frequency in the flanking regions of negative HCV epitopes. Those amino acids may hinder the human proteasom and prohibit processing. Six amino acids had a significantly increased frequency in the flanking regions of negative HCV epitopes in the comparison with human epitopes.

Table 7.2: Significant differences in a Wilcoxon test between the amino acid numbers in the flanking regions between human, HCV^+ and HCV^- datasets. Compared were all amino acids without using the position. No correction was applied to the p-values.

type	amino acid	p-value
HCV^+ & HCV^-	A	0.019
	G	0.002
	H	0.022
$HUMAN^+$ & HCV^-	A	< 0.001
	D	0.021
	E	< 0.001
	G	0.008
	K	< 0.001
	N	0.002
	Q	0.002
	S	0.041
	T	< 0.001
	V	0.008
	W	< 0.001
$HUMAN^+$ & HCV^+	A	< 0.001
	C	0.008
	D	0.004
	E	< 0.001
	G	< 0.001
	H	0.004
	K	< 0.001
	N	0.033
	Q	< 0.001
	R	0.007
	T	< 0.001
	V	0.002
	W	0.003

Fourteen amino acids showed a significantly increased frequency in the flanking regions of negative HCV epitopes in the comparison with human epitopes (see Figure 7.6).

A comparison between HCV^-/HCV^+ and HCV^-/Human shows two positions in the flanking regions with the same amino acids: at position 10 both HCV^+ have an increased amount of glutamic acid and at position 22 both HCV^- have an increased amount of leucine. The former indicates a positively charged binding pocket in one of the processing enzymes, the later a neutral binding site.

Table 7.3: Top ten p-values from comparison between positive and negative epitopes with an Fisher's exact test. Three datasets were used all for HLA-A*02. The HCV positive and negative datasets were generated from HCV T cell data from IEDB in combination with MHC predicted data out of the HCV reference genome H77. The first dataset is a comparison between both HCV datasets, the second one is a comparison between human T cell data from IEDB and the negative HCV epitopes, the third one was a comparison between human T cell data and the positive HCV epitopes. FDR was applied to the p-values to correct for multiple testing errors. The p-values were taken from the analysis of the positive dataset generated with an in IC_{50} < 5000nM. The last column shows if the data points show up in the comparison with the positive dataset generated with an in IC_{50} < 500nM.

	position	Fisher p-value	amino acid	in $IC_{50} < 500\ nM$
HCV$^+$ & HCV$^-$	4	0.021	S	
	5	0.024	K	
	9	0.020	L	
	10	0.024	H	
	10	0.039	E	
	11	0.028	L	
	17	0.020	N	x
	19	0.020	S	
	23	0.028	H	
	28	0.024	C	
HUMAN$^+$ & HCV$^-$	1	0.045	W	
	10	0.014	E	x
	10	0.028	Q	
	11	0.019	E	x
	12	0.024	W	x
	15	0.045	M	
	22	0.024	E	x
	25	0.039	A	
	30	0.024	G	x
	30	0.028	E	x
HUMAN$^+$ & HCV$^+$	3	0.005	E	x
	8	0.014	T	x
	15	0.004	E	
	16	0.000	E	x
	16	0.020	A	x
	21	0.020	A	x
	22	0.004	E	
	25	0.012	A	x
	28	0.021	A	x
	30	0.002	A	x

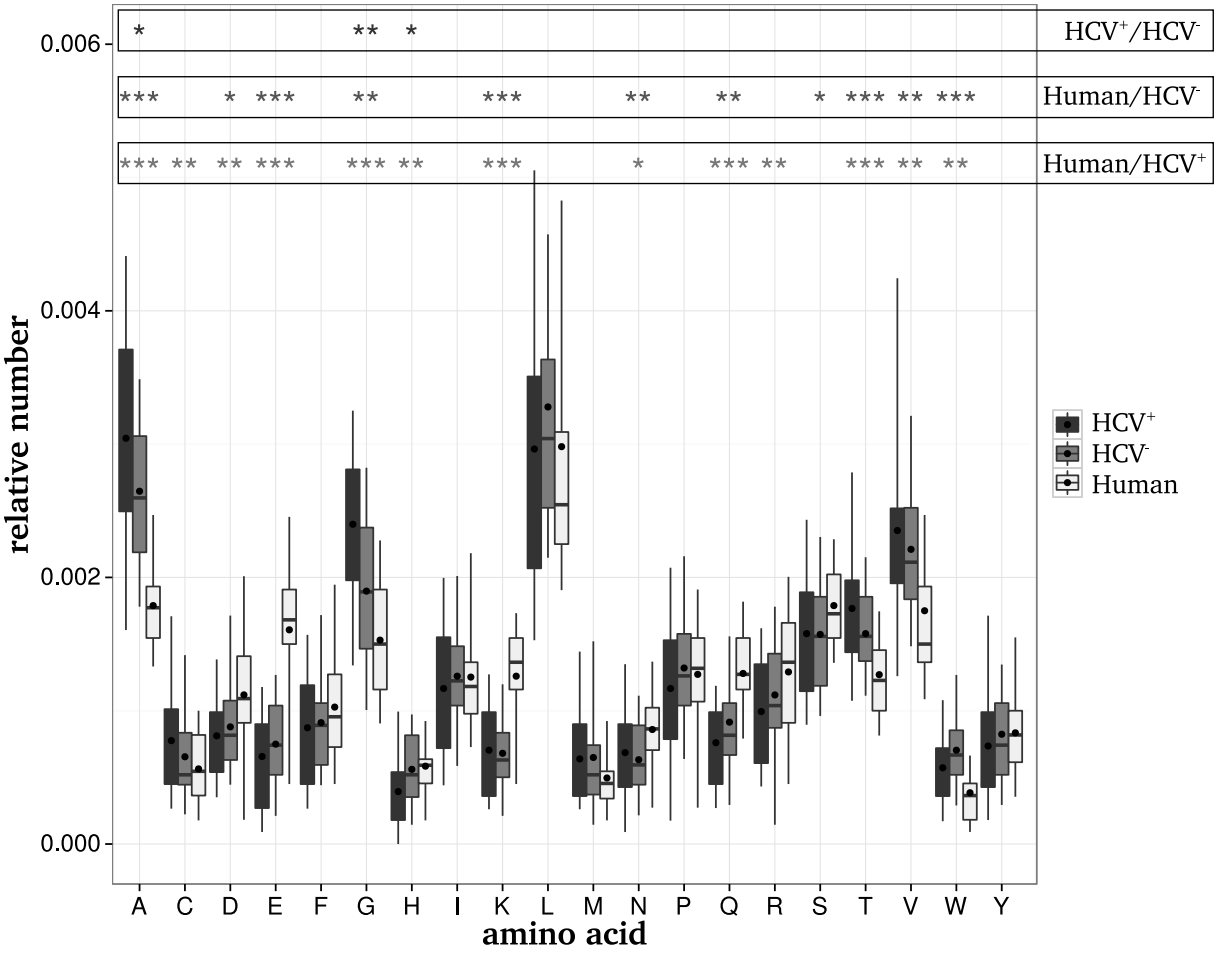


Figure 7.5: Amino acid numbers in HCV and human extended epitope. Shown is the relative frequency for every amino acid inside human an HCV extended epitope. The negative epitopes were selected with an IC $<$ 50 nM, the positive epitopes were selected with IC $<$ 500 nM. Wilcoxon test: * $<$ 0.05, ** $<$ 0.015, *** $<$ 0.001

A combination of these results with the results from the general frequencies analysis leads to a reduced set of significant differences (see Table 7.4). Most of the significantly different amino acids between the human and one of the HCV datasets have to be treated carefully, because there is a general difference and one cannot tell if there is a underlying position based difference. Only two amino acids, leucine at position 22 and methionine at position 15, are still a real difference inside the flanking regions of negative HCV epitopes in the comparison with human epitopes and only one amino acid, phenylalanine at position 3 is still a real difference inside the flanking regions of positive HCV epitopes in the comparison with human epitopes. In the comparison between the flanking regions of both HCV epitope datasets only one amino acid had to be excluded, histidine at position 10 and 23.

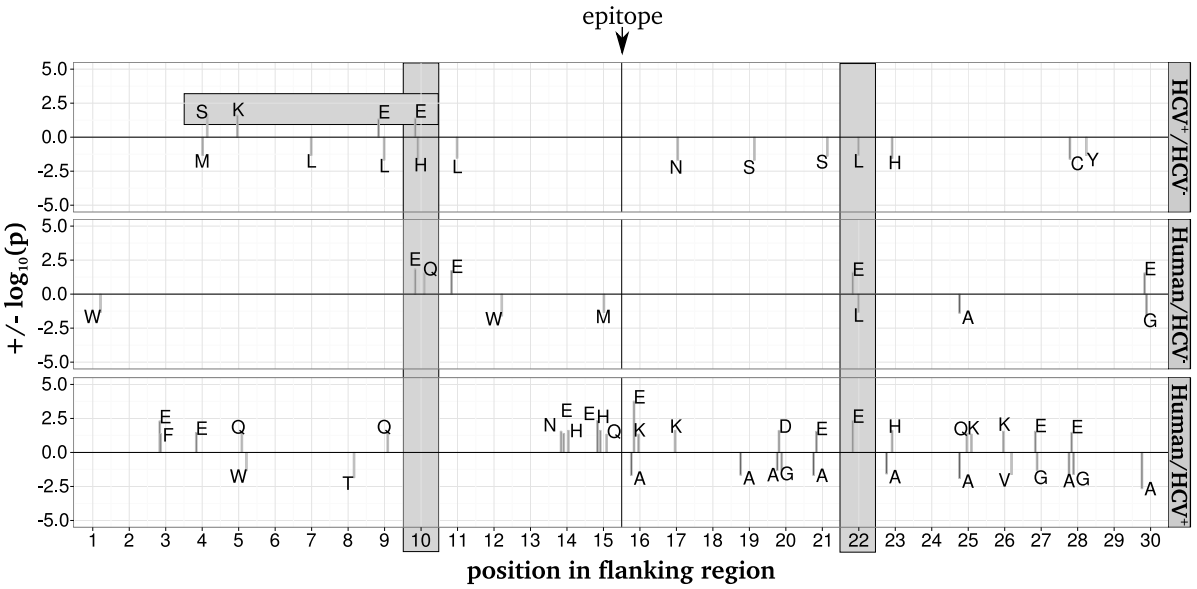


Figure 7.6: Significant amino acids and their positions between positive and negative epitopes. All those amino acids and positions are shown which had a significant difference in the frequency between either human and HCV or positive and negative HCV dataset. The direction of the bar is according to the Odds ratio value and indicates in which set the amino acid is more frequent. Marked in horizontal gray bar are those amino acids, which could be used in a *in vitro* test, to render positive epitopes negative. Marked in vertical gray bar are two positions with an increased amino acid in either HCV^+ in HCV^+/HCV^- and HCV^-/Human or HCV^- in HCV^-/HCV^+ and HCV^-/Human which points to a possible connection between sequence position and processing.

7.3.5 Epitopes for verification

We selected twelve epitopes from the list of known HLA-A*02 epitopes from the Los Alamos database to be further processed *in vitro*. An epitope was selected if it had the correct amino acid at the correct position for a positive epitope, so that a modification into another amino acid should render this epitope negative (see Table 7.5). Or if it did not have the correct amino acid for a negative epitope. A modification of this amino acid into the amino acid more frequent in the negative epitopes should render this epitope negative.

Table 7.4: Overview of the number of significant amino acid changes between human and positive/ negative HCV epitopes. T cell epitopes from IEDB combined with predicted MHC binder were compared position wise and amino acid-wise with a Fisher's exact test. Noted is also if there were significant differences in the amino acid frequencies between both extended epitopes, which was also calculated with a Fisher's exact test. Marked with gray are those fields, that had a significant amino acid change, but no significant different amino acid frequency.

Amino acid	HCV^+ & HCV^- #significant	sda*	HN^+ & HCV^- #significant	sda*	HN^+ & HCV^+ #significant	sda*
A		✓	1	✓	8	✓
C	1			✓		
D				✓	1	✓
E	2		4	✓	9	✓
F					1	
G		✓	1	✓	3	✓
H	2	✓	3	✓		
K	1			✓	4	✓
L	4		1			
M	1		1			
N	1			✓	1	✓
Q			1	✓	4	✓
R				✓		
S	3			✓		
T				✓	1	✓
V				✓	1	✓
W			2	✓	1	✓
Y	1					

*sda: significant difference in the amino acid frequencies. HN: Human

7.4 Discussion

Our data was composed from MHC-binding and T cell assays. MHC-binding assays test if a given short amino acid chain binds to a certain MHC molecule. A T cell assay is a test for (memory) T cell proliferation after a second exposure to a given short amino acid chain. These tests react positively with an increased cell proliferation, if the T cell recognizes the molecule. This is no guarantee per se that the recognition of the short amino acid chain is the result of a previous antigen processing that resulted in this amino acid. One possibility would be a cross reaction [46]. Therefore it is not sure if the basic data from IEDB is appropriate for the task of finding differences in the amino acid chain between those regions which are processed by the proteasom and those which are not. To eliminate this factor of uncertainty, we used the T cell assay dataset to select from

Table 7.5: HLA-A*02 epitopes with labeled amino acids. Shown are extended epitopes with a significant difference between the flanking regions of positive and negative HCV epitopes. Flanking regions were taken from the reference sequence H77, epitopes from Los Alamos database.

Amino acid	flanking region	epitope	flanking region	protein
4S	WLL**S**PRGSRPSWGPT	ADLMGYIPLV	GKVIDTLTCGFADLM	Core
	TLC**S**ALYVGDLCGSV	FLVSQLFTF	SPRRHWTTQDCNCSI	E1
	SIA**S**WAIKWEYVVLL	FLLLADARV	CSCLWMMLLISQAEA	E2
	GSR**S**LTPCTCGSSDL	YLVTRHADV	IPVRRRGDSRGSLLS	NS3
	CSG**S**WLRDIWDWICE	VLSDFKTWL	KAKLMPQLPGIPFVS	NS5a
5K	RNLG**K**VIDTLTCGFA	ALMGYIPLV	GAPLGGAARALAHGV	Core
9E	MTCMSADL**E**VVTSTW	VLVGGVLAA	LAAYCLSTGCVVIVG	NS4a
	MGGNITRV**E**SENKVV	ILDSFDPLV	AEEDEREVSVPAEIL	NS5a
10E	SIASWAIKW**E**YVVLL	FLLLADARV	CSCLWMMLLISQAEA	E2
	AVQTNWQKL**E**VFWAK	HMWNFISGI	QYLAGLSTLPGNPAI	NS4b
22X	LIFCHSKKKCDELAA	KLVALGINAV	AYYRGL**D**VSVIPTSG	NS3
	MGGNITRVESENKVV	ILDSFDPLV	AEEDER**E**VSVPAEIL	NS5a

MHC binding assay results epitopes which are recognized and those which only bind the MHC molecule. In this way we could be sure that the epitopes used are at least able to be presented by the cell.

Our findings indicate that over the whole length of the 15 amino acids in front of the epitope and the 15 amino acids behind it, not only the immediate vicinity may be important for processing, but also positions further away. Previous studies of this topic are rare and mostly findings are made accidentally in the search for escape mutations inside the epitope, as in the study of Timm et al., in which a change of alanine to threonine in the COOH-terminal flanking region showed no altered IFN-γ secretion [51]. Seifert et al. found a Y/F mutation in HLA-A*02 patients at position one of the carboxy flanking region, which prevented proteasomal cleavage at this position [44]. In the N-terminal region, amino acid frequencies are increased in both data sets, but only in the negative dataset in the C-terminal region. The results also show that (basic or acidic) polar amino acids are better for epitope processing, than non-polar ones. This conforms with the results from Livingston et al., who found that non-polar amino acids perform poorly at least at the C_1 terminus and positively charged amino acids (K or R) are most frequently associated with optimal CTL responses [26]. Godkin et al. found in 2001 for MHC-class II that phenylalanine, isoleucine, valine, leucine, proline, lysine, and glutamic acid at the N-terminal and arginine, histidine, and lysine at the C-terminal flanking region were most frequent, but MHC-class II molecules can present larger epitopes with still attached flanking regions

[17]. However, Steers et al. found a different result for HIV epitopes: aliphatic amino acids with a non-polar side chain and a neutral side chain charge, and neutral amino acids with non-polar and neutral side chain charge at the N-terminal flanking region, and aliphatic amino acid with a polar side chain, and a neutral side chain charge or basic amino acid with a (non) polar and neutral side chain charge at the C-terminal flanking region [49].

It is notable that on the N-terminal site of the flanking region amino acid changes are in the positive and negative HCV dataset, but on the C-terminal site only in the negative. The amino acids which are more frequent in the positive flanking regions are simultaneously less frequent in the negative ones, which indicates a selection pressure against those amino acids and vice versa with those amino acids which are more frequent inside the negative flanking regions. The proteasom is said to cleave only the C-terminus of the MHC ligand, so that the ligand with its N-terminal flanking region will then enter the ER, where this precursor is clipped. This means that those amino acids on the C-terminal site may be important for the cleavage by the proteasome, whereas the amino acids at the N-terminal flanking region may be more important for the cleavage inside the ER. This cleavage is mainly managed through the ERAP (ERAAP) proteins, but not by that alone. Experiments have shown that ERAAP deficient mice show poor MHC-I presentation for certain peptides, but normal or even increased amounts for other peptides [45]. Wild-type mice immunized with ERAAP-ko mice cells showed a potent immune response and analysis of those bound peptides revealed large changes in the composition and the proportion of N-terminally extended peptides [2, 20].

Leucine is the preferred amino acid for HLA-A*02 binding of epitopes at position 2 [30] and the preferred amino acid for the aminopeptidase ERAP1. Our findings show a higher amount of leucine in the N-terminal flanking region of negative epitopes. The reason for this could be that epitopes overlap and those positions are also inside another epitope.

The comparison between HCV and human data proved to be extremely difficult, because almost all amino acids showed huge differences in the amount. This huge discrepancy may mask real effects at certain positions, for example: if there is generally more alanine in HCV genomes than in human genomes, an increased amount of alanine at a certain position in humans, leads to a reduced significance at this position. This could indicate the wrong assumption that every missing significant difference between human flanking regions and HCV flanking regions for those amino acids, which showed a general difference

in the amount of amino acids, is more interesting than those which had a significant difference. One could not analyze such a problem with our techniques.

Our results denote that there is not "'the best amino acid"' to evade proteasomal cleavage at every position in the flanking region, but only a good one, which has to be also suitable for the original function of the protein. Throughout the field, there are a lot of amino acid mutations known that destroy the function of the protein. Those changes, despite their increased chance to evade the proteasom, would not remain in a stable viral population [10, 11]. Therefore it is difficult to estimate if there is a ranking inside amino acids at a basis of escape from proteasomal cleavage.

Although an escape mutation in the flanking region is not directly in connection with the HLA/ T cell specific recognition, the HLA type may influence those evasive mechanisms [12], although the changes in the sequence are small compared to the actual epitope [13]. Such mutations are only necessary if the epitope is recognized and a mutation inside the epitope leads to a reduced viral fitness. In this study, all epitopes were included, regardless of the HLA specificity. This may have led to less significant data. A second problem could be that it was not checked whether the patients behind those epitopes received any kind of treatment or had co-infections for which they received treatment. For HIV infections Kourjian et al. found that certain antiviral drugs change the peptides produced by the aminopeptidases, e.g. a reduced trimming of the N-terminus or enhanced cleavage of acidic residues [24]. Ghany et al. proclaims that up to 8% of those with chronic HCV infection may be HIV co-infected [16], and thus we may have compared epitopes generated through antiviral drugs with those generated by the uninfluenced proteasome.

References

[1] S. F. Altschul et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Res* 25 (1997), pp. 3389–3402. URL: http://blast.ncbi.nlm.nih.gov/Blast.cgi (cit. on p. 102).

[2] N. Blanchard et al. "Endoplasmic Reticulum Aminopeptidase Associated with Antigen Processing Defines the Composition and Structure of MHC Class I Peptide Repertoire in Normal and Virus-Infected Cells". In: *The Journal of Immunology* 184.6 (2010), pp. 3033–3042. DOI: 10.4049/jimmunol.0903712 (cit. on p. 115).

[3] K. J. Blight et al. "Efficient Replication of Hepatitis C Virus Genotype 1a RNAs in Cell Culture". In: *Journal of Virology* 77.5 (2003), pp. 3181–3190. DOI: 10.1128/JVI.77.5.3181-3190.2003. eprint: http://jvi.asm.org/content/77/5/3181.full.pdf+html. URL: http://jvi.asm.org/content/77/5/3181.abstract (cit. on p. 102).

[4] P. Cascio et al. "26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide". In: *The EMBO Journal* 20.10 (2001), pp. 2357–2366. DOI: 10.1093/emboj/20.10.2357. URL: http://emboj.embopress.org/content/20/10/2357 (cit. on p. 101).

[5] C. M. Chang et al. "GBV-C Infection and Risk of NHL among U.S. Adults". In: *Cancer Research* 74.19 (2014), pp. 5553–5560. DOI: 10.1158/0008-5472.CAN-14-0209 (cit. on p. 100).

[6] J. Choi, Z. Xu, and J.-h. Ou. "Triple Decoding of Hepatitis C Virus RNA by Programmed Translational Frameshifting". In: *Molecular and Cellular Biology* 23.5 (2003), pp. 1489–1497. DOI: 10.1128/MCB.23.5.1489-1497.2003. eprint: http://mcb.asm.org/content/23/5/1489.full.pdf+html. URL: http://mcb.asm.org/content/23/5/1489.abstract (cit. on p. 102).

[7] Q. Choo et al. "Isolation of a cDNA derived from a blood-borne non-A, non-B hepatitis genome". In: *Science* 244 (1989), pp. 359–362 (cit. on p. 100).

[8] T. U. Consortium. "Activities at the Universal Protein Resource (UniProt)". In: *Nucleic Acids Research* 42.D1 (2014), pp. 191–198. DOI: 10.1093/nar/gkt1140 (cit. on p. 103).

[9] S. Cook and E. C. Holmes. "A multigene analysis of the phylogenetic relationships among the flaviviruses (Family: Flaviviridae) and the evolution of vector transmission". In: *Arch Virol* 151.2 (Sept. 2005), pp. 309–325. DOI: 10.1007/s00705-005-0626-6. URL: http://dx.doi.org/10.1007/s00705-005-0626-6 (cit. on p. 100).

[10] S. Crotty and R. Andino. "Implications of high {RNA} virus mutation rates: lethal mutagenesis and the antiviral drug ribavirin". In: *Microbes and Infection* 4.13 (2002), pp. 1301–1307. ISSN: 1286-4579. DOI: http://dx.doi.org/10.1016/S1286-4579(02)00008-4 (cit. on p. 116).

[11] S. Crotty, C. Cameron, and R. Andino. "Ribavirin's antiviral mechanism of action: lethal mutagenesis?" English. In: *Journal of Molecular Medicine* 80.2 (2002), pp. 86–95. ISSN: 0946-2716. DOI: 10.1007/s00109-001-0308-0. URL: http://dx.doi.org/10.1007/s00109-001-0308-0 (cit. on p. 116).

[12] R. Draenert et al. “Immune Selection for Altered Antigen Processing Leads to Cytotoxic T Lymphocyte Escape in Chronic HIV-1 Infection”. In: *The Journal of Experimental Medicine* 199.7 (2004), pp. 905–915. DOI: 10.1084/jem.20031982 (cit. on p. 116).

[13] A. L. Erickson et al. “The Outcome of Hepatitis C Virus Infection Is Predicted by Escape Mutations in Epitopes Targeted by Cytotoxic T Lymphocytes”. In: *Immunity* 15.6 (2001), pp. 883–895. ISSN: 1074-7613. DOI: http://dx.doi.org/10.1016/S1074-7613(01)00245-X (cit. on p. 116).

[14] F. Förster et al. “Unveiling the Long-Held Secrets of the 26S Proteasome”. In: *Structure* 21.9 (2013), pp. 1551–1562. URL: http://www.cell.com/structure/abstract/S0969-2126(13)00302-X (cit. on p. 101).

[15] S. L. George et al. “Interactions between GB virus type C and HIV”. In: *Current infectious disease reports* 4.6 (2002), pp. 550–558 (cit. on p. 100).

[16] M. G. Ghany et al. “Diagnosis, management, and treatment of hepatitis C: An update”. In: *Hepatology* 49.4 (2009), pp. 1335–1374. ISSN: 1527-3350. DOI: 10.1002/hep.22759. URL: http://dx.doi.org/10.1002/hep.22759 (cit. on p. 116).

[17] A. J. Godkin et al. “Naturally Processed HLA Class II Peptides Reveal Highly Conserved Immunogenic Flanking Region Sequence Preferences That Reflect Antigen Processing Rather Than Peptide-MHC Interactions”. In: *The Journal of Immunology* 166.11 (2001), pp. 6720–6727. DOI: 10.4049/jimmunol.166.11.6720 (cit. on p. 115).

[18] A. Hattori et al. “Characterization of Recombinant Human Adipocyte–Derived Leucine Aminopeptidase Expressed in Chinese Hamster Ovary Cells”. In: *Journal of Biochemistry* 128.5 (2000), pp. 755–762 (cit. on p. 101).

[19] T. Kanaseki et al. “ERAAP and Tapasin Independently Edit the Amino and Carboxyl Termini of MHC Class I Peptides”. In: *The Journal of Immunology* 191.4 (2013), pp. 1547–1555. DOI: 10.4049/jimmunol.1301043 (cit. on p. 101).

[20] T. Kanaseki et al. “ERAAP Synergizes with MHC Class I Molecules to Make the Final Cut in the Antigenic Peptide Precursors in the Endoplasmic Reticulum”. In: *Immunity* 25.5 (Nov. 2006), pp. 795–806. DOI: 10.1016/j.immuni.2006.09.012. URL: http://dx.doi.org/10.1016/j.immuni.2006.09.012 (cit. on p. 115).

[21] S. I. van Kasteren et al. “Chemical biology of antigen presentation by {MHC} molecules”. In: *Current Opinion in Immunology* 26 (2014). Innate immunity * Antigen processing, pp. 21–31. ISSN: 0952-7915. DOI: http://dx.doi.org/10.1016/j.coi.2013.10.005. URL: http://www.sciencedirect.com/science/article/pii/S0952791513001829 (cit. on p. 101).

[22] Y. Kim et al. “Immune epitope database analysis resource”. In: *Nucleic Acids Research* 40.W1 (2012), W525–W530. DOI: 10.1093/nar/gks438 (cit. on p. 102).

[23] A. A. Kolykhalov et al. “Transmission of Hepatitis C by Intrahepatic Inoculation with Transcribed RNA”. In: *Science* 277.5325 (1997), pp. 570–574. DOI: 10.1126/science.277.5325.570. URL: http://www.sciencemag.org/content/277/5325/570.abstract (cit. on p. 102).

[24] G. Kourjian et al. “Sequence-Specific Alterations of Epitope Production by HIV Protease Inhibitors”. In: *The Journal of Immunology* 192.8 (2014), pp. 3496–3506. DOI: 10.4049/jimmunol.1302805 (cit. on p. 116).

[25] J.-J. Lefrère et al. “GBV-C/hepatitis G virus (HGV) RNA load in immunodeficient individuals and in immunocompetent individuals”. In: *Journal of medical virology* 59.1 (1999), pp. 32–37 (cit. on p. 100).

[26] B. D. Livingston et al. “Optimization of epitope processing enhances immunogenicity of multiepitope {DNA} vaccines”. In: *Vaccine* 19.32 (2001), pp. 4652–4660. ISSN: 0264-410X. DOI: http://dx.doi.org/10.1016/S0264-410X(01)00233-X (cit. on p. 114).

[27] C. Lundegaard et al. “NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11”. In: *Nucleic Acids Research* 36.suppl 2 (2008), W509–W512. DOI: 10.1093/nar/gkn202 (cit. on p. 102).

[28] M. E. Major and S. M. Feinstone. “The molecular virology of hepatitis C”. In: *Hepatology* 25.6 (1997), pp. 1527–1538. ISSN: 1527-3350. DOI: 10.1002/hep.510250637. URL: http://dx.doi.org/10.1002/hep.510250637 (cit. on p. 107).

[29] H. B. Mann and D. R. Whitney. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60 (cit. on p. 105).

[30] M. Matsui, C. E. Hioe, and J. A. Frelinger. “Roles of the six peptide-binding pockets of the HLA-A2 molecule in allorecognition by human cytotoxic T-cell clones.” In: *Proceedings of the National Academy of Sciences* 90.2 (1993), pp. 674–678. DOI: 10.1073/pnas.90.2.674 (cit. on p. 115).

[31] H. Matsumoto et al. “Characterization of a recombinant soluble form of human placental leucine aminopeptidase/oxytocinase expressed in Chinese hamster ovary cells”. In: *European Journal of Biochemistry* 267.1 (2000), pp. 46–52. ISSN: 1432-1033. DOI: 10.1046/j.1432-1327.2000.00949.x. URL: http://dx.doi.org/10.1046/j.1432-1327.2000.00949.x (cit. on p. 101).

[32] A. Milicic et al. “CD8+ T Cell Epitope-Flanking Mutations Disrupt Proteasomal Processing of HIV-1 Nef”. In: *The Journal of Immunology* 175.7 (2005), pp. 4618–4626. DOI: 10.4049/jimmunol.175.7.4618 (cit. on p. 101).

[33] J. Neefjes et al. “Towards a systems understanding of MHC class I and MHC class II antigen presentation”. In: *Nat Rev Immunol* 11.12 (Dec. 2011), pp. 823–836. ISSN: 1474-1733. URL: http://dx.doi.org/10.1038/nri3084 (cit. on p. 101).

[34] P. K. Nelson et al. “Global epidemiology of hepatitis B and hepatitis C in people who inject drugs: results of systematic reviews”. In: *The Lancet* 378.9791 (Aug. 2011), pp. 571–583. DOI: 10.1016/s0140-6736(11)61097-0. URL: http://dx.doi.org/10.1016/S0140-6736(11)61097-0 (cit. on p. 100).

[35] M. Nielsen et al. “Reliable prediction of T-cell epitopes using neural networks with novel sequence representations”. In: *Protein Science* 12.5 (2003), pp. 1007–1017. ISSN: 1469-896X. DOI: 10.1110/ps.0239403. URL: http://dx.doi.org/10.1110/ps.0239403 (cit. on p. 102).

[36] M. S. Panter et al. “Dynamics of Major Histocompatibility Complex Class I Association with the Human Peptide-loading Complex”. In: *Journal of Biological Chemistry* 287.37 (2012), pp. 31172–31184. DOI: 10.1074/jbc.M112.387704 (cit. on p. 101).

[37] P. Paul et al. “A Genome-wide Multidimensional RNAi Screen Reveals Pathways Controlling MHC Class II Antigen Presentation”. In: *Cell* 145.2 (Apr. 2011), pp. 268–283. DOI: 10.1016/j.cell.2011.03.023. URL: http://dx.doi.org/10.1016/j.cell.2011.03.023 (cit. on p. 101).

[38] B. Peters and A. Sette. “Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method”. In: *BMC Bioinformatics* 6.1 (2005), p. 132. DOI: 10.1186/1471-2105-6-132. URL: http://www.biomedcentral.com/1471-2105/6/132 (cit. on p. 102).

[39] O. G. Pybus et al. “Genetic History of Hepatitis C Virus in East Asia”. In: *Journal of Virology* 83.2 (2009), pp. 1071–1082. DOI: 10.1128/JVI.01501-08 (cit. on p. 100).

[40] K. E. Reed, A. E. Gorbalenya, and C. M. Rice. “The NS5A/NS5 Proteins of Viruses from Three Genera of the Family Flaviviridae Are Phosphorylated by Associated Serine/Threonine Kinases”. In: *Journal of Virology* 72.7 (1998), pp. 6199–6206 (cit. on p. 100).

[41] C. M. Rice. “Flaviviridae: the viruses and their replication”. In: ed. by H. P. M. Fields B. N. Knipe D. M. 3rd. Vol. Fields virology. Lippincott-Raven Publishers, Philadelphia, Pa, 1996, 1:931–960 (cit. on p. 100).

[42] M. Sarwar et al. “NS4A protein as a marker of HCV history suggests that different HCV genotypes originally evolved from genotype 1b”. In: *Virology Journal* 8.1 (2011), p. 317. DOI: 10.1186/1743-422X-8-317 (cit. on p. 100).

[43] L. Saveanu et al. “IRAP Identifies an Endosomal Compartment Required for MHC Class I Cross-Presentation”. In: *Science* 325.5937 (2009), pp. 213–217. DOI: 10.1126/science.1172845. URL: http://www.sciencemag.org/content/325/5937/213.abstract (cit. on p. 101).

[44] U. Seifert et al. “Hepatitis C virus mutation affects proteasomal epitope processing”. In: *The Journal of Clinical Investigation* 114.2 (July 2004), pp. 250–259. DOI: 10.1172/JCI20985 (cit. on p. 114).

[45] N. Shastri et al. “Monitoring peptide processing for {MHC} class I molecules in the endoplasmic reticulum”. In: *Current Opinion in Immunology* 26 (2014), pp. 123–127. ISSN: 0952-7915. DOI: http://dx.doi.org/10.1016/j.coi.2013.11.006 (cit. on p. 115).

[46] Z. T. Shen et al. “Disparate Epitopes Mediating Protective Heterologous Immunity to Unrelated Viruses Share Peptide–MHC Structural Features Recognized by Cross-Reactive T Cells”. In: *The Journal of Immunology* 191.10 (2013), pp. 5139–5152. DOI: 10.4049/jimmunol.1300852 (cit. on p. 113).

[47] P.-Y. Shi. *Molecular Virology and Control of Flaviviruses*. Horizon Scientific Press, 2012 (cit. on p. 100).

[48] J. Sidney et al. “Quantitative peptide binding motifs for 19 human and mouse {MHC} class I molecules derived using positional scanning combinatorial peptide libraries”. In: *Immunome Res* 4.1 (2008), p. 2. DOI: 10.1186/1745-7580-4-2. URL: http://dx.doi.org/10.1186/1745-7580-4-2 (cit. on p. 102).

[49] N. J. Steers et al. “Designing the epitope flanking regions for optimal generation of {CTL} epitopes”. In: *Vaccine* 32.28 (2014), pp. 3509–3516. DOI: http://dx.doi.org/10.1016/j.vaccine.2014.04.039 (cit. on pp. 101, 115).

[50] T. Tanioka et al. “Human Leukocyte-derived Arginine Aminopeptidase: The third member of the oxytocinase subfamily of aminopeptidases”. In: *Journal of Biological Chemistry* 278.34 (2003), pp. 32275–32283. DOI: 10.1074/jbc.M305076200. URL: http://www.jbc.org/content/278/34/32275.abstract (cit. on p. 101).

[51] J. Timm et al. “CD8 Epitope Escape and Reversion in Acute HCV Infection”. In: *The Journal of Experimental Medicine* 200.12 (2004), pp. 1593–1604. DOI: 10.1084/jem.20041006 (cit. on p. 114).

[52] R. Vita et al. “The immune epitope database 2.0.” In: *Nucleic Acids Res.* 38 (Database issue) (Jan. 2010), pp. D854–62 (cit. on p. 102).

[53] E. H. Warren et al. “An Antigen Produced by Splicing of Noncontiguous Peptides in the Reverse Order”. In: *Science* 313.5792 (2006), pp. 1444–1447. DOI: 10.1126/science.1130660. eprint: http://www.sciencemag.org/content/313/5792/1444.full.pdf. URL: http://www.sciencemag.org/content/313/5792/1444.abstract (cit. on p. 101).

[54] E. Westaway et al. “Flaviviridae”. In: *Intervirology* 24.4 (1985), pp. 183–192. DOI: 10.1159/000149642. URL: http://dx.doi.org/10.1159/000149642 (cit. on p. 100).

8

Additional publications - only abstracts

I have contributed in the following papers with certain tools written in R (as for example SeqFeatR) or with my expertise:

8.1 Differential selection in HIV-1 *gp120* between subtype B and East Asian variant B'

Dang S, Wang Y, Budeus B, Verheyen J, Yang R, Hoffmann D. **Virologica Sinica (2014)**

HIV-1 evolves strongly and undergoes geographic differentiation as it spreads in diverse host populations around the world. For instance, distinct genomic backgrounds can be observed between the pandemic subtype B, prevalent in Europe and North-America, and its offspring clade B' in East Asia. Here we ask whether this differentiation affects the selection pressure experienced by the virus. To answer this question we evaluate selection pressure on the HIV-1 envelope protein *gp120* at the level of individual codons using a simple and fast estimation method based on the ratio k a /k s of amino acid changes to synonymous changes. To validate the approach we compare results to those from a state-of-the-art mixed-effect method. The agreement is acceptable, but the analysis also demonstrates some limitations of the simpler approach. Further, we find similar distributions of codons under stabilizing and directional selection pressure in *gp120* for subtypes B and B' with more directional selection pressure in variable loops and more stabilizing selection in the constant regions. Focusing on codons with increased k a /k s values in B', we show that these codons are scattered over the whole of *gp120*, with remarkable clusters of higher density in regions flanking the variable loops. We identify

a significant statistical association of glycosylation sites and codons with increased k a /k s values.

Link: http://link.springer.com/article/10.1007%2Fs12250-014-3389-y

8.2 Adaptation of the hepatitis B virus core protein to CD8$^+$ T cell selection pressure

Helenie Kefalakes, Bettina Budeus, Andreas Walker, Christoph Jochum, Gudrun Hilgard, Andreas Heinold, Falko Heinemann, Guido Gerken, Daniel Hoffmann, Joerg Timm **Hepatology (2015)**

Background & Aims: Chronic infections with the hepatitis B virus (HBV) are worldwide an enormous public health problem. Effective suppression of viral replication can be achieved with inhibitors of the viral polymerase. However, in most cases lifelong treatment is required to avoid recurrence of viremia. Activation of HBV-specific CD8 T cells by therapeutic vaccination may promote sustained control of viral replication by clearance of cccDNA from infected hepatocytes. Importantly, little is known about the exact targets of the CD8 T cell response and the extent of selection pressure on the virus. Here, it was hypothesized that CD8 T cell responses associated with strong selection pressure on the virus can be identified by viral sequence analysis. **Methods:** The HBV core gene was amplified and sequenced from 148 patients with chronic HBV infection and the HLA class I genotype (A and B locus) was determined. Residues under selection pressure in the presence of particular HLA class I alleles were identified by a statistical approach utilizing the novel analysis package 'SeqFeatR'. Candidate CD8 T cell epitopes were confirmed by analysis of PBMCs from patients with chronic infection. **Results:** With a false discovery rate set to 0.2 a total of 9 residues under selection pressure in the presence of 10 different HLA class I alleles were identified. Additional immunological experiments confirmed that 7 of the residues were located inside epitopes targeted by patients with chronic HBV infection carrying the relevant HLA class I-allele. Consistent with viral escape some of the selected substitutions impaired recognition by HBV-specific CD8 T cells. **Conclusion:** Viral sequence analysis allows identification of HLA class I-restricted epitopes under reproducible selection pressure in HBV core. The possibility of viral adaptation to CD8 T cell immune pressure needs attention in the context of therapeutic vaccination.

Link: http://onlinelibrary.wiley.com/doi/10.1002/hep.27771/abstract

8.3 AmpliconDuo: A Split-Sample Filtering Protocol for High-Throughput Amplicon Sequencing of Microbial Communities

Anja Lange, Steffen Jost, Dominik Heider, Christina Bock, Bettina Budeus, Elmar Schilling, Axel Strittmatter, Jens Boenigk, Daniel Hoffmann **PLoS One (2015)**

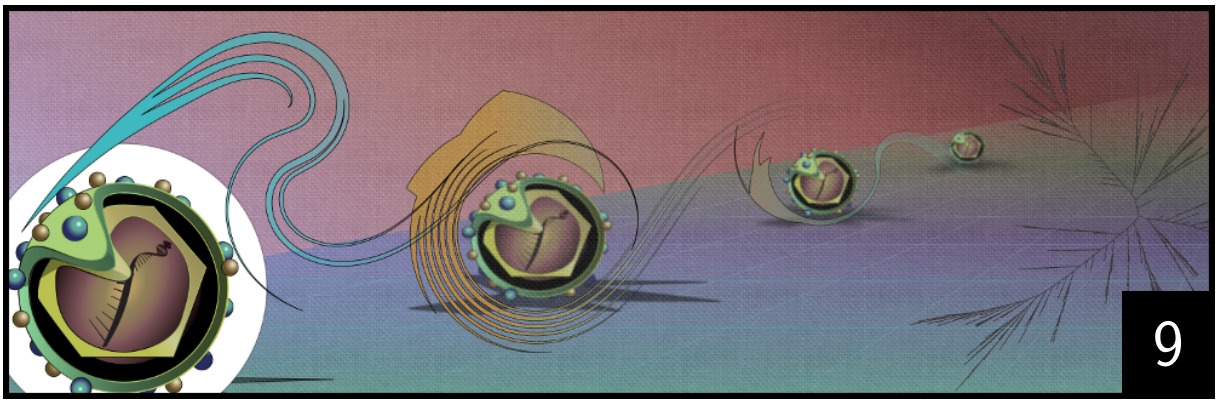
High throughput sequencing (HTSeq) of small ribosomal subunit amplicons has the potential for a comprehensive characterization of microbial community compositions, down to rare species. However, the error-prone nature of the multi-step experimental process requires that the resulting raw sequences are subjected to quality control procedures. These procedures often involve an abundance cutoff for rare sequences or clustering of sequences, both of which limit genetic resolution. Here we propose a simple experimental protocol that retains the high genetic resolution granted by HTSeq methods while effectively removing many low abundance sequences that are likely due to PCR and sequencing errors. According to this protocol, we split samples and submit both halves to independent PCR and sequencing runs. The resulting sequence data is graphically and quantitatively characterized by the discordance between the two experimental branches, allowing for a quick identification of problematic samples. Further, we discard sequences that are not found in both branches ("'AmpliconDuo filter"'). We show that the majority of sequences removed in this way, mostly low abundance but also some higher abundance sequences, show features expected from random modifications of true sequences as introduced by PCR and sequencing errors. On the other hand, the filter retains many low abundance sequences observed in both branches and thus provides a more reliable census of the rare biosphere. We find that the AmpliconDuo filter increases biological resolution as it increases apparent community similarity between biologically similar communities, while it does not affect apparent community similarities between biologically dissimilar communities. The filter does not distort overall apparent community compositions. Finally, we quantitatively explain the effect of the AmpliconDuo filter by a simple mathematical model.

Link: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141590

PART

IV

PHYLOGENETIC SEQUENCES ANALYSIS



9

Phylogenetic analysis on HCV infection chains

Chaos was the law of nature; Order was the dream of man.

THE EDUCATION OF HENRY ADAMS

Abstract

Around 170 million people worldwide are infected with HCV. Depending on the origin of the infection a patient might get compensations for health, moral and economic distresses he has to suffer from. An example for such a compensation is the Anti-D law in Germany for women and their offspring, who were infected by contaminated blood. Since HCV is an RNA-virus and has a high mutation rate, direct infections would be visible in the similarity of the patient's viral sequences compared to the potential source.

We studied HCV sequences from the Anti-D cohort and HCV sequences from the Hepatitis C Virus (HCV) Database Project to find regions in the genome where closer related sequences could be distinguished from the farther related ones. We used R to build a bootstrap-like method, in which the distance of two random sequences from the same genotype is noted. This estimation is repeated Θ times for a given window length λ. If all pairs from the Θ repetitions were direct neighbors in a phylogenetic tree and therefore had the smallest distance in the tree, λ would be appropriate as a genome region to identify infection chains. We found two genome regions to fit our requirements: A part around the hypervariable region 2 and a part of the NS5-B sequence.

9.1 Introduction

50% to 80% of HCV infections are chronic and lead eventually to cirrhosis and hepatocellular carcinoma. It is not uncommon that an infected person searches for the source of his or her infection because there may be compensations paid for health, moral and economic distress [24]. Often these sources are intravenous injected drugs or blood transfusions [4, 36, 48], before blood donor screening by second-generation antibody test was introduced [3], but even now there are cases of unsafe injection practices, which can lead to health-care–associated hepatitis C [5, 12, 53], either through transmission from health care worker to patient or from patient to patient [8, 14]. Also four cases of homicide with HCV are known [17, 20, 40, 42]. In around 10% of cases the infection chain is unknown [13].

In large outbreaks such as the Anti-D cohort in east Germany [19, 52], where women were immunized with infected blood, a lot of compensation is being paid to the victims through an insurance for injury from immunization [2, 18, 22, 23] and every year new patients come forward who claim to have been infected in such situations. Such discussions, if the patient is right, are often enervating for both sides. The blood tragedy in Canada is even thematized in a novel [41]. So it is vital not only for the victims, but also for the potential source to estimate if an infection chain exists and if compensation should be paid by the people responsible.

An analysis of the degree is needed to identify if the HC-virus from two subjects is related. Since RNA viruses such as HIV and HCV evolve very fast (0.13 to $2.24 * 10^{-1}$/genome site/yr [26, 33]) and HCV employs an RNA-dependent RNA polymerase, which has no proof-reading function to correct for errors in the copy, changes in the basic genetic material are often unique for a subspecies. Viral populations accumulate enough such changes in a short time period, that molecular phylogenetic inference methods can be used to analyze the short-term history of these viruses.

Molecular phylogenetic analyses are widely used to examine the degree of nucleotide or amino acid sequences. The first methods for such an analysis were distance based in the 1960s. Later on maximum parsimony and maximum likelihood [21, 47] - both being character based methods - were introduced. With increasing computational power, bigger datasets and lager problems could be analyzed. For both kinds of analysis more complex models and methods were found like Bayesian Inference [28, 43], which are included in

a maximum likelihood method [50]. Often bootstrapping is used to get a qualitative value of the correctness of a single branch. Bootstrapping is a statistical method (for random sampling with replacement) which allows estimation of the sampling distribution of almost any statistic. So it is possible to analyze if a set of samples from an infected patient derived from a given source, if there is a precondition fulfilled: there have to be enough phylogenetic signals for reconstructing their evolutionary history [30]. A low signal can occur mainly through two situations. Analyzing a relatively slow evolving region, which results in too few genetic differences or the saturation of substitutions in a very fast evolving region. Neither will result in a reliable estimation of the degree of relationship. The first situation is more common, and therefore hypervariable regions in the sequences like E1/E2 from HCV are often used [24, 25], but there is also one case where the core region of HCV was used as evidence in court [44]. For HIV there are more cases of molecular phylogenetic analyses used in court to provide evidence that would lead to compensations being paid [6, 37].

We examined HCV sequences from different (sub)genotypes and different countries to find regions where we could distinguish closer related sequences from the rest of the sequences. These regions could then be used to identify if two sequences are closely related, and the result could be used as evidence in court.

9.2 Methods

9.2.1 Data

We took 45 known Anti-D HCV sequences and made a sequence set with all of the sequences from the Hepatitis C Virus Database Project [32] (4th October 2012). This sequence set was heavily purified for duplicated sequences and sequences which differ massively from the rest of the set (sequence similarity below 60 percent), since we want to identify regions in similar sequences. Amino acids in front of the starting codon were removed. The remaining set contained 1883 sequences (see Table 9.1).

The sequences were aligned with the reference sequence H77 [7, 16, 31]. Most of the 1883 sequences were of subtype 1 (1479 sequences) and 2 (162 sequences). The rest of the known subtypes 3 - 6 were relatively few (242) and their subtypes were merged.

Table 9.1: Overview of the composition of the used HCV sequence set.

subtype	1?	1a	1b	1c	1e	1g	2?	2a	2b	2c	2d	2i
number	20	679	776	2	1	1	12	31	89	8	1	5

subtype	2j	2k	2l	2m	2q	2r	3	4	5	6	?
number	5	4	1	3	2	1	66	64	9	78	25

9.2.2 Computational Analysis

To find the best region in the sequence for the estimation of related sequences, we created a bootstrap like procedure. We generated 100 trees (M) of a set of 20 sequences (N), each 200 nucleotides long, and counted the number of trees in which two test sequences with the same genotype were direct neighbors. We used the R-packages ape [39] and phangorn [49] to generate the trees for this method.

Data: sequences in FASTA format
Result: list with bootstrap values for a range of sequence positions

```
initialization (N, M);
for region in sequence regions do
    b = 0;
    sequence 1 = select random sequence;
    sequence 2 = select random sequence;
    find subtype for this two sequences;
    if subtypes are the same then
        selected sequences = get N sequences of same subtype;
        model = use model test on selected sequences;
        while m <M+1 do
            set of sequences = random number of sequences out of N;
            create tree from model and set of sequences;
            if sequence 1 and sequence 2 are neighbors then
                increase b by 1;
            end
        end
        write position and b to result file;
    end
end
```

Algorithm 9.1: find_with_trees algorithm

Table 9.2: HCV positions with the longest stretch of 100% bootstraps.

		Position start	Position end
First position	Alignment	1554	1760
	H77	1575	1833
Second position	Alignment	8945	9150
	H77	8471	8850

9.3 Results

We got bootstrap-like values (which we call pseudo-bootstraps) from the 'find_with_-trees' algorithm for regions of each 200*bp* over the whole genome. These values indicate where the tree-building function in the algorithm can easily differ two sequences of a certain subtype (e.g. two sequences from the Anti-D-cohort out of a set of sequences with subtype 1b). The analysis showed two larger regions of high discriminatory power, one around the hypervariable region 2 and one in NS5B (see Figure 9.1). For a better overview we combined every 5 single values together and displayed the mean as a line plot.

In the list output, we could identify those regions more easily. We chose those regions which had the longest stretch of pseudo-bootstraps above 95. Around the hypervariable region 2 it is in the H77 reference genome 1575-1833 and for NS5B it is 8471-8850 (see Table 9.2).

9.4 Discussion

We could identify two regions in the HCV genome with our bootstrap-like algorithm. Those regions were: a part around the hypervariable region 2 (HVR2), and a part of the non-structural protein 5B (NS5B). This reflects other paper in the field. In 2002 Salemi and Vandamme used maximum likelihood methods to analyses whole HCV genomes and also found the NS5B region to give the best results for their method followed by NS3 [45]. In 2007 Kato et al. analyzed blood samples from injecting drug users and found a highly similar nucleotide sequence inside NS5B and they proclaimed that it could demonstrate not only in-house or iatrogenic infection, but also be used for forensic applications such as identifying accomplices [29]. Pagani et al. found in 2012 that subtyping of HCV genomes is better done with NS5B sequencing instead of a commercial PCR/hybridization method

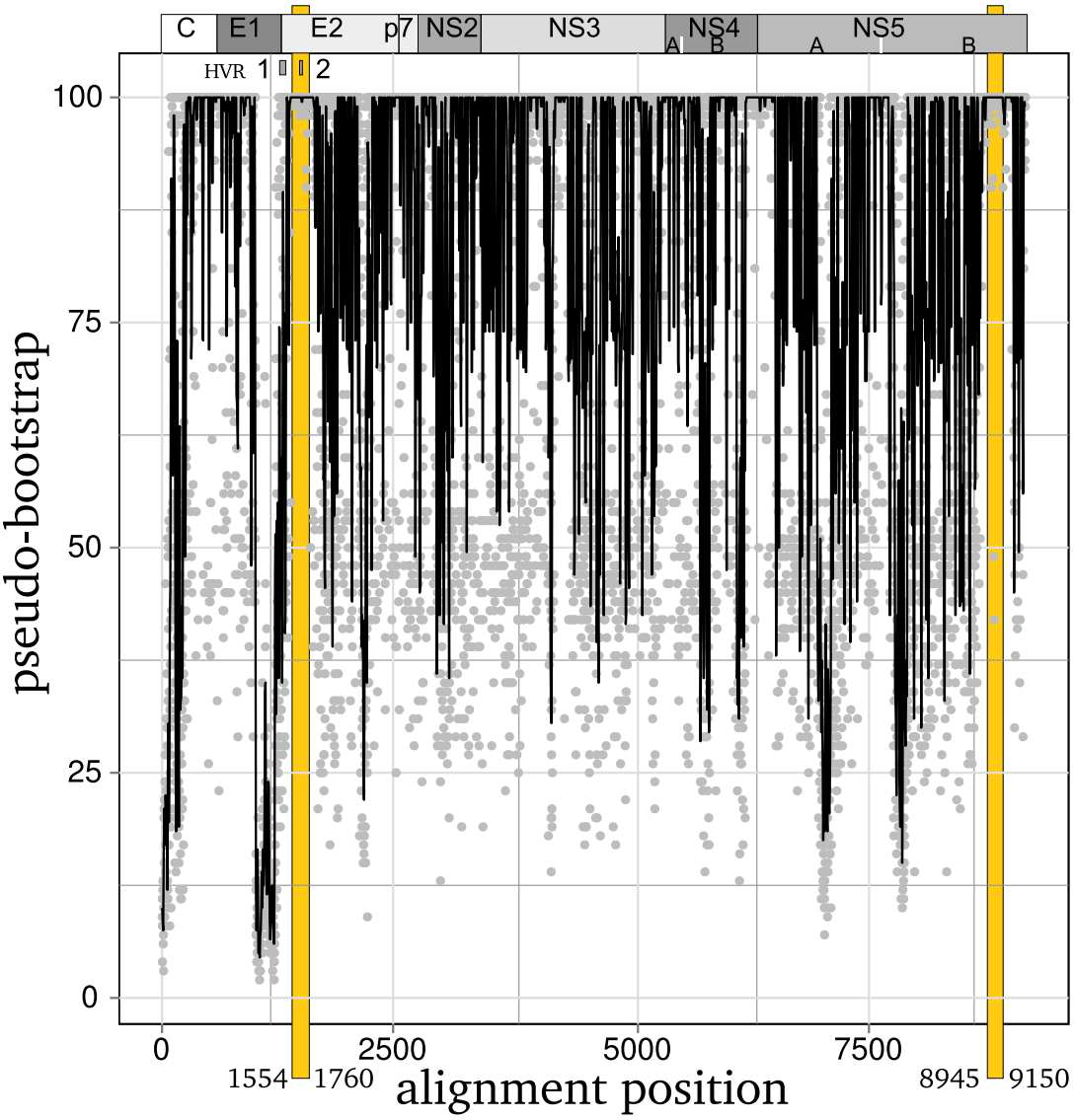


Figure 9.1: Bootstrap values for the whole HCV genome. Every point is the pseudo-bootstrap value from a comparison of a region of 200 *bp*. As black line plot is the mean value of every 5 single points. The HCV genome is noted above.

based on the conserved 5'-untranslated region [38]. But up until now there is no consensus about which region should be used. Other studies tried NS3, core, and envelope proteins [11, 15, 27, 35]. DeCarvalho-Mello et al. found in 2010 that sequences from different origins showed similar sequence patterns inside NS5B and NS3 [9].

Most of the previous studies mentioned analyzed sequences from certain outbreaks or medical sets, where a relationship between the sequences was known or could be guessed. Our dataset consists only of sequences from the Hepatitis C Virus (HCV) Database Project [32]

and the Anti-D sequence set up until October 2012, which is of course specific for Germany. For this sequence set sequence based relationships are easy to recognize, since Anti-D sequences clearly separate in phylogenetic analyses with other subtype 1b sequences [51]. An early analysis in 1998 of sequences from the Anti-D cohort by McAllister had some difficulties on PCR level and found that the HVR region of Anti-D sequence are very diverse [34]. Casino et al. showed one year later that no phylogenetic relationship exists between those sequences inside the HVR [10].

Since we found HVR2, not NS3, to be the second region of interest a combination of these two region should be made to investigate the relationship of HCV sequences. This combination would increase the chance to better identify the infection chain, as DeCarvalho-Mello et al. suggested, too [9]. NS5B sequence similarity gives a quick overview over the relationship of the sequences, whereas the HVR enables to take a deeper look inside, because this region has a stronger selection against the immune system [1, 34, 46], but also a high risk of errors through PCR. NS5B alone is not diverse enough, HVR too much, but together they should give a clear picture. The found results should be analyzed in vitro to confirm that NS5B and HVR are suitable for use in forensics.

References

[1] I. Abbate et al. "HVR-1 quasispecies modifications occur early and are correlated to initial but not sustained response in HCV-infected patients treated with pegylated- or standard-interferon and ribavirin". In: *Journal of Hepatology* 40.5 (2004), pp. 831–836. ISSN: 0168-8278. DOI: http://dx.doi.org/10.1016/j.jhep.2004.01.019 (cit. on p. 132).

[2] Y.-S. Ahn and H.-S. Lim. "Occupational Infectious Diseases among Korean Health Care Workers Compensated with Industrial Accident Compensation Insurance from 1998 to 2004". In: *Industrial Health* 46.5 (2008), pp. 448–454. DOI: 10.2486/indhealth.46.448 (cit. on p. 127).

[3] H. Alter and M. Houghton. "Hepatitis C virus and eliminating post-transfusion hepatitis". In: *NATURE MEDICINE* 6.10 (Oct. 2000), pp. 1082–1086. ISSN: 1078-8956. DOI: 10.1038/80394 (cit. on p. 127).

[4] M. J. Alter. "Epidemiology of hepatitis C". In: *Hepatology* 26.S3 (1997), 62S–65S. ISSN: 1527-3350. DOI: 10.1002/hep.510260711. URL: http://dx.doi.org/10.1002/hep.510260711 (cit. on p. 127).

[5] M. J. Alter. "Healthcare should not be a vehicle for transmission of hepatitis C virus". In: *Journal of Hepatology* 48.1 (2008), pp. 2–4. ISSN: 0168-8278. DOI: http://dx.doi.org/10.1016/j.jhep.2007.10.007 (cit. on p. 127).

[6] E. Bernard et al. "HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission*". In: *HIV Medicine* 8.6 (2007), pp. 382–387. ISSN: 1468-1293. DOI: 10.1111/j.1468-1293.2007.00486.x. URL: http://dx.doi.org/10.1111/j.1468-1293.2007.00486.x (cit. on p. 128).

[7] K. J. Blight et al. "Efficient Replication of Hepatitis C Virus Genotype 1a RNAs in Cell Culture". In: *Journal of Virology* 77.5 (2003), pp. 3181–3190. DOI: 10.1128/JVI.77.5.3181-3190.2003. eprint: http://jvi.asm.org/content/77/5/3181.full.pdf+html. URL: http://jvi.asm.org/content/77/5/3181.abstract (cit. on p. 128).

[8] J.-P. Bronowicki et al. "Patient-to-Patient Transmission of Hepatitis C Virus during Colonoscopy". In: *New England Journal of Medicine* 337.4 (1997). PMID: 9227929, pp. 237–240. DOI: 10.1056/NEJM199707243370404. eprint: http://www.nejm.org/doi/pdf/10.1056/NEJM199707243370404. URL: http://www.nejm.org/doi/full/10.1056/NEJM199707243370404 (cit. on p. 127).

[9] I. M. V. G. de Carvalho-Mello et al. "Molecular evidence of horizontal transmission of hepatitis C virus within couples". In: *Journal of General Virology* 91.3 (2010), pp. 691–696. DOI: 10.1099/vir.0.015594-0 (cit. on pp. 131, 132).

[10] C. Casino et al. "Variation of hepatitis C virus following serial transmission: multiple mechanisms of diversification of the hypervariable region and evidence for convergent genome evolution." In: *Journal of General Virology* 80.3 (1999), pp. 717–25 (cit. on p. 132).

[11] N. d. P. Cavalheiro et al. "Hepatitis C: sexual or intrafamilial transmission? Epidemiological and phylogenetic analysis of hepatitis C virus in 24 infected couples". en. In: *Revista da Sociedade Brasileira de Medicina Tropical* 42 (June 2009), pp. 239–244. ISSN: 0037-8682 (cit. on p. 131).

[12] C. f. D. C. CDC. "Acute Hepatitis C Virus Infections Attributed to Unsafe Injection Practices at an Endoscopy Clinic - Nevada, 2007". In: *Morbidity and Mortality Weekly Report* 57(19) (2008), pp. 513–517. URL: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5719a2.htm (cit. on p. 127).

[13] C. f. D. C. CDC. "Recommendations for Prevention and Control of Hepatitis C Virus (HCV) Infection and HCV-Related Chronic Disease". In: *Morbidity and Mortality Weekly Report* 47(RR19) (1998), pp. 1–39. URL: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5719a2.htm (cit. on p. 127).

[14] C. f. D. C. CDC. "Transmission of Hepatitis B and C Viruses in Outpatient Settings — New York, Oklahoma, and Nebraska, 2000–2002". In: *Morbidity and Mortality Weekly Report* 52(38) (2003), pp. 901–906. URL: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5719a2.htm (cit. on p. 127).

[15] K. Chayama et al. "Molecular analysis of intraspousal transmission of hepatitis C virus". In: *Journal of Hepatology* 22.4 (1995), pp. 431–439. ISSN: 0168-8278. DOI: http://dx.doi.org/10.1016/0168-8278(95)80106-5 (cit. on p. 131).

[16] J. Choi, Z. Xu, and J.-h. Ou. "Triple Decoding of Hepatitis C Virus RNA by Programmed Translational Frameshifting". In: *Molecular and Cellular Biology* 23.5 (2003), pp. 1489–1497. DOI: 10.1128/MCB.23.5.1489-1497.2003. eprint: http://mcb.asm.org/content/23/5/1489.full.pdf+html. URL: http://mcb.asm.org/content/23/5/1489.abstract (cit. on p. 128).

[17] S. H. Cody et al. "Hepatitis C Virus Transmission From an Anesthesiologist to a Patient". In: *Archives of Internal Medicine* 162.3 (Feb. 2002), p. 345. DOI: 10.1001/archinte.162.3.345. URL: http://dx.doi.org/10.1001/archinte.162.3.345 (cit. on p. 127).

[18] M. Cornberg et al. "A systematic review of hepatitis C virus epidemiology in Europe, Canada and Israel". In: *Liver International* 31 (2011), pp. 30–60. ISSN: 1478-3231. DOI: 10.1111/j.1478-3231.2011.02539.x. URL: http://dx.doi.org/10.1111/j.1478-3231.2011.02539.x (cit. on p. 127).

[19] S. Dittmann et al. "Long-term persistence of hepatitis C virus antibodies in a single source outbreak". In: *Journal of Hepatology* 13.3 (1990), pp. 323–327. URL: http://www.journal-of-hepatology.eu/article/0168-8278(91)90076-N/abstract (cit. on p. 127).

[20] J. I. Esteban et al. "Transmission of Hepatitis C Virus by a Cardiac Surgeon". In: *New England Journal of Medicine* 334 (1996). PMID: 8569822, pp. 555–561. eprint: http://www.nejm.org/doi/pdf/10.1056/NEJM199602293340902. URL: http://www.nejm.org/doi/full/10.1056/NEJM199602293340902 (cit. on p. 127).

[21] J. Felsenstein. *Confidence limits on phylogenies: an approach using the bootstrap*. 1985. DOI: 10.2307/2408678 (cit. on p. 127).

[22] D. Fitzsimons et al. "Prevention of viral hepatitis in the Nordic countries and Germany". In: *Scandinavian Journal of Infectious Diseases* 37 (2005), pp. 549–560 (cit. on p. 127).

[23] *Gesetz über die Hilfe für durch Anti-D-Immunprophylaxe mit dem Hepatitis–C–Virus infizierte Personen (Anti-D-Hilfegesetz – AntiDHG)*. Vol. I. 38. Bundesgesetzblatt, Aug. 2000 (cit. on p. 127).

[24] F. González-Candelas. "Molecular Phylogenetic Analyses in Court Trials". In: *eLS*. John Wiley & Sons, Ltd, 2010. DOI: 10.1002/9780470015902.a0021575. URL: http://dx.doi.org/10.1002/9780470015902.a0021575 (cit. on pp. 127, 128).

[25] F. González-Candelas et al. "Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source". In: *BMC Biology* 11.1 (2013), p. 76. ISSN: 1741-7007. DOI: 10.1186/1741-7007-11-76. URL: http://www.biomedcentral.com/1741-7007/11/76 (cit. on p. 128).

[26] R. Gray et al. "The mode and tempo of hepatitis C virus evolution within and among hosts". In: *BMC Evolutionary Biology* 11.1 (2011), p. 131. ISSN: 1471-2148. DOI: 10.1186/1471-2148-11-131. URL: http://www.biomedcentral.com/1471-2148/11/131 (cit. on p. 127).

[27] M. Honda et al. "Risk of hepatitis C virus infections through household contact with chronic carriers: Analysis of nucleotide sequences". In: *Hepatology* 17.6 (1993), pp. 971–976. ISSN: 1527-3350. DOI: 10.1002/hep.1840170605. URL: http://dx.doi.org/10.1002/hep.1840170605 (cit. on p. 131).

[28] J. P. Huelsenbeck et al. "Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology". In: *Science* 294.5550 (2001), pp. 2310–2314. DOI: 10.1126/science.1065889. eprint: http://www.sciencemag.org/content/294/5550/2310.full.pdf. URL: http://www.sciencemag.org/content/294/5550/2310.abstract (cit. on p. 127).

[29] H. Kato et al. "Identification and phylogenetic analysis of hepatitis C virus in forensic blood samples obtained from injecting drug users". In: *Forensic Science International* 168.1 (2007), pp. 27–33. ISSN: 0379-0738. DOI: http://dx.doi.org/10.1016/j.forsciint.2006.06.007 (cit. on p. 130).

[30] P. Keim et al. "Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales". In: *Infection, Genetics and Evolution* 4.3 (2004). 6th International Meeting on Microbial Epidemiological Markers, pp. 205–213. ISSN: 1567-1348. DOI: http://dx.doi.org/10.1016/j.meegid.2004.02.005. URL: http://www.sciencedirect.com/science/article/pii/S1567134804000346 (cit. on p. 128).

[31] A. A. Kolykhalov et al. "Transmission of Hepatitis C by Intrahepatic Inoculation with Transcribed RNA". In: *Science* 277.5325 (1997), pp. 570–574. DOI: 10.1126/science.277.5325.570. URL: http://www.sciencemag.org/content/277/5325/570.abstract (cit. on p. 128).

[32] C. Kuiken et al. "HCV sequence database". In: *Bioinformatics* 21.3 (2005), pp. 379–84 (cit. on pp. 128, 131).

[33] M. Kurosaki et al. "Rapid sequence variation of the hypervariable region of hepatitis C virus during the course of chronic infection". In: *Hepatology* 18.6 (1993), pp. 1293–1299. ISSN: 1527-3350. DOI: 10.1002/hep.1840180602. URL: http://dx.doi.org/10.1002/hep.1840180602 (cit. on p. 127).

[34] J. McAllister et al. "Long-Term Evolution of the Hypervariable Region of Hepatitis C Virus in a Common-Source-Infected Cohort". In: *Journal of Virology* 72.6 (1998), pp. 4893–4905 (cit. on p. 132).

[35] K. Nakashima et al. “Sexual Transmission of Hepatitis C Virus among Female Prostitutes and Patients with Sexually Transmitted Diseases in Fukuoka, Kyushu, Japan”. In: *American Journal of Epidemiology* 136.9 (1992), pp. 1132–1137 (cit. on p. 131).

[36] M. Orsini. “The Politics of Naming, Blaming and Claiming: HIV, Hepatitis C and the Emergence of Blood Activism in Canada”. In: *Canadian Journal of Political Science/Revue canadienne de science politique* 35 (03 Sept. 2002), pp. 475–498. ISSN: 1744-9324. DOI: 10.1017/S0008423902778323. URL: http://journals.cambridge.org/article_S0008423902778323 (cit. on p. 127).

[37] C.-Y. Ou et al. “Molecular Epidemiology of HIV Transmission in a Dental Practice”. In: *Science* 256.5060 (1992), pp. 1165–1171. DOI: 10.1126/science.256.5060.1165. eprint: http://www.sciencemag.org/content/256/5060/1165.full.pdf (cit. on p. 128).

[38] E. Pagani et al. “Comparison of hepatitis C virus subtyping by ns5b sequencing with 5’utr based methods”. In: *Minerva medica* 103.4 (Aug. 2012), pp. 293–297. ISSN: 0026-4806 (cit. on p. 131).

[39] E. Paradis, J. Claude, and K. Strimmer. “APE: analyses of phylogenetics and evolution in R language”. In: *Bioinformatics* 20 (2004), pp. 289–290 (cit. on p. 129).

[40] J. L. Perry, R. D. Pearson, and J. Jagger. “Infected health care workers and patient safety: A double standard”. In: *American Journal of Infection Control* 34.5 (2006), pp. 313–319. ISSN: 0196-6553. DOI: http://dx.doi.org/10.1016/j.ajic.2006.01.004 (cit. on p. 127).

[41] A. Picard. *The Gift of Death: Confronting Canada’s Tainted-Blood Tragedy*. Harpercollins Canada; Revised edition, 1997 (cit. on p. 127).

[42] “Radiology Technician Convicted of Infecting Patients with Hepatitis C”. In: *Biotechnology Law Report* 31.4 (Aug. 2012), pp. 382–382. DOI: 10.1089/blr.2012.9813. URL: http://dx.doi.org/10.1089/blr.2012.9813 (cit. on p. 127).

[43] B. Rannala and Z. Yang. “Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference”. English. In: *Journal of Molecular Evolution* 43.3 (1996), pp. 304–311. ISSN: 0022-2844. DOI: 10.1007/BF02338839. URL: http://dx.doi.org/10.1007/BF02338839 (cit. on p. 127).

[44] R. Ross et al. “Inquiries on intraspousal transmission of hepatitis C virus: benefits and limitations of genome sequencing and phylogenetic analysis”. In: *Forensic Science International* 100.1–2 (1999), pp. 69–76. ISSN: 0379-0738. DOI: http://dx.doi.org/10.1016/S0379-0738(98)00200-X. URL: http://www.sciencedirect.com/science/article/pii/S037907389800200X (cit. on p. 128).

[45] M. Salemi and A.-M. Vandamme. “Hepatitis C Virus Evolutionary Patterns Studied Through Analysis of Full-Genome Sequences”. English. In: *Journal of Molecular Evolution* 54.1 (2002), pp. 62–70. ISSN: 0022-2844. DOI: 10.1007/s00239-001-0018-9 (cit. on p. 130).

[46] V. Saludes et al. “Evolutionary dynamics of the E1–E2 viral populations during combination therapy in non-responder patients chronically infected with hepatitis C virus subtype 1b”. In: *Infection, Genetics and Evolution* 13 (2013), pp. 1–10. ISSN: 1567-1348. DOI: http://dx.doi.org/10.1016/j.meegid.2012.09.012 (cit. on p. 132).

[47] L. Sanderson and J. Michael. "Confidence limits on phylogenies: The bootstrap revisited". In: *Cladistics* 5.2 (1989), pp. 113–129. ISSN: 1096-0031. DOI: 10.1111/j.1096-0031.1989.tb00559.x. URL: http://dx.doi.org/10.1111/j.1096-0031.1989.tb00559.x (cit. on p. 127).

[48] E. P. o. H. C. E. (C. H. C. Schabas Richard. *Report on the meeting of the Expert Panel on Hepatitis C Epidemiology, June 17-18, 1998*. [Ottawa?]: Health Canada, 1999 (cit. on p. 127).

[49] K. Schliep. "phangorn: phylogenetic analysis in R". In: *Bioinformatics* 27.4 (2011), pp. 592–593 (cit. on p. 129).

[50] E. Suárez-Díaz and V. H. Anaya-Muñoz. "History, objectivity, and the construction of molecular phylogenies". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 39.4 (2008), pp. 451–468. ISSN: 1369-8486. DOI: http://dx.doi.org/10.1016/j.shpsc.2008.09.002. URL: http://www.sciencedirect.com/science/article/pii/S1369848608000745 (cit. on p. 128).

[51] J. Timm et al. "Human leukocyte antigen–associated sequence polymorphisms in hepatitis C virus reveal reproducible immune responses and constraints on viral evolution". In: *Hepatology* 46.2 (2007), pp. 339–349. ISSN: 1527-3350. DOI: 10.1002/hep.21702 (cit. on p. 132).

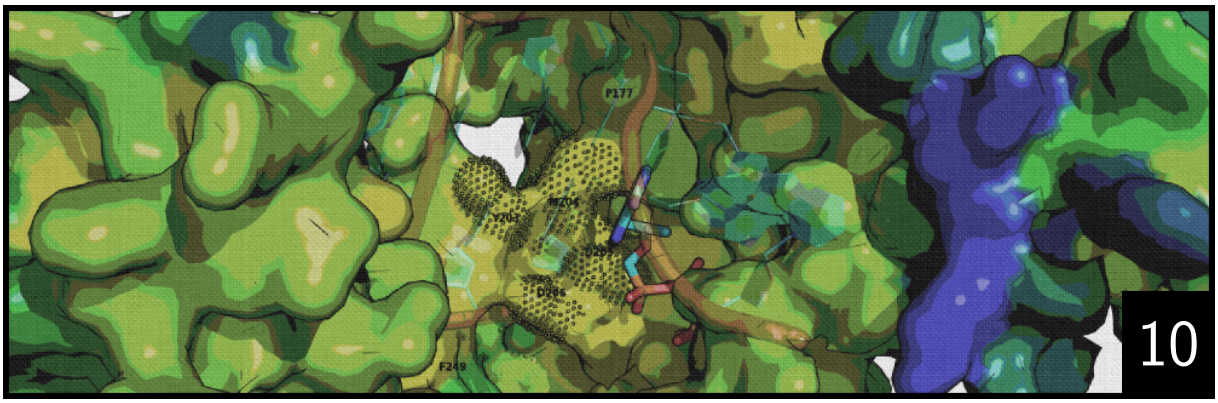
[52] M. Wiese et al. "Low Frequency of Cirrhosis in a Hepatitis C (Genotype 1b) Single-Source Outbreak in Germany: A 20-Year Multicenter Study". In: *Hepatology* 32.1 (2000), pp. 91–96. ISSN: 0270-9139. DOI: http://dx.doi.org/10.1053/jhep.2000.8169. URL: http://www.sciencedirect.com/science/article/pii/S0270913900040635 (cit. on p. 127).

[53] I. T. Williams, J. F. Perz, and B. P. Bell. "Viral Hepatitis Transmission in Ambulatory Health Care Settings". In: *Clinical Infectious Diseases* 38.11 (2004), pp. 1592–1598. DOI: 10.1086/420935 (cit. on p. 127).

PART

V

HOMOLOGY MODELING



Mutations rtP177G and rtF249A in the reverse transcriptase domain of hepatitis B virus polymerase confer resistance to tenofovir

Time flies like an arrow; fruit flies like a banana

GILBERT BURCK

Abstract

Long term antiviral therapy with nucleoside/nucleotide analogs have been routinely used to treat chronic hepatitis B virus (HBV) infection but may lead to the emergence of drug-resistant viral mutants. However, the HBV resistance mutations for tenofovir (TDF) remain controversial. It is speculated that the genetic barrier for TDF resistance may be high for HBV. We asked whether selected amino acid substitutions in HBV polymerase may reduce susceptibility to TDF. A series of amino acids in HBV polymerase were selected based on bioinformatics analysis for mutagenesis. The replication competence and susceptibility to TDF of the mutated HBV clones were determined both in vitro and in vivo. nineteen mutations in HBV polymerase were included and impaired the replication competence of HBV genome in different degrees. The mutations at rtL77F (sS69C), rtF88L (sF80Y), and rtP177G (sR169G) also significantly affected HBsAg expression. The HBV mutants with rtP177G and rtF249A were found to have reduced susceptibility to TDF in vitro with a resistance index of 2.53 and 12.16, respectively. The testing in in vivo model based on the hydrodynamic injection revealed the antiviral effect of

TDF against wild type and mutated HBV genomes and confirmed the reduced the susceptibility of mutant HBV to TDF.

This chapter is based on the following publication:

Bo Qin, Bettina Budeus, Liang Cao, Chunchen Wua, Yun Wang, Xiaoyong Zhang, Simon Rayner, Daniel Hoffmann, Mengji Lu, Xinwen Chen (2013). **The amino acid substitutions rtP177G and rtF249A in the reverse transcriptase domain of hepatitis B virus polymerase reduce the susceptibility to tenofovir.** Antiviral Research 97.

http://www.sciencedirect.com/science/article/pii/S0166354212002872

Antiviral Research 97 (2013) 93–100

Contents lists available at SciVerse ScienceDirect

Antiviral Research

journal homepage: www.elsevier.com/locate/antiviral

The amino acid substitutions rtP177G and rtF249A in the reverse transcriptase domain of hepatitis B virus polymerase reduce the susceptibility to tenofovir

Bo Qin [a,b,c], Bettina Budeus [d], Liang Cao [a], Chunchen Wu [a,b], Yun Wang [a], Xiaoyong Zhang [b], Simon Rayner [a], Daniel Hoffmann [d], Mengji Lu [a,b,*], Xinwen Chen [a,*]

[a] State Key Lab of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China
[b] Institute of Virology, University Hospital of Essen, University of Duisburg-Essen, Essen, Germany
[c] Shaoxing Center for Disease Control and Prevention, Shaoxing 312071, Zhejiang Province, China
[d] Department of Bioinformatics, University of Duisburg-Essen, Essen, Germany

ARTICLE INFO

Article history:
Received 22 August 2012
Revised 4 December 2012
Accepted 6 December 2012
Available online 20 December 2012

Keywords:
Hepatitis B virus
Drug resistance
Tenofovir
Nucleoside/nucleotide analogs
Replication competence
Hydrodynamic injection mouse model

ABSTRACT

Long term antiviral therapy with nucleoside/nucleotide analogs have been routinely used to treat chronic hepatitis B virus (HBV) infection but may lead to the emergence of drug-resistant viral mutants. However, the HBV resistance mutations for tenofovir (TDF) remain controversial. It is speculated that the genetic barrier for TDF resistance may be high for HBV. We asked whether selected amino acid substitutions in HBV polymerase may reduce susceptibility to TDF. A series of amino acids in HBV polymerase were selected based on bioinformatics analysis for mutagenesis. The replication competence and susceptibility to TDF of the mutated HBV clones were determined both *in vitro* and *in vivo*. nineteen mutations in HBV polymerase were included and impaired the replication competence of HBV genome in different degrees. The mutations at rtL77F (sS69C), rtF88L (sF80Y), and rtP177G (sR169G) also significantly affected HBsAg expression. The HBV mutants with rtP177G and rtF249A were found to have reduced susceptibility to TDF *in vitro* with a resistance index of 2.53 and 12.16, respectively. The testing in *in vivo* model based on the hydrodynamic injection revealed the antiviral effect of TDF against wild type and mutated HBV genomes and confirmed the reduced the susceptibility of mutant HBV to TDF.

1. Introduction

Patients with chronic hepatitis B (CHB) have been successfully treated with interferon and nucleoside/nucleotide analogs including lamivudine (LMV), adefovir dipivoxil (ADV), entecavir (ETV), telbivudine (LdT), and tenofovir (TDF) (Zoulim and Locarnini, 2009). The nucleoside/nucleotide analogs inhibit reverse transcription of the hepatitis B virus (HBV) polymerase, but do not directly interfere with the formation of covalently closed circular DNA (cccDNA). Therefore, long term antiviral therapy is necessary, which usually induces the emergence and selection of drug-resistant mutations in the viral polymerase (Zoulim and Locarnini, 2009). Since the HBV polymerase gene is overlapped by the surface protein gene, the mutations in HBV polymerase may also result in amino acid (aa) substitutions in HBsAg that potentially result in immune escape or modification of viral fitness (Torresi, 2002; Villet et al., 2009).

Several HBV polymerase gene mutations have been reported to account for drug resistance. In particular, mutations at rtM204I or rtM204V in the YMDD motif within the reverse transcript (RT) domain of the HBV polymerase lead to LMV and LdT resistance (Brunelle et al., 2005). These mutations are usually accompanied by compensatory mutations of rtL180M and/or rtV173L which restore HBV replication capacity (Pallier et al., 2006). In addition to the substitutions at position rt204, a combination of mutations in the B, C, or D domain of HBV-RT could result in resistance to ETV (Zoulim and Locarnini, 2009).

TDF is widely used to treat HBV patients in the US and Europe. To date, only a few aa substitutions like rtA194T, rtV214A, and rtQ215S associated with TDF resistance have been reported and still need to be further confirmed (Liu et al., 2009; Zoulim and Locarnini, 2009). It has been shown rtA181V+rtN236T double mutants are resistant to TDF *in vitro*; however, clinical data suggested patients with rtA181 or rtN236T remain susceptible to TDF (Qi et al., 2007). On the other hand, TDF, as a first-line antiretroviral drug for human immunodeficiency virus (HIV) since 2001 (Soler-Palacin et al., 2011), does induce resistant mutations in HIV RT with the K65R mutation of HIV RT reducing TDF susceptibility about 2-fold

* Corresponding authors. Address: Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China. Tel./fax: +86 027 87199106.
E-mail addresses: mengji.lu@uni-due.de (M. Lu), chenxw@wh.iov.cn (X. Chen).

0166-3542/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.antiviral.2012.12.007

(Whitcomb et al., 2003). Substitution K70E within HIV-RT was selected in HIV patients with virological failure under TDF therapy (Delaugerre et al., 2008). Co-mutations M41L, L210W, and T215Y reduce TDF susceptibility about 4-fold. However, one or two of them retain partial TDF susceptibility (Miller et al., 2004). The crystal structures of two related complexes of HIV-1 RT with template primer and TDF were determined (Tuske et al., 2004). The availability of HIV RT structural information makes it possible to determine the relative position of the aa residues with respect to the active center of the enzyme as well as the distance to the substrate if TDF is used for the modeling.

In the present study, we used another strategy to assess the potential of HBV to develop TDF resistance. Given the homology between HIV RT and HBV polymerases (Bartholomeusz et al., 2004), we aligned their primary aa sequences and identified corresponding positions of aa residues in HBV polymerase and HIV RT. Based on this alignment, the distances of a given aa residue in HBV polymerase to the TDF substrate could be estimated by comparison with HIV RT. Therefore we selected a number of aa residues according to their distances to TDF, predicted on the basis of a bioinformatics approach. We assumed that an aa substitution at a given position, especially with a change to an aa residue with a large or complex side chain, may influence the binding of the natural substrates and nucleoside/nucleotide analogs and leads to changed replication competence and nucleoside/nucleotide analogs resistance of mutant HBV genomes in some cases. To test this hypothesis, a series of replication-competent HBV constructs harboring different mutations were then constructed according the predictions. The replication capacity and resistance phenotype were analyzed both *in vitro* and *in vivo*. The results demonstrated that two mutations, rtP177G and rtF249A, significantly reduced HBV susceptibility to TDF and could be potentially used as a reference for TDF-resistance. Interestingly, the positions rtP177 and rtF249 may be not in direct contact with TDF molecule but make contacts to the template and primer oligonucleotides, respectively. Our approach could generally contribute to the understanding of HBV drug resistance.

2. Materials and methods

2.1. Plasmid constructs

For construction of HBV mutant plasmids with aa substitutions, the plasmid pHBV1.3 containing a replication-competent wild-type (WT) HBV 1.3-fold over-length genotype A genome (GenBank accession No. U95551, ayw) was used as a backbone (Gan et al., 1987; Lei et al., 2006). Mutations were introduced into the HBV polymerase gene by PCR-based mutagenesis using the primer pairs Primer-F, Primer-R and primers with the specific mutations listed in Table S1. Plasmids pHBV1.3-rtXs (X stands for the mutation, for example "P177G") were produced and the aa substitutions are presented in Fig. 1B. For *in vivo* assay, pAAV-HBV-1.3 and pAAV-HBV1.3-rtXs were constructed based on plasmid pHBV1.3 and pHBV1.3-rtXs respectively. The plasmids pHBV1.3/pHBV1.3-rtXs and pAAV were digested by Sac I, then end-filled with T4 and Klenow DNA polymerase, respectively. The recovered products were digested by HindIII and ligated by T4 ligase to generate pAAV-HBV-1.3 or pAAV-HBV1.3-rtXs.

2.2. Cells and mice

Hepatoma cell line Huh7 cells were cultured in Dulbecco's modified Eagle medium (DMEM; Invitrogen) supplemented with 10% fetal bovine serum (FBS; Gibco), 2 mM/L of glutamine, 100 IU/mL of penicillin and 100 IU/mL of streptomycin at 37 °C in a 5% CO_2 atmosphere. Female BALB/c mice (6–8 weeks of age; H-2^d) and female C57BL/6 (6–8 weeks of age; H-2^b) were held under specific pathogen-free conditions in the Central Animal Laboratory of Wuhan Institute of Virology, Chinese Academy of Sciences and were handled following the guidelines of the animal ethical standards (Meng et al., 2008).

2.3. Nucleoside analogs

Lamivudine (LMV) (Glaxo Smith Kline), adefovir dipivoxil (ADV) (Gilead Sciences), entecavir (ETV) (Bristol-Myers Squibb Co), telbivudine (LdT) (Novartis), and tenofovir (TDF) (Gilead Sciences) were dissolved in appropriate solution according to the manufactures' instructions and used at the indicated concentrations.

2.4. Enzyme-linked immunosorbent assay (ELISA)

HBsAg and HBeAg in mouse sera and culture supernatants of Huh7 cells transfected with the plasmids pHBV1.3 and pHBV1.3-rtXs were detected by using a commercial routine diagnostic assay for HBsAg and HBeAg (Kehua, Shanghai) according to the manufacturer's instructions.

2.5. Western blot analysis

Huh7 cells transfected with indicated plasmids were harvested at 72 hour post transfection (hpt). The protein concentrations were determined with a Bio-Rad protein assay kit (Bio-Rad). Total cell lysates (50 μg) were submitted to Western-blot assay by probing with anti-HBcAg antibody (Dako) and anti-β-actin (Beyotime), with appropriate secondary antibody. Densitometry assays were performed simultaneously.

2.6. Immunohistochemical (IHC) staining

Liver tissue was collected from mice sacrificed at the indicated time points. Intrahepatic HBcAg was detected by IHC staining of formalin-fixed paraffin-embedded liver tissue sections using rabbit anti-HBc antibodies (DAKO) with appropriate HRP-conjugated secondary antibody, and visualized by the Envision System (Huang et al., 2006). The liver sections were also stained with hematoxylin and eosin.

2.7. Detection of encapsidated HBV DNA

Huh7 cells (1×10^6) were transfected with 2 μg of the plasmids pHBV1.3 and pHBV1.3-rtXs by using lipofectamine 2000 (Invitrogen) and cultured in the presence of the different nucleoside/nucleotide analogs at the indicated concentrations. To control the transfection efficiency, a SEAP expression vector was cotransfected to monitor the transfection efficiency. Encapsidated HBV replicative intermediates were purified and subjected to Southern blot analysis as described previously (Meng et al., 2008; Qiu et al., 2011).

HBV DNA was quantified by real-time PCR using the primers (RC-sense and RC-antisense, Table S1) specially designed for the detection of HBV relaxed circular (rc) genomes in Sybr green reaction mix (Roche). Plasmid pAAV-HBV1.3 was used as a standard. All samples were analyzed in triplicate.

2.8. HBV challenge by hydrodynamic injection (HI) in mice

C57BL/6 mice were challenged by hydrodynamic injection of replication-competent pAAV-HBV1.3 and pAAV-HBV-rtXs respectively, as described previously (Huang et al., 2006). Ten micrograms of each plasmid in a volume of 0.9% NaCl solution

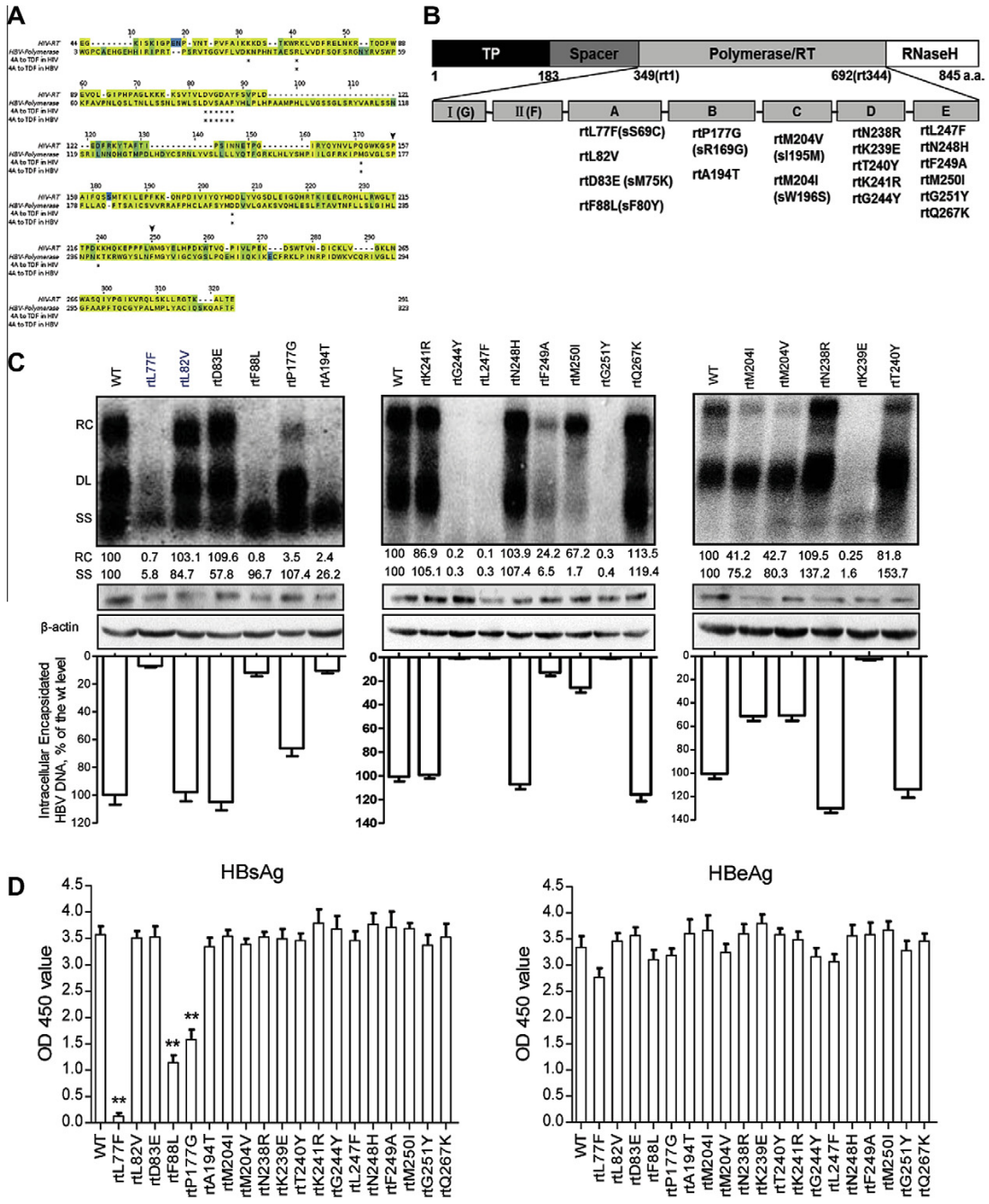


Fig. 1. (A) Alignment of amino acid sequences of HBV-RT genotypes A–H and HIV-RT. Coloring indicates sequence conservation across all nine sequences (see Section 2). For clarity, only the HBV-RT wild type sequence (V01460.1) is shown together with the HIV-RT sequence. The two sites P177 and F249 where TDF resistance mutations have been found in HBV-RT are marked by arrowheads, and asterisks mark residues with distances of up to 4 Å from any heavy atom of TDF in the X-ray structure of HIV-RT or the corresponding homology model of HBV-RT. (B) Structure of HBV reverse transcriptase and location of substitution sites. HBV P protein is divided into four regions, terminal protein, spacer, reverse transcriptase and RNaseH. Reverse transcriptase is subdivided into seven regions, G, F, A, B, C, D, and E. All the amino acid substitutions examined in this present study are located in A–E. (C) Detection of HBV replication intermediates by Southern blot. HBV replication intermediates relaxed circular (RC), double stranded linear (DL) and single stranded (SS) HBV DNAs are indicated (upper panel). The HBcAg was detected using Western blot with mouse polyclonal antibody (Dako, Carpinteria, CA). Beta-actin was used as a loading control (Middle panel). The intensity of RC and SS form of HBV replication intermediates were analyzed and compared with the wild type (set as 100, the numbers below the Southern blot). Intracellular encapsidated HBV DNA levels of each construct were compared with that of wild type genome (set as 100%, lower panel). (D) The expression of HBsAg and HBeAg were measured by commercial ELISAs (Kehua, Shanghai). Each value is the mean of three independent experiments. The error bars represent the standard deviation. Statistically significant differences between the groups are displayed as $^{*}(p < 0.05)$ or $^{**}(p < 0.01)$.

equivalent to 0.08 mL/g of the mouse body weight were injected into the tail veins of mice within 8 s. After 1 day post hydrodynamic injection (dphi), every mouse of the treatment group was treated with 500 μg TDF daily through intraperitoneal injection.

2.9. Multiple alignment of HIV RT and HBV polymerase sequences

Sequences were taken from GenBank, and translated from nucleotide to amino acid sequences using UGENE. For pairwise sequence alignments, the EMBOSS program "water" was used. Multiple sequence alignments were performed with T-COFFEE (Taly et al., 2011) and MUSCLE (Edgar, 2004). In all alignment procedures, default settings were employed. A homology model of HBV-RT, the reverse transcriptase domain of HBV polymerase, based on the complex of HIV-RT with TDF (PDB entry 1T03:B), and the mentioned multiple sequence alignment was prepared with modeller (Eswar et al., 2008) and visualized with pymol (Bramucci et al., 2012). Figures showing alignment and homology model, respectively, were consistently color-coded according to the degree of sequence conservation in each column of the alignment from yellow (=100% identity, i.e. complete conservation in column) to blue (=13% identity, i.e. all residues in column different). This sequence conservation was computed with R-package bio3d (Grant et al., 2007).

2.10. Statistical analysis

The statistical analysis was carried out using GraphPad (GraphPad Software). Differences in multiple comparisons were determined for statistical significance using the Student's *t*-test. $p < 0.05$ was considered as statistically significant. Results were presented as Means ± S.D.

3. Results

3.1. The estimation of the distances of the aa residues of HBV polymerase to TDF bound to the active center

So far, TDF-resistance-associated aa substitutions of HBV-RT remain elusive and controversial. Therefore, we attempted to answer the question whether specific aa substitutions may lead to a reduced susceptibility of HBV to TDF.

Sequences of HBV-RT of genotypes A through H and HIV-RT were submitted to pairwise and multiple sequence alignments (Fig. 1A). The alignment of the RT domains of both enzymes as shown in Fig. 1A was essentially the same for all alignment algorithms used. In particular the catalytic YMDD motif and several positions known to interact with TDF in HIV-RT were conserved between both enzymes. Based on the sequence alignment of HBV polymerase (GenBank accession No. CAA48354.1) and HIV-RT (HIV-1 polymerase protein gene, GenBank accession No. HQ718313), 19 aa residues in the HBV polymerase were selected to test their potential role in the development of TDF-resistance. These aa residues are distributed in the RT domains A, B, C, D, and E of the HBV polymerase (Fig. 1B). However, when matched to their counterparts in the HIV RT, they may have different distances to the bound substrate TDF according to the structural information of HIV RT (Fig. S1).

For the introduction of aa substitutions, three criteria were considered: (1) the side chain of the mutated aa residues should be significantly different to the wild type if possible; (2) the introduced aa substitution in HBV polymerase should have no or little influence in the coding sequence of HBsAg; (3) The number of nucleotide need to be mutated should be as less as possible. A series of point mutations on HBV replication-competent pHBV1.3 were designed and constructed using fusion PCR (Fig. 1B). Among these substitutions, rtL77F, rtD83E, rtF88L, rtP177G, rtM204I and rtM204I resulted in the aa substitutions sS69C, sM75K, sF80Y, sR169G, sI195M and sW196S in HBsAg, respectively.

3.2. Replication of pHBV1.3-rtXs in Huh7 cells

To determine the replication competence of HBV genomes with aa substitutions in HBV-RT, pHBV1.3-rtXs were transfected into Huh7 cells and the intracellular encapsidated viral genomes were extracted and subjected to Southern-blot analysis (Fig. 1C, upper panel).

For pHBV1.3-rtL77F, -rtK239E, -rtG244Y, -rtL247F, and -rtG251Y, no HBV replication intermediates were detected in Southern blot (Fig. 1C), indicating that the substitution mutants resulted in a complete loss of replication competence. These aa residues may be functionally essential for both RT and DNA polymerase activities of HBV polymerase or the aa substitutions prevented binding of the substrates, considering the large chains of the aa substitutions. For pHBV1.3-rtF88L, rtP177G, rtA194T, -rtM204I and -rtM204V, single-stranded (SS) DNA bands were synthesized, whereas the relaxed circular (RC) DNA bands were relatively weak or invisible. Compared with pHBV1.3, the levels of intracellular encapsidated HBV DNA were about 11.9%, 66.3%, 10.4%, 51.2%, and 50.8% for pHBV-rtF88L, -rtP177G, -rtA194T, -rtM204I, and -rtM204V, respectively, indicating a reduced replication competence. In contrast, pHBV-rtF249A and -rtM250I only produced RC DNA bands and their levels were 12.9% and 25.8% as compared to that of pHBV1.3, respectively (Fig. 1C, lower panel), which suggest that these two aa substitutions may impair the RT activity of HBV polymerase. The remaining 7 aa substitutions, rtL82V, rtD83E, rtN238R, rtT240Y, rtK241R, rtN248H, and rtQ267K, had no obvious effect on HBV replication capacity.

In all cases, the HBcAg expression levels were comparable (Fig. 1C, middle panel), indicating that HBcAg expression is not associated with HBV replication *in vitro*. Thus, the failure of the detection of HBV replication intermediates was not related to HBcAg expression.

The levels of HBsAg and HBeAg in culture supernatants of transfected Huh7 cells were measured by ELISA. HBsAg was absent from the supernatant of pHBV1.3-rtL77F transfected cells, while pHBV-rtF88L and pHBV-rtP177G produced significantly lower amounts of HBsAg compared to pHBV1.3 (Fig. 1D). The aa substitutions rtL77F, rtF88L, and rtP177G in HBV polymerase lead to the corresponding aa substitutions sS69C, sF80Y and sR169G in HBsAg which significantly affected HBsAg expression and/or secretion. The remaining mutated HBV genomes expressed similar levels of HBsAg compared with the wild type, although some of them exhibited changes in the HBsAg sequence. Nevertheless, in all cases, HBeAg expression levels were comparable, indicating that the plasmid transfection efficiency was the same.

3.3. TDF-resistance assay in vitro

To analyze the influence of the aa substitutions in HBV polymerase described above on the susceptibility to nucleoside/nucleotide analogs, we first determined the concentration range of nucleoside/nucleotide analogs that inhibited the replication of wild type HBV in Huh7. Huh7 cells were transfected with pHBV1.3 and treated with different concentration of the nucleoside/nucleotide analogs 6 h later. Encapsidated HBV DNA was then extracted at 96 h and detected by Southern blot. All tested drugs LMV, ADV, LdT, ETV, and TDF inhibited HBV DNA replication and the half maximal effective concentration (EC_{50}) of these drugs were 1.15, 1.38, 11.56, 0.79, and 0.19 μM, respectively (Fig. S2A–E). As reported previously, all nucleoside/nucleotide analogs did not affect HBcAg, HBsAg and HBeAg expression (Figs. S2 and S3).

Subsequently, the sensitivity to TDF of the replication competent HBV genomes with aa substitutions was tested in Huh7 cells. Southern blot and subsequent densitometry analysis demonstrated that pHBV-rtP177G and -rtF249A displayed reduced sensitivity to TDF, as manifested by sustained viral replication upon TDF treatment (Fig. 2A). This phenotype is in contrast to pHBV1.3, for which viral DNA production decreased sharply as TDF concentration increased. Other mutants, including pHBV-rtL82V, -rtD83E, -rtN238R, -rtT240Y, -rtK241R, -rtN248H, -rtM250I, and -rtQ267K exhibited a similar TDF sensitivity to pHBV1.3 (Fig. S4). HBV RC DNA was quantified by real-time PCR and the results showed that the antiviral effect of TDF to pHBV-rtP177G and -rtF249A were significantly compromised compared to pHBV1.3 (Fig. 2B), as manifested by the observation that EC_{50} of TDF to pHBV-rtP177G and -rtF249A were 0.48 and 2.31 μM, respectively, which were significantly higher than that for pHBV1.3 (0.19 μM) (Fig. 2C). The resistance indexes of the HBV genomes with rtP177G and rtF249A substitutions to TDF were 2.53 and 12.16, respectively.

In addition, pHBV-rtP177G and -rtF249A remained sensitive to LMV, ADV, ETV, and LdT (Fig. 2D).

3.4. rtP177G and -rtF249A compromise antivirus effect of TDF in mice

The aa substitutions rtP177G and -rtF249A in HBV polymerase conferred a certain degree of resistance to TDF *in vitro*. However, it is not clear whether such HBV mutants are able to replicate *in vivo* and show a different susceptibility to TDF compared with the wild type HBV genome. Then on, we explored the mouse model based on HI to characterize HBV replication upon TDF treatment. C57BL/6 mice were respectively challenged with pAAV-rtP177G, -rtF249A, -rtA194T, or -HBV1.3 by HI and treated with 500 μg TDF or PBS daily. HBV DNA in the serum and liver from mice was measured by real-time PCR targeted to RC at the indicated time points.

HBV DNA and HBsAg became detectable in mouse sera after HI with pAAV-HBV1.3 (Fig. 3), comparable with the previous published data (Yin et al., 2011). Serum HBV DNA reached the peak

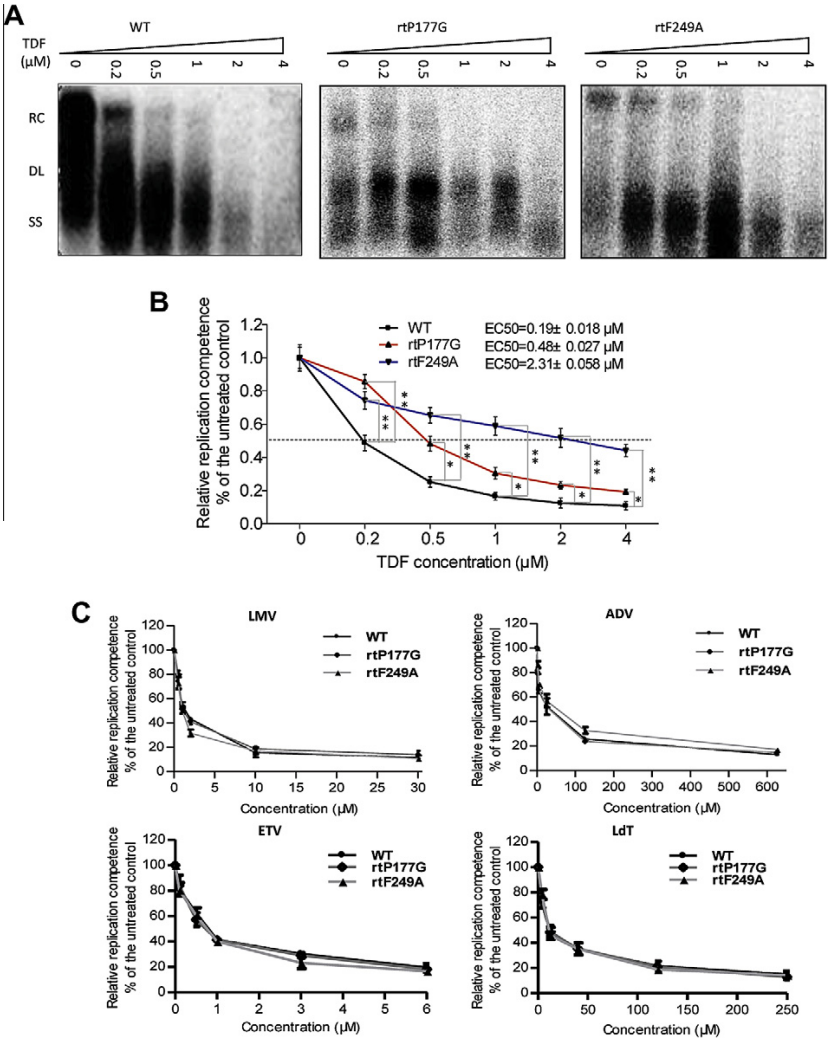


Fig. 2. TDF-resistance assay *in vitro*. Huh7 cells were transfected with pHBV1.3, -rtP177G, and -rtF249A, respectively, and then treated with the different concentrations of TDF. Encapsidated HBV DNAs were purified from intracellular core particles for Southern-blot and real-time PCR analysis. (A) Southern blot for detection of HBV replication intermediates: wt (left), rtP177G (middle), rtF249A (right). (B) Analysis of HBV RC DNA by specific real-time PCR. The relative replication competence pHBV1.3, -rtP177G, and -rtF249A was given as the percentage of the RC DNA compared with the untreated control. The EC50s are indicated. (C) Real-time PCR to analyze the susceptibility of pHBV1.3, rtP177G, and rtF249A to LMV, ADV, ETV and LdT. Each value is the mean of at least 3 independent experiments. The error bars represent the standard deviation. Statistically significant differences between the groups are displayed as $^{*}(p < 0.05)$ or $^{**}(p < 0.01)$.

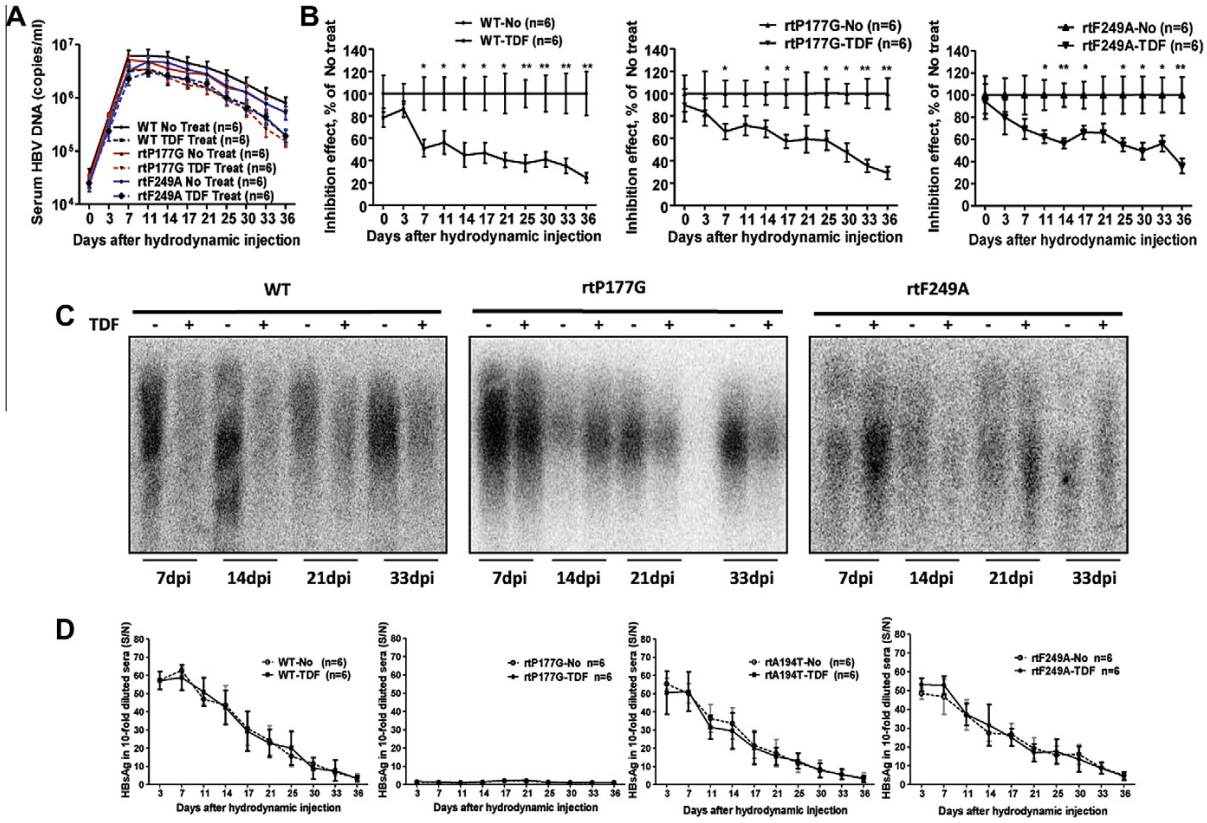


Fig. 3. TDF-resistance assay *in vivo*. C57BL/6 (H-2b) mice were challenged with pAAV-HBV1.3, -rtP177G, and -rtF249A by tail vein HI. After 1 d, mice were treated with 500 μg TDF per day. At the indicated time points, HBV DNA and proteins in the sera and liver were measured by qRT-PCR targeted to HBV RC DNA, Southern blot, ELISA, and IHC, respectively. (A) The analysis of HBV DNA in mouse sera of HBV1.3, -rtP177G, and -rtF249A treated with TDF or vesicle by real time RT-PCR. (B) The antiviral effect of TDF *in vivo* was shown as the inhibition rate of serum HBV DNA in TDF-treated mice compared with that of untreated control mice. The average HBV DNA copy numbers of the untreated control group at each indicated time points were set as 100%. (C) To detect the HBV DNA in the liver, mice from each group were killed at the indicated time points. Total DNA was isolated from the liver tissue and subjected to Southern blot. (D) Kinetics of HBsAg expression. The HBsAg level was determined by a commercial ELSA and given as optical density value (OD 450). S/N means samples/Negative.

level at 7 dphi, decreased gradually thereafter, but remained positive for the complete observation period of 5 weeks (Fig. 3A). Daily treatment with TDF led to a significant reduction of the HBV DNA levels from 7 dphi on and suppressed the HBV DNA level below 50% of the control on 13 dphi (Fig. 3B). The serum HBsAg titers in mice were determined by ELISA. Results showed that HBsAg in all the mice injected with pAAV-HBV1.3 was detected at 3 dphi, and declined gradually from 7 to 36 dphi (Fig. 3D). HBsAg levels were comparable in mice irrespective of TDF treatment. We further detected intrahepatic HBcAg by IHC staining with specific antibody (Fig. S5). HBcAg was strongly expressed in hepatic cells in pAAV-HBV-1.3 injected mice for 3 weeks and persisted weakly thereafter, at least up to 32 dphi. TDF treatment significantly decreased the hepatic HBcAg expression compared with the corresponding untreated mice, indicating TDF inhibits the formation of HBV nucleocapsids. Taken together, the antiviral activity of TDF could be demonstrated in the mouse model based on HI.

Similarly, serum HBV DNA reached the highest level at 7 dphi in pAAV-HBV1.3-rtP177G injected mice, in contrast to rtF249A that displayed a delayed viral DNA maximum level at 11 dphi (Fig. 3A). The result of Southern blot with the liver tissue also confirmed pAAV-HBV1.3 possessed a higher replication potential than pAAV-rtP177G and -rtF249A *in vivo* (Fig. 3C), in accordance with *in vitro* results (Fig. 1C). Despite the reduced replication competence of HBV genomes with the aa substitution in HBV polymerase, high serum HBV DNA levels could be established and maintained in mice after HI with both mutant HBV genomes for the experimental period (Fig. 3A). In pAAV-rtP177G and -rtF249A injected mice, TDF suppression of HBV DNA replication appeared to be less effective. HBV DNA levels in TDF treated mice were lower than that in control mice however and remained over the 50% marker for a prolonged time over 20 d (Fig. 3A), consistent with the *in vitro* data obtained in the previous experiments.

In mice receiving injections with pAAV-HBV1.3-rtF249A, HBsAg levels were comparable in mice irrespective of TDF treatment, comparable with previous findings. In contrast, HBsAg from pAAV-rtP177G injected mice, either TDF treated or not, was always undetectable (Fig. 3D), and was consistent with the *in vitro* assay (Fig. 1D). HBcAg expression was not detected in liver sections of mice that received HI with pAAVHBV1.3-rtA194T, -rtP177G, and -rtF249A (data not shown), most likely due to the decreased replication competence of mutated HBV genomes in mice.

3.5. The aa substitution rtP177G and rtF249A in the structure model of HBV RT

A homology model of the HBV-RT structure was built based on the multiple sequence alignment of HIV-RT with eight HBV-RT sequences representing HBV genotypes A through H. Fig. 4 shows that the sequence in the putative TDF-binding region of HBV-RT

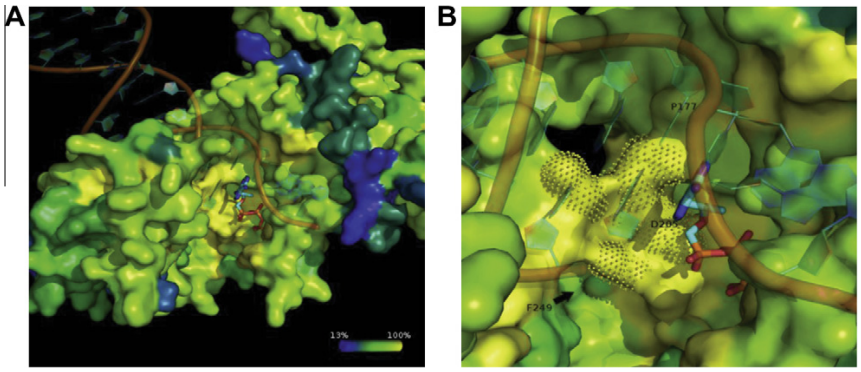


Fig. 4. A homology model of HBV-RT based on HIV-RT. Coloring according to the degree of conservation as described in Section 2 from yellow (complete conservation) to blue (high diversity). (A) The overview showing HBV-RT with bound primer and template oligonucleotides, and TDF molecule (sticks). (B) Close-up around TDF with the catalytic YMDD motif indicated by dots above molecular surface. The central catalytic D205 and the two resistance relevant positions P177 and F249 are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is well-conserved and similar to that of HIV-RT. In the model, the two positions rtP177 and rtF249 are close to the bound TDF molecule though not in direct contact with it (Fig. 4A). However, rtP177 and rtF249 apparently do make contacts to the template and primer oligonucleotides, respectively (Fig. 4B).

4. Discussion

Viral mutants with RT mutations resistant to nucleoside/nucleotide analogs may be selected during long-term antiviral therapy. However, to date, TDF-associated resistance mutations have seldom been found in patients (Zoulim and Locarnini, 2009). For the first time, we combined the bioinformatics analysis, phenotypic assay for drug resistance, and *in vivo* analysis based on HI. Such an approach may be generally useful to understand the structure–function relationship of aa substitutions involved in drug resistance and their biological relevance. Interestingly, the mutated HBV genomes with aa substitutions rtP177G and rtF249A in HBV RT showed reduced replication capacity but enhanced resistance to TDF in both *in vitro* and *in vivo* analysis. The alignment of HBV RT with HIV RT and bioinformatics analysis suggested that rtP177 and rtF249 may not in directly interact with TDF molecule but with the template and primer oligonucleotides. Thus, the accurate molecular mechanisms leading to TDF resistance by these aa substitutions remain to be determined. The important implication of our results is that the aa substitutions in HBV RT may reduce the susceptibility to TDF. It needs to be investigated whether such or similar aa substitutions may play a role in long term TDF treatment.

HBV polymerase is a multi-functional protein with the activity of RNA- and DNA-dependent DNA polymerization. The conserved RT domain catalyzes reverse transcription using the TP domain as a protein primer to initiate the process, and is also involved in the packaging of pgRNA into viral nucleocapsids (Beck and Nassal, 2007; Lee, 1997). Here, the aa substitutions rtL77F, rtK239E, rtG244Y, rtL247F, and rtG251Y may strongly disturb the polymerase function and led to the complete loss of replication capability of HBV mutants. The YMDD catalytic motif is shared by both HBV RT and HIV RT (Zoulim, 2004) and the substitutions rtM204V and rtM204I result in HBV resistance to LMV and in the reduction of HBV replication competence, due to the lower affinities of the YMDD-mutant polymerases for the natural dNTP substrates (Gaillard et al., 2002). Southern blot analysis indicated that beside rtM204V and rtM204I, HBV with rtF88L, rtP177G, and rtA194T substitutions also produced significantly less RC forms of HBV replication intermediates. The effect of the substitution rtA194T on HBV replication competence has been documented in an earlier publication, whereas the rtF88L and rtP177G substitutions have not been found in patients so far. Notably, substitutions rtF249A and rtM250I lead to no or less SS DNA, but still produced RC DNA, likely due to a change of the relative activities of the RT and DNA-dependent DNA polymerase. Thus, the different aa substitutions affected HBV replication through very different molecular mechanisms. Further experiments are needed to reveal the mechanisms associated with the rtF88L, rtP177G, and rtA194T substitutions. The aa residues that are not located at the catalytic pocket or substitutions at sites which do not impact HBV polymerase conformation may maintain similar replication capacity as the wild type; these aa substitutions comprise rtL82V, rtD83E, rtN238R, rtT240Y, rtK241R, rtN248H, and rtQ267K.

It has been reported that the rtA194T mutation is associated with TDF resistance found in two HBV-HIV-coinfected patients (Sheldon et al., 2005), and this was confirmed in cell lines (Amini-Bavil-Olyaee et al., 2009), although a subsequent report failed to confirm this finding (Delaney et al., 2006). Here, we also failed to detect sufficient replication capacity of HBV in the presence of the rtA194T substitution both *in vitro* and *vivo*, so the susceptibility to TDF cannot be evaluated. Drug-resistance mutations often result in reduction of replication capacity and adaptive mutations occur in an attempt to restore the replication capacity, like the compensatory mutations rtL80I and/or rtV173L and/or rtL180M for rtM204V/I. This could be the reason for the low replication capacity of the HBV genome with mutation rtA194T in our study. Consistently, the aa substitutions rtP177G and rtF249A impaired the replication competence of HBV *in vitro*. In the background of a potential adaptive co-substitution mutant HBV genomes with rtP177G and rtF249A may recover their replication capacity.

It is desirable to establish *in vivo* models with prolonged persistence of drug-resistant HBV genomes, mimicking chronic HBV infection in patients. The HBV mouse model based on HI has been used to study HBV replication (Giladi et al., 2003). Here, we demonstrated the usefulness of the HI mouse model to study drug resistant HBV mutants. The HBV genomes with rtP177G and rtF249A substitutions, which are resistant to TDF *in vitro*, also have reduced sensibility to TDF in C57BL/6 mice. Thus, this model could be refined and better standardized in the future.

Acknowledgments

We are grateful to Xuefang An, Yuan Zhou, and Xue Hu for excellent technical support. This work was supported in part by the National Basic Research Priorities Program of China (2011CB

106303) and Grants from the Deutsche Forschungsgemeinschaft (DFG Transregio TRR60 and GRK1045/2).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.antiviral.2012.12.007.

References

Amini-Bavil-Olyaee, S., Herbers, U., Sheldon, J., Luedde, T., Trautwein, C., Tacke, F., 2009. The rtA194T polymerase mutation impacts viral replication and susceptibility to tenofovir in hepatitis B e antigen-positive and hepatitis B e antigen-negative hepatitis B virus strains. Hepatology 49, 1158–1165.

Bartholomeusz, A., Tehan, B.G., Chalmers, D.K., 2004. Comparisons of the HBV and HIV polymerase, and antiviral resistance mutations. Antivir. Ther. 9, 149–160.

Beck, J., Nassal, M., 2007. Hepatitis B virus replication. World J. Gastroenterol. 13, 48–64.

Bramucci, E., Paiardini, A., Bossa, F., Pascarella, S., 2012. PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL. BMC Bioinformatics 13 (Suppl. 4), S2.

Brunelle, M.N., Jacquard, A.C., Pichoud, C., Durantel, D., Carrouee-Durantel, S., Villeneuve, J.P., Trepo, C., Zoulim, F., 2005. Susceptibility to antivirals of a human HBV strain with mutations conferring resistance to both lamivudine and adefovir. Hepatology 41, 1391–1398.

Delaney, W.E., Ray, A.S., Yang, H., Qi, X., Xiong, S., Zhu, Y., Miller, M.D., 2006. Intracellular metabolism and in vitro activity of tenofovir against hepatitis B virus. Antimicrob. Agents Chemother. 50, 2471–2477.

Delaugerre, C., Flandre, P., Marcelin, A.G., Descamps, D., Tamalet, C., Cottalorda, J., Schneider, V., Yerly, S., LeGoff, J., Morand-Joubert, L., Chaix, M.L., Costagliola, D., Calvez, V., 2008. National survey of the prevalence and conditions of selection of HIV-1 reverse transcriptase K70E mutation. J. Med. Virol. 80, 762–765.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Eswar, N., Eramian, D., Webb, B., Shen, M.Y., Sali, A., 2008. Protein structure modeling with MODELLER. Methods Mol. Biol. 426, 145–159.

Gaillard, R.K., Barnard, J., Lopez, V., Hodges, P., Bourne, E., Johnson, L., Allen, M.I., Condreay, P., Miller, W.H., Condreay, L.D., 2002. Kinetic analysis of wild-type and YMDD mutant hepatitis B virus polymerases and effects of deoxyribonucleotide concentrations on polymerase activity. Antimicrob. Agents Chemother. 46, 1005–1013.

Gan, R.B., Chu, M.J., Shen, L.P., Qian, S.W., Li, Z.P., 1987. The complete nucleotide sequence of the cloned DNA of hepatitis B virus subtype adr in pADR-1. Sci. Sin. B 30, 507–521.

Giladi, H., Ketzinel-Gilad, M., Rivkin, L., Felig, Y., Nussbaum, O., Galun, E., 2003. Small interfering RNA inhibits hepatitis B virus replication in mice. Mol. Ther. 8, 769–776.

Grant, B.J., McCammon, J.A., Caves, L.S., Cross, R.A., 2007. Multivariate analysis of conserved sequence-structure relationships in kinesins: coupling of the active site and a tubulin-binding sub-domain. J. Mol. Biol. 368, 1231–1248.

Huang, L.R., Wu, H.L., Chen, P.J., Chen, D.S., 2006. An immunocompetent mouse model for the tolerance of human chronic hepatitis B virus infection. Proc. Nat. Acad. Sci. U.S.A. 103, 17862–17867.

Lee, W.M., 1997. Hepatitis B virus infection. N. Engl. J. Med. 337, 1733–1745.

Lei, Y.C., Hao, Y.H., Zhang, Z.M., Tian, Y.J., Wang, B.J., Yang, Y., Zhao, X.P., Lu, M.J., Gong, F.L., Yang, D.L., 2006. Inhibition of hepatitis B virus replication by APOBEC3G in vitro and in vivo. World J. Gastroenterol. 12, 4492–4497.

Liu, Y., Wang, C.M., Cheng, J., Liang, Z.L., Zhong, Y.W., Ren, X.Q., Xu, Z.H., Zoulim, F., Xu, D.P., 2009. Hepatitis B virus in tenofovir-naive Chinese patients with chronic hepatitis B contains no mutation of rtA194T conferring a reduced tenofovir susceptibility. Chin. Med. J. (Engl.) 122, 1585–1586.

Meng, Z., Xu, Y., Wu, J., Tian, Y., Kemper, T., Bleekmann, B., Roggendorf, M., Yang, D., Lu, M., 2008. Inhibition of hepatitis B virus gene expression and replication by endoribonuclease-prepared siRNA. J. Virol. Methods 150, 27–33.

Miller, M.D., Margot, N., Lu, B., Zhong, L., Chen, S.S., Cheng, A., Wulfsohn, M., 2004. Genotypic and phenotypic predictors of the magnitude of response to tenofovir disoproxil fumarate treatment in antiretroviral-experienced patients. J. Infect. Dis. 189, 837–846.

Pallier, C., Castera, L., Soulier, A., Hezode, C., Nordmann, P., Dhumeaux, D., Pawlotsky, J.M., 2006. Dynamics of hepatitis B virus resistance to lamivudine. J. Virol. 80, 643–653.

Qi, X., Xiong, S., Yang, H., Miller, M., Delaney, W.E., 2007. In vitro susceptibility of adefovir-associated hepatitis B virus polymerase mutations to other antiviral agents. Antivir. Ther. 12, 355–362.

Qiu, J., Qin, B., Rayner, S., Wu, C.C., Pei, R.J., Xu, S., Wang, Y., Chen, X.W., 2011. Novel evidence suggests Hepatitis B virus surface proteins participate in regulation of HBV genome replication. Virol. Sin. 26, 131–138.

Sheldon, J., Camino, N., Rodes, B., Bartholomeusz, A., Kuiper, M., Tacke, F., Nunez, M., Mauss, S., Lutz, T., Klausen, G., Locarnini, S., Soriano, V., 2005. Selection of hepatitis B virus polymerase mutations in HIV-coinfected patients treated with tenofovir. Antivir. Ther. 10, 727–734.

Soler-Palacin, P., Melendo, S., Noguera-Julian, A., Fortuny, C., Navarro, M.L., Mellado, M.J., Garcia, L., Uriona, S., Martin-Nalda, A., Figueras, C., 2011. Prospective study of renal function in HIV-infected pediatric patients receiving tenofovir-containing HAART regimens. AIDS 25, 171–176.

Taly, J.F., Magis, C., Bussotti, G., Chang, J.M., Di Tommaso, P., Erb, I., Espinosa-Carrasco, J., Kemena, C., Notredame, C., 2011. Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. Nat. Protoc. 6, 1669–1682.

Torresi, J., 2002. The virological and clinical significance of mutations in the overlapping envelope and polymerase genes of hepatitis B virus. J. Clin. Virol. 25, 97–106.

Tuske, S., Sarafianos, S.G., Clark Jr., A.D., Ding, J., Naeger, L.K., White, K.L., Miller, M.D., Gibbs, C.S., Boyer, P.L., Clark, P., Wang, G., Gaffney, B.L., Jones, R.A., Jerina, D.M., Hughes, S.H., Arnold, E., 2004. Structures of HIV-1 RT-DNA complexes before and after incorporation of the anti-AIDS drug tenofovir. Nat. Struct. Mol. Biol. 11, 469–474.

Villet, S., Billioud, G., Pichoud, C., Lucifora, J., Hantz, O., Sureau, C., Deny, P., Zoulim, F., 2009. In vitro characterization of viral fitness of therapy-resistant hepatitis B variants. Gastroenterology 136 (168–176), e162.

Whitcomb, J.M., Parkin, N.T., Chappey, C., Hellmann, N.S., Petropoulos, C.J., 2003. Broad nucleoside reverse-transcriptase inhibitor cross-resistance in human immunodeficiency virus type 1 clinical isolates. J. Infect. Dis. 188, 992–1000.

Yin, Y., Wu, C., Song, J., Wang, J., Zhang, E., Liu, H., Yang, D., Chen, X., Lu, M., Xu, Y., 2011. DNA immunization with fusion of CTLA-4 to hepatitis B virus (HBV) core protein enhanced Th2 type responses and cleared HBV with an accelerated kinetic. PLoS One 6, e22524.

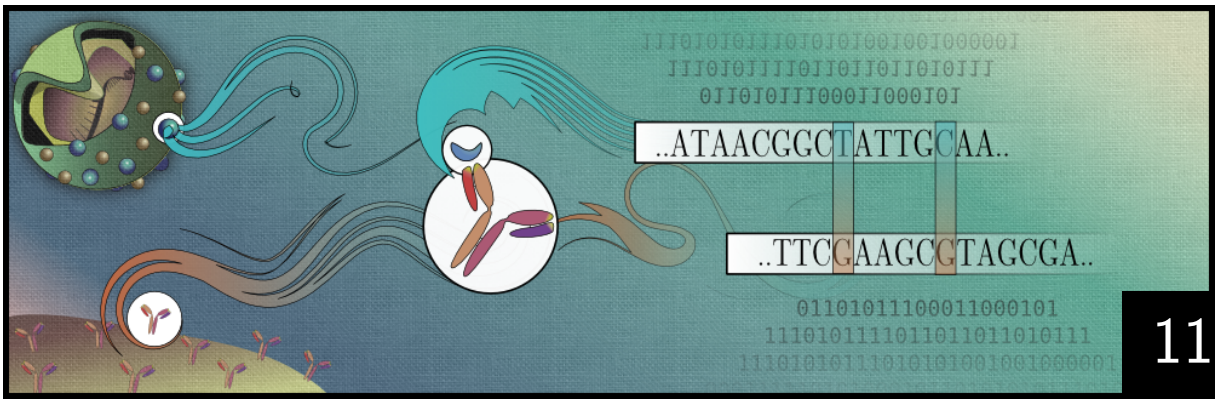
Zoulim, F., 2004. Mechanism of viral persistence and resistance to nucleoside and nucleotide analogs in chronic hepatitis B virus infection. Antiviral Res. 64, 1–15.

Zoulim, F., Locarnini, S., 2009. Hepatitis B virus resistance to nucleos(t)ide analogues. Gastroenterology 137 (1593–1608), e1591–1592.

PART

VI

DISCUSSION AND OUTLOOK



Discussion and outlook

If you want a guarantee, buy a toaster

CLINT EASTWOOD - THE ROOKIE

The internal struggle against infectious diseases is an important evolutionary process [5] with two sides of the same coin. Mainly, the host tries to evade the viral effects, if they cumber the host, but sometimes viral sequences end up engulfed in the hosts sequences and their proteins usefully assimilated into the internal processes of the host cell. So far there are dozens of examples known, some of them as astonishing as the evolution of the mammal placenta, which incorporates different viral envelope proteins in different classes of mammals and suggests therefore that mammal and with it human evolution was partly made possible by viruses [3, 16]. But those useful in-corporations are rare, and much of evolution is driven by antiviral techniques [25]. All living species use some kind of immune system to avert being destroyed by parasites [18, 20]. The mammal and especially the human immune system consists of complex processes to evade viruses and other unwanted intruders [9, 12, 21] and the human MHC-system is an example of disease-driven selective pressure [8, 26]. The research of this multilayer system - be it on the viral side or the side of the immune system - is important to understand, process, and to develop new useful drugs, which interact between virus and immune system.

Many aspects of the interaction between viruses and the immune system were highlighted in this thesis through sequence analyses. (Nucleotide-) Sequences are the basis of everything in biology and medicine [1]. They contain the basic information regarding the building blocks of viruses and living organisms [1]. Analysis of sequences enable researchers to get a basic idea of the interaction of the products of those sequences and

the underlying systems of evolution and selection pressure. These kinds of analyses are possible since Fred Sanger invented sequencing of DNA in the 1950s [22]. The invention of the next-generation sequencing methods in the 1990s enables researchers to dive even deeper into the sequence space and quasispecies of viruses and immune system cells. New third generation sequencing techniques (TGS)[15] will increase the amount of sequences further and be able to grant new insights into the mechanisms of polymerases[23]. An example for these new techniques is Nanopore Based Sequencing in which a sequence is guided through a nanopore and each nucleotide passing measured. Up until now this technique is not advanced enough and has a high error rate, but it is cheap, since no expensive chemicals are necessary, and fast [4, 6, 11]. A second kind of improvement is the cost of sequencing, which will decline further in the near future. As more and more next-generation sequencing techniques became available, a race for the 1,000 $ per genome started, and was finished by different companies. The newest inventions are Illuminas new sequencer named HiSeq X Ten Sequencer[13, 24] and Pacific Biosciences SMRT, which enables the customer to carry out next-generation sequencing of long fragments up to 20,000 nucleotides without a PCR [2, 10, 14, 17]. With only 1,000$ or less for sequence reads from a single whole genome with good quality, and the new possibility to get long reads, sequencing will become an important tool in medical research.

There are many sequences available from the viral big three: HIV, HBV, and HCV. Those sequences, together with further information like patients HLA type or known epitopes, may contain everything needed as a basis to develop new vaccines or drugs against infections, which are necessary to decrease the amount of infections. Razavi et al. predicted the total number of HCV infections should decline or remain flat, whereas the number of liver diseases will increase, but they only analyzed first world countries in Europe, Australia, and America [19], while in Asia and/or third world countries it may be worse. The development of a drug against HCV or a vaccine would significantly decrease both infection rate and disease effects. A patient with HIV however can never be fully cured. For those patients early antiretroviral therapy to sustain suppression of viral replication, and therefore lower the HIV transmission rate, is needed [7]. The development of such drugs which are safer, simpler, and better tolerated is still a developing field of research.

The amount of sequences is huge even now, but will only increase with more experiments made and therefore tools for sequence analysis are becoming more and more important. Be it the implications of infection chains, viral tropism, or the diversity of quasispecies, all

these things can be analyzed through sequences. The implementation of such a tool was accomplished with the R-package `SeqFeatR`, which was up until now used in four different groups in the medical field of research, but is available to all interested researcher of the sequence space, and will provide further insights into the relationship of viruses and the immune system in the future.

References

[1] B. Alberts et al. *Molecular Biology of the Cell*. 4th. New York: Garland Science, 2002 (cit. on p. 150).

[2] H. Buermans and J. den Dunnen. "Next generation sequencing technology: Advances and applications". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* (2014). ISSN: 0925-4439. DOI: http://dx.doi.org/10.1016/j.bbadis.2014.06.015 (cit. on p. 151).

[3] G. Cornelis et al. "Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora". In: *Proceedings of the National Academy of Sciences* 109.7 (2012), E432–E441. DOI: 10.1073/pnas.1115346109 (cit. on p. 150).

[4] M. Eisenstein. "Oxford Nanopore announcement sets sequencing sector abuzz". In: *Nat Biotech* 30.4 (Apr. 2012), pp. 295–296. ISSN: 1087-0156. URL: http://dx.doi.org/10.1038/nbt0412-295 (cit. on p. 151).

[5] J. B. S. Haldane. "Disease and evolution". In: *Ric Sci [Suppl]* 19 (1949), pp. 2–11 (cit. on p. 150).

[6] E. C. Hayden. "Pint-sized DNA sequencer impresses first users". In: *Nature* 521.7550 (May 2015), pp. 15–16. DOI: 10.1038/521015a. URL: http://dx.doi.org/10.1038/521015a (cit. on p. 151).

[7] R. S. Hogg et al. "HIV treatment strategies that can weather future challenges". In: *The Lancet Infectious Diseases* 14.7 (2014), pp. 534–535. DOI: http://dx.doi.org/10.1016/S1473-3099(14)70768-6 (cit. on p. 151).

[8] J. C. Howard. "Disease and evolution". In: *Nature* 352.6336 (Aug. 1991), pp. 565–567. DOI: 10.1038/352565a0 (cit. on p. 150).

[9] C. J. Janeway et al. "Immunobiology: The Immune System in Health and Disease". In: 5th. New York: Garland Science, 2001. Chap. Principles of innate and adaptive immunity. (Cit. on p. 150).

[10] J. T. Ladner et al. "Standards for Sequencing Viral Genomes in the Era of High-Throughput Sequencing". In: *mBio* 5.3 (2014). DOI: 10.1128/mBio.01360-14 (cit. on p. 151).

[11] A. H. Laszlo et al. "Decoding long nanopore sequencing reads of natural DNA". In: *Nat Biotech* 32.8 (Aug. 2014), pp. 829–833. ISSN: 1087-0156. URL: http://dx.doi.org/10.1038/nbt.2950 (cit. on p. 151).

[12] G. W. Litman, J. P. Rast, and S. D. Fugmann. "The origins of vertebrate adaptive immunity". In: *Nature Reviews Immunology* 10.8 (Aug. 2010), pp. 543–553. DOI: 10.1038/nri2807 (cit. on p. 150).

[13] E. Mardis. "Anticipating the $1,000 genome". In: *Genome Biology* 7.7 (2006), p. 112. ISSN: 1465-6906. DOI: 10.1186/gb-2006-7-7-112. URL: http://genomebiology.com/2006/7/7/112 (cit. on p. 151).

[14] E. R. Mardis. "Next-Generation DNA Sequencing Methods". In: *Annual Review of Genomics and Human Genetics* 9.1 (2008). PMID: 18576944, pp. 387–402. DOI: 10.1146/annurev.genom.9.081307.164359 (cit. on p. 151).

[15] S. El-Metwally, O. Ouda, and M. Helmy. "New Horizons in Next-Generation Sequencing". English. In: *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*. Vol. 7. SpringerBriefs in Systems Biology. Springer New York, 2014, pp. 51–59. ISBN: 978-1-4939-0714-4. DOI: 10.1007/978-1-4939-0715-1_6. URL: http://dx.doi.org/10.1007/978-1-4939-0715-1_6 (cit. on p. 151).

[16] S. Mi et al. "Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis". In: *Nature* 403.6771 (Feb. 2000), pp. 785–789. ISSN: 0028-0836 (cit. on p. 150).

[17] R. Mitra et al. "Statistical Analyses of Next Generation Sequencing Data: An Overview". English. In: *Statistical Analysis of Next Generation Sequencing Data*. Ed. by S. Datta and D. Nettleton. Frontiers in Probability and the Statistical Sciences. Springer International Publishing, 2014, pp. 1–24. ISBN: 978-3-319-07211-1. DOI: 10.1007/978-3-319-07212-8_1. URL: http://dx.doi.org/10.1007/978-3-319-07212-8_1 (cit. on p. 151).

[18] T. Nürnberger et al. "Innate immunity in plants and animals: striking similarities and obvious differences". In: *Immunological Reviews* 198.1 (2004), pp. 249–266. ISSN: 1600-065X. DOI: 10.1111/j.0105-2896.2004.0119.x (cit. on p. 150).

[19] H. Razavi et al. "The present and future disease burden of hepatitis C virus (HCV) infection with today's treatment paradigm". In: *Journal of Viral Hepatitis* 21 (2014), pp. 34–59. ISSN: 1365-2893. DOI: 10.1111/jvh.12248. URL: http://dx.doi.org/10.1111/jvh.12248 (cit. on p. 151).

[20] J. Rimer, I. R. Cohen, and N. Friedman. "Do all creatures possess an acquired immune system of some sort?" In: *BioEssays* 36.3 (2014), pp. 273–281. ISSN: 1521-1878. DOI: 10.1002/bies.201300124 (cit. on p. 150).

[21] Rinkevich. "Invertebrates versus Vertebrates Innate Immunity: In the Light of Evolution ('Nothing in biology makes sense except in the light of evolution' T. Dobzhansky, Amer Biol Teacher 1973;35:125–9)". In: *Scandinavian Journal of Immunology* 50.5 (1999), pp. 456–460. ISSN: 1365-3083. DOI: 10.1046/j.1365-3083.1999.00626.x (cit. on p. 150).

[22] F. Sanger and A. Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of Molecular Biology* 94.3 (1975), pp. 441–448. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1016/0022-2836(75)90213-2 (cit. on p. 151).

[23] E. E. Schadt, S. Turner, and A. Kasarskis. "A window into third-generation sequencing". In: *Human Molecular Genetics* 19.R2 (2010), R227–R240. DOI: 10.1093/hmg/ddq416 (cit. on p. 151).

[24] R. F. Service. "The Race for the $1000 Genome". In: *Science* 311.5767 (2006), pp. 1544–1546. DOI: 10.1126/science.311.5767.1544. eprint: http://www.sciencemag.org/content/311/5767/1544.full.pdf. URL: http://www.sciencemag.org/content/311/5767/1544.short (cit. on p. 151).

[25] L. M. Van Blerkom. "Role of viruses in human evolution". In: *American Journal of Physical Anthropology* 122.S37 (2003), pp. 14–46. ISSN: 1096-8644. DOI: 10.1002/ajpa.10384 (cit. on p. 150).

[26] R. M. Zinkernagel, H. Hengartner, and L. Stitz. "On the role of viruses in the evolution of immune responses". In: *British Medical Bulletin* 41.1 (1985), pp. 92–97 (cit. on p. 150).

APPENDICES

A

Supplementary Material: SeqFeatR for the discovery of feature-sequence associations

S1 Alignment

V3 amino acid sequences of CCR5- and CXCR4-tropic HIV-1. Figure A.2 was produced by SeqFeatR with this input. All sequences (84 from CXCR4-tropic and from 928 CCR5-tropic virus) have the same length of 35 amino acids and have not been submitted to an extra alignment step. Note that the feature labels "X4" (for CXCR4-tropic) and "R5" (for CCR5-tropic) have been added at the end of the FASTA headers after a semicolon.

S2 Alignment

Alignment of SSU nucleotide sequences from Chlamydomonas. Alignment of RNA sequences of small ribosomal subunit sequences: 9 from *Chlamydomonas applanata*, 10 from *Chlamydomonas reinhardtii*. Figure A.3 was generated by SeqFeatR with this input. Note again that the last element of the FASTA header stands for the feature, here: RH for *reinhardtii* and AP for *applanata*.

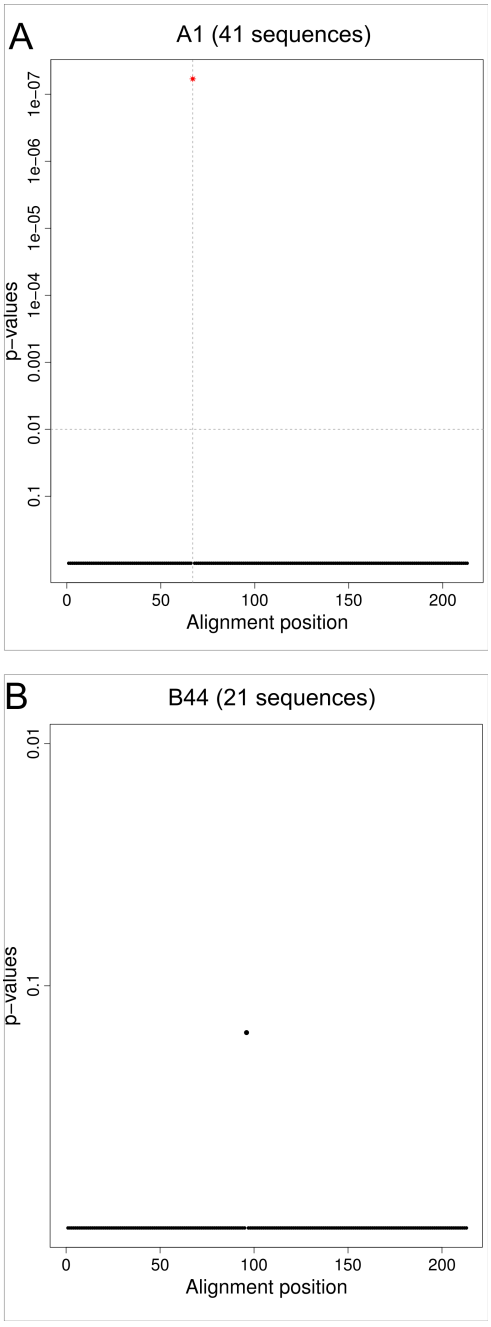


Figure A.1: Frequentist approach with correction for multiple testing. Association of alignment positions of HBV core protein with patient HLA types A*01 (**A**) and B*44 (**B**). Sequence numbers in panel titles are feature-carrying fractions of the total of 148 sequences included in the alignment. Association of sequences with feature HLA were analyzed with Fisher's exact test, and resulting p values were corrected for multiple testing with FDR option.

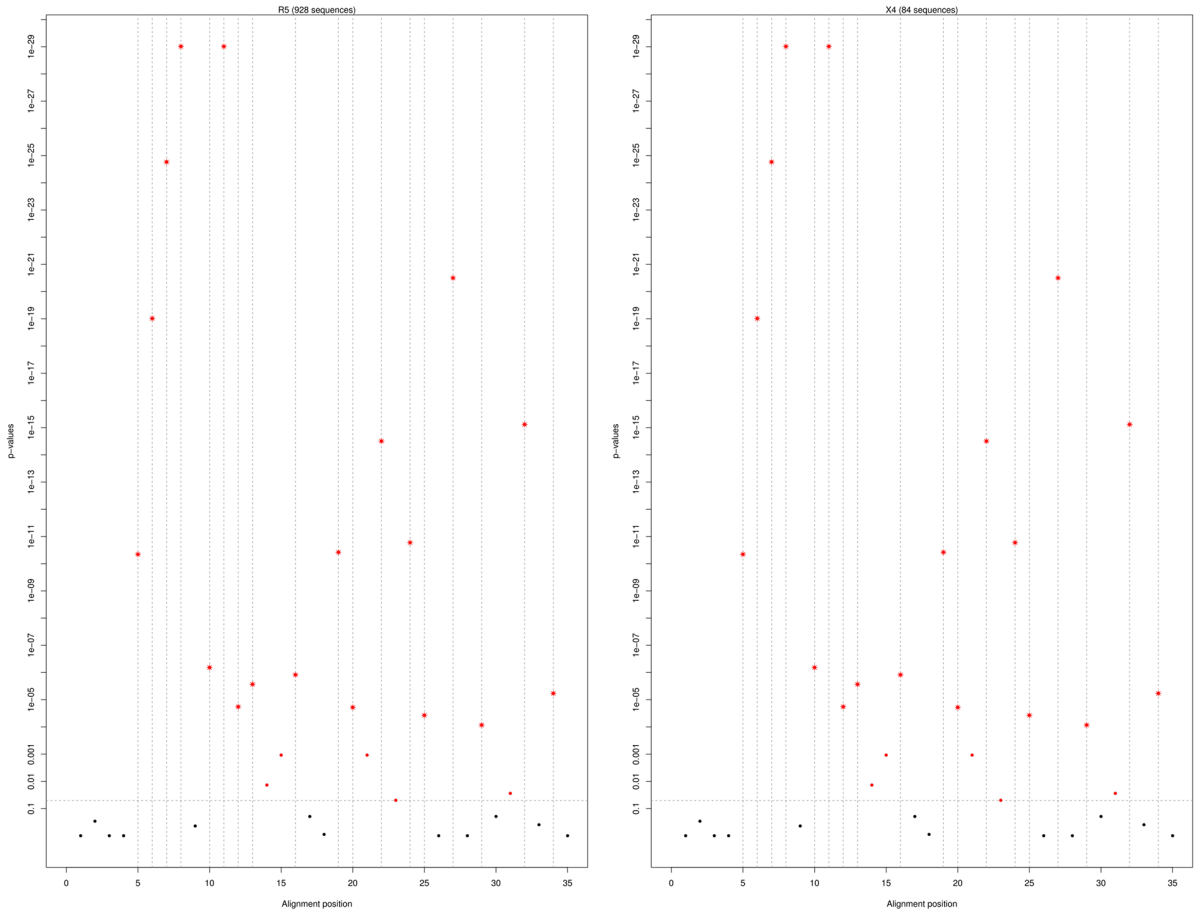


Figure A.2: Association of V3 sequence positions with HIV-1 co-receptor tropism. Manhattan plot output of SeqFeatR showing sites in the V3 amino acid sequences Figure A that are significantly associated with co-receptor tropism.

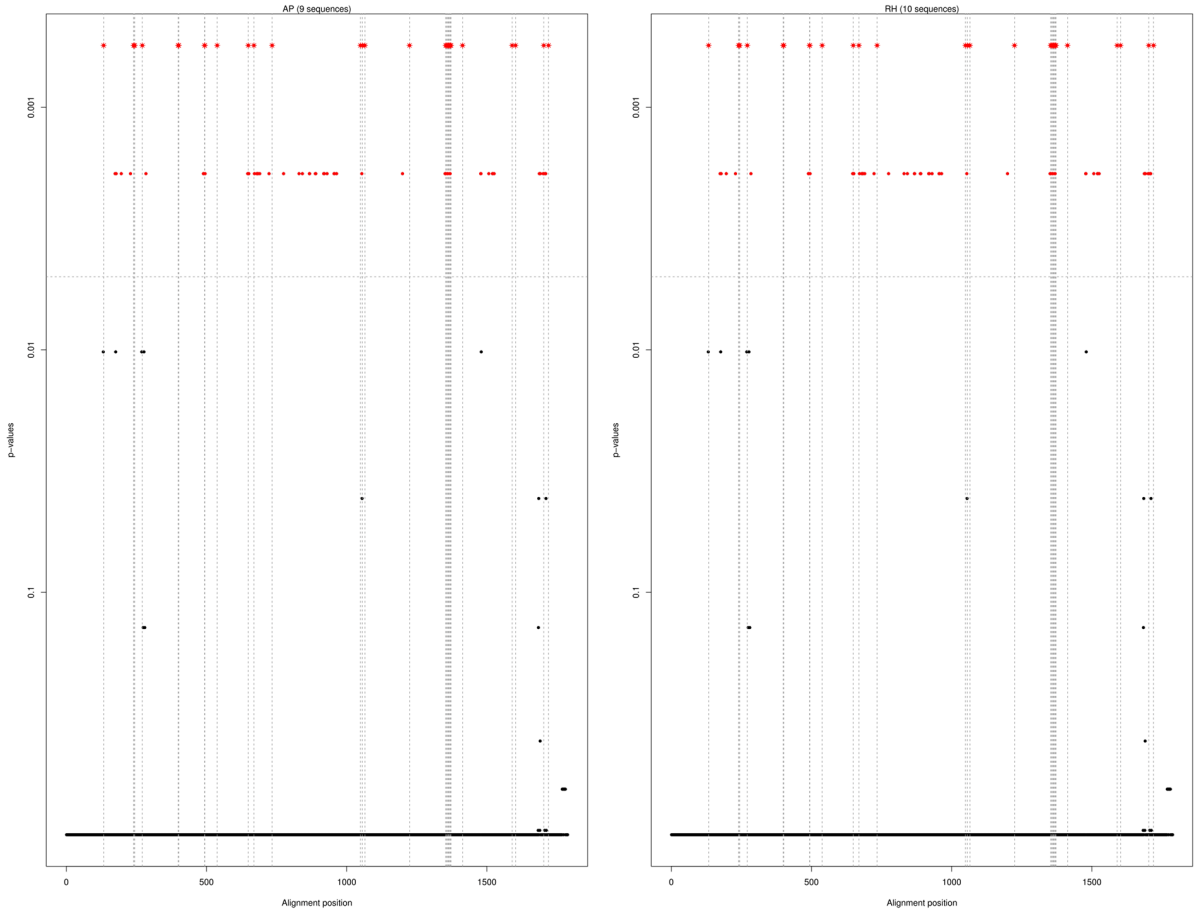


Figure A.3: Association of Chlamydomonas SSU nucleotide sequence position with species. Manhattan plot output of SeqFeatR showing sites in the SSU nucleotide sequence alignment Figure A that are significantly associated with *Chlamydomonas* species, here: *Chlamydomonas reinhardtii* (RH) vs *Chlamydomonas applanata* (AP).

B

Supplementary Material: Tutorial for R-package `SeqFeatR`

This tutorial gives you some of the technical background underlying SeqFeatR that should enable you to understand what SeqFeatR does and also how to use it and how to interpret the output. If you are solely interested in a HowTo, you may find the following two video tutorials helpful:

- For the GUI in the SeqFeatR R-package: https://www.youtube.com/watch?v=-CYidGPE6dw
- For the SeqFeatR web-server: https://www.youtube.com/watch?v=3z4Smk3mI18

B.1 SeqFeatR discovers statistically significant feature - sequence associations

Imagine the following alignment[1] of amino acid sequences in FASTA format, taken from 14 patients that either have a certain feature ("f") or do not have that feature ("n"). The feature may for instance be an HLA type[2], a genetic disease, etc. In the following FASTA formatted input for SeqFeatR you can see a letter (f or n) indicating feature or not-feature at the end of each comment line:

```
>P01_HLA_A01_00_B01_02;f
LPDIQGNENMGYQPSWIFCGMETNGSQCLEEMFHCCWINC
>P02_HLA_A01_00_B01_02;f
MPDWNQKWGNDHLASINLD-WLKTIQQPGIEKHLRFYENW
>P03_HLA_A01_02_B01_02;f
VPDASGKHGIIGMDVTSSMERRHGMVQLPWPAMVWGRPHW
>P04_HLA_A01_00_B01_02;f
MPDVRGVGCARRDCLIVHRFCMPFNNQVYCKVWIVYWTYK
>P05_HLA_A01_00_B01_02;f
QPDTPKITRKEATAIHKCGIHWQTNCQKLSTVHPFHHQVD
>P06_HLA_A01_02_B01_02;f
```

[1] The "alignment" shown here does not look like a good alignment. Usually, alignments show much more columns where many sequences have the same amino acids, and there may also be gaps, indicated by "-" in some sequences. This bad alignment was chosen for one reason only: it demonstrates that it can be difficult to spot relationships between features and sequence positions.

[2] Human leukocyte antigen (https://en.wikipedia.org/wiki/Human_leukocyte_antigen), a classification scheme of human immune systems.

```
SWDDFSDFTMVHQWYAQGTLGPYKAMQLKMIFQGVSIMEV
>P07_HLA_A01_02_B01_02;f
IPDEPCYCCVKNKILTVEIGVHHAKSQVRRNIDNIRRKTE
>P08_HLA_A04_03_B04_03;n
HFST-ICPYIWKMYFTWMGQKLVIQKVNGRTPPHCDECNQ
>P09_HLA_A04_03_B04_03;n
SNFT-TTKLRDQHNLYPAGLQEIEHKVDHQILGIYGQIWY
>P10_HLA_A04_03_B04_03;n
ETSTALRTQDQTFMLALRANYMVMLKVLDCISVKLFICWR
>P11_HLA_A04_03_B04_00;n
DSSTMDAECSTLQRFIWWHAHYAWIRVAKKPYCLDCPYAV
>P12_HLA_A04_03_B04_03;n
KKSTLGIARGIQRSHGWYWRQTHCVMVLTPSQHKMGEKSW
>P13_HLA_A04_03_B04_00;n
ICSTELCGCLINWPPMQWIVFAHMDDVNDSQTNTCDMRSQ
>P14_HLA_A04_03_B04_03;n
GPSTNARTMGGQDCAYMTHTLTKHIWVILAFDPIMIVHKP
```

Can you discover statistically significant associations of the feature with the presence or absence of certain amino acids at certain sequence[3] positions? It is difficult to spot such associations with the naked eye, but they are there:

- There is a strong association of feature f with amino acid P at the second position, though patient 14 is an exception as she is n and still has P at second position.
- There is a strong association of n (i.e. not having feature f) with amino acid T at fourth position, though patient 5 is an exception as he has f and still has a T at fourth position.

In its basic application, `SeqFeatR` tests *all* sequence positions and quickly identifies the second and fourth positions as being statistically significantly associated with the feature (f) or its absence (n). `SeqFeatR` shows these associations graphically in two ways, as Manhattan plot and as odds ratio (OR) plot.

[3]Strictly, we are not dealing with *sequence* positions but with *sequence alignment* positions.

B.2 The core of SeqFeatR: Fisher's exact test

B.2.1 An example: association of a feature with sequence

We have mentioned that in the above alignment there is seemingly a strong association of the occurrence of amino acid P at position 2 with the feature f. The probability and strength of this association can be quantified, respectively, by a p-value computed with Fisher's exact test, a well-known statistical test for association, and by an odds ratio (OR). At its core, SeqFeatR does exactly this.

In the above example of the association of P at position 2 with feature f, SeqFeatR internally would first collect occurrences in a frequency table and then compute from that frequency table p-value and OR:

- 6 sequences with feature f *and* P at position 2
- 1 sequence with feature f *and not* P at position 2
- 1 sequence with feature n (= *not* f) *and* P at position 2
- 6 sequences with feature n *and not* P at position 2

SeqFeatR collects these data in a frequency table:

		Proline	
		+	-
feature f	+	6	1
	-	1	6

Submitting this table to Fisher's exact test yields a p-value of 0.0291. At a significance level of 0.05 we therefore *reject* the null hypothesis (= *no* association of f and P at position 2) and rather assume an association of f and P at position 2.

The *strength* of the association is quantified by the odds ratio (OR) that is computed from the elements of the above contingency table:

$$OR = \frac{N_{f,P2}/N_{f,not\,P2}}{N_{not\,f,P2}/N_{not\,f,not\,P2}} = \frac{6/1}{1/6} = 36.$$

(Note: there are several methods to estimate the odds ratio. The simple one shown here is called Wald's method. The one used by SeqFeatR yields about 23.5.)

An OR much greater than 1 ($OR \gg 1$) as we have it here ($OR = 36$) means that we have a *strong positive association* of the feature f with $P2$: f and $P2$ occur much more often together than expected if we had no association.

If we have *a weak or no association*, the OR lies around 1. Then f and $P2$ would occur together and not together in the same ratios.

If we have a *strong negative association*, $0 < OR \ll 1$. In case of f and $P2$, a negative association means that f and $P2$ occur *less* often together than expected if we had no association.

B.2.2 Another example: association of HLA type and sequence

In the above set of FASTA formatted sequences, we had ended each sequence header with the name of a feature, either f or n, separated from the rest of the header line by a semicolon. A specific type of feature that is often used in SeqFeatR analyses is the *HLA type*. (For the HLA type there is an optional way of telling SeqFeatR about this specific feature by giving the positions of the HLA information in the FASTA header, see section B.4 of this tutorial and first tab of SeqFeatR graphical user interface.) For instance, SeqFeatR will automatically discover in the sequences above a significant association between HLA*B01 and amino acid D at third sequence position:

- 7 sequences with HLA*B01 *and* D3
- 0 sequences with *not* HLA*B01 *and* with D3
- 0 sequences with HLA*B01 *and not* D3
- 7 sequences with *not* HLA*B01 *and not* D3

Thus, we obtain the following contingency table:

		D	
		+	-
HLA B*01	+	7	0
	-	0	7

Fisher's exact test yields a p-value < 0.001 and we have an OR of infinity. Thus we have a *significant and strongly positive* association of HLA*B01 and D3.

B.3 Graphical output

B.3.1 $-\log_{10}$ p-value plot ("Manhattan plot")

The so-called Manhattan plot, i.e. a plot of $-\log_{10}$ p-values along the sequence, is a convenient means to discover significant associations of sequence alignment positions with features. `SeqFeatR` produces Manhattan plots consisting of two separate plots (Figure B.1): The top half of the plot focuses on complete epitopes or putative epitopes comprising "windows" of several sequence positions (e.g. windows of 9 positions), while the bottom half gives a more detailed picture of the same data at the level of single sequence positions. The x-axis for both plots is the same, namely the positions in the input sequence alignment.

Let us start the discussion with the more fundamental bottom part of Figure B.1, the simple Manhattan plot. In this plot, `SeqFeatR` can mark a significance level α (here: $\alpha = 0.01$) with a horizontal line. Associations with $-\log_{10} p$-values above that line (i.e. p-values $< \alpha$) are shown with a special symbol (here: red stars) and considered significant. To ease the visual localization of the highly significant positions, they are additionally marked with vertical lines that hit the sequence axis at the corresponding positions. (The resolution of the x-axis is usually to coarse to show single positions, but sufficient to localize significantly associated positions in the fully resolved csv-file which is given as second output file.)

Now the top part of the Manhattan plot of Figure B.1, focusing not on single positions but on complete putative epitopes. If the features in your `SeqFeatR` input have been HLA types, and if you then see several sequence positions in close proximity showing up with high $-\log_{10} p$ values in the Manhattan plot, you may have found a HLA epitope. The top part of the Manhattan plot highlights such position clusters. It can potentially show three different curves, a black, a red, and a yellow curve, each indicating potential epitopes. The red and yellow curves are optional.

The black curve directly relates to the bottom part of the plot: it shows the number of sequence positions with significant feature association in a window of 9 amino acids (or

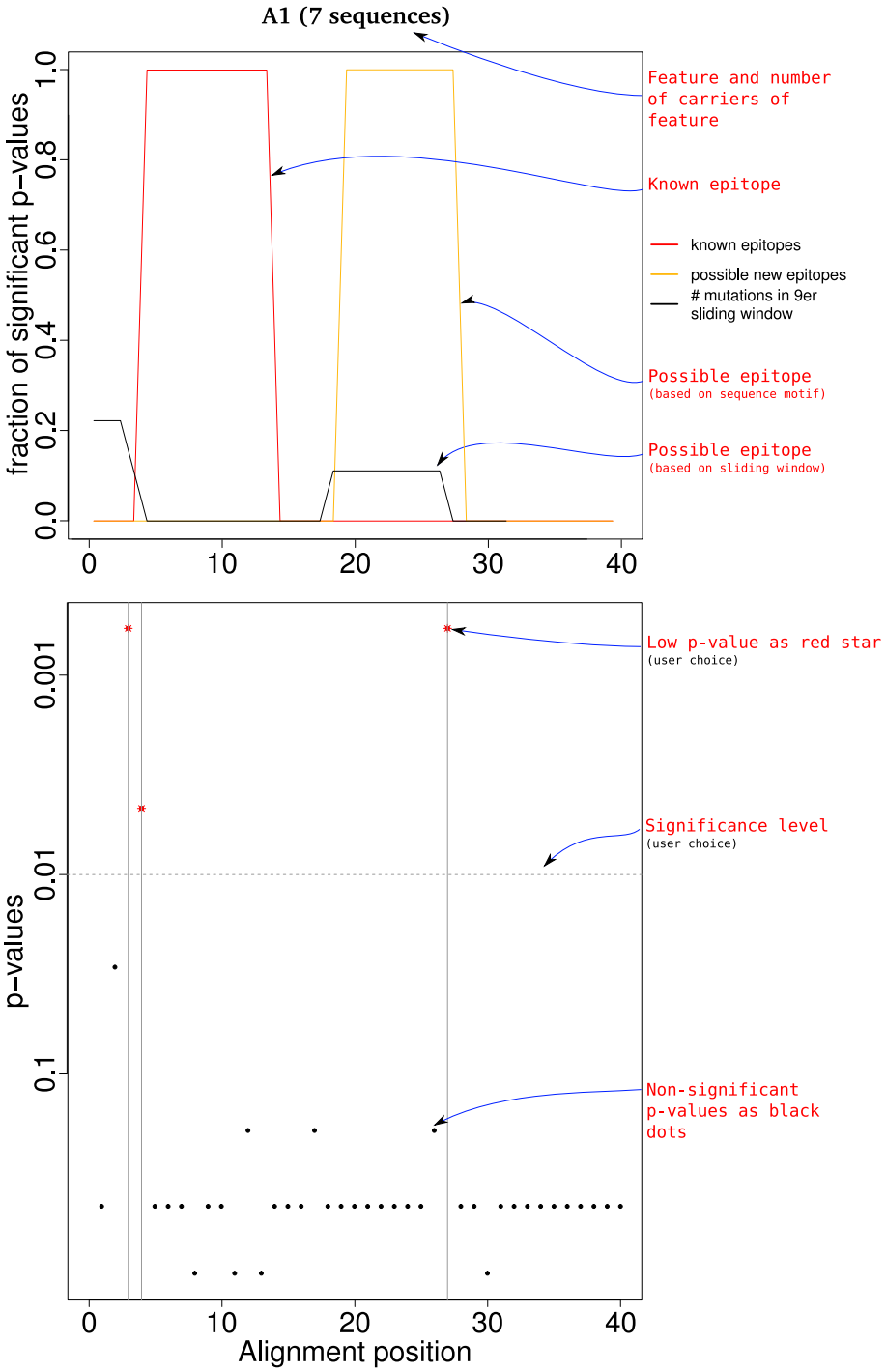


Figure B.1: Manhattan plot (p-values along sequence). The y-axis scales with the $-\log_{10}$(p-value), i.e. the higher the point, the more significant the association. Top half of figure combines three different ways of showing possible epitopes: (1) possible epitopes from a window-wise statistical analysis of your data (black line), (2) known epitopes, e.g. from the literature (red line), (3) pieces of sequence alignment that conform with certain sequence patterns (yellow line). Bottom half of figure: Manhattan plot with the p-values for each sequence position. There are additional annotations in this graphic to explain what you see in red. Those are *not* in the real output from `SeqFeatR`.

any other window length given by the user), divided by the window width (e.g. 9). The default window width of 9 amino acids corresponds to a typical length of a MHC I binding peptide. This window is shifted over the whole sequence and the fraction of significantly associated position computed for each window position and plotted as y-value at that position. An HLA epitope will show up as bump of the black curve to high values, similar to the bump in the top part of Figure B.1.

The red curve (optional) allows the user to mark known epitopes, e.g. published in the literature or in a database. This can be helpful for comparisons. You can enter data for the red curve in an extra csv file ("Known epitopes" in `SeqFeatR` GUI and web interface). Here an example of what you could put into such a csv file[4]:

```
4;12;A1
2;9;A3
```

This example marks two known epitopes with two lines of the form `EpitopeStart; EpitopeStop; HLAtype`. The first line (`4;12;A1`) corresponds to the epitope shown as bump of the red line in the top of Figure B.1. The level of the bump is always at a value of 1.0 for these known epitopes. The bump is only shown if the HLA type in the csv file matches the HLA type in the sequence alignment.

With the yellow curve (optional) the user can mark alignment regions that are conforming with given sequence patterns. Such patterns can be defined in another csv file ("Known binding motifs" in `SeqFeatR` GUI and web interface) following this format[5].:

```
Genotype;Motif;Reference
A*01;x[PV]xxxx[DENQ]EN;SYFPEITHI
```

The header line (`Genotype; Motif; Reference`) describes the structure of the following lines. All elements in one row are separated by semicolons. The first element is the HLA type (here: `A*01`), the second is the actual definition of the motif (here: `x[PV]xxxx[DENQ]EN`), and the third gives a reference to the origin of this motif [6].

[4] A similar example is part of the R-package `SeqFeatR` and there called `Example_epitopes_aa.csv`.

[5] A similar example is part of the R-package `SeqFeatR` and there called `Example_HLA_binding_motifs_aa.csv`

[6] SYFPEITHI is just an example of a possible reference – the motif was not really taken from SYFPEITHI

The definition of the motif requires a bit of explanation. The motif shown here covers nine amino acid positions. Here the nine amino acid positions are shown as indices:

$$\texttt{x}_1\underbrace{[PV]}_{2}\texttt{x}_3\texttt{x}_4\texttt{x}_5\texttt{x}_6\underbrace{[DENQ]}_{7}\texttt{E}_8\texttt{N}_9$$

The letter `x` stands for: "this could be any amino acid". The two square brackets at positions 2 and 7 show which amino acids are allowed at these two positions, e.g. proline (P) or valine (V) at position 2. At position 8 there has to be a glutamate (E), and at position 9 must be an asparagine (N). Many sequences conform with this particular motif description, e.g. `APYEILDEN` or `SVRKTSQEN`. In general, motifs should be expressed in the same way using the elements `x`, `[]`, and capitals `ACDE...Y`. `SeqFeatR` shows a bump of the yellow line to 1.0 if several conditions are fulfilled simultaneously: (a) the HLA type in the sequence alignment matches the HLA type in the csv file, (b) the motif occurs in one of the aligned sequences, and (c) in the sequence window covered by the motif there is at least one significant association of an alignment position with the HLA.

B.3.2 Advanced `SeqFeatR` plotting, e.g. odds ratio plot

The R-package `SeqFeatR` offers a number of more advanced commands that are not yet available through the web interface of `SeqFeatR`. A nice example is the odds ratio plot that requires the use of the function `orPlot` of the R-package `SeqFeatR`. This is why you could be interested in the odds ratio plot:

While the Manhattan plot is a useful means to gain an overview over the distribution of significant sequence–feature associations along the alignment, there is still important information missing: Which amino acid is characteristic for the positions with low p-value? Is an amino acid overrepresented or underrepresented at such a position in sequences with a certain feature? All this information can be extracted from the csv file produced by `SeqFeatR`, but the program also provides a new plot, that we call *odds ratio plot*, to visualize this information (Figure B.2).

We had introduced the odds ratio (OR) in section B.2.1 as a way to quantify the strength of the association. The OR tells us whether a feature is over- or underrepresented at a position ($OR > 1$ or $OR < 1$, respectively), with $OR = 1$ indicating vanishing association. For easier visual recognition, we show the logarithms ($\log_{10} OR$) of the OR values along the sequence. To given an example: using this logarithmic representation, a tenfold

overrepresentation of an amino acid at a sequence position in sequences with a certain feature (e.g. HLA type) shows up in the OR plot as *upwards* pointing bar of length 1, a tenfold underrepresentation as *downwards* pointing bar of length 1.

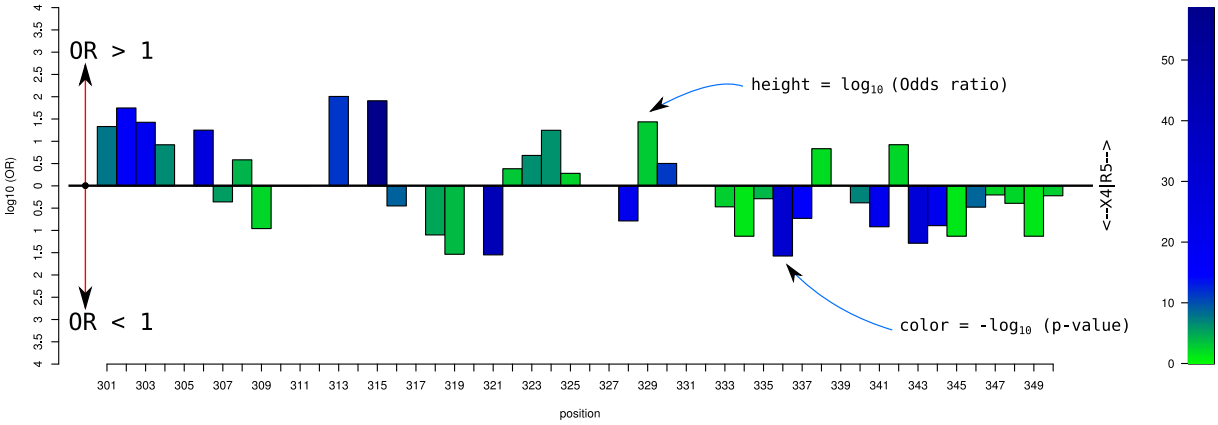


Figure B.2: Example odds ratio plot. Here we analyze amino acid sequences of HIV-1 gp120 protein variants, and we have as feature the so-called co-receptor tropism of HIV-1, which can be "R5" or "not R5" (the latter is often called "X4"). The odds ratio (OR) plot shows for each sequence alignment position the association strength $\log_{10}(OR)$ as bar height and the p-value as bar color. The plot demonstrates that high values of $\log_{10} OR$ (long bars) and high statistical significance (blue color) are not the same.

There is one important caveat: even if a long OR bar indicates strong association, it still may not be *statistically significant*. This may be confusing at first, but think about a situation in which you have a small set of sequences, say two sequences, with a certain feature, and one without the feature. In the sequences with the feature we have amino acid A at the first alignment position, in the sequences without the feature we have G at that position. Thus A is overrepresented and G underrepresented in the sequences with the feature, right? In fact your OR would be infinity (see example in section B.2.2). But would you believe this? Probably not, since your sequence set is so small that the p-value from the Fisher's exact test is 0.33.

Therefore the odds ratio plot allows you to combine OR information and p-value information. In Figure B.2 OR bars with low p-values (highly significant) are filled with a blue color, while OR bars with higher p-values (not so highly significant) are filled with a green hue. As you can see, by far not all long bars are blue, i.e. only a subset of positions may have strong and significant associations with the feature.

B.4 Input: sequences and features

The whole `SeqFeatR` analysis is critically dependent on the input. Therefore we summarize here how to prepare the input data properly.

First, before you submit sequences to `SeqFeatR`, you have to *align* them, i.e. do not submit sequences to `SeqFeatR`, but a multiple sequence alignment (MSA). There exist several popular tools for this task, for `mafft` (http://www.ebi.ac.uk/Tools/msa/mafft/). Make sure that the resulting MSA is in FASTA format (or Pearson/FASTA) (for a description of FASTA format, see https://en.wikipedia.org/wiki/FASTA_format, an example is shown on page A-9). Usually, this output format can be chosen as option in the input form of `mafft` and other MSA programs, so you do not have to do this manually.

In the FASTA formatted MSA, we have one block for each sequence. Such a block consists of a header line (starting with >) that can be used to describe the sequence, and the sequence itself on the following lines. `SeqFeatR` expects in each FASTA header line a label that tells it which feature this sequence carries. There are two different types of features that can be put into FASTA headers:

1. A feature can be given by a letter or word at the end of each FASTA header after a semicolon, as in:

   - >some information;feature
   - >patient 1;f
   - >HCVA;n
   - etc.

   Anything after a semicolon in the FASTA-header is interpreted as name of a feature.

2. HLA types are a special case of features accepted by `SeqFeatR`. Here an example snippet from a FASTA file with the encoded HLA type information in the FASTA headers (see also example on page A-9):

```
>P1 HLA_A0403_B0403 donor 1
SNFT-TTKLRDQHNLYPAGLQEIEHKVDHQILGIYGQIWY
ETSTALRTQDQTFMLALRANYMVMLKVLDCISVKLFICWR
DSSTMDAECSTLQRFIWWHAHYAWIRVAKKPYCLDCPYAV
```

```
>P2 HLA_A0403_B0404 donor 2
...
```

Now let us focus on the first header and for orientation write numbers `123...` beneath the header that give us the positions of the characters in that line:

```
>P1 HLA_A04_03_B04_03 donor 1
123456789.123456789.123456789.1234
```

Here, `HLA_A04_03` corresponds to HLA-A*04:03[7] with locus A, group number 04, and variant 03. The group number "04" covers positions 10 and 11 of the line, the variant "03" is at positions 13 and 14. Analogously, for the B locus we have group "04" at positions 17 and 18, and variant "03" at positions 20 and 21. We give `SeqFeatR` these four position intervals of A and B groups and variants, as shown in Figure B.3. Importantly, the HLA types in *all* FASTA blocks in one MSA file have to take the *same* positions in their respective lines to be properly recognized by `SeqFeatR`.

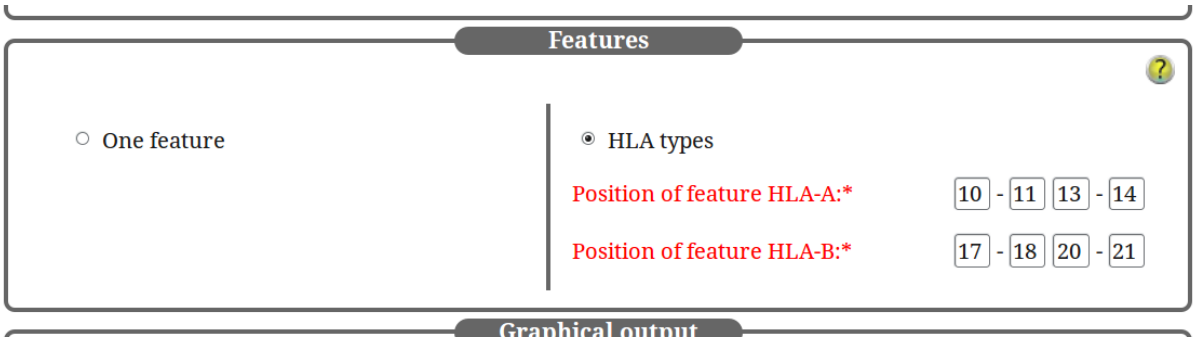


Figure B.3: Screenshot of `SeqFeatR` web interface with filled HLA positions.

[7] See e.g. https://en.wikipedia.org/wiki/Human_leukocyte_antigen#Nomenclature

B.5 Multiple comparison correction

`SeqFeatR` applies the same statistical test to *many* sequence alignment positions. Under these conditions it is likely that some of the statistical test yield a low p-value just by chance, and not because of a real association.

An example may illustrate this so-called *multiple testing problem* (or *multiple comparison problem*): imagine you toss a fair coin (fair = fifty-fifty chance for head and tail) four times. The most probably result is that you have 2 times head and 2 times tail. The probability to have 4 times head is only 1 in 16 (1/2 for the first toss times 1/2 for the second ... $= (1/2)^4 = 1/16$). Now let us do 16 of these 4-toss experiments, i.e. we toss the coin 16×4 times. Then we expect that in one of these 4-toss experiments we will see 4 times head, just by chance. If we only consider the experiment with the outcome of 4 times head, we have the wrong impression that the coin is *not* fair. In the same way, a set of sequence alignment positions corresponds to a set of random experiments, and if we carry out many association tests, some of them seemingly indicate an association, but that association is not real.

`SeqFeatR` offers various types of corrections (graphical user interface/web interface: P-value correction) as available through the standard R-stats package function `p.adjust` (in R call help page of p.adjust for more information). This includes e.g. very conservative Bonferroni correction or the probably more useful False Discovery Rate correction.

B.6 Hints

- Do **not** use a word processor such as Word or LibreOffice/ OpenOffice to prepare the sequence alignment input. These programs likely destroy the FASTA format by (invisible) extra characters and invalidate the input. Instead, use an editor for raw text such as `notepad` in Windows systems, or `gedit`, `vim` etc. on Linux systems.
- `SeqFeatR` can only understand letters from canonical DNA/acrshortRNA (A, C, G, T, U) and amino acid alphabets (A, C, D, E, ...) in the FASTA sequences. Do not use special characters (?!; etc.) or characters for wobbles (R, K, Y, ...) in sequences. You can however use X or B for unidentified nucleic acids or amino acids.

B.7 Advanced Feature: Bayes Factor in SeqFeatR

If for any reason you do not want to work with p-values, you can use `SeqFeatR` with Bayes Factor (BF) instead of the above mentioned Fisher's exact test. The BF for two hypotheses H_0 and H_1, given sequence and feature data D, is the ratio of posterior odds and the corresponding prior odds: $BF = (p(H_1|D)/p(H_0|D)) \,/\, (\pi_1/\pi_0)$. In other words, the BF equals the posterior odds if the prior probabilities π_0, π_1 are equal and thus the prior odds is 1.

Imagine the following table:

		certain amino acid	
		+	-
feature f	+	p_{11}	p_{12}
	-	p_{21}	p_{22}

In our case, H_1 is the hypothesis that a feature is associated with an amino acid or nucleotide at an alignment position, and H_0 is the hypothesis that there is no such association. The higher the BF, the more likely H_1 (association) and the less likely H_0 (no association).

Here we use a BF for the hypothesis H_1 that feature and amino acid at an alignment position are *close* to independence vs. H_0 that they are independent. Albert and Gupta presented a 'close to independence' model which is now used in `SeqFeatR` [8]. The prior belief in the independence is expressed by a user chosen K: the higher this hyperparameter, the more dominant the independence structure will be in comparison to the observed counts, and for $K \to \infty$ complete independence is achieved.

While `SeqFeatR` allows for setting an explicit K value, it may not be easy to specify an appropriate value of K that is applicable to all alignment positions. Therefore, `SeqFeatR` also offers an empirical Bayes variant of this BF. In this variant, an individual value of K is estimated from each contingency table itself.

[8] Albert J. Bayesian Computation with R. Springer Verlag; 2009.

B.8 Advanced Feature: Discovering associations between mutation tuples and features

If you have discovered single positions with feature associations, you can use `SeqFeatR` to discover if those positions correlate and are as a combination associated with a feature. These associations are far from visible in the sequences itself, but they are there:

- There is a strong association of feature $HLA*A01$ with amino acid pair PQ at the 3 position and 27 position, though patient 6 is an exception as she is $HLA*A01$ and has not PQ but WQ.
- There is an association of feature $HLA*A02$ with amino acid pair DH at the 3 and 23 position, though patient 6 is an exception as she is $HLA*A02$ and has not DH but DY.

`SeqFeatR` tests *all* sequence position pairs below a given p-value and quickly identifies the 3+27 pair as being statistically significantly associated with the feature ($HLA*A01$) or its absence. `SeqFeatR` shows these associations graphically as a heat map and in combination with another feature as a "Tartan plot".

B.8.1 An example: association of HLA type and pairs

We have noted that in the above alignment there is seemingly a strong association of the occurrence of amino acids P at position 3 and amino acid Q at position 27 with the feature $HLA*A01$. The probability and strength of this association can again be quantified, by a p-value computed with Fisher's exact test and by an odds ratio.

In the above example of the association of P at position 3 and Q at position 27 with feature $HLA*A01$, `SeqFeatR` internally would first collect occurrences in a frequency table and then compute from that frequency table p-value and OR like in the case of a single position B.2.1:

- 6 sequences with HLA*A01 *and* P3Q27
- 0 sequences with *not* HLA*A01 *and* with P3Q27
- 1 sequences with HLA*A01 *and not* P3Q27

- 7 sequences with *not* HLA*A01 *and not* P3Q27

Thus, we obtain the following contingency table:

		DH	
		+	-
HLA A*01	+	6	1
	-	0	7

Fisher's exact test yields a p-value < 0.005 and we have an OR of Infinity. Thus we have a *significant and strongly positive* association of HLA*A01 and P3Q27.

B.9 Tartan plot: visual comparison of different associations of features and sequence position tuples

The Tartan plot is a way to visualize the comparison of two different types of associations between pairs of sequence alignment positions (lower left vs. upper right triangle). Association strengths are color coded (color legend on the right) and depend e.g. on the p-value of the association of this pair and the feature. For orientation, axes can be annotated and sequence substructures can be indicated by lines. The different types of associations can be for example two different HLA types, or HLA type and distance in the protein of the sequences. Here we show an example of two HLA types, HLA*A01 and HLA*B03 B.4.

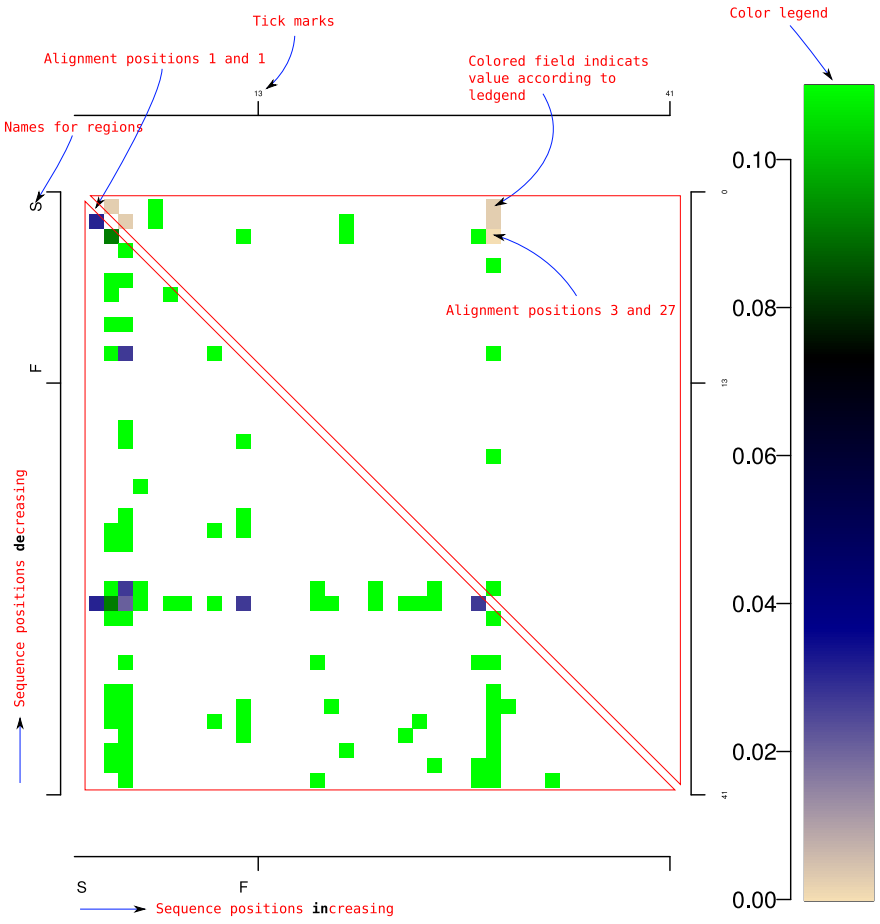


Figure B.4: Tartan plot. This plot combines the information of two position-pair and feature association. The first feature (upper triangle) is HLA*A01, the second one (lower triangle) is HLA*B03. A color-filled field represents the value of a position-pair according to the color legend. There are additional annotations in this graphic to explain what you see in red. Those are *not* in the real output from `SeqFeatR`.

C

Supplementary Material: Proteasomal selection pressure on hepatitis C virus epitopes

Table C.1: Significant p-values from comparison between positive and negative epitopes with an Fisher's exact test. Three datasets were used all for HLA-A*02. The HCV positive and negative datasets were generated from HCV T cell data from IEDB in combination with MHC predicted data out of the HCH reference genome H77. This dataset was a comparison between both HCV datasets. FDR was applied to the p-values to correct for multiple testing errors.

	position	Fisher p-value	amino acid	MW p-value
HCV^+ and HCV^-	4	0.0453318905	M	
	4	0.0212912181	S	
	5	0.0239590308	K	
	7	0.0453318905	L	
	9	0.0453318905	E	
	9	0.0203958021	L	
	10	0.0390251451	E	
	10	0.0239590308	H	
	11	0.0282174805	L	
	17	0.0203958021	N	
	19	0.0203958021	S	
	21	0.0398472595	S	
	22	0.0460295601	L	
	23	0.0275357767	H	
	28	0.0239590308	C	
	28	0.0453318905	Y	

Table C.2: Significant p-values from comparison between positive and negative epitopes with an Fisher's exact test. Three datasets were used all for HLA-A*02. The HCV positive and negative datasets were generated from HCV T cell data from IEDB in combination with MHC predicted data out of the HCH reference genome H77. This dataset was a comparison between human T cell data from IEDB and the negative HCV epitopes. FDR was applied to the p-values to correct for multiple testing errors.

	position	Fisher p-value	amino acid	MW p-value
$HUMAN^{+}$ and HCV^{-}	1	0.0453318905	W	
	10	0.0282174805	Q	
	10	0.0139489695	E	
	11	0.018770049	E	
	12	0.0242838185	W	
	15	0.0453318905	M	
	22	0.0239590308	E	
	22	0.0454598032	L	
	25	0.0394199305	A	
	30	0.0239590308	G	
	30	0.0275357767	E	

Table C.3: Significant p-values from comparison between positive and negative epitopes with an Fisher's exact test. Three datasets were used all for HLA-A*02. The HCV positive and negative datasets were generated from HCV T cell data from IEDB in combination with MHC predicted data out of the HCH reference genome H77. This dataset was a comparison between human T cell data and the positive HCV epitopes. FDR was applied to the p-values to correct for multiple testing errors.

	position	Fisher p-value	amino acid	MW p-value
HUMAN$^+$ and HCV$^+$	3	0.0048242837	E	
	3	0.0409545159	F	
	4	0.0340193345	E	
	5	0.0347537615	Q	
	5	0.0460295601	W	
	8	0.0139489695	T	
	9	0.0340193345	Q	
	14	0.0239590308	N	
	14	0.0275357767	E	
	14	0.0398472595	H	
	15	0.0239590308	H	
	15	0.0454598032	Q	
	15	0.0044454765	E	
	16	0.0001687577	E	
	16	0.0203958021	A	
	16	0.0453318905	K	
	17	0.0239590308	K	
	19	0.0212912181	A	
	20	0.0453318905	G	
	20	0.0464373392	A	
	20	0.0239590308	D	
	21	0.0203958021	A	
	21	0.0282174805	E	
	22	0.0044454765	E	
	23	0.0239590308	H	
	23	0.0261121552	A	
	25	0.0122052311	A	
	25	0.0398472595	K	
	25	0.0454598032	Q	
	26	0.0239590308	V	
	26	0.0275357767	K	
	27	0.0275357767	E	
	27	0.0479014747	G	
	28	0.0212836442	A	
	28	0.0239590308	G	
	28	0.0331642414	E	
	30	0.002149528	A	

List of Figures

List of Tables

List of Algorithms

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Own contribution, co-authors and collaborators for the published manuscripts

`SeqFeatR` for the discovery of feature-sequence associations

This chapter is based on the following publication:

Bettina Budeus, Jörg Timm, and Daniel Hoffmann (2015). **SeqFeatR for the discovery of feature-sequence associations.**

http://www.plosone.org/article/related/info%3Adoi%2F10.1371%2Fjournal.pone.0146409

BB developed the tool and R-package and wrote parts of the manuscript and tutorial. DH wrote most parts of the manuscript and tutorial and supervised the development. JT established the tool in his lab and verified the results in vitro.

Complexity of the human memory B cell compartment is determined by the versatility of clonal diversification in germinal centres

This chapter is based on the following publication:

Bettina Budeus, Stefanie Schweigle, Martina Przekopowitz, Daniel Hoffmann, Marc Seifert, and Ralf Küppers (2015). **Complexity of the human memory B-cell compartment is de-termined by the versatility of clonal diversification in germinal centers.**

http://www.pnas.org/content/early/2015/08/28/1511270112.long

BB processed the next generation sequences, analyzed the results, created most of the figures and tables except Fig. S1A and B and Tab. S2 and Tab. S4 and contributed to the manuscript. SS purified the blood samples, sorted the cells, did a PCR and created the tree files as model for the tree figures. MP analyzed BCL6. DM provided ideas and contributed to the manuscript. MS and RK wrote most parts of the manuscript and provided analytical insights and support.

Mutations rtP177G and rtF249A in the reverse transcriptase domain of hepatitis B virus polymerase confer resistance to tenofovir

This chapter is based on the following publication:

Bo Qin, Bettina Budeus, Liang Cao, Chunchen Wua, Yun Wang, Xiaoyong Zhang, Simon Rayner, Daniel Ho˙mann, Mengji Lu, Xinwen Chen (2013). **The amino acid substitutions rtP177G and rtF249A in the reverse transcriptase domain of hepatitis B virus polymerase reduce the susceptibility to tenofovir**. Antiviral Research 97.

http://www.sciencedirect.com/science/article/pii/S0166354212002872

BB created the model of the HBV polymerase, analyzed it regarding the chosen mutations and contributed to manuscript preparation.

Erklärung:
Hiermit erkläre ich, gem.§ 9 der Promotions ordnung der Math.-Nat. Fakultäten zur Erlangung der Dr.rer.nat., dass die oben genannten Beteiligungen an den wissenschaftlichen Veröffentlichungen der Wahrheit entsprechen.

Plankstadt, den _____

Unterschrift der Doktorandin

Unterschrift d. wissenschaftl. Betreuers/ Mitglieds der Universität Duisburg-Essen

Erklärung:

Hiermit erkläre ich, gem.§ 9 der Promotions ordnung der Math.-Nat. Fakultäten zur Erlangung der Dr.rer.nat., dass von mir die urheber- und lizensrechtliche Seite (Copyright) geklärt wurde und Rechte Dritter der Publikation nicht entgegenstehen.

Plankstadt, den_____

Unterschrift der Doktorandin

Erklärung:

Hiermit erkläre ich, gem.§6 Abs.2, f der Promotions ordnung der Math.-Nat. Fakultäten zur Erlangung der Dr.rer.nat., dass ich das Arbeitsgebiet, dem das Thema "**Statistical analysis of sequence populations in virology and immunology**" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Bettina Budeus befürworte.

Plankstadt, den _____

_____	_____
Name des wissenschaftlichen Betreuers/ Mitglieds der Universität Duisburg-Essen	Unterschrift d. wissenschaftl. Betreuers/ Mitglieds der Universität Duisburg-Essen

Erklärung:

Hiermit erkläre ich, gem.§7 Abs.2, c und e der Promotionsordnung der Math. -Nat. Fakultäten zur Erlangung des Dr.rer.nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient habe und alle wörtlich oder inhaltlich übernommenen Stellen als solche gekennzeichnet habe.

Plankstadt, den _____

_____

Unterschrift der Doktorandin

Erklärung:

Hiermit erkläre ich, gem.§7 Abs.2, d und f derPromotionsordnung der Math. -Nat. Fakultäten zur Erlangung des Dr.rer.nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe, dass diese Arbeit von keiner anderen Fakultät abgelehnt worden ist, und dass ich die Dissertation nur in diesem Verfahren einreiche.

Plankstadt, den _____

_____

Unterschrift der Doktorandin