Computational Approaches for Improving Treatment and Prevention of Viral Infections

Dissertation zur Erlangung des Grades des
Doktors der Naturwissenschaften
an der Fakultät für Mathematik und Informatik der
Universität des Saarlandes

von
Matthias Döring

Saarbrücken
2019

Tag des Kolloqiums	30.04.2019
Dekan	Prof. Dr. Sebastian Hack
Vorsitzender der Prüfungskommission	Prof. Dr. Gerhard Weikum
Erster Berichterstatter	Prof. Dr. Nico Pfeifer
Zweiter Berichterstatter	Prof. Dr. Dr. Thomas Lengauer
Dritter Berichterstatter	Prof. Dr. Olga Kalinina
Akademischer Mitarbeiter	Dr. Peter Ebert

Abstract

The treatment of infections with HIV or HCV is challenging. Thus, novel drugs and new computational approaches that support the selection of therapies are required. This work presents methods that support therapy selection as well as methods that advance novel antiviral treatments.

geno2pheno[ngs-freq] identifies drug resistance from HIV-1 or HCV samples that were subjected to next-generation sequencing by interpreting their sequences either via support vector machines or a rules-based approach. geno2pheno[coreceptor-hiv2] determines the coreceptor that is used for viral cell entry by analyzing a segment of the HIV-2 surface protein with a support vector machine. openPrimeR is capable of finding optimal combinations of primers for multiplex polymerase chain reaction by solving a set cover problem and accessing a new logistic regression model for determining amplification events arising from polymerase chain reaction.

geno2pheno[ngs-freq] and geno2pheno[coreceptor-hiv2] enable the personalization of antiviral treatments and support clinical decision making. The application of openPrimeR on human immunoglobulin sequences has resulted in novel primer sets that improve the isolation of broadly neutralizing antibodies against HIV-1. The methods that were developed in this work thus constitute important contributions towards improving the prevention and treatment of viral infectious diseases.

Zusammenfassung

Die Behandlung von HIV- oder HCV-Infektionen ist herausfordernd. Daher werden neue Wirkstoffe, sowie neue computerbasierte Verfahren benötigt, welche die Therapie verbessern. In dieser Arbeit wurden Methoden zur Unterstützung der Therapieauswahl entwickelt, aber auch solche, welche neuartige Therapien vorantreiben.

geno2pheno[ngs-freq] bestimmt, ob Resistenzen gegen Medikamente vorliegen, indem es Hochdurchsatzsequenzierungsdaten von HIV-1 oder HCV Proben mittels Support Vector Machines oder einem regelbasierten Ansatz interpretiert. geno2pheno[coreceptor-hiv2] bestimmt den HIV-2 Korezeptorgebrauch dadurch, dass es einen Abschnitt des viralen Oberflächenproteins mit einer Support Vector Machine analysiert. openPrimeR kann optimale Kombinationen von Primern für die Multiplex-Polymerasekettenreaktion finden, indem es ein Mengenüberdeckungsproblem löst und auf ein neues logistisches Regressionsmodell für die Vorhersage von Amplifizierungsereignissen zurückgreift.

geno2pheno[ngs-freq] und geno2pheno[coreceptor-hiv2] ermöglichen die Personalisierung antiviraler Therapien und unterstützen die klinische Entscheidungsfindung. Durch den Einsatz von openPrimeR auf humanen Immunoglobulinsequenzen konnten Primersätze generiert werden, welche die Isolierung von breit neutralisierenden Antikörpern gegen HIV-1 verbessern. Die in dieser Arbeit entwickelten Methoden leisten somit einen wichtigen Beitrag zur Verbesserung der Prävention und Therapie viraler Infektionskrankheiten.

Acknowledgements

This dissertation would not have been possible without the guidance and support from several people that I would like to acknowledge in the following paragraphs. I am grateful to Nico Pfeifer and Thomas Lengauer for the opportunity of pursuing a PhD at the Max Planck Institute for Informatics. It was a privilege to work in an environment that allowed me to fully immerse myself in research, free from other concerns. The expert guidance and support provided by Nico and Thomas were integral to shaping my doctoral research. I am particularly thankful for Nico's kind mentorship and continual support.

I would like to express my appreciation for many people at the Max Planck Institute for Informatics. For providing helpful advice before and during my time as a PhD student, I would like to thank Bastian Beggel, Glenn Lawyer, and Markus List. For the fruitful and stimulating collaboration on the geno2pheno web service, I would like to express my gratitude to Joachim Büch and Georg Friedrich. For their great technical support, I am indebted to Achim, Georg, and everyone working in the Information Services and Technology group. I would like to thank Alejandro Pironti and Prabhav Kalaghatgi for their contributions to my work; geno2pheno[ngs-freq] would not have been possible without the groundwork that was laid by Alejandro. For her administrative work, I am thankful to Ruth Schneppen-Christmann who always helped me with organizing my conference trips.

I am grateful to a multitude of people outside Max Planck Society. I am indebted to Rolf Kaiser for his openness, inventiveness, and the organization of platforms such as the AREVIR meeting and the Rettenstein symposium. For their input to my scientific work, I would like to highlight Eva Heger, Elena Knops, Florian Klein, and Christoph Kreer. Additionally, I would like to acknowledge the contributions from Pedro Borrego, Ricardo Camacho, Martin Däumer, Josef Eberle, Meryem Seda Ercanoglu, Nathalie Lehnen, Andreia Martins, Martin Obermeier, Philipp Schommers, Saleta Sierra-Aragon, Simone Susser, Alexander Thielen, and Nuno Taveira. I would also like to thank the following organizations that funded my doctoral research: Max Planck Society, the German Ministry of Health (MASTER-HIV/HEP), the European HIV Coreceptor Study Panel (EucoHIV), and the German Center for Infection Research (HCV Treatment optimization). Without governmental funding, this work would not have been possible.

Heartful thanks go to all my peers at the Center for Bioinformatics in Saarbrücken who always were a source of inspiration, motivation and joyfulness; most notably Peter Ebert, Anna Hake, Lisa Handl, Sivarajan Karunanithi, Tim Kehl, Fabian Müller, Sarvesh Nikumbh, Alejandro Pironti, Florian Schmidt, Lara Schneider, Nora Katharina Speicher, and Thorsten Will. I would like to highlight Peter Ebert, Anna Hake, Eva Heger, Edith Heiter, Elena Knops, and Nora Katharina Speicher for providing feedback on this dissertation. For their companionship during my studies I am deeply thankful to Miriam Bah, Max Fischer, Alexander Junge, Sebastian Keller, and Andreas Mohr. Above all, I would like to thank my family. I am indebted to my parents for their continual support and to my brother for sparking my interest in computers. Finally, I would like to thank Edith Heiter for always being there for me. I cannot imagine anything better than coming home and seeing your smiling face.

Contents

List of Figures 13

List of Tables 17

List of Algorithms 19

1 *Introduction* 21

I BACKGROUND 27

2 *Virological and Immunological Foundations* 29

2.1 *Viruses and Viral Pathogenesis* 30

2.2 *Defense Mechanisms against Viruses* 32

2.2.1 Components of the Immune System 32
2.2.2 Overview of the Adaptive Immune System 33
2.2.3 Cells of the Adaptive Immune System. 33
2.2.4 Antibody Structure and Function 35
2.2.5 Neutralizing Antibodies 36

2.3 *Human Immunodeficiency Virus* 37

2.3.1 Introduction to HIV 38
2.3.2 Structure and Genome Organization 38
2.3.3 Life Cycle 40
2.3.4 Coreceptor Usage 42

2.3.5 Transmission and Course of Infection 45
2.3.6 Treatment of HIV Infection 46
2.3.7 Neutralizing Antibodies and Treatment 51
2.4 *Hepatitis C Virus* 54
2.4.1 Introduction to HCV 54
2.4.2 Structure and Genome Organization 55
2.4.3 Life Cycle 55
2.4.4 Transmission and Course of Infection 57
2.4.5 Treatment of HCV Infection 58
2.5 *Molecular Techniques* 60
2.5.1 Phenotypic HIV Coreceptor Testing 60
2.5.2 Phenotypic Resistance Testing. 62
2.5.3 Sequencing of Viral Genomes 66
2.5.4 Polymerase Chain Reaction. 69
3 *Methodological Foundations* 73
3.1 *Overview of Machine Learning* 73
3.2 *Supervised Learning* 74
3.2.1 Preliminaries 74
3.2.2 Supervised Learning as Function Estimation . . . 75
3.2.3 The Bias-Variance Decomposition 76
3.2.4 Training and Test Errors 77
3.2.5 Estimating Model Errors. 78
3.2.6 Limiting Model Complexity via Regularization . . 81
3.2.7 Feature Selection 81
3.3 *Measures of Predictive Performance* 83
3.3.1 The Confusion Matrix. 83
3.3.2 Performance Measures for Non-Scoring Classifiers . 84
3.3.3 Performance Measures for Scoring Classifiers . . . 85
3.4 *Models for Supervised Learning* 86
3.4.1 Logistic Regression. 86
3.4.2 Support Vector Machines 87
3.5 *Clustering* 93
3.5.1 *K*-Means Clustering 93
3.5.2 Hierarchical Clustering 95
3.6 *Statistical Significance Tests* 96
3.6.1 Introduction to Hypothesis Testing 97
3.6.2 McNemar's Test 98
3.6.3 Fisher's Exact Test 98

3.6.4 Wilcoxon Rank-Sum Test 99
3.6.5 Multiple Hypothesis Testing 99

3.7 *Optimization with Linear Programs* 101

3.7.1 Linear and Integer Linear Programming 101
3.7.2 Branch and Bound 102
3.7.3 The Set Cover Problem 102

II Contributions 105

4 *Interpreting Drug Resistance from Next-Generation Sequencing Data* 107

4.1 *Genotypic Resistance Testing* 109

4.1.1 Overview 109
4.1.2 Resistance Testing with geno2pheno[resistance] . . 111
4.1.3 Resistance Testing with geno2pheno[hcv] 116
4.1.4 Resistance Testing with NGS Data 117

4.2 *Approach of geno2pheno[ngs-freq]* 119

4.2.1 Format of Input Files 120
4.2.2 Prevalence Cutoffs 121
4.2.3 Workflow 121
4.2.4 Quality Control Mechanisms 126
4.2.5 Visualization of Drug Resistance 127
4.2.6 Validation 128
4.2.7 Technical Details 129

4.3 *Case Studies* 130

4.3.1 HIV-1 Resistance Interpretation 130
4.3.2 HCV Resistance Interpretation 132

4.4 *Discussion* 133

5 *Predicting HIV-2 Coreceptor Usage* 139

5.1 *Genotypic Methods for Coreceptor Identification* 141

5.1.1 Identification of HIV-1 Coreceptor Usage 141
5.1.2 Identification of HIV-2 Coreceptor Usage 143

5.2 *Model Development* 143

5.2.1 Phenotypic Measurements 144
5.2.2 Data Collection and Processing 145
5.2.3 Characteristics of the V3 Data Set 148

5.2.4 Model Selection and Validation 148
5.2.5 Development of geno2pheno[coreceptor-hiv2] . . . 149
5.3 *Results* 152
5.3.1 Model Selection and Performance 152
5.3.2 Validation of the Rules-Based Approach 153
5.3.3 Comparison of Predictions from SVMs and Rules . 154
5.3.4 Learning Discriminatory Features from SVMs . . . 155
5.3.5 Separation of R5 and X4-Capable Variants 156
5.3.6 The geno2pheno[coreceptor-hiv2] Web Server . . . 157
5.3.7 Validation of the Web Service 158
5.4 *Discussion* 159
6 *Designing Multiplex PCR Primers for Human Immunoglobulins* 163
6.1 *Multiplex Primer Design* 165
6.1.1 Biophysical Characteristics of PCR 165
6.1.2 Related Work 167
6.1.3 Technical Details of openPrimeR 170
6.2 *Evaluation of Primers* 170
6.2.1 Preliminaries 171
6.2.2 Detection of Amplification Events 172
6.2.3 Physicochemical Constraints 174
6.2.4 Metrics for Physicochemical Constraints 176
6.3 *Primer Design and Selection Algorithms* 179
6.3.1 Primer Initialization 179
6.3.2 Primer Filtering 181
6.3.3 Primer Set Optimization 183
6.3.4 Local Primer Design 191
6.3.5 Global Primer Design 192
6.3.6 Feasibility of Primer Design 192
6.3.7 Optimal Subset Selection 197
6.4 *Primer Design for Human Immunoglobulin Sequences* 200
6.4.1 Data for Immunoglobulin Primer Design. 200
6.4.2 *In Silico* Validation 200
6.4.3 *In Vitro* Validation 202
6.5 *Discussion* 208
7 *Predicting PCR Amplification Events* 213
7.1 *PCR Data, Features, and Statistical Models* 215

7.1.1 Template Design and PCR Measurements 215
7.1.2 Data Set Construction. 216
7.1.3 Feature Encoding 217
7.1.4 Logistic Regression Models. 218
7.1.5 Validation of Models and Classifiers 219
7.2 *Results* 220
7.2.1 Properties of the Data Set 221
7.2.2 Logistic Regression Models. 223
7.2.3 Comparison of Model and Classifier Performance . 223
7.2.4 Interpretation of *TMM* 225
7.3 *Discussion* 226

III Synthesis 229

8 *Conclusions* 231

9 *Outlook* 235

IV Appendix 241

A *Supplementary Material* 243

B *Publications* 253

C *Plagiarism Prevention* 255

Glossaries 257
General Terms 257
Drugs against HIV 259
Drugs against HCV 261
Acronyms 263

Bibliography 269

List of Figures

2.1 Size of viruses 31
2.2 Cytopathic effect of herpes simplex virus infection . 31
2.3 B cell activation 33
2.4 Structure of an antibody 35
2.5 Mechanisms of antibodies 37
2.6 Structure of HIV. 39
2.7 Organization of the HIV-1 genome 39
2.8 Life cycle of HIV 41
2.9 HIV cell entry 42
2.10 Course of HIV infection 45
2.11 An example of Kaposi's sarcoma. 46
2.12 Development of drug resistance 47
2.13 Timeline of HIV drugs 49
2.14 Distribution of HIV neutralizers 52
2.15 Envelope epitopes of HIV-1 antibodies 53
2.16 HCV genome organization 55
2.17 HCV Life Cycle 56
2.18 A person with jaundice 57
2.19 Characteristics of a cirrhotic liver 58
2.20 The resistance factor 64
2.21 Sequencing via the Sanger method 67
2.22 Chemistry of Illumina/Solexa sequencing 68
2.23 Polymerase chain reaction 70

3.1 Training and test errors 77
3.2 Hold-out validation 80
3.3 Cross validation 80
3.4 Nested cross validation 80
3.5 Example for the AUC 86
3.6 Hyperplanes of SVMs. 88
3.7 The epsilon-insensitive tube in SVR. 90
3.8 *K*-means clustering. 93
3.9 The elbow test 94

3.10 Hierarchical clustering 95
3.11 A 2-dimensional polytope 101

4.1 From frequency files to consensus sequences . . . 120
4.2 Workflow of geno2pheno[ngs-freq] 122
4.3 Translation of codons to amino acids 124
4.4 Resistance table of geno2pheno[resistance] 127
4.5 Architecture of geno2pheno[ngs-freq] 129
4.6 Visualizing HIV-1 drug resistance using NGS data . 131
4.7 Table of resistance mutations for HIV-1 132
4.8 Visualizing HCV drug resistance using NGS data . 137
4.9 Table of resistance mutations for HCV. 137

5.1 X4-probabilities for V3 sequences (CXCR4 markers). 155
5.2 Distribution of X4-probabilities 156
5.3 Visualization of V3 amino acid features 157

6.1 Variability in different antibody regions 164
6.2 The set cover problem 183
6.3 Greedy state 0. 185
6.4 Greedy state 1. 185
6.5 Greedy state 2. 185
6.6 Greedy state 3. 185
6.7 Greedy state 4. 185
6.8 Reference beta distributions 198
6.9 Representation of primer design difficulty 199
6.10 IGHV primer binding regions 202
6.11 IGHV primer set constraints and coverage 203
6.12 Validation on germline IGHVs 204
6.13 Validation on IGHVs from single B cells 206
6.14 Validation on IGHVs of an HIV-1-infected person . 208

7.1 Construction of PCR data set 216
7.2 Encodings for primer-template 3′ mismatches . . . 218
7.3 Impact of primers on the amplification status . . . 221
7.4 Relationship between variables in LR_2 223
7.5 Performance of models that predict amplification . 224
7.6 Visualization of *TMM*. 226

A.1 Decision values for HIV-1 coreceptor prediction . . 244
A.2 Subset coverage of IGHV primers 247
A.3 Constraint deviation of IGHV primers. 248
A.4 IGKV primer set constraints and coverage 249
A.5 IGKV primer binding regions 250
A.6 IGLV primer set constraints and coverage 251

A.7 IGLV primer binding regions 252

List of Tables

2.1 Drugs against HCV 59

3.1 Structure of a confusion matrix 83
3.2 Structure of a contingency table 98

5.1 V3s with discordant annotations of coreceptor use . 146
5.2 Distribution of data set observations 148
5.3 Performance of SVMs for HIV-2 coreceptor testing . 153
5.4 Performance of the rules-based approach 154
5.5 Predictive performance of individual rules 154
5.6 Predictive features of the linear SVM 156
5.7 Performance of geno2pheno[coreceptor-hiv2] . . . 159

6.1 Overview of multiplex primer design tools 169
6.2 Default coverage settings 173
6.3 Permissible properties of PCR primers 174
6.4 PCR settings of openPrimeR 175
6.5 Constraint fulfillment distribution 177
6.6 Structure of the constraint fulfillment matrix . . . 177
6.7 Overview of designed IGHV primers 201
6.8 Overview of designed IGKV primers 201
6.9 Overview of designed IGLV primers 201
6.10 Overview of immunoglobulin templates 202

7.1 Primers used in the PCRs 215
7.2 Distribution of data set labels 217
7.3 Primer set properties 222
7.4 Rates of amplification 222
7.5 Significance of features 225
7.6 Cutoffs for coverage models 225

A.1 IUPAC nucleotide ambiguity codes 243
A.2 Primer design settings for IGH, IGK, and IGL . . . 244
A.3 Overview of available IGHV primer sets 245

A.4 Overview of available IGKV primer sets 246
A.5 Overview of available IGLV primer sets 247

C.1 Results of plagiarism detection 256

List of Algorithms

1 Nondegenerate primer initialization 180
2 Degenerate primer initialization 182
3 Filtering of primers 184
4 Greedy primer optimization 188
4 Greedy primer optimization (continued) 189
4 Greedy primer optimization (continued) 190
5 ILP primer optimization 191
6 Primer design for a single strand 193
7 Generalized primer design procedure 194
8 Global primer design 195
9 Classifying the feasibility of a primer design task 198

1
Introduction

> Information is a source of learning. But unless it is organized, processed, and available to the right people in a format for decision making, it is a burden, not a benefit.
>
> William Pollard

In 2017, there were more than 36 million people living with human immunodeficiency virus (HIV)[1]. If left untreated, persons in the final stage of infection develop acquired immunodeficiency syndrome (AIDS), which is lethal. In the year 2017 alone, just below 1 million people died due to AIDS-related causes[2]. In order to reduce the disease burden of HIV infection[3], it is necessary to reinforce our efforts in preventing, diagnosing, and treating the infection. Antiretroviral treatment is important for two reasons. First, effective treatments enable HIV-positive persons to live long and healthy lives rather than succumbing to the infection[4]. Second, since infected persons with durably suppressed viral loads[5] cannot transmit the infection[6], successful treatments also prevent further transmissions[7]. The aim of this dissertation is the development of novel computational methods that can improve treatment and prevention of viral infections. The developed approaches fall into two categories. The first category comprises approaches that can guide the selection of antiviral treatments and the second category deals with approaches that can advance the development of novel antiviral agents.

Treatment decisions are typically informed by sequencing of the viral genome, which allows for inferences about the properties of the virus. A common application of sequencing is the genotypic identification of viral drug resistance. Human immunodeficiency virus type 1 (HIV-1) drug resistance testing is crucial because it can allow for the selection of effective antiretroviral drugs even when resistance mutations are detected[8]. Although HIV is the most prominent virus for which drug resistance is a concern, there are also other prevalent viral infections such as the hepatitis C virus (HCV)[9], where drug resistance is an issue[10]. Thus, this dissertation is not only concerned with drug resistance of the HIV-1 but also of HCV. The identification of drug resistance can be informed by expert opinions. In order to determine which mutations are associated with drug resistance,

[1] UNAIDS 2018

[2] UNAIDS 2018

[3] Lopez et al. 2006

[4] Antiretroviral Therapy Cohort Collaboration et al. 2017

[5] A viral load is suppressed if the plasma concentration of HIV RNA is below the level of detection.

[6] Cohen et al. 2015; Rodger et al. 2016

[7] Cohen et al. 2011

[8] Lucas 2005

[9] There are approximately 170 million chronic HCV infections according to Mohd Hanafiah et al. (2013).

[10] Holmes and Thompson 2015

physicians and virologists regularly form panels to discuss evidence from the literature and to exchange their personal experiences. These meetings give rise to rules-based approaches, which rely on expertly crafted sets of rules in order to identify whether a viral sequence is associated with drug resistance or not.

The combination of increasing data availability and technological innovation has ushered in another approach that is based on supervised learning, a type of machine learning that is concerned with learning from data in order to fit predictive models.[11] Statistical models for the identification of drug resistance have two advantages over rules-based approaches. First, they can identify novel associations between viral genotype and resistance phenotype from the data. Second, statistical models can be updated with little human intervention. However, new prediction scenarios also require new computational approaches. Predictive models that identify drug resistance were previously only applicable to viral sequences obtained through Sanger sequencing. Currently, next-generation sequencing (NGS) is in the process of supplanting Sanger sequencing for the purpose of drug resistance testing. In this work, a drug resistance test for NGS samples from HIV-1 or HCV was developed, which tackles two challenges. The first challenge involves processing the thousands of reads that are generated by NGS. The second challenge concerns the interpretation of drug resistance from NGS data. Since resistance mutations may be found only in a small number of reads, it is necessary to interpret drug resistance under consideration of the read prevalence.

A further, HIV-specific problem that can inform treatment decisions is the identification of coreceptor usage. Since HIV needs to interact with a coreceptor before it can enter human immune cells, drugs blocking the coreceptor — so-called coreceptor antagonists — are used to prevent HIV cell entry. Because HIV can use different types of coreceptors, it is necessary to identify which coreceptor is used before prescribing coreceptor antagonists.[12] Although several models for identifying HIV-1 coreceptor usage are available[13], human immunodeficiency virus type 2 (HIV-2) has been overlooked for a long time. Therefore, the provision of a model for identifying HIV-2 coreceptor usage was another aim of this dissertation.[14] The main challenge in this project was the selection of features and models that afford high predictive accuracies.

Besides supporting the selection of antiviral treatments, this dissertation also aims at advancing the development of novel antiviral agents. Although the autologous immune response against HIV-1 infection is insufficient for controlling the infection, certain heterologous antibodies enable promising new treatment strategies.[15] The

[11] Predictive models that are based on machine learning are called statistical models in the following.

[12] HIV coreceptor usage can also be considered a resistance phenotype as the use of a certain coreceptor may preclude the use of available coreceptor antagonists.

[13] Pillai et al. 2003; Jensen et al. 2003; Lengauer et al. 2007; Pfeifer and Lengauer 2012; Thielen and Lengauer 2012

[14] HIV-2 drug resistance was not investigated in this work due to the scarcity of the data. However, a rules-based approach is available in form of the HIV2-EU resistance interpretation engine, which can be accessed via HIV-GRADE.

[15] An antibody is called *autologous* if it targets only autologous viruses, that is, viruses found in the host where the antibody emerged. An antibody is called *heterologous* if it can inhibit heterologous viruses, that is, variants that are not present in the viral population found in the host that generated the antibody.

isolation of novel antibodies hinges on our ability to amplify immunoglobulin transcripts using multiplex polymerase chain reaction (mPCR). The success of mPCR critically depends on the used combination of primers, oligomers that are complementary to the template sequences, which are required for polymerase attachment and elongation. Available sets of primers for the amplification of immunoglobulin sequences may preclude the amplification of highly mutated HIV-1-specific antibody sequences. Since established mPCR primer design tools were found incapable of producing suitable primers, a further aim of this work was to develop a new primer design approach for the design of mPCR primers that improve the amplification of highly mutated antibodies to HIV-1. An important aspect of this work was to ensure that all of the germline immunoglobulin sequence variants can be amplified using the designed primers. For this purpose, it was necessary to develop optimization algorithms for selecting minimum sets of primers maximizing the coverage of templates.

Outline

This dissertation is structured into four parts. Part I provides the biological and methodological background via Chapters 2 and 3, respectively. Chapter 2 introduces the virological and immunological foundations of this work. This background chapter begins with an introduction to viruses (Section 2.1) and host defense mechanisms (Section 2.2). In the following sections, the characteristics of HIV (Section 2.3) and HCV (Section 2.4) are described. The chapter concludes with the molecular techniques that are essential for virological research (Section 2.5).

In Chapter 3, I offer the methodological foundations of my work. After providing an overview of machine learning (Section 3.1), I introduce important concepts from supervised learning (Section 3.2), various measures of predictive performance (Section 3.3), and models for prediction such as logistic regression and support vector machines (Section 3.4). The unsupervised learning technique of clustering is illustrated in terms of K-means and hierarchical clustering (Section 3.5). Since significance tests are particularly useful for the comparison of predictive approaches, I describe the principles of statistical hypothesis testing (Section 3.6). At the end of the chapter (Section 3.7), the reader is familiarized with optimization through linear programs, which are exemplified through the set cover problem.

Part II comprises the four original scientific contributions of this dissertation, which are structured into Chapters 4, 5, 6, and 7. Chapter 4 introduces geno2pheno[ngs-freq], a web service for the identification of drug resistance from NGS samples of HIV-1 or HCV.

After an introduction to genotypic resistance testing (Section 4.1), I describe the three main contributions of geno2pheno[ngs-freq]: decoupling of data preprocessing and interpretation, provision of access to the models of geno2pheno[resistance] and geno2pheno[hcv] in the NGS setting, and improvement of model interpretability (Section 4.2). The clinical usefulness of the developed approach is exemplified through two cases studies (Section 4.3). The chapter concludes with a critical discussion of the benefits and limitation of using geno2pheno[ngs-freq] for the detection of drug resistance in minor viral populations (Section 4.4).

Chapter 5 is concerned with the prediction of HIV-2 coreceptor usage. The chapter begins with an overview of established genotypic methods for the identification of HIV-1 and HIV-2 coreceptor usage (Section 5.1). I then describe the development of a support vector machine for the prediction of HIV-2 coreceptor usage (Section 5.2), which has been made available in terms of geno2pheno[coreceptor-hiv2], the first web service for this task (Section 5.3). The discussion (Section 5.4) focuses on differences between the molecular markers of HIV-1 and HIV-2 coreceptor usage and tries to explain the conflicting results that were obtained from phenotypic coreceptor assays.

Chapter 6 describes the work that led to openPrimeR, a program for evaluating and designing primer sets for mPCR. To establish the background, I first describe requirements and established approaches to mPCR primer design (Section 6.1). Next, I formulate methods for the evaluation of primers (Section 6.2) and offer algorithms for the design and selection of mPCR primers (Section 6.3). The usefulness of openPrimeR is demonstrated through its application on human immunoglobulin sequences (Section 6.4). The chapter concludes with a discussion (Section 6.5), in which the obtained results are interpreted and the limitations of openPrimeR are presented.

The molecular determinants of amplification events resulting from polymerase chain reaction are examined in Chapter 7. Novel polymerase chain reaction data were used to construct a logistic regression model for the prediction of amplification events (Section 7.1). The characteristics of successful amplification events are analyzed and the favorable performance of the statistical model is shown (Section 7.2). The discussion (Section 7.3) comprises a commentary on the use of predictive models for primer design.

Part III contains Chapter 8 and Chapter 9, the final chapters of the main matter. Chapter 8 discusses the relevance of the presented approaches and indicates possibilities for future work. Chapter 9 provides the outlook.

The appendix can be found in Part IV. It includes Chapters A, B, and C. Chapter A provides supplemental information. Chapter B lists the scientific contributions that I made during my doctoral phase. Chapter C summarizes the results of running plagiarism detection software on this dissertation. The back matter comprises several glossaries, the index, and the bibliography.

Part I:

Background

This work is at the junction of life and computer sciences. For the appreciation of the developed methods, it is important to have a firm grasp of concepts from several fields. Chapter 2 intends to familiarize the reader with the most important virological and immunological concepts. Chapter 3 introduces concepts from machine learning, statistics, and optimization.

2

Virological and Immunological Foundations

Nature is not human-hearted.

Lao Tzu

2008 was the first year in which noncommunicable diseases caused more deaths than infectious diseases[1] and, in 2013, UNAIDS reported a 33% reduction in new HIV infections compared to 2001[2]. These data signify the progress that has been made in the prevention, diagnosis, and treatment of viral infectious diseases. Still, viral infections remain a major source of disease burden. In 2016, there were more than 35 million HIV-infected people worldwide and 1 million people died of AIDS-related diseases that year alone[3]. HIV and AIDS are particularly prevalent in developing countries: In 2016, 7% of East and Southern African adults were infected with HIV and just under half a million people in this region died of AIDS-related illnesses that year[4]. Moreover, despite years of efforts, there is still no vaccine against HIV[5]. Additionally, since public health efforts have focused on the containment of the HIV epidemic, other viral infections have been overlooked. For example, the number of HCV-related deaths increased from 0.89 million in 1990 to 1.45 million in 2013[6].

HIV and HCV share commonalities with respect to disease progression and treatment. Both infections can remain without symptoms for several years but when symptoms appear, they are often life-threatening and acquired damage may be irreversible. The timely diagnosis of HIV or HCV infection is therefore particularly important. Once diagnosed, infections with these viruses can be treated with a diverse arsenal of potent drugs. A major difference between HIV and HCV infection is that HIV infection is cleared neither spontaneously nor through treatment because the HIV provirus persists in reservoirs of host cells that can be reactivated at any point in time. Thus, HIV-infected persons have to submit to life-long treatments, while HCV-infected persons clear the infection either spontaneously or through an antiviral treatment that typically last twelve weeks.

Since drug resistance can impair the treatment of an HIV or HCV

[1] Organization et al. 2008

[2] UNAIDS 2013

[3] UNAIDS 2016

[4] AVERT 2016

[5] Johnston and Fauci 2008

[6] Wiktor and Hutin 2016

infection, drug resistance testing is relevant in two ways. First, since resistant viral strains can be transmitted, selection of the initial treatment can be guided by the identification of drug resistance. Second, in case of treatment failure due to the emergence of resistant variants during therapy, a new treatment can be selected on the basis of drug resistance testing.

The purpose of this chapter is to summarize the current body of knowledge on HIV and HCV, with a focus on the aspects that are relevant for their treatment and prevention. In order to appreciate the commonalities of viral infections, an introduction to viruses and viral pathogenesis is provided in Section 2.1. Since defense mechanisms of the host immune system are involved in the prevention and pathogenesis of viral infections, Section 2.2 describes the components of the immune system with a focus on the adaptive immune response. Note that the information provided in the aforementioned sections is based on Flint (2004) if not stated otherwise. Section 2.3 introduces HIV and deals with the treatment-relevant aspects of drug resistance, viral coreceptor usage, and the antibody response. The properties of HCV and the consequences of HCV infection are discussed in Section 2.4. The chapter concludes with Section 2.5, which introduces molecular techniques that are crucial for the investigation of viruses and host immune responses such as phenotypic methods for the determination of coreceptor usage and drug resistance, sequencing, and polymerase chain reaction (PCR).

2.1 *Viruses and Viral Pathogenesis*

Viruses are small obligate parasites (i.e. are dependent on a living host organism for reproduction). Lacking the energy-generating and biosynthetic systems that are required for independent existence, viruses are simpler than even the smallest microorganisms. With diameters as small as 100 nm, they are typically much smaller than other pathogens such as bacteria whose diameters are on the μm scale (Figure 2.1).[7]

Viral genomes are composed of either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). All viruses share a replication cycle consisting of three stages. In the first stage, the viral genome is replicated in an appropriate host cell and viral proteins are expressed through the cellular translation machinery. Second, newly synthesized viral proteins assemble to form progeny virions that bud from the host cell. Third, progeny viruses infect new cells by entering an appropriate cell and releasing their genome into the cytoplasm.

Viral Pathogenesis Viral infection is associated with cellular damage[8].

[7] Of course, there are exceptions. For instance, viruses of the *Mimivirdae* family have a relatively large diameter of 400 nm. Klosneuvirus, which belongs to this family, even has a comprehensive translational machinery (Schulz et al., 2017).

[8] Cell damage is defined as any cellular change that limits cellular function.

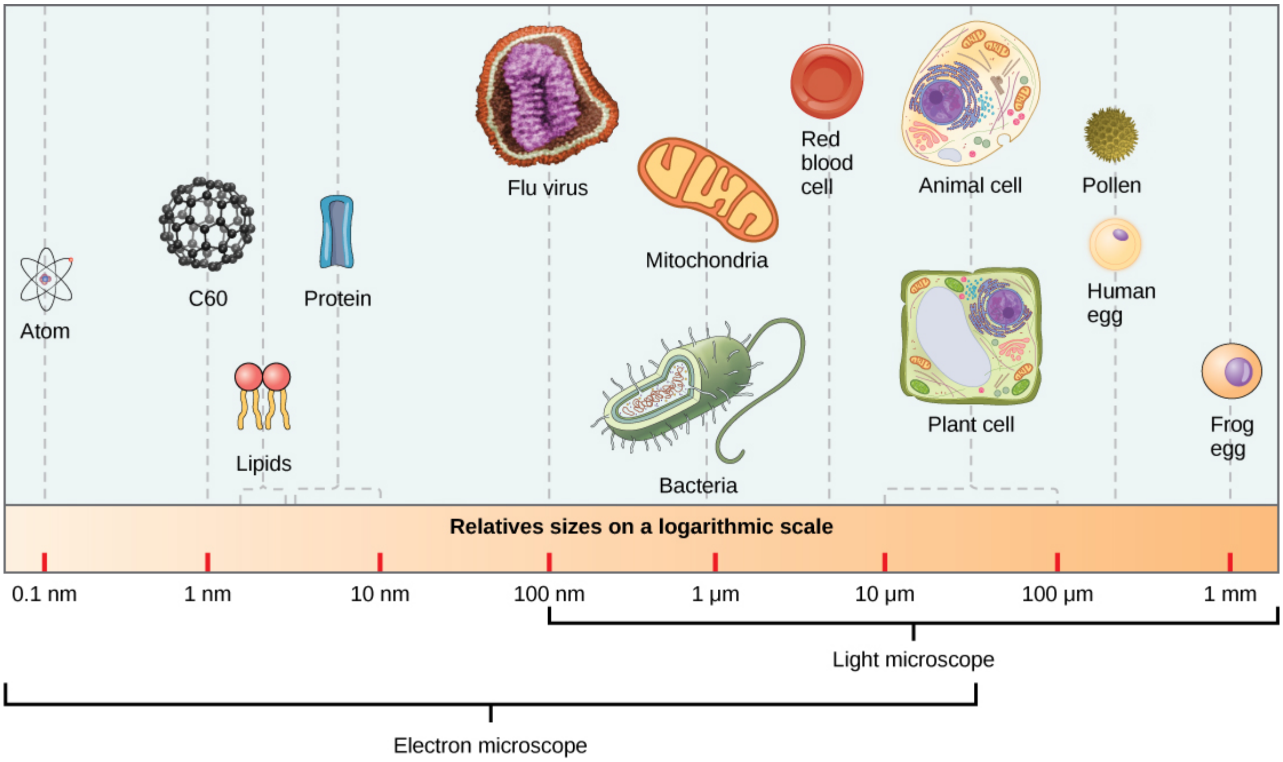


Figure contributed to LibreTexts by OpenStax, licensed under CC BY-NC-SA.

Figure 2.1: Size of viruses.

We can distinguish damage that is a direct consequence of viral infection (*direct damage*) and self-inflicted damage resulting from the immune response (*indirect damage*). Direct damage is often the result of cytopathic effects (structural changes in infected cells, see Figure 2.2 for an example), which are associated with apoptosis[9].

By hijacking (reprogramming) infected cells, viral infection can impede essential cellular processes such as translation, synthesis of DNA and RNA, as well as vesicular transport. Alterations in these processes can lead to the autolytic digestion of infected cells by increasing membrane permeability[10]. Although less common, direct damage can also be a consequence of noncytolytic effects. For example, lymphocytic choriomeningitis virus is a noncytolytic virus that infects cells in the pituitary gland. The pathogenicity of the virus is solely based on reducing the production of growth hormone in infected cells, which can be fatal for the host[11].

The pathology of viruses is largely based on indirect damage that results from the immune response of the host, for example through immunopathological lesions. These lesions can be caused by CD8$^+$ cells, CD4$^+$ cells, and B cells. CD8$^+$ cells can cause lesions due to secretion of perforin, their major cytolytic protein. CD4$^+$ cells can damage cells by recruiting other cells of the immune system such

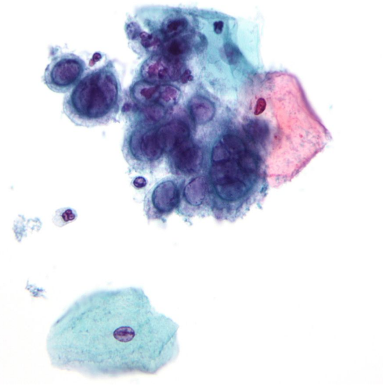


Figure 2.2: Formation of multinucleated giant cells during herpes simplex virus infection.
Adapted from Wikipedia, licensed under CC BY 3.0.

[9] Laurent-Crawford et al. 1991

[10] Note that cell death is not to the advantage of the virus but merely a side effect of viral proliferation.

[11] Rodriguez et al. 1983

as macrophages or neutrophils, which induce cell lysis and cause inflammations. B cells can cause pathological effects when extensive viral replication occurs in sites that are inaccessible to the immune system or when the immune response is unable to clear the infection.

2.2 *Defense Mechanisms against Viruses*

There are several layers of defense mechanisms against viral infections. Before a person can become infected by a virus, it needs to overcome the physical and chemical defense mechanisms of the human body. Only once the virus is inside the body does the innate immune response commence (Section 2.2.1). When the innate immune response cannot contain the infection, the adaptive immune response (Section 2.2.2) begins to develop specific B and T cells (Section 2.2.3). If the adaptive immune response is successful, it will produce antibodies (Section 2.2.4) that are capable of eliminating the pathogen (Section 2.2.5).

2.2.1 *Components of the Immune System*

The primary defense mechanisms against viruses comprise physical and chemical barriers such as the skin. These barriers are assisted by surface-cleansing mechanisms such as blinking (eyelids), swallowing (throat), and the flow of mucus (mucosal system). If intact, the skin is virtually impervious to viral infections as it consists of keratinized cells that are continually shed. Thus, regions of the human body that are not covered by skin such as the respiratory or urogenital tract are at greatest risk of infection. Once a virus breaches the primary defense mechanisms, the immune system begins to intervene.

The immune system relies on innate (non-specific) and adaptive (specific) components. The innate immune system represents the first line of defense as it immediately targets any type of pathogen. The following components are major contributors to the innate immune system: interferons, protein activators that stimulate the immune system; the complement system, a collection of serum proteins that is capable of destroying infected cells as well as pathogens; natural killer (NK) cells, cytolytic lymphocytes that recognize and destroy infected cells; neutrophils, phagocytic granulocytes that migrate to sites of infection during (acute) inflammations; and macrophages, phagocytes that eliminate pathogens during (chronic) inflammations.

The following sections provide more details on the adaptive immune response.

2.2.2 *Overview of the Adaptive Immune System*

The adaptive immune response can be differentiated into the humoral and cell-mediated response. The humoral response involves serum and lymph proteins, while the cell-mediated response requires the action of effector cells. An important characteristic of the adaptive immune system is its ability to discern infected from uninfected cells. This function is facilitated by the major histocompatibility complex (MHC) molecules that are present on the surfaces of cells[12].

Binding of antigens by B-cell and T-cell receptors triggers a cascade of reactions, which involves the production of cytokines, the differentiation of immune cells, the production of antibodies and, eventually, the elimination of the pathogen or pathogen-infected cell. Upon initial infection, it usually takes days or weeks until the adaptive response becomes fully active. However, if a pathogen has been previously encountered, the adaptive response can occur within hours even if the previous infection has been months or years in the past. For this purpose, a subset of lymphocytes called memory cells is maintained after each encounter with a foreign antigen.

The following section sheds light on the cells that are essential for the adaptive immune response.

2.2.3 *Cells of the Adaptive Immune System*

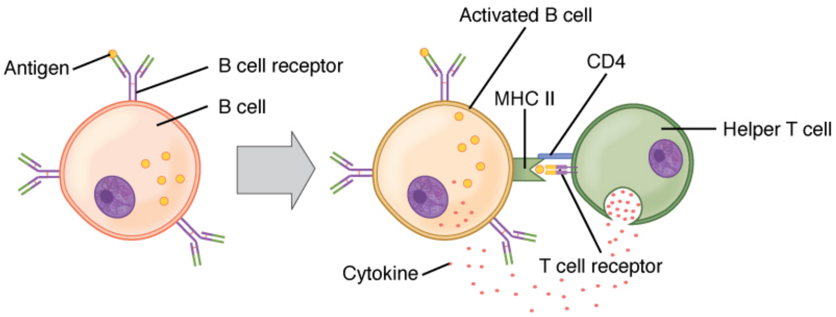


Figure 2.3: B cell activation.

Figure reproduced from *The Adaptive Immune Response: B-lymphocytes and Antibodies* by Rice University, licensed under CC BY 4.0.

The adaptive immune response is initiated by the interplay of lymphocytes and antigen-presenting cells (APCs). Lymphocytes are white blood cells that circulate in the blood and the lymphatic system, often settling in the lymphoid organs that are present throughout the body. Lymph nodes are the centers at which antigen presentation takes place. Almost all cells are APCs because they present endogenous antigens via major histocompatibility complex class I (MHC I). The subset of APCs that performs exogenous antigen presentation through major histocompatibility complex class II (MHC II)

[12] In humans, MHC receptors are encoded through the human leukocyte antigen (HLA) gene complex.

makes up the set of professional APCs. These cells ingest extracellular proteins either via endocytosis (e.g. dendritic cells) or via surface antibodies (e.g. B cells). The major classes of lymphocytes are B cells and T cells.

B Cells B cells originate in the bone marrow and travel to lymph and lymphoid organs where they are activated by binding their specific antigens. Interaction of MHC II with the T-cell receptor of a T helper cell leads to the secretion of T-cell cytokines that stimulate the proliferation and differentiation of B cells into plasma and memory B cells (Figure 2.3). While memory cells are long-lived and express membrane-bound antibodies on their surface, plasma cells exist only a few days and secrete antibodies instead of exposing them on their membranes.

Over the course of an immune response, affinity maturation leads to the development of B cells with increased affinities to antigens. Affinity maturation occurs in the germinal centers of the secondary lymphoid organs and encompasses two processes: somatic hypermutation (SHM) and clonal expansion under affinity-based selection[13].

SHM describes the development of mutations in the variable regions of immunoglobulin genes, particularly in the complementarity-determining regions (CDRs)[14]. During clonal expansion in the germinal centers, B cells that have undergone SHM are competing for resources such as antigen presented by dendritic cells as well as proliferation/survival signals from T helper cells. As a result of this competition, only B cells exhibiting receptors with high antigen affinities are retained.

T Cells T-cell precursors are produced in the bone marrow and mature in the thymus. The maturation process entails positive and negative selection pressure. Positive selection pressure ensures the proliferation of T cells that bind to the specific MHC molecules of the individual, while negative selection pressure eliminates T cells that recognize host peptides. Only approximately 1% to 2% of immature T cells entering the thymus emerge as mature T cells. When the T-cell receptors of a naive T cell interact with peptide fragments presented by MHC, the T cell becomes activated and differentiates into long-lived memory and short-lived effector T cells (e.g. helper and cytotoxic cells, respectively).

Differentiation of naive T cells into effector T cells is determined by their surface antigens, the cluster of differentiation (CD) markers. CD4$^+$ cells that interact with antigen presented through MHC II differentiate into T helper (Th) cells, while CD8$^+$ cells that interact with antigens presented by MHC I differentiate into cytotoxic T lympho-

[13] Victora and Nussenzweig 2012

[14] Yaari et al. 2013

cytes (CTLs). These types of cells perform contrasting functions. Th cells produce growth factors and cytokines that stimulate the recruitment of specific lymphocytes. CTLs recognize antigens presented via MHC I receptors and subsequently destroy the peptide-presenting cell.

There are two types of Th cells: Th1 and Th2 cells. These helper cells release different types of cytokines that influence the activation and proliferation of specific immune cells. Th1 cells promote the cell-mediated proinflammatory response as they enable CTLs to mature. Th2 cells increase the antibody response by stimulating the maturation of immature B cells and resting macrophages. Additionally, they reduce the inflammatory response by producing specific interleukins. The two types of Th cells act like a seesaw — the cytokines produced by one class tend to suppress the generation of cytokines from the other class.

2.2.4 *Antibody Structure and Function*

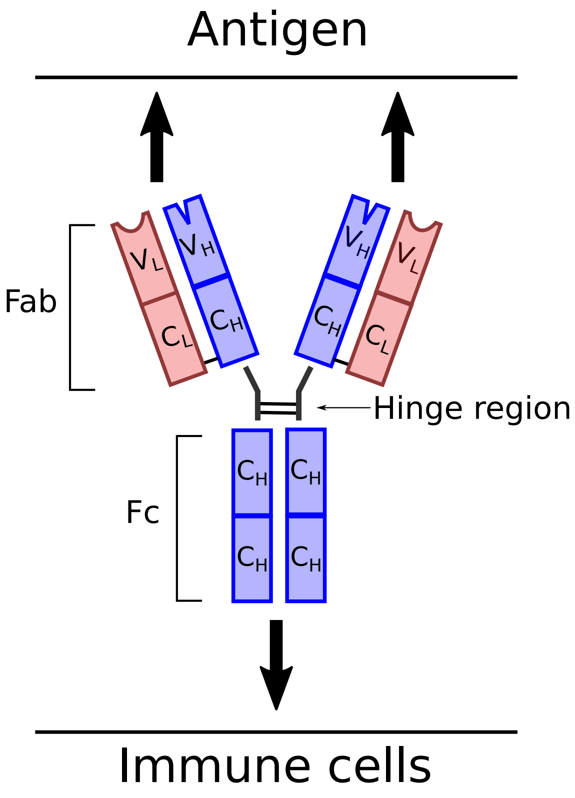


Figure 2.4: Structure of an antibody. The variable regions of the heavy (blue) and light (red) chains are indicated by V_H and V_L, respectively. The corresponding constant regions are indicated by C_H and C_L. The geometric inlets at the end of the variable chains indicate the complementarity-determining regions.

Antibodies (immunoglobulins) are glycoproteins with Y-shaped structures (Figure 2.4) that provide a link between the adaptive and innate components of the immune system[15]. Each antibody consists of two heavy and two light polypeptide chains whose segments can be differentiated into constant and variable regions. The glycosylated constant regions interact with the host immune system and are highly conserved. The variable regions, on the other hand, are involved in antigen recognition and are typically highly mutated.

The two functional components of an antibody can be localized with respect to the central disulfide bonds that make up the hinge region. The constant region that lies on one side of the hinge region is termed the fragment crystallizable (Fc) region. The region that lies on the other side of the hinge region is called the fragment antigen-binding (Fab) region because each arm of Fab contains three CDRs[16]. Due to their involvement in antigen binding, the CDRs of Fab, are particularly adapted to the molecular structure of the antigen that is recognized by the antibody.

There are five classes of immunoglobulins: IgA, IgG, IgD, IgE, and IgM. They are defined by their characteristic heavy chain constant regions (α, γ, δ, ϵ, and μ, respectively). Each antibody class is associated with a specific range of functions. For example, IgG, IgA, and IgM are often produced as a result of viral infections. While IgA has a more localized role as it is important for mucosal immunity, both IgG and IgM are associated with systemic immune responses. IgG is the most abundant antibody and associated with memory responses. IgM is the largest antibody and is involved in the primary immune response[17]. Different classes of antibodies arise by *switching*, that is, recombination of constant region genes during plasma cell differentiation.

2.2.5 *Neutralizing Antibodies*

Neutralization describes the ability of an antibody to prevent the cellular entry of pathogens[18]. Neutralization can only occur when antibody has bound to a sufficient number of pathogen receptors via the CDRs of Fab (Figure 2.5). Mechanisms that allow for neutralization include steric hindrance, target dissociation, and promotion of structural inflexibility in the pathogen's surface proteins[19]. Neutralization by itself can already prevent pathogenic effects[20]. For example, an antibody targeting the viral cluster of differentiation antigen 4 (CD4) binding site can prevent HIV cell entry by disrupting the interaction between CD4 and glycoprotein 120 (gp120).

Since neutralization is a reversible process that does not eliminate pathogens, controlling an infection also requires cell-mediated effector functions on whose basis even non-neutralizing antibodies can support the immune response[21]. The cell-dependent effector

[15] Panda and Ding 2015

[16] The antigen binding site of an antibody is also called paratope.

[17] A primary immune response occurs when the immune system encounters an antigen for the first time.

[18] Klasse 2014

[19] Morris and Moody 2017

[20] Lu et al. 2017

[21] Trkola 2014; Platt et al. 2012

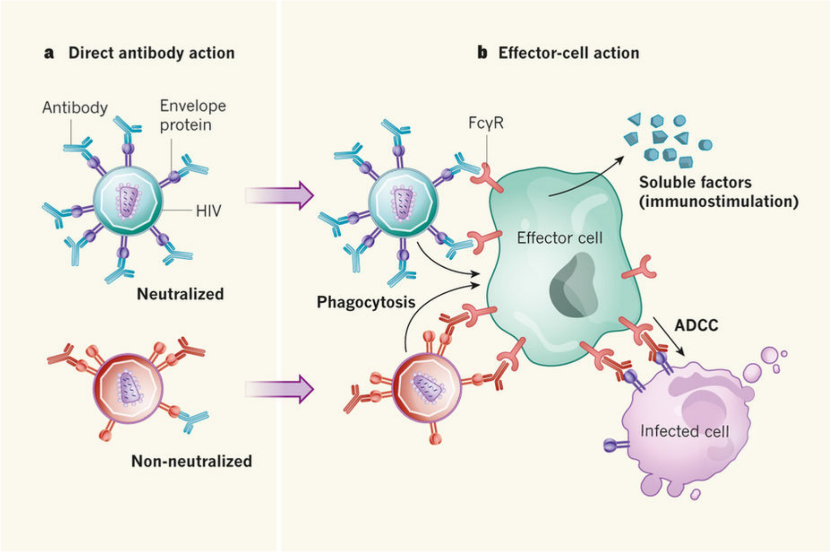


Figure 2.5: Mechanisms of antibodies. Pathogens that have been neutralized are prevented from entering cells. Once a pathogen has been bound by antibodies, cell-dependent effector functions enable pathogen elimination.

Figure reproduced with permission from Springer Nature (Trkola, 2014).

functions of antibodies (Figure 2.5) are complement-dependent cytotoxicity (CDC) and antibody-dependent cell-mediated cytotoxicity (ADCC)[22].

CDC is mediated by the complement system. The interaction of Fc with components of the complement system triggers the complement cascade, which leads to the formation of the membrane attack complex (MAC). Attachment of MAC to the cellular surface of an infected cell results in cell lysis. ADCC is an instance of opsonization, a process in which pathogens are marked for digestion by phagocytes. In ADCC, pathogens are marked by antibodies. The interaction between Fc and the Fcγ receptor of an effector cell such as an NK cell leads to cell lysis through the release of cytotoxins.

2.3 *Human Immunodeficiency Virus*

HIV is a retrovirus that infects human immune cells and causes AIDS. There are two types of HIV, HIV-1 and HIV-2, as introduced in Section 2.3.1. HIV is a parasite whose structural constituents are of viral and human origin (Section 2.3.2). In order to replicate, the virus must hijack suitable host cells (Section 2.3.3). Through the use of different cellular coreceptors, HIV can enter distinct subsets of human immune cells (Section 2.3.4). The spread of HIV is mainly due to its high rate of sexual transmission, while its lethality is due to AIDS, which develops when the infection is not treated (Section 2.3.5). A diverse arsenal of potent antiretroviral agents is capable of controlling the infection

[22] Forthal 2014

(Section 2.3.6). Additionally, novel types of antibodies unlock new treatment and prevention strategies (Section 2.3.7).

2.3.1 Introduction to HIV

HIV, which was discovered[23] in 1984, is a lentivirus[24] of group VI that belongs to the *Retroviridae* family. It infects human immune cells and is the cause of AIDS. HIV is classified into HIV-1 and HIV-2 based on genetic differences. There are two groups of HIV-1: group M, which includes most HIV-1 isolates, and group O, which represents relatively rare outliers. There are at least seven group M subtypes, which are associated with distinct geographic areas. For example, subtype B (10% of global infections) is most common in Europe and North America, while subtype C (50% of global infections) is prevalent in Southern Africa and India[25].

While HIV-1 is a global epidemic, HIV-2 is mainly prevalent in Western Africa as well as European countries with colonial ties to these countries such as France and Portugal[26]. The smaller spread of HIV-2 compared to HIV-1 is a consequence of its reduced infectivity[27], lower replicative capacity[28], and increased susceptibility to neutralization by the immune system[29]. With an estimated prevalence of one to two million cases in Western Africa alone [30], HIV-2 is still a global health concern. HIV-2 is genetically diverse and there exist eight phylogenetic groups, named A-H. The phylogenetic groups A and B are the most prevalent genotypes: They are causative of almost all cases of clinical disease and are therefore considered the epidemic groups of HIV-2[31]. The other groups (C-H) are considered nonepidemic[32] because viruses from these groups are rarely transmitted between humans.

Note that, when the term HIV is used in the following, I refer to properties that are shared by the two types of immunodeficiency viruses.[33]

2.3.2 Structure and Genome Organization

HIV is a spherical virus with a diameter of roughly 145 nm[34]. It contains two copies of positive single-stranded (ss) RNA. The genetic information carried by the virus is enclosed by a conical capsid, which consists of thousands of copies of the capsid protein p24. The capsid encloses viral RNA bound to the nucleocapsid protein p7 as well as the enzymes involved in viral replication: RT, PR, and IN.

The integrity of the virion is ensured by a matrix consisting of p17 proteins that surrounds the capsid. The matrix is enclosed by a membrane comprising two layers of host phospholipids that are appropriated by the virus when budding from a cell. The viral

[23] Gallo et al. 1984

[24] Lentiviruses (*lente-*, Latin for slow) are characterized by their long incubation periods. The term *incubation period* refers to the time between infection with a pathogen and the appearance of the first symptoms.

[25] Hemelaar et al. 2006

[26] Reeves et al. 1999

[27] Gilbert et al. 2003

[28] Popper et al. 1999

[29] Blaak et al. 2006

[30] Gottlieb et al. 2008

[31] Gao et al. 1994; Marlink 1996; Aguchi et al. 2000

[32] Gao et al. 1994; Chen et al. 1997

[33] This is different to the commonplace usage of the term HIV, which is frequently used as a short hand for HIV-1.

[34] Briggs et al. 2003

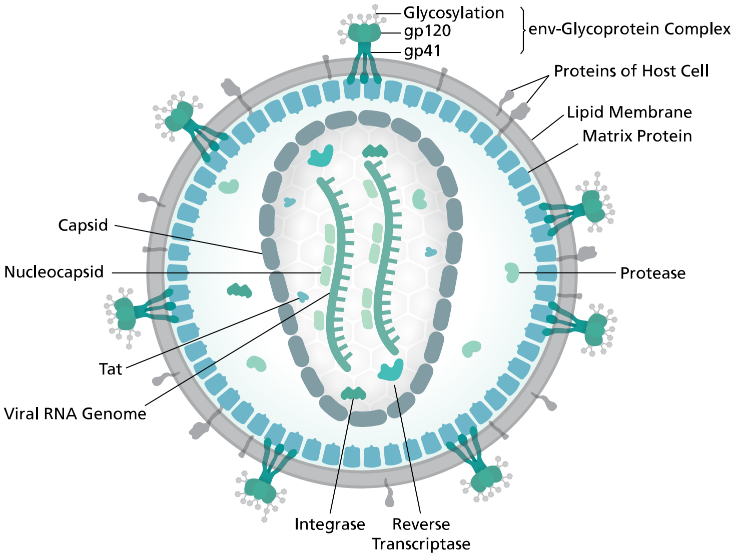
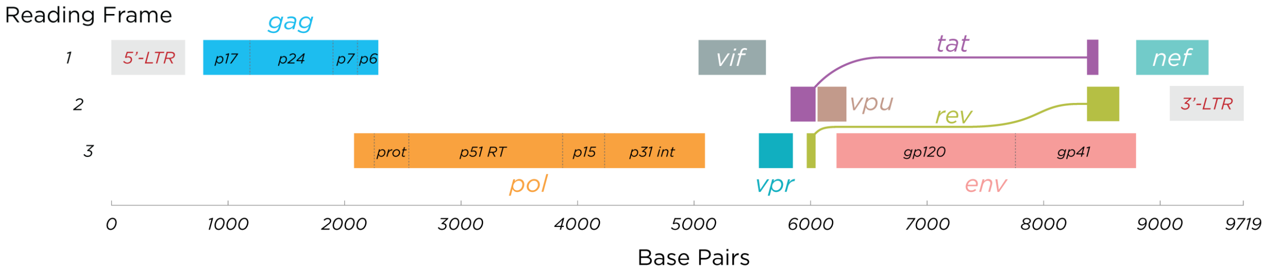


Diagram of the HIV virion by Thomas Splettstoesser, licensed under CC-BY-SA 4.0.

Figure 2.6: Structure of HIV.

membrane serves two purposes. It masks the virus from the immune system and displays the envelope glycoprotein, which facilitates viral cell entry.



Structure of the HIV-1 genome by Thomas Splettstoesser, licensed under CC-BY-SA 4.0.

Figure 2.7: The HIV-1 genome.

Genome Organization Due to the presence of insertions or deletions, the length of the HIV genome can vary. The HIV-1 reference strain HXB2 has a genome comprising 9 719 bases[35]. The HIV-1 genome consists of nine genes that are grouped into three classes: genes encoding structural proteins (*gag*, *pol*, *env*), genes encoding regulatory proteins (*tat*, *rev*)[36], and genes for accessory proteins (*nef*, *vpr*, *vif*, *vpu*)[37]. The length of the HIV-2 reference sequence SIVMM239 is 10 278 bases[38]. With respect to genome organization, the only difference between HIV-1 and HIV-2 is that HIV-2 codes for *vpx* instead of *vpu*[39]. Since the genes that encode the structural proteins of HIV are

[35] Korber et al. 1998

[36] Debaisieux et al. 2012; Hope 1999

[37] Malim and Bieniasz 2012

[38] Calef et al. 2005

[39] Vicenzi and Poli 2013

particularly relevant for this work, they are described in more detail in the following.

gag (group-specific antigen) encodes the precursor gag polyprotein. It consists of matrix (MA) protein p17, capsid (CA) protein p24, spacer peptide 1 (SP1), nucleocapsid (NC) protein p7, spacer peptide 2 (SP2), and protein p6. *pol* (polymerase) codes for the HIV protease (PR), the reverse transcriptase (RT), and the integrase (IN). *env* (envelope) codes for gp160 that is cleaved by a host protease within the endoplasmatic reticulum (ER) in order to obtain the surface protein gp120 and the transmembrane glycoprotein 41 (gp41).

Albeit structurally similar, the proteins of HIV-2 have different identifiers than those of HIV-1 as they exhibit contrasting molecular weights. For example, the HIV-2 precursor protein gp140 is processed to obtain gp125 and gp36, which correspond to gp120 and gp41 in HIV-1[40]. In the following, the protein identifiers of HIV-1 are used when referring to mechanisms that are shared between HIV-1 and HIV-2. Otherwise, the respective protein identifiers for HIV-1 or HIV-2 are used.

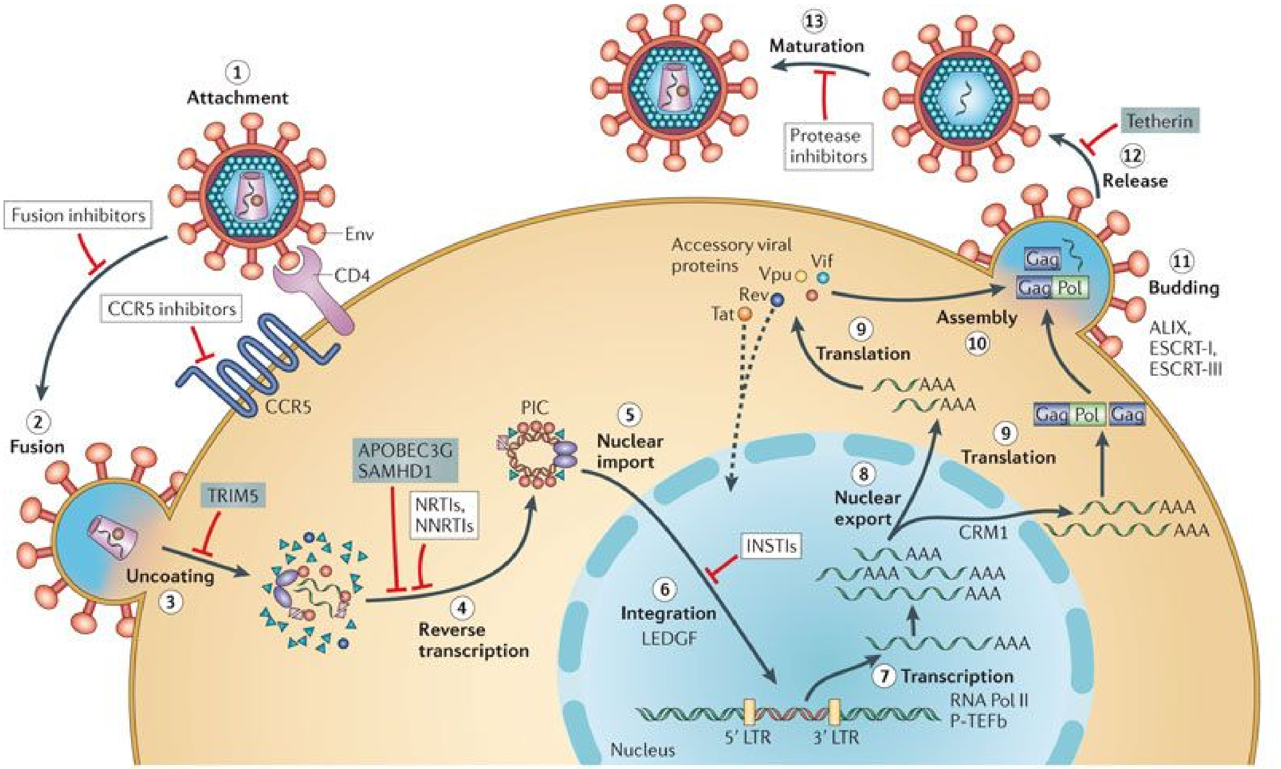
The next section summarizes the life cycle of HIV and indicates the roles of the aforementioned viral and host proteins.

2.3.3 *Life Cycle*

The life cycle of HIV (Figure 2.8) begins with its entry into a human $CD4^+$ cell. HIV cell entry involves binding of the CD4 receptor, coreceptor binding, and finally virus-cell fusion (Figure 2.9). The interaction of cellular CD4 and viral gp120 leads to a conformational change in the envelope protein that allows for binding a coreceptor on the cellular surface. The tight binding afforded by the interactions with CD4 and a coreceptor enables the fusion of viral and host membrane via gp41.

Once the viral capsid has entered the cell, it dissolves and releases the formerly enclosed RNA into the cell. Exploiting cellular nucleotides, RT transcribes viral RNA to complementary deoxyribonucleic acid (cDNA), which migrates to the nucleus where IN catalyzes the integration of viral DNA into the host genome. Once viral DNA is integrated into the host genome, it is called proviral DNA, or, in short, provirus. Next, cellular enzymes transcribe proviral DNA into messenger RNA (mRNA) that is transported out of the nucleus for translation in the ribosomes. After translation, the polypeptides of the virus are cleaved by PR and a new virion is assembled. During budding of the newly assembled virion, PR cleaves the gag-pol polyprotein at nine sites rendering each fragment into a functional protein. This processing completes the viral life cycle, giving rise to

[40] Marcelino et al. 2010



Nature Reviews | Microbiology

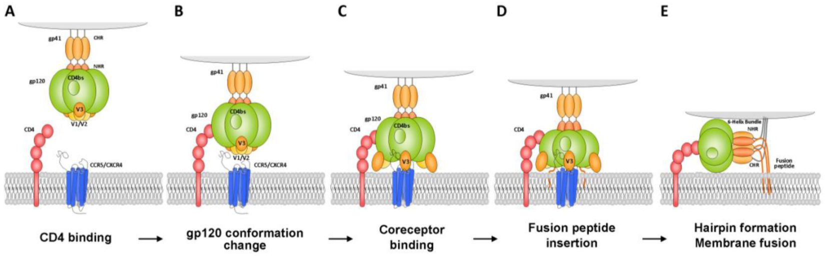
Figure reproduced with permission from Springer Nature (Engelman and Cherepanov, 2012).

Figure 2.8: Life cycle of HIV. Arrows indicate steps in the life cycle, while red lines indicate potential sites for inhibition.

another infectious viral particle.

As illustrated by Figure 2.8, each step in the viral life cycle offers unique opportunities for inhibiting viral replication. An overview of mechanisms that can be exploited for antiretroviral therapy (ART) is provided in Section 2.3.6. The next section sheds light on the coreceptors that are used during HIV cell entry and introduces how different variants of HIV can be differentiated based on their coreceptor usage.

2.3.4 *Coreceptor Usage*



The figure was created by Delhalle et al. (2012) and is licensed under CC-BY-SA 4.0.

Figure 2.9: The stages of HIV cell entry. (A) Binding of the CD4 receptor to the viral envelope protein. (B) Opening of the gp120 structure allows for the interaction of the variable loops with the cellular coreceptor. (C) Binding of the coreceptor increases proximity between HIV and the cell. (D) Insertion of the viral fusion peptide into the cellular membrane. (E) Fusion of viral and human membrane allows for cell entry.

The coreceptors that are required for the cellular entry of HIV (Figure 2.9) are G-protein coupled receptors with a 7-transmembrane structure that are located on human $CD4^+$ cells. Their natural ligands are chemokines, a group of cytokines that is involved in chemotaxis (i.e. cellular movement). The four classes of chemokine receptors (CXC, CC, CX3C, and XC) are defined by the chemokine subfamilies with which they interact. The main coreceptors that are used by HIV are the C-C chemokine receptor type 5 (CCR5), a beta-chemokine receptor, and the C-X-C chemokine receptor type 4 (CXCR4), an alpha-chemokine receptor. While naive $CD4^+$ cells express mainly CXCR4[41], monocyte-derived macrophages express higher quantities of CCR5 than CXCR4[42]. On the basis of differential coreceptor expression, two types of viral strains can be identified: macrophage-tropic (M-tropic) and T-cell tropic (T-tropic) HIV. While M-tropic strains typically use CCR5 and predominantly infect macrophages and memory T-cells, T-tropic strains enter cells via CXCR4 and predominantly infect naive $CD4^+$ cells in addition to macrophages.

In order to refer to specific types of coreceptor usage in this work, an adjusted version of the notation that was established by Berger et al. (1998) is used. According to this notation, a virus that can use only CCR5 is called *R5*. If a virus can use CXCR4 but not CCR5, it is

[41] Bleul et al. 1997

[42] Tuttle et al. 1998

called *X4*. If a viral population can use both CCR5 and CXCR4, it is called *R5X4* or *dual-tropic*[43]. Since routinely used assays for tropism determination (Section 2.5.1) cannot distinguish R5X4 viruses from mixed populations of R5 and X4 variants, the term *dual/mixed* is used in these cases. In this work, I use the term *X4-capable* to indicate the presence of any virus capable of using CXCR4 (i.e. either an X4 virus, an R5X4 virus, or a mixed population) and *R5* for viruses that can use only the CCR5 coreceptor.

HIV-1 Coreceptor Usage HIV-1 uses the CCR5 and CXCR4 coreceptors. In the initial stages of the infection, the CCR5 coreceptor is predominantly used. A switch to the CXCR4 coreceptor is correlated with disease progression[44]. Patients infected with X4-capable variants exhibit increased viral loads[45]. The variable loop 3 (V3) of the surface glycoprotein gp120 is the main molecular marker for the use of the CXCR4 coreceptor[46]. The variable loop 1 (V1) and variable loop 2 (V2) regions have a minor influence on HIV-1 coreceptor usage[47].

HIV-2 Coreceptor Usage Although *in vitro* experiments have demonstrated that there are HIV-2 strains capable of infecting cells independent of the CD4 receptor[48], HIV-2 enters cells *in vivo* by first binding to CD4 and then interacting with a coreceptor[49]. HIV-2 is known to use a wide range of coreceptors *in vitro*[50] but exhibits a limited coreceptor usage *in vivo*. As demonstrated by experiments with primary blood mononuclear cells (PBMCs), HIV-2 predominantly uses CCR5 and CXCR4[51] — variants independent of CCR5 and CXCR4 are infrequent and have been described only in asymptomatic patients[52]. The use of CXCR4 is often associated with broad coreceptor usage[53], which is particularly prevalent in late stage patients[54]. X4-capable strains of HIV-2 are highly virulent, associated with progressed disease[55], and are less susceptible to antibody neutralization than R5 strains[56].

There are three loops on the HIV-2 surface glycoprotein (gp125) that are important for coreceptor binding: V1, V2, and V3. Of these loops, V3 has the greatest impact on coreceptor usage and has been studied most extensively. It has been shown that specific substitutions in V3 and an increased V3 net charge[57] are associated with X4-capability. With regard to the impact of V1 and V2 on HIV-2 coreceptor usage, the following is known. Changes in the V1 and V2 of HIV-2 influence the usage of the CCR8 coreceptor[58] and substitutions at the base and the tip of V1/V2 affect CXCR4 usage[59]. Additionally, insertions in the V1/V2 region seem to be associated with reduced rates of disease progression[60].

More information on the molecular markers of HIV-2 coreceptor

[43] *R5X4* denotes a viral population that consists of R5 and X4 viruses, while *dual-tropic* denotes a virus that can use both CCR5 and CXCR4.

[44] Connor et al. 1997

[45] Weiser et al. 2008

[46] Hwang et al. 1991

[47] Clapham and McKnight 2002; Doms and Trono 2000

[48] Reeves et al. 1999; Chen et al. 1997; Dumonceaux et al. 1998

[49] Clapham and McKnight 2002

[50] Bron et al. 1997

[51] Mörner et al. 2002; Zhang et al. 2000

[52] Azevedo-Pereira et al. 2003

[53] Coreceptor usage is called broad or promiscuous when a variety of coreceptors rather than a single coreceptor can be used effectively.

[54] Blaak et al. 2005; van Der Ende et al. 2000; Owen et al. 1998

[55] Blaak et al. 2005; Mörner et al. 1999

[56] Marcelino et al. 2012

[57] Owen et al. 1998; Mörner et al. 1999; Isaka et al. 1999; Skar et al. 2010

[58] Santos-Costa et al. 2014

[59] Santos-Costa et al. 2014

[60] Skar et al. 2010

usage are available in Chapter 5.

Natural Resistance to R5-tropic HIV Some individuals are carriers of the CCR5Δ32 mutation, a deletion of 32 base pairs (bp) in the CCR5 gene. This deletion results in a premature stop codon. As a consequence, expressed CCR5 coreceptors are non-functional. Both heterozygous and homozygous carriers of the CCR5Δ32 mutation exhibit an increased protection against infection with HIV. While heterozygosity has been shown to reduce the risk of HIV infection by 70% compared to individuals not carrying the mutation[61], carriers of the homozygous variant are nearly fully resistant against infection with HIV[62] as initial infection typically occurs through viruses using CCR5. The CCR5Δ32 mutation (either hetero- or homozygous) has an average prevalence of 10% in Caucasians but is almost absent in native African, Asian, and American Indian populations[63]. The prevalence of the homozygous variant ranges between 1% and 2% in Caucasian populations[64].

Coreceptor Antagonists Coreceptor antagonists can impede HIV cell entry by blocking the molecular interaction of coreceptors and viral surface proteins. Over the years, several compounds targeting CCR5 have been developed, most notably maraviroc[65], TAK-779[66], vicriviroc[67], and aplaviroc[68]. However, the CCR5 antagonist maraviroc is the only coreceptor antagonist that has obtained approval by the federal drug administration (FDA) at the present time[69]. The development of CXCR4 antagonists was less successful: none of the researched compounds such as AMD3100[70], AMD070[71], and AMD3451[72] have obtained FDA approval.[73]

Use of Maraviroc Maraviroc is typically used in salvage treatments of heavily treatment-experienced patients for which few other treatment options remain. It has been intensively studied for treating HIV-1-infected persons[74] but less is known about its use in HIV-2 infection [75]. The first successful application of maraviroc in an HIV-2-infected patient was reported in 2010[76]. In 2012, two independent studies demonstrated the inhibitory effect of maraviroc *in vitro*[77], and, in the same year, the first long-term use of maraviroc was reported in a salvage patient[78]. HIV can evade the drug pressure from maraviroc by switching to the CXCR4 coreceptor. Additionally, HIV-1 can become resistant against maraviroc through an alanine insertion (G310_P311insA) in V3[79]. No similar resistance mutation has been described for HIV-2 yet.

[61] Marmor et al. 2001

[62] Huang et al. 1996

[63] Dean et al. 1996; Samson et al. 1996; Stephens et al. 1998; Libert et al. 1998

[64] Samson et al. 1996; Marmor et al. 2001; Muxel et al. 2008

[65] Dorr et al. 2005

[66] Baba et al. 1999

[67] Strizki et al. 2005

[68] Nakata et al. 2005

[69] Sax 2007

[70] Donzella et al. 1998

[71] Skerlj et al. 2010

[72] Princen et al. 2004

[73] For example, research into AMD3100 was discontinued since the compound can only be administered parenterally and its use is associated with cardiac toxicity.

[74] van Lelyveld et al. 2012

[75] Borrego et al. 2012; Peterson and Rowland-Jones 2012

[76] Armstrong-James et al. 2010

[77] Borrego et al. 2012; Visseaux et al. 2012

[78] Caixas et al. 2012

[79] Garcia-Perez et al. 2015

2.3.5 *Transmission and Course of Infection*

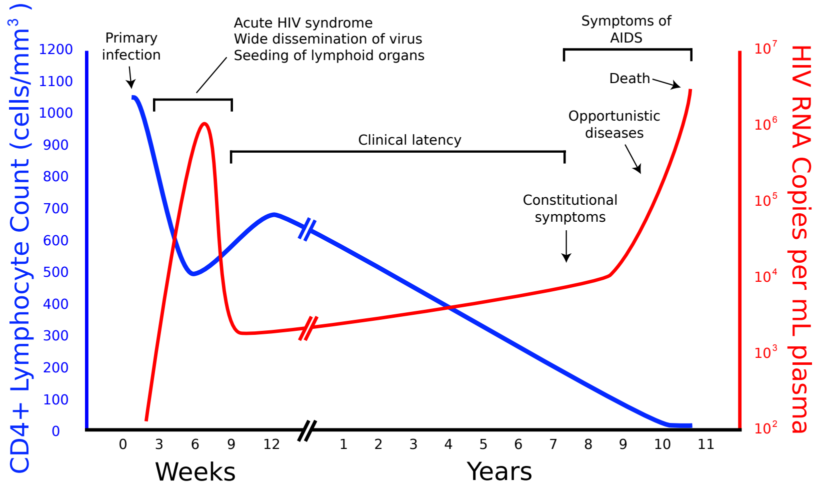


Figure 2.10: Course of HIV infection.

Retrieved from Wikipedia, licensed under CC0 1.0 Universal Public Domain Dedication.

The major mode of HIV transmission is sexual intercourse. One of the reasons why the sexual transmission of HIV is so efficient is that semen-derived enhancer of viral infection (SEVI) drastically increases the risk of sexual transmission[80]. Since the gut-associated lymphoid tissue (GALT) contains the majority of T lymphocytes[81], HIV is most prevalent among men who have sex with men (MSM), particularly those practicing receptive anal sex with frequently changing partners[82]. For example, in the United States, MSM make up 70% of all new HIV infections each year[83]. A comparatively small risk group (10% of global infections) in which HIV is transmitted via blood involves intravenous drug use (IVDU)[84]. In the past, outbreaks of HIV also occurred in clinical settings[85]. For example, in 1988, 99 children were infected with HIV-1 in two Russian hospitals due to blood-contaminated needles[86]. However, improved hygiene standards and routine testing of blood products have rendered the transmission of HIV in this context highly unlikely.

The course of HIV infection (Figure 2.10) can be separated into three phases that are characterized by distinct viral loads (VLs) and CD4 cell counts. In the acute phase, the virus rapidly replicates and the VL reaches its peak. The majority of infected persons develops acute retroviral syndrome (flu-like symptoms) roughly two weeks after initial infection[87]. In the remainder of persons, the infection remains asymptomatic[88]. Since antibodies against HIV appear four to six weeks after infection[89], antibody-based HIV tests are still negative during acute infection. Therefore, the risk of HIV transmission

[80] Münch et al. 2007

[81] Guadalupe et al. 2003

[82] Beyrer et al. 2012

[83] CDC 2017

[84] Aceijas et al. 2004

[85] Bobkov et al. 1994a

[86] Bobkov et al. 1994b

[87] Schacker et al. 1996

[88] Cohen et al. 2010

[89] The point in time at which antibodies against HIV are detectable is called seroconversion.

is particularly high during this brief stage of infection in which diagnosis is not possible and VLs are high[90].

The second stage is the asymptomatic (clinically latent) stage. In this phase of the infection, HIV plasma VLs are low because replicating viruses are being concentrated in the lymphoid organs[91]. Without treatment, the asymptomatic phase of HIV-1 infection culminates in AIDS after roughly ten years of infection. AIDS is characterized by a CD4 count less than 200 cells per mL of plasma or the acquisition of an AIDS-defining condition. AIDS-defining conditions are opportunistic infections and and certain cancers such as Kaposi's sarcoma[92], see Figure 2.11. An opportunistic infection is an infection that is more frequent or more severe in immunocompromised persons than in control patients. Examples for opportunistic infection include pneumocystis pneumonia or tuberculosis. Without treatment, people with AIDS typically survive only a few years.

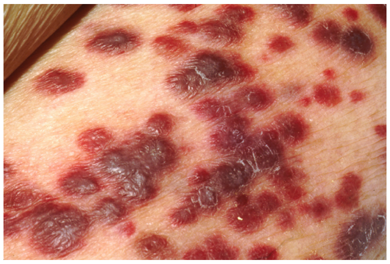


Figure 2.11: Kaposi's sarcoma is a type of cancer that can manifest via purple skin lesions. Retrieved from Wikipedia, licensed under CC BY 3.0.

The rate of CD4 depletion during the course of an HIV-2 infection is distinctly slower in HIV-1 infection. Therefore, the clinically latent phase can last for decades[93]. While HIV-1-infected persons without detectable VLs are a rarity ($< 1\%$)[94], undetectable VLs were found in about 40% of untreated HIV-2 infected persons in Western Africa[95]. Despite the slower course of HIV-2 disease progression, infection with HIV-2 eventually leads to AIDS if left untreated[96].

2.3.6 *Treatment of HIV Infection*

An infection with HIV is treated through a combination of antiretrovirals (ARVs) that interferes with distinct molecular events in the viral life cycle. To understand the aims of antiretroviral treatment, I first give an overview of markers for treatment success. Thereafter, I describe how drug resistance emerges and why it is one of the main challenges of HIV treatment. Finally, I give an overview of antiviral agents against HIV infection and their use for the treatment of HIV-1 or HIV-2 infection.

Treatment Goals and Markers Treatment of HIV infection can be evaluated with respect to its virological, immunological, and clinical success. The primary goal of treatment is the reduction of the VL to an undetectable level ($\leq$ 50 copies per mL). Since persons with undetectable VLs are highly unlikely to transmit the virus[97], reaching undetectability is critical for the implementation of treatment as prevention (TasP)[98,99]. Moreover, long periods of undetectable VLs ensure immune reconstitution, which is measured in terms of the CD4 cell count. HIV-infected persons with normal CD4 cell counts ($\geq$ 500 cells per mL) are typically in good health[100]. The main con-

[90] Pinkerton 2007

[91] Pantaleo et al. 1993; Schwartz and Nair 1999

[92] Schneider et al. 2008

[93] Azevedo-Pereira et al. 2005

[94] Tiemessen et al. 2012

[95] Popper et al. 1999; Berry et al. 2002

[96] Campbell-Yesufu and Gandhi 2011

[97] Cohen et al. 2016; Rodger et al. 2016

[98] Cohen et al. 2011

[99] The idea of TasP is that the HIV incidence rate can be reduced by increasing the proportion of HIV-1 infected persons that have undetectable viral loads.

[100] Sax 2013; Gill et al. 2002

sideration for evaluating the clinical success of a treatment is the absence of immunodeficiency symptoms [101] and the tolerability of the treatment[102]. Based on these considerations, HIV treatment failure can be differentiated into virological, immunological, and clinical failure.

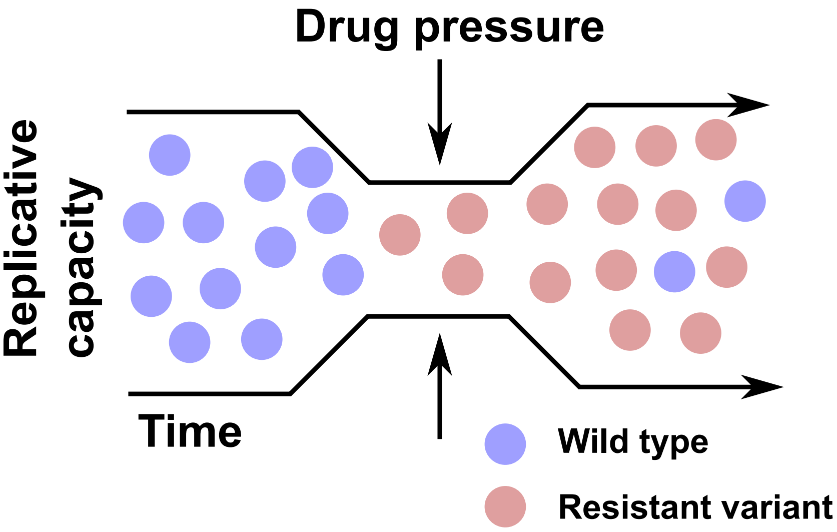


Figure 2.12: Development of drug resistance. If drug pressure insufficiently inhibits viral replication, resistant variants emerge that can replicate even under drug pressure. If drug pressure subsides, reversions to the wild type are possible.

Drug Resistance Similarly to almost all RNA viruses, HIV has a polymerase that lacks proofreading ability[103]. The virus exhibits a replication rate of roughly 10^{10} virions per day[104,105] and a mutation rate of 3.4×10^{-5} per cycle[106]. Due to the high rate at which HIV can mutate its genome, it can quickly adapt to changing conditions in the host.

Viral evolution can even be accelerated by the host itself. Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) is a protein that can hypermutate the provirus, rendering it replication-incompetent. However, mutations induced by APOBEC do not always prevent viral replication. The action of APOBEC may therefore support viral escape from the immune response[107] and lead to the development of drug resistance[108].

For these reasons, resistance mutations are quickly selected when sub-inhibitory drug levels allow for viral replication in the presence of drug, as illustrated in Figure 2.12. HIV drug resistance is one of the major challenges of ART because strains carrying drug resistance mutations are archived in the latent viral reservoir and can emerge at a later point in time[109]. Thus, drug resistance mutations may persist for an indefinite amount of time. Due to the threat that is posed by drug resistant strains, resistance testing is recommended prior to

[101] Hirsch et al. 2004

[102] Carr and Cooper 2000a

[103] Sanjuán and Domingo-Calap 2016; Roberts et al. 1988

[104] Perelson et al. 1996

[105] Note that this high rate of replication is possible even though it takes slightly more than two days to complete the HIV life cycle (Murray et al., 2011), which is described in Section 2.3.3. Bacterial life cycles are much shorter. For example, the replication time of the Escherichia coli bacterium is at 60–90 minutes (Fossum et al., 2007).

[106] Mansky and Temin 1995

[107] Kim et al. 2014

[108] Noguera-Julian et al. 2016

[109] Noë et al. 2005

initiating ART or in case of treatment failure[110].

Drugs for Treating HIV Infection The arsenal of antiretroviral drugs against HIV consists of more than 20 compounds (Figure 2.13). These compounds can be differentiated into six classes: nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (PIs), entry inhibitors (EIs), integrase strand-transfer inhibitors (INSTIs), and antibody-based treatments (Section 2.3.7).

Both NRTIs and NNRTIs prevent reverse transcription of viral RNA. When viral RT incorporates an NRTI into newly synthesized DNA, cDNA synthesis fails because strand elongation is prematurely interrupted[111]. NNRTIs, on the other hand, allosterically inhibit RT by binding to a hydrophobic pocket that induces conformational changes in the active site of the enzyme[112].

PIs interfere with the maturation of immature HIV particles that have just budded from a host cell by preventing the action of viral protease, which is responsible for cleaving the gag polyprotein[113]. Viral particles whose PR has been inhibited remain immature (i.e. non-infectious) because they lack critical structural components such as the capsid or matrix proteins. PIs are typically combined with boosters such as RTV[114] or cobicistat[115]. Boosters improve the bioavailability of PIs by reducing the metabolism of PIs through the inhibition of cytochrome P450 3A, the main driver of PI metabolism in the intestine and liver[116].

EIs reduce viral replication by prevention of viral cell entry. There are two types of EIs: attachment and fusion inhibitors. Attachment inhibitors such as maraviroc compete for binding to the CCR5 coreceptor[117], while fusion inhibitors such as enfuvirtide interact with the transmembrane protein gp41[118].

INSTIs prevent the integration of viral cDNA into the host genome[119]. Finally, recently introduced treatments based on antibodies such as ibalizumab[120] allow for the neutralization of viral particles and the subsequent elimination of infected cells through functions of the immune system.

History of Antiretroviral Treatment The development of antiretroviral treatments can be grouped into four phases. Until the early 1990s, only NRTIs were available. In this period, treatments were monotherapies consisting of a single NRTI or combinations of two NRTIs[121]. Since these treatments used a single mechanism for inhibiting viral replication, early treatment failure due to the emergence of drug resistance was common.

The introduction of NNRTIs and PIs in the mid 1990s heralded

[110] Hirsch et al. 2008

[111] Nikolenko et al. 2005

[112] Schauer et al. 2014

[113] Huff 1991

[114] Zeldin and Petruschke 2003

[115] Gallant et al. 2013

[116] Deeks 2014

[117] Dorr et al. 2005

[118] Matthews et al. 2004

[119] Hare et al. 2011

[120] Bruno and Jacobson 2010

[121] Larder et al. 1996

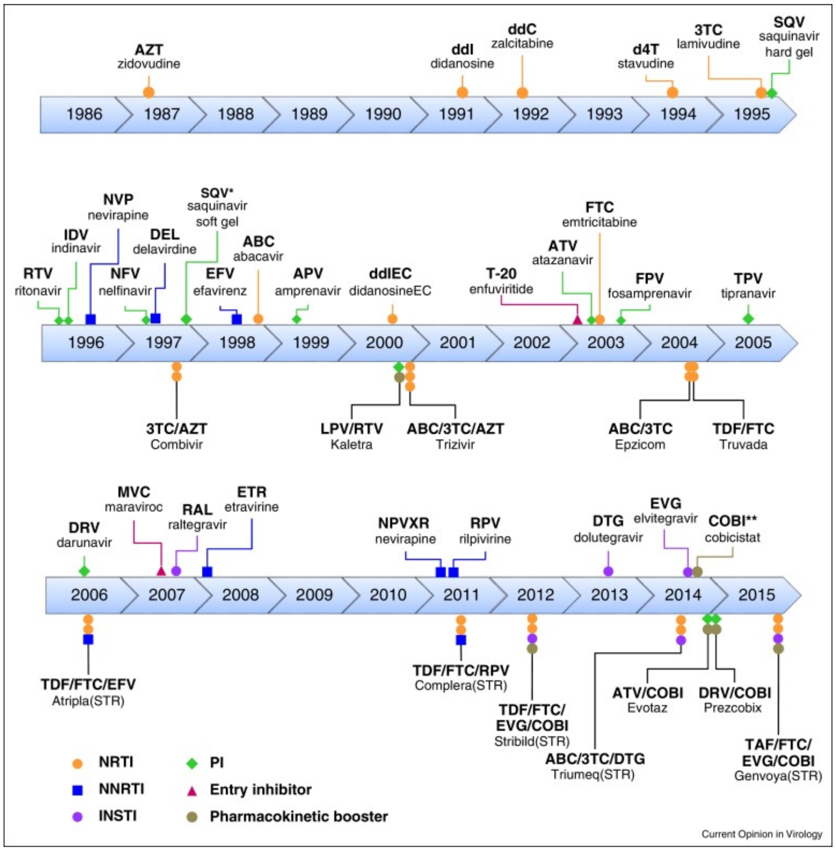


Figure 2.13: Timeline of HIV drugs.

The figure was created by Cihlar and Fordyce (2016) and is licensed under CC BY-NC-ND 4.0.

the era of highly active antiretroviral therapy (HAART). HAART critically improved treatment success[122] by introducing triple therapies consisting of three drugs from at least two different classes. With the increasing number of available antiretroviral drugs, it became challenging to select suitable drug combinations. Therefore, resistance testing became an essential tool for the selection of effective, personalized treatment regimens in the early 2000s[123].

The advent of INSTIs at the end of the 2000s marked the inception of an even more effective era of ART. Because INSTIs were much more potent than previous drugs, they quickly became part of recommended first-line treatments[124]. Treatments with INSTIs reduced VLs quickly and proved to be very durable, giving rise to fewer resistance mutations than other drugs[125]. Since effective therapies of HIV were commonplace now, the term HAART was supplanted by ART.

Current Treatment Strategies The current era of ART is mainly concerned with the simplification of treatments and the prevention of

[122] Palella et al. 2006

[123] Weinstein et al. 2001; Beerenwinkel et al. 2003; Hirsch et al. 2000

[124] Rockstroh et al. 2012

[125] Markowitz et al. 2007; Raffi et al. 2013; Chang et al. 2016

new infections. One aspect of treatment simplification is reducing the pill burden. An increasing number of one-pill once a day regimens is becoming available[126]. These drugs can critically improve patient adherence, which is an important factor for treatment success[127]. Another aspect of treatment simplification is the implementation of dual rather than triple therapies for maintenance therapy. By relying on fewer drugs, these treatments can reduce side effects, save costs, and retain future treatment options[128].

Regarding prevention, a major concern is the implementation of pre-exposure prophylaxis (PrEP)[129]. After the effectiveness of PrEP had been demonstrated by the PROUD[130] and IPERGAY studies[131], the global availability of PrEP has been strongly recommended by the World Health Organization[132]. In the future, long-acting, injectable formulations could improve the effectiveness of PrEP even further[133]. Moreover, TasP still plays a major role: A variety of tolerable and robust treatments are currently under investigation[134].

Treatment of HIV-1 Infection The guidelines for the treatment of HIV-1 infection have changed considerably through the years. In previous years, treatment was initiated only when the infection was sufficiently severe (e.g. at a CD4 count less than 350 cells per mL). However, recent studies such as the START trial have shown that early therapy improves treatment outcomes and reduces the risk of HIV transmission[135]. Therefore, recent European AIDS Clinical Society (EACS) guidelines recommend that an infection with HIV-1 should be treated immediately irrespective of CD4 counts[136]. All of the aforementioned antiretroviral drugs can be used for treating HIV-1.

Treatment of HIV-2 Infection Since HIV-2 infection is characterized by a slower disease progression than HIV-1 infection, treatment can be initiated later (e.g. when the CD4 cell count reaches a value less than 500 cells per mL[137]). However, since the immune reconstitution of patients infected with HIV-2 is slower than in patients that are infected with HIV-1[138], treatment should not be started too late. Available treatments for individuals infected with HIV-2 are limited because the development of antiretrovirals is directed towards HIV-1 rather than HIV-2. Only few drugs that are effective against HIV-1 can be used for treating HIV-2 infection. HIV-2 is intrinsically resistant to NNRTIs[139], to the fusion inhibitor enfuvirtide[140], and exhibits reduced susceptibility to PIs other than LPV, DRV, and SQV[141]. Moreover, drug resistance can emerge particularly rapidly in HIV-2-infected persons[142].

[126] Gubernick et al. 2017

[127] Airoldi et al. 2010

[128] Soriano et al. 2017

[129] PrEP advocates the use of antiretroviral drugs before initiating in high-risk sexual contacts in order to prevent HIV infection.

[130] McCormack et al. 2016

[131] Molina et al. 2015

[132] WHO 2015

[133] Spreen et al. 2013; Margolis et al. 2017

[134] Cihlar and Fordyce 2016

[135] INSIGHT START Study Group 2015

[136] Behrens et al. 2017

[137] Thiébaut et al. 2011

[138] Drylewicz et al. 2008

[139] Tuaillon et al. 2004; Witvrouw et al. 1999

[140] Witvrouw et al. 2004

[141] Raugi et al. 2016

[142] Smith et al. 2009; Menéndez-Arias and Tözsér 2008

2.3.7 *Neutralizing Antibodies and Treatment*

Antibodies to HIV develop within the first few weeks of infection. However, these early antibodies are non-neutralizing antibodies that do not have a detectable effect on viremia[143]. Indeed, it usually takes several months until neutralizing antibodys (NAbs) against envelope glycoprotein (Env) are elicited via affinity maturation[144].

The development of effective NAbs is hindered by the high mutational rate of the surface glycoprotein and its defensive mechanisms. For example, antibodies often develop against easily accessible epitopes such as V1/V2/V3. These epitopes act as decoy epitopes because HIV-1 can quickly escape neutralization by antibodies that bind to these hypervariable regions[145]. Moreover, the presence of the glycan shield[146] often prevents the development of antibodies against more conserved viral epitopes such as the CD4 binding site (CD4bs)[147].

The majority of persons (65%) develops antibodies that are effective against viral strains from a single HIV-1 clade[148] but ineffective against strains from other clades[149]. Broadly neutralizing antibodies (bNAbs) can neutralize HIV-1 strains from several clades and typically appear after two to three years post-seroconversion[150]. According to Simek et al. (2009), an antibody exhibits broad neutralization if its half maximal inhibitory concentration (IC50) titer[151] is at least 100 across at least four viral clades (Figure 2.14). Although bNAbs are elicited in approximately 34% of HIV-1 infected individuals, merely 1% of HIV-1 infected persons produce antibodies that are both broad and potent, with serum titers of at least 300 across four viral clades[152]. These rare persons, who are decisive for researching antibody responses, are called elite neutralizers.

The interaction between HIV-1 and the adaptive immune system is an evolutionary arms race in which HIV-1 always outpaces the immune response because emerging antibodies cannot prevent viral escape[153]. Therefore autologous antibodies NAbs are ineffective *in vivo* although they are effective against wild-type strains *in vitro*. Despite their ability to broadly neutralize HIV-1, bNAbs also do not alter the disease progression of HIV-1-infected persons[154]. Even elite neutralizers progress towards AIDS if left untreated[155], except in rare cases[156]. While autologous antibodies cannot control HIV infection, treatments with heterologous antibodies are being successfully used.

Overview of Broadly Neutralizing Antibodies The first generation of bNAbs was identified in the early and mid 1990s. Although these antibodies were not very potent, they revealed many potential antibody target sites (Figure 2.15) including V3[157], CD4bs[158], a CD4-induced

[143] Overbaugh and Morris 2012; Burton et al. 2011; Ackerman and Alter 2013

[144] Levesque et al. 2009

[145] Karlsson Hedestam et al. 2008

[146] The Env glycan shield, which makes up half the mass of the protein, consists of N-linked glycans that are usually not recognized by antibodies and conceal other potential epitopes (Doores, 2015).

[147] Dubrovskaya et al. 2017

[148] Clades represent the subtypes (A to K) of HIV-1 from group M.

[149] Simek et al. 2009

[150] Euler and Schuitemaker 2012

[151] The IC50 titer of an antibody corresponds to the smallest serum dilution at which viral replication is inhibited by 50%.

[152] Simek et al. 2009

[153] Dingens et al. 2017

[154] Euler et al. 2010; Gray et al. 2011; Piantadosi et al. 2009

[155] Euler et al. 2012

[156] Freund et al. 2017

[157] Gorny et al. 1993

[158] Delwart et al. 1993

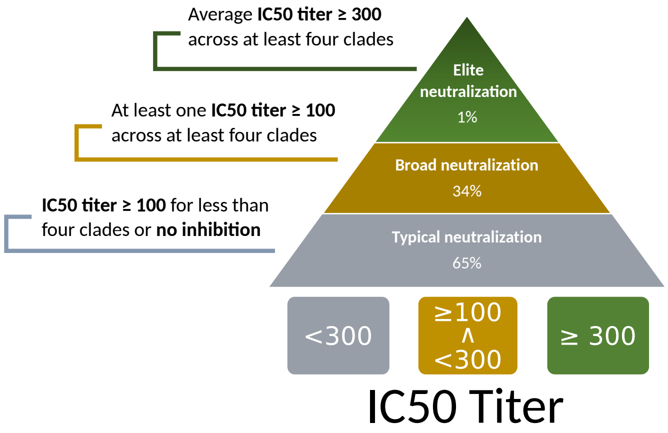

Average **IC50 titer ≥ 300** across at least four clades
At least one **IC50 titer ≥ 100** across at least four clades
IC50 titer ≥ 100 for less than four clades or **no inhibition**
Elite neutralization
1%
Broad neutralization
34%
Typical neutralization
65%
<300
≥100 ∧ <300
≥ 300
IC50 Titer

Figure 2.14: Distribution of HIV neutralizers and their definition according to Simek et al. (2009). The percentages below each type of neutralization indicates the prevalence in HIV-1 infected persons.

site[159], glycans on gp120[160], and the membrane proximal external region (MPER) of gp41[161]. However, interest in bNAbs subsided in the following years because it was thought that their low potency rendered them ineligible for treatment purposes[162].

Two developments invigorated research into bNAbs once again. The development of single cell cloning techniques allowed for the systematic screening of antibodies from elite neutralizers with respect to both neutralization and breadth[163]. Moreover, the RV144 human vaccine trial revealed that a reduced risk of infection was associated with the presence of antibodies targeting the viral envelope spike[164]. Thus, in the following years further bNAbs were isolated. The main target sites of second-generation bNAbs are CD4bs[165], the N160 glycan-dependent site associated with the V1/V2 loops[166], the N332 glycan-dependent site at the base of V3[167], and the MPER of gp41[168]. Note that some persons elicit neutralization through multiple antibodies targeting several epitopes on the envelope spike[169], while others rely on a single clonal antibody[170].

Antibody-based Treatments Naturally emerging antibodies against HIV cannot control the infection because viral evolution outpaces the adaptive processes of the immune system. However, potent, heterologous antibodies that are administered as a part of antiviral treatment, are capable of neutralizing HIV. The relevance of swift and potent immune responses is underlined by studies in simians[171], which suggest that immunization can be achieved if antibody treatments are started directly after infection[172].

Treatments based on bNAbs are different from treatments that rely on conventional ARVs with respect to toxicity, drug delivery, frequency of treatment, and emergence of drug resistance. While frequently used ARVs such as RTV and TDF are associated with hepatoxicity[173] and nephrotoxicity[174], respectively, NAbs exhibit

[159] Thali et al. 1993

[160] Trkola et al. 1996

[161] Muster et al. 1993

[162] Mascola and Nabel 2001

[163] Scheid et al. 2009

[164] Haynes et al. 2012

[165] Scheid et al. 2011; Zhou et al. 2010; Diskin et al. 2011; West et al. 2012

[166] Walker et al. 2009, 2011; Bonsignori et al. 2012

[167] Walker et al. 2011; Mouquet et al. 2012; Pejchal et al. 2011

[168] Huang et al. 2012

[169] Scheid et al. 2009

[170] Scheid et al. 2011

[171] Simians comprise monkeys and apes.

[172] Nishimura et al. 2017

[173] Sulkowski 2004

[174] Fernandez-Fernandez et al. 2011

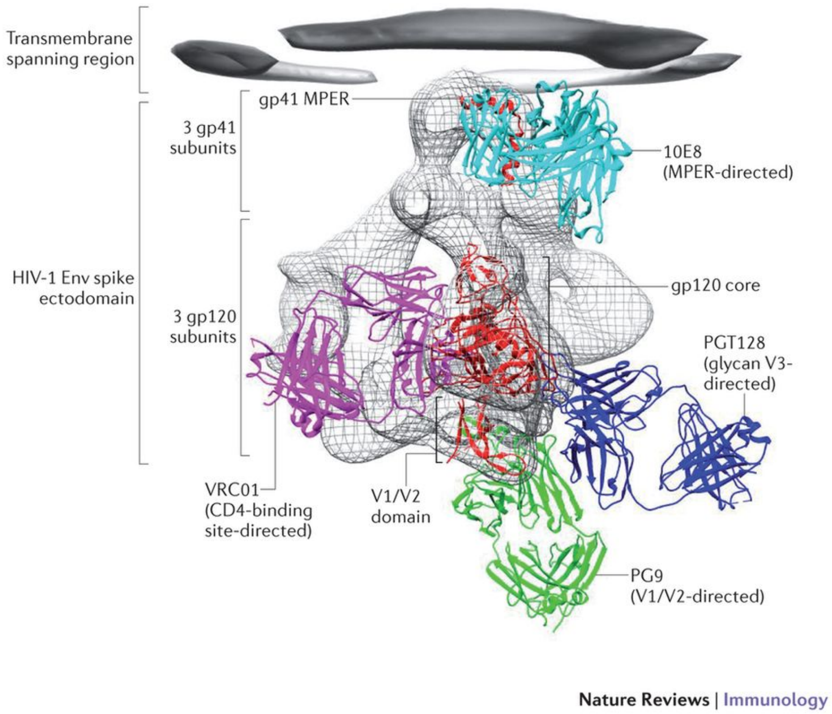

Transmembrane spanning region
HIV-1 Env spike ectodomain
3 gp41 subunits
3 gp120 subunits
gp41 MPER
10E8 (MPER-directed)
gp120 core
PGT128 (glycan V3-directed)
VRC01 (CD4-binding site-directed)
V1/V2 domain
PG9 (V1/V2-directed)
Nature Reviews | Immunology

Figure reproduced with permission from Springer Nature (Kwong et al., 2013).

Figure 2.15: Envelope epitopes targeted by antibodies against HIV-1. The grey grid shows an image of the envelope obtained by cryo-electron microscopy.

different side effects such as immune reconstitution inflammatory syndrome (IRIS) or injection site reactions[175]. Thus, antibody-based treatments may be well-suited for HIV-1-infected persons that are prone to organ damage (e.g. those that are coinfected with HCV).

In contrast to conventional ART, which requires the daily oral intake of drugs, antibody-based treatments require infusions[176]. What first appears as a disadvantage is compensated by the long half-life of antibodies, which enables less frequent dosing, for example only every few weeks[177]. Moreover, since antibodies target different epitopes than conventional ARVs, they are particularly useful for patients exhibiting multi-class drug resistance to established drugs.

At the current point in time, there is only one FDA-approved antibody, the CD4 binding site inhibitor ibalizumab[178]. The drug is being used in combination with other ARVs in order to treat patients exhibiting high levels of drug resistance[179]. A plethora of other antibodies are currently under investigation in clinical studies. One of the most promising bNAbs of the second generation is VRC01[180], a bNAb targeting the CD4bs. VRC01 is currently being investigated in the antibody-mediated prevention study (AMP), a clinical trial in phase 2b. Here, study participants receive infusions of VRC01 every 8 weeks and are followed for 22 months. The goal of this study

[175] Hansel et al. 2010

[176] The only conventional antiretroviral that has to be injected is the fusion inhibitor T-20 (Matthews et al., 2004).

[177] Huang et al. 2017

[178] Bruno and Jacobson 2010

[179] Iacob and Iacob 2017

[180] Lynch et al. 2015; Ledgerwood et al. 2015; Wu et al. 2010a

is to determine whether periodic injections with VRC01 safely and effectively protect high-risk persons from infection with HIV-1. If VRC01 is shown to protect against HIV-1 infection, the drug may be an interesting candidate for the use in PrEP.

In the future, combination therapies consisting of several antibodies targeting distinct epitopes could reduce the risk of viral escape[181]. The recent development of bispecific bNAbs[182], which simultaneously target two distinct epitopes, may considerably improve the robustness and effectiveness of antibody-based treatments. Finally, improving our understanding of how potent bNAbs are elicited could have profound ramifications for the development of an HIV-1 vaccine[183].

2.4 Hepatitis C Virus

HCV infects human hepatocytes and is a leading cause of chronic liver disease (Section 2.4.1). The HCV genome consists of structural and non-structural proteins (Section 2.4.2). In contrast to HIV, HCV does not integrate into the genome of host cells in order to replicate (Section 2.4.3). HCV is highly prevalent and can cause lethal diseases such as liver cirrhosis or hepatocellular carcinoma (Section 2.4.4). The continual use of highly potent drugs over several weeks can typically clear the infection (Section 2.4.5).

2.4.1 Introduction to HCV

HCV, which was discovered[184] in 1989, is a group IV hepacivirus that belongs to the family of *Flaviviridae*. HCV infects human hepatocytes and is a leading cause of chronic liver disease, cirrhosis, and hepatocellular carcinoma (HCC). It is also the most common indication for liver transplantation in many countries[185]. There are approximately 170 million people living with chronic HCV infection. Each year, 350 000 people die due to HCV-related causes[186]. In contrast to hepatitis A and hepatitis B virus, there is no vaccine against HCV[187].

HCV is genetically highly diverse. There are seven HCV genotypes (designated as 1–7) that differ in their nucleotide sequence by at least 30%[188]. HCV genotypes are further differentiated into subtypes (designated by letters a, b, and so on) that exhibit a nucleotide divergence of at least 20%. The globally most prevalent genotypes[189] are genotype 1 (46.2% of global and 70% of European infections[190]) and genotype 3 (30.1% of global infections).

[181] Bolton et al. 2016

[182] Steinhardt et al. 2018

[183] Burton et al. 2004

[184] Choo et al. 1989

[185] Chen and Morgan 2006

[186] Mohd Hanafiah et al. 2013

[187] Bell 2002; Mast et al. 2004; Houghton and Abrignani 2005

[188] Smith et al. 2014; Simmonds 2004

[189] Messina et al. 2015

[190] Petruzziello et al. 2016

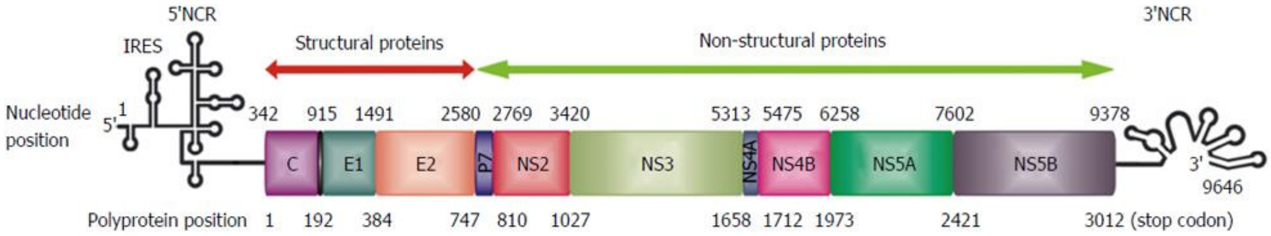


Image adapted from Echeverría et al. (2015), licensed under CC BY-NC 4.0.

Figure 2.16: HCV genome organization.

2.4.2 *Structure and Genome Organization*

The HCV virion consists of a nucleocapsid that is surrounded by a host-derived membrane containing the glycoproteins E1 and E2. The (+)ssRNA genome of HCV (Figure 2.16) has a length of about 9.6 kb and is defined by a single open-reading frame that is translated into a precursor polyprotein[191]. The polyprotein can be divided into two regions: the 5′ region, which contains the structural proteins, and the 3′ region, which contains the non-structural proteins.

There are three structural proteins: C (core), E1, and E2. C is involved in the nucleocapsid, while E1 and E2 form the surface glycoproteins. The non-structural proteins of HCV are NS1, the transmembrane protein NS2, the protease NS3, the protease cofactor NS4A, the intracellular membrane protein NS4B, the viral replication factor NS5A[192], and the RNA polymerase NS5B. Research is focused on the non-structural proteins NS3, NS5A, and NS5B because these proteins are the target of direct-acting antivirals (DAAs).

2.4.3 *Life Cycle*

The life cycle of HCV (Figure 2.17) begins with its attachment to specific receptors on hepatocytes[193]. Thereafter, HCV is internalized and the nucleocapsid is released into the cytoplasm where viral RNA is uncoated. Positive-stranded viral RNA is used as a template for both translation and generation of a complementary RNA intermediate. The negative stranded RNA is used as the template for the synthesis of additional positive RNA strands that are required to produce new virions.

Using the internal ribosome entry site (IRES) (Figure 2.16), HCV RNA directly recruits the cellular translation apparatus and initiates the translation of viral proteins[194]. The precursor protein is then processed by both host and viral proteases in order to obtain mature viral proteins[195]. HCV replication occurs within the ER-membrane bound replication complex, which contains replicating viral RNA and non-structural proteins as well as cellular proteins[196]. After the virus

[191] Kato 2000

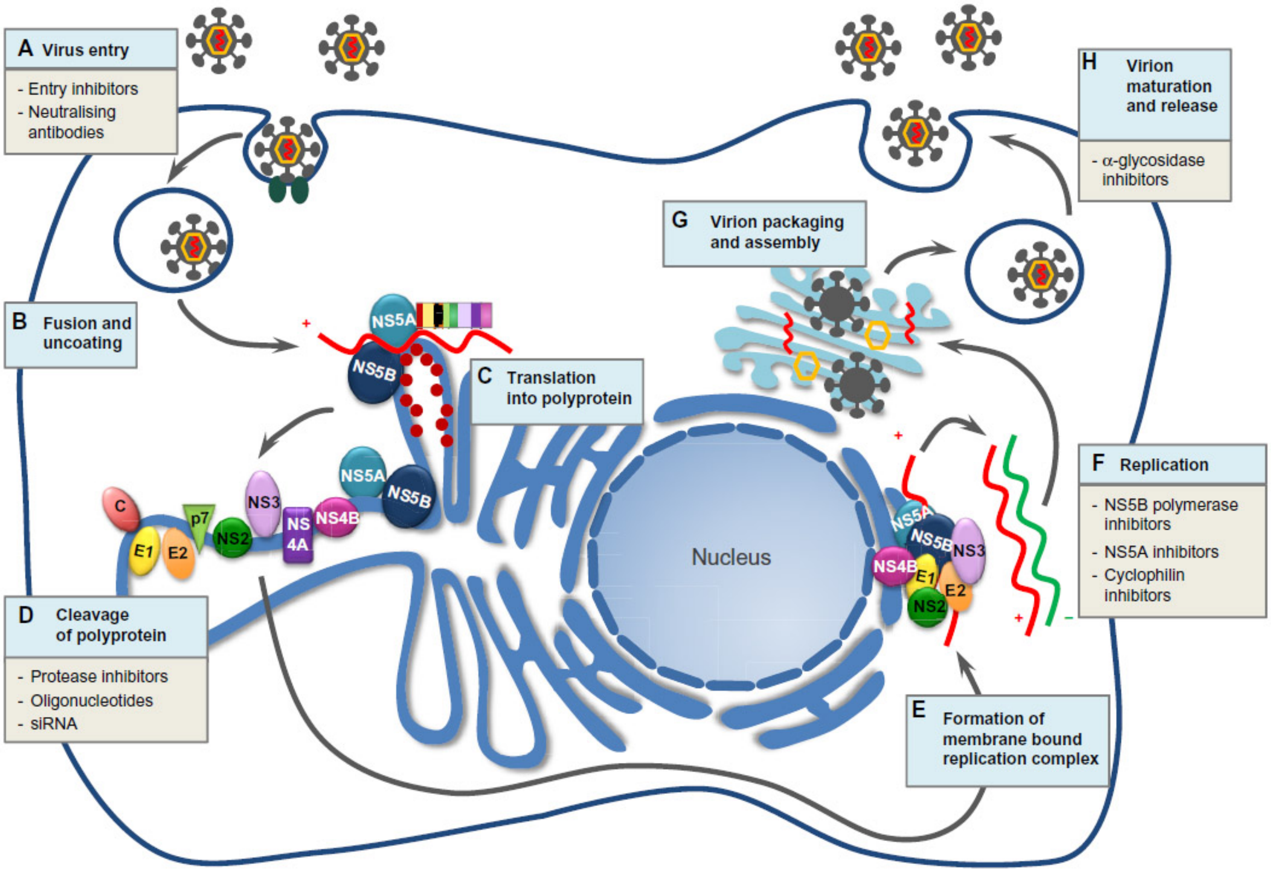
[192] Macdonald and Harris 2004

[193] Zhu et al. 2014

[194] Niepmann 2013

[195] Carrère-Kremer et al. 2004

[196] Gosert et al. 2003


A Virus entry
- Entry inhibitors
- Neutralising antibodies
B Fusion and uncoating
C Translation into polyprotein
D Cleavage of polyprotein
- Protease inhibitors
- Oligonucleotides
- siRNA
E Formation of membrane bound replication complex
F Replication
- NS5B polymerase inhibitors
- NS5A inhibitors
- Cyclophilin inhibitors
G Virion packaging and assembly
H Virion maturation and release
- α-glycosidase inhibitors
Nucleus
NS5A
NS5B
NS3
NS4B
NS 4A
NS2
p7
C
E1
E2
+
-

This figure was created by Holmes and Thompson (2015) and is licensed under CC-BY-NC 3.0.

Figure 2.17: The HCV life cycle.

has been assembled in the ER, it is released from the cell[197].

2.4.4 *Transmission and Course of Infection*

HCV is a blood-borne infection. Currently, the main route of infection with HCV is IVDU[198] although sexual contact with HIV-1-positive persons or MSMs seems to be a contributing factor[199]. However, there is almost no risk of sexual transmission in groups of persons that neither belong to IVDU nor MSM[200]. In the past, blood transfusions also played an important role. Nowadays, routine screening of blood samples renders infection via blood products highly unlikely (roughly 1 in 2 million)[201]. The high prevalence of HCV infection can be explained by the following three factors. First, HCV was as discovered as recently as 1989. Until then, HCV could disseminate freely. Second, blood products were previously treated less carefully than nowadays, which suggests that clinics could have been a major driver of HCV spread. For example, the high prevalence of HCV in Egypt is largely due to a parenteral antischistosomal therapy campaign (1960s to 1980s), in which unsterilized injection material was reused[202]. Third, since HCV infection often remains asymptomatic for many years, this increases the likelihood that a person unknowingly transmits the infection during his lifetime.

Between 70% and 80% of HCV infections follow an asymptomatic course[203]. The remainder of infections leads to clinical symptoms (e.g. malaise, weakness, anorexia, or jaundice, see Figure 2.18) within three to twelve weeks after infection. During the acute phase, considerable hepatocyte necrosis takes place, which is evidenced by exceedingly high levels of serum alanine aminotransferase. During the first weeks of infection, HCV RNA levels peak at 10^5 to 10^7 international units (IU) per mL of blood[204] Antibodies to HCV become detectable one to three months after exposure.

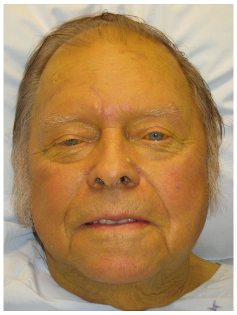


Figure 2.18: A person showing symptoms of jaundice (high bilirubin levels). Jaundice is associated with diseases of the liver.

Retrieved from Wikipedia, licensed under CC BY 3.0.

HCV infection is considered to be chronic if HCV RNA persists for at least six months after the onset of acute infection. The majority of HCV infected persons (75% to 85%) cannot spontaneously clear the infection[205]. HCV disease progression is measured in terms of liver fibrosis, which is graded according to liver biopsy. Here, the number of mononuclear inflammatory cells and the number of dead or dying hepatocytes are considered. Liver cirrhosis describes the advanced stage of fibrosis in which the function of the liver is limited (Figure 2.19). HCV infected persons can unknowingly develop compensated cirrhosis, which is not associated with any symptoms. Decompensated cirrhosis, whose treatment necessitates a liver transplant, is associated with life-threatening conditions including bleeding varices[206], ascites[207], encephalopathy, or jaundice.[208]

[197] Jones and McLauchlan 2010

[198] Garfein et al. 1996

[199] van de Laar et al. 2007; Tohme and Holmberg 2010

[200] Marincovich et al. 2003; Vandelli et al. 2004

[201] Stramer et al. 2004

[202] Frank et al. 2000

[203] Chen and Morgan 2006

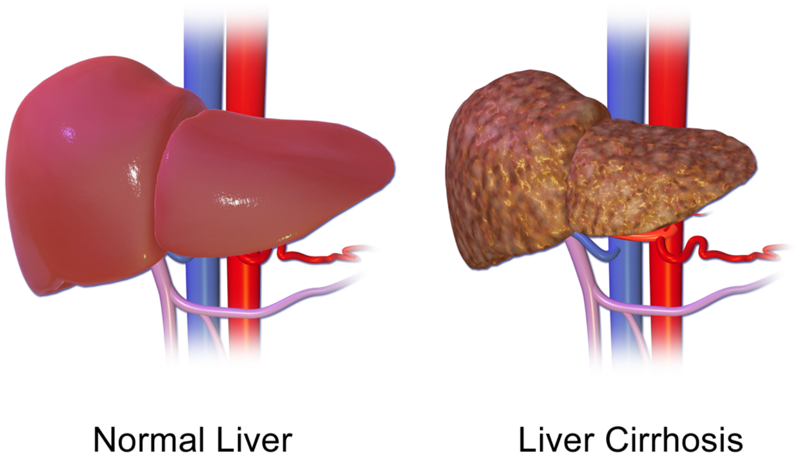
[204] HCV viral loads are measured in terms of international units because different laboratories had used different standards for counting the number of HCV RNA copies before. A viral load below 800 000 IU per mL is considered to be low, while a load above 800 000 IU per mL is considered to be high.

[205] Chen and Morgan 2006

[206] A varix is an abnormally dilated vessel with a tortuous course.

[207] Ascites refers to the abnormal buildup of fluid in the abdomen.

[208] Chen and Morgan 2006



Retrieved from Wikipedia, licensed under CC BY 4.0.

Figure 2.19: Comparison of a normal and a cirrhotic liver.

After ten years of HCV infection, the cumulative probability of acquiring decompensated cirrhosis reaches approximately 30%. Once decompensated cirrhosis is present, the 5-year survival drops to 50%. In cirrhotic patients, HCC develops at a rate of 1% to 4%. The mean time to development of cirrhosis and HCC is 24 years and 27 years, respectively[209]. Note that HCV infection can also cause extrahepatic disease manifestations, which typically occur in the form of autoimmune disorders, most notably cryoglobulinemia[210].[211]

2.4.5 *Treatment of HCV Infection*

The success of HCV treatment is measured in terms of the long-lasting disappearance of HCV-RNA from serum, which is called sustained virologic response (SVR). Previously, treatment of HCV infection was based on the administration of interferon and ribavirin.[212] For persons with acute symptomatic hepatitis C, an immediate monotherapy with interferon could clear the infection in 90% of patients within 24 weeks[213]. The treatment of chronic infections, however, proved more challenging: combinations of interferon and ribavirin attained SVR rates of only 50% after 24 or 48 weeks of treatment[214].

The way in which HCV infection is treated has changed drastically since the arrival of DAAs in 2011. In contrast to interferon and ribavirin, DAAs target specific steps in the HCV replication cycle in a similar fashion as antiretroviral drugs against HIV. There are three classes of DAAs available (Table 2.1): inhibitors of NS3, NS5A, and NS5B.[215] Modern treatments, which are based on these drugs, are

[209] Chen and Morgan 2006

[210] Cryoglobulinemia is a condition in which the blood contains large amounts of immunoglobulins that become insoluble at reduced temperatures.

[211] Agnello and De Rosa 2004

[212] Manns et al. 2006

[213] Jaeckel et al. 2001; Santantonio et al. 2005; Wiegand et al. 2006

[214] Poynard et al. 1998; McHutchison et al. 1998; Manns et al. 2001; McHutchison et al. 2002

[215] Interestingly, all FDA-approved drugs against HIV target structural proteins, while all FDA-approved drugs against HCV target non-structural proteins.

Date	Drug	Genotype
NS3 Inhibitors		
May 14, 2011	Bocprevir (BOC)	1
May 23rd, 2011	Telaprevir (TVR)	1
November 22, 2013	Simeprevir (SMV)	1,4
July 2014	Asunaprevir (ASV)	1
January 28, 2016	Grazoprevir (GZR)	1,4
July 22, 2016	Paritaprevir (PTV)	1,4
July 18, 2017	Voxilaprevir (VOX)	Pangenotypic
August 3, 2017	Glecaprevir (GLE)	Pangenotypic
NS5A Inhibitors		
October 10, 2014	Ledipasvir (LDV)	1,4,5,6
July 24, 2015	Daclatasvir (DCV)	1,3
July 24, 2015	Ombitasvir (OBV)	1,4
January 28, 2016	Elbasvir (EBR)	1,4
June 28, 2016	Velpatasvir (VEL)	Pangenotypic
August 3, 2017	Pibrentasvir (PIB)	Pangenotypic
NS5B Inhibitors		
December 6, 2013	Sofosbuvir (SOF)	1,2,3,4
July 22, 2016	Dasabuvir (DSV)	1

Table 2.1: Drugs for the treatment of HCV infection. The *Genotype* column indicates the genotypes for which drugs are approved and the *Date* column indicates the date of FDA approval. Pangenotypic are approved for all HCV genotypes.

all-oral since interferon infusions are no longer necessary[216].

With the advent of DAAs, the required treatment duration was reduced to 12 weeks and the rate of SVR was raised to 90% for most patients[217]. However, because the first DAAs were effective only against specific HCV genotypes[218], the effectiveness of HCV treatments such as SOF+LDV or GZR+EBR varied in dependence on the genotype[219]. Recently introduced pangenotypic treatments against HCV such as SOF+VEL or GLE+PIB are eligible for all HCV genotypes and ensure SVR rates of at least 95%[220].

Treatment failure with DAAs can be caused by the selection of resistance-associated variants (RAVs)[221]. RAVs in NS3 and NS5A are particularly problematic as they can persist for months[222] or even years after treatment cessation[223], respectively. Currently, it is not recommended to perform systematic resistance testing prior to the initiation of treatments in DAA-naive patients[224]. Initial drug resistance testing is not recommended for three reasons. First, 99% of patients without cirrhosis that are treated with current pangenotypic DAAs reach SVR after twelve weeks[225]. Second, the presence of resistance-associated substitutions (RASs) leads to treatment failure in only a small fraction of patients[226]. Third, the clinical consequences of treat-

[216] Holmes and Thompson 2015

[217] Holmes and Thompson 2015

[218] European Association for the Study of the Liver 2018

[219] Holmes and Thompson 2015

[220] Rockstroh 2018

[221] Pawlotsky 2011

[222] Susser et al. 2011; Pawlotsky 2016

[223] Yoshimi et al. 2015

[224] European Association for the Study of the Liver 2018

[225] Puoti et al. 2018

[226] Wang et al. 2018

ment failure are small because organ damage resulting from HCV accumulates over several years; requiring an additional twelve weeks to eliminate the infection is of little clinical consequence. Thus, drug resistance testing is mostly used in case of treatment failure. Based on the information of resistance testing, the treating clinician can either select another treatment or continue the previous treatment for an extended period of time. Of course, it is also important to monitor the epidemiology of HCV drug resistance for surveillance purposes.

2.5 *Molecular Techniques*

This section introduces molecular techniques that are useful for the study of viruses. Phenotypic methods for the identification of HIV coreceptor usage (Section 2.5.1) are employed to investigate the molecular characteristics of viruses using different coreceptors. Results from phenotypic assays inform genotypic methods for the identification of HIV coreceptor usage (Chapter 5), which are used to guide the prescription of the CCR5 coreceptor antagonist maraviroc. Phenotypic methods for the determination of HIV drug resistance (Section 2.5.2) can offer insights into the mutations that confer resistance. Viral genomes are routinely sequenced (Section 2.5.3) for genotypic resistance testing (Chapter 4), in which viral genomes are analyzed with respect to drug resistance mutations in order to aid treatment selection. Polymerase chain reaction methods (Section 2.5.4) are important for a multitude of biological techniques. mPCR is particularly important for the elucidation of antiviral immune responses (Chapter 6).

2.5.1 *Phenotypic HIV Coreceptor Testing*

Phenotypic approaches for the identification of HIV coreceptor usage rely on assays that are based either on $CD4^+$ cell lines or on primary cells. The $CD4^+$ cell lines are engineered such that they express only certain coreceptors and elicit a specific signal upon viral infection. Assays based on primary cells, on the other hand, are typically evaluated according to the concentration of capsid antigen (p24 for HIV-1 and p27 for HIV-2), which provides a surrogate marker for the number of infected cells[227].

Primary Cells vs Cell Lines Assays based on primary cells have a higher agreement with *in vivo* coreceptor usage than assays based on cell lines. Since engineered cell lines often express higher concentrations of $CD4^+$ receptors and HIV coreceptors than naturally occurring cells[228], they are often hypersensitive to coreceptor usage.

[227] Older assays differentiated between X4-capable- and R5 variants by considering the extent to which syncytia (multinucleated cells due to fusion) are formed. While X4-capable variants tend to induce syncytia, this is not the case for R5-tropic strains (Björndal et al., 1997; Owen et al., 1998).

[228] Choudhry et al. 2006

Assays that are based on such cell lines therefore typically yield a greater number of false positive results (i.e. reporting a virus as X4-capable although it is incapable of infecting a CD4 cell exhibiting only the CXCR4 coreceptor *in vivo*). Still, assays based on cell lines are more frequently used than assays based on primary cells because they are standardized, less work-intensive, and are more suitable for detecting coreceptors other than CCR5 or CXCR4.

Two Types of Assays There are two approaches for measuring coreceptor usage phenotypically. The first approach relies on comparing the infection status of two types of cells that express different coreceptors. The second approach uses coreceptor antagonists to block specific coreceptors and subsequently measures the infection status. Assays based on U87 cells and GHOST (3) cells[229] follow the first approach, while assays based on TZM-bl cells[230] follow the second approach. PBMCs can be used for both approaches since PBMCs with homozygous CCR5Δ32 can be used for the first approach, while conventional PBMCs can be used for the second approach[231]. In the next paragraphs, I first deal with assays that follow the first approach and then discuss assays pursuing the second strategy.

The Trofile assay is the standard assay for identifying HIV-1 coreceptor usage and relies on U87 cells[232]. In the first step, a replication-defective pseudovirus carrying the envelope region of the target virus is constructed via transfection[233]. The use of a pseudovirus prevents multiple rounds of infection that could bias the results of the assay[234]. Two types of U87 cells that provide different luminescent signals upon infection are used. U87:CXCR4 expresses only CXCR4, while U87:CCR5 expresses only CCR5. If a signal is measured only from CCR5- or CXCR4-expressing cells, the virus is identified as R5 or X4, respectively. Otherwise, if signals from both types of cells are found, the virus is identified as dual/mixed. An improved version of the original Trofile assay, the enhanced-sensitivity Trofile assay (ESTA), was shown to obtain sensitivities of 100% for detecting envelope sequences from X4-capable variants even at a prevalence of only 0.3%[235].

Assays based on TZM-bl cells follow the second strategy since TZM-bl cells exhibit both CCR5 and CXCR4 simultaneously. Upon infection, these cells express firefly luciferase enzyme under the control of the HIV promoter such that HIV infection can be detected. By blocking one of the coreceptors with excessive amounts of coreceptor antagonists and then measuring luminescence, R5 and X4-capable variants can be discriminated[236].

[229] Vödrös et al. 2001

[230] Platt et al. 1998; Wei et al. 2002

[231] Zhang et al. 2000

[232] Whitcomb et al. 2007; Low et al. 2009; Reeves et al. 2009

[233] Bilska et al. 2017

[234] Lin et al. 2010

[235] Trinh et al. 2008

[236] Davis et al. 2008; Borrego et al. 2012

Interpretation of Results Interpreting the results of phenotypic assays can be challenging. For example, the fluorescence readouts from GHOST (3) cells should be compared to a control cell line expressing only the CD4 receptor to account for background noise levels. If no reference cell line is available, a threshold for discerning true signal and noise has to be chosen judiciously. For TZM-bl cells, it is necessary to interpret low levels of residual replication after the application of coreceptor antagonists. Furthermore, care should be taken when comparing results from assays relying on the use of coreceptor antagonists because these assays may yield different results for the same isolate in case that different types of coreceptor antagonists were used (e.g. the CCR5 antagonists maraviroc and TAK-779).

2.5.2 *Phenotypic Resistance Testing*

Phenotypic resistance testing relies on experimental monitoring of the extent to which viral replication is inhibited at varying concentrations of an antiretroviral drug[237]. If viral replication is suppressed at relatively low drug concentrations, a viral strain is considered to be susceptible to a drug. Otherwise, if relatively high drug concentrations are required for inhibiting a viral strain, it is considered to be resistant[238].

Phenotypic assays determine drug resistance through an *in vitro* binding assay. An advantage of phenotypic assays over genotypic approaches (Chapter 4) is that they directly assess the efficiency with which the drug binds to its target rather than searching for genomic footprints that are indicative of resistance. Moreover, they measure the effect of individual drugs rather than the resistance patterns of multiple drugs, which is the case for genotypic approaches that rely on clinical data[239]. However, since phenotypic tests are performed *in vitro*, it is necessary to interpret their results carefully when these tests are used for *in vivo* applications (e.g. guiding treatment choices). Due to the accuracy of phenotypic measurements, they have formed the basis for genotypic approaches for resistance testing such as geno2pheno[resistance][240]. Since phenotypic tests are time- and cost-intensive, they are typically not applied in clinical settings and are more relevant for basic research. For example, phenotypic methods haven been used to identify the resensitizing effect of the RT mutations M184V and L74V[241] or to determine patterns of co-occurring resistance mutations[242].

There are two generations of phenotypic resistance tests. The first generation is based on the cultivation of PBMCs[243]. Since phenotypic resistance testing based on PBMCs infected with HIV-1 was found

[237] Mayer et al. 2001

[238] Drugs approved for treating HIV inhibit susceptible strains at therapeutic concentrations but the inhibition of resistant strains would require concentrations associated with toxicity.

[239] Pironti et al. 2017a

[240] Beerenwinkel et al. 2003

[241] Larder et al. 1995; Miranda et al. 2005

[242] Shafer 2006

[243] Japour et al. 1993

to be very time-consuming, new assays that can be performed in 8–10 days were developed. These PCR-based assays utilize recombinant viruses[244] and are performed in the following way. After amplifying the genomic regions of interest (typically PR, RT, or IN of the *pol* gene) from viral plasma RNA through reverse transcription polymerase chain reaction (RT-PCR), the corresponding segments are inserted into a modified HIV-1 vector. Finally, drug resistance is determined by measuring the IC50 as described in Section 2.5.2.

Antivirogram and PhenoSense Based on this second-generation approach, two commercial phenotypic assays were developed: the Antivirogram assay (formerly Virco, later Janssen Pharmaceutica)[245] and the PhenoSense assay (formerly ViroLogic, later Monogram Biosciences)[246]. The Antivirogram assay requires at least 1000 HIV-1 RNA copies per mL and includes PR and the majority of RT (up to codon 482). The PhenoSense assay requires at least 500 HIV-1 RNA copies per mL and also considers protease cleavage sites in *gag* in addition to PR. RT, however, is included only until codon 313[247].

The PhenoSense assay performs only a single cycle of viral replication and captures the extent of replication with a luminescent marker. This is achieved through a vector containing a defective *env* gene, which ensures that budding virions are not replication-competent. Antivirogram, on the other hand, compares the cytopathic effect of HIV in the presence and absence of an ARV. Since this is a more indirect marker that relies on multiple cycles of replication, the results that are produced by the Antivirogram assay are typically less accurate than those of the PhenoSense assay.

The observed correlation between the PhenoSense and the Antivirogram assay depends on the distribution of observed levels of resistance in the study population. In one study where the samples had either very low or very high levels of resistance, the assays agreed well with a concordance of 91.5%[248]. Another study in which a large number of samples showed intermediate levels of resistance reported an agreement of only 36%[249], which suggests that the Antivirogram assay is less accurate for samples with intermediate resistance[250]. Due to the greater commercial success of the PhenoSense assay, the Antivirogram assay was discontinued in 2010.

Interpretation of Phenotypic Resistance Tests

In the following paragraphs, I summarize methods for the interpretation of IC50 values obtained from phenotypic resistance tests. For this purpose, I first introduce the notion of resistance factors and then discuss how these values can be transformed to discrete,

[244] Kellam and Larder 1994; Shi and Mellors 1997; Hertogs et al. 1998; Martinez-Picado et al. 1999; Petropoulos et al. 2000

[245] Hertogs et al. 1998

[246] Petropoulos et al. 2000

[247] The reverse primer of the assay ends at position 313 where it introduces a PinAI restriction site.

[248] Qari et al. 2002

[249] Wang et al. 2004b

[250] Zhang et al. 2005

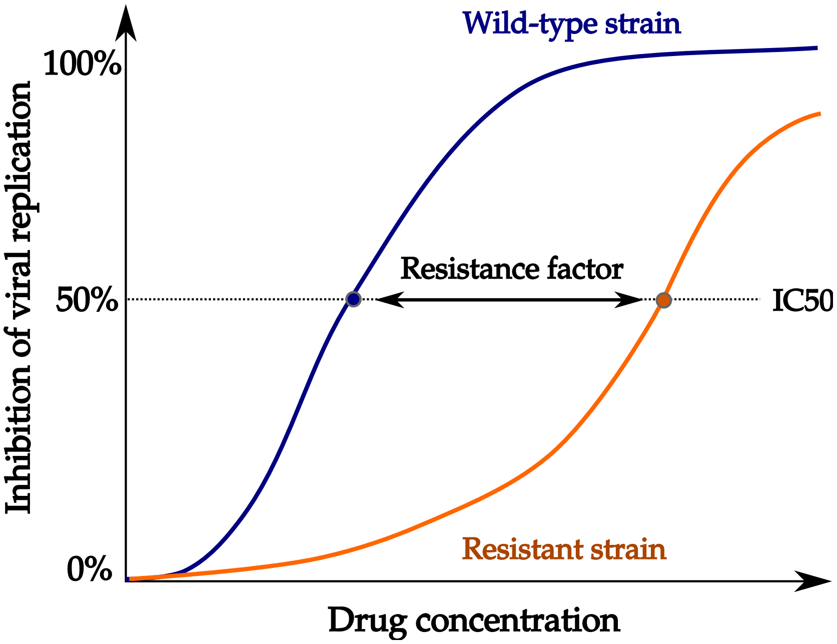


Figure 2.20: Determination of the resistance factor from a dose-response curve. The blue curve shows the inhibition of a wild-type virus by an antiretroviral drug, while the orange curve shows the inhibition of a mutated virus. The IC50 is the drug concentration at which viral replication is inhibited by 50%.

interpretable levels of resistance.

The Resistance Factor The resistance factor (RF)[251] is determined by comparing the IC50 of a mutated viral strain ($\mathrm{IC50}_{\mathrm{MT}}$) with the IC50 from the corresponding wild-type strain ($\mathrm{IC50}_{\mathrm{WT}}$):

$$RF = \frac{\mathrm{IC50}_{\mathrm{MT}}}{\mathrm{IC50}_{\mathrm{WT}}}$$

The interpretation of RF is straightforward. $RF > 1$ indicates a relative increase in resistance, $RF < 1$ indicates a relative increase in drug susceptibility, and $RF = 1$ indicates susceptibility at the level of the wild-type strain.

The RF is limited in that it reflects only a single point along the dose-response curve of a drug (Figure 2.20) and thereby discards information that may be obtained from considering the shape of the sigmoidal dose-response curve. Two drugs with the same RF may behave very differently when varying their dose[252]. This means that the IC50 underestimates the impact of HIV mutations that do not change the RF but influence the slope of the dose-response curve. Therefore, the use of measures other than the IC50 may be warranted when interpreting drug resistance. For example, the instantaneous inhibitory potential, which gives the log inhibition of single-round infectivity at clinical concentrations, may be favorable because it considers both slope and IC50[253]. Note that the approaches that are discussed in the following use only the IC50.

[251] The resistance factor is also called fold change.

[252] Prinz 2010

[253] Sampah et al. 2011

There are two challenges with respect to the RF. First, how to compare RFs across drugs? Second, how to make RFs more interpretable? The following paragraphs are concerned with answering these questions.

Standardization of Resistance Factors Before answering the first question, it is important to understand that RFs vary considerably depending on the tested ARV. This is intuitively clear because *RF* is defined relative to IC50_{WT}. A drug whose IC50_{WT} is low is generally associated with greater RFs than a drug whose IC50_{WT} is high. For example, the mean IC50 for subtype B isolates from untreated HIV-1 infected persons is 0.01 μM for ZDV, while the corresponding mean IC50 for TDF is 1.1 μM[254]. Hence, viral RFs relating to different drugs should never be compared directly but only after standardization to z-scores.

The approach of geno2pheno[resistance][255] uses the mean, μ, and standard deviation, σ, from the normal distribution of RFs from therapy-naive HIV-1 infected persons in order to compute the z-score of an estimated RF x, which is defined by

$$z_{\text{RF}} = \frac{x - \mu}{\sigma} \,.$$

Relative to therapy-naive persons, $z_{\text{RF}} < 0$ and $z_{\text{RF}} > 0$ indicate strains that are either more susceptible or less susceptible to a drug, respectively.

Discretization of Resistance Factors In order to improve the interpretability of the RF, it is necessary to find cutoffs that give rise to meaningful, discrete levels of resistance. Drug resistance is typically described using the SIR classification scheme, which considers three levels of resistance: *susceptible*, *intermediate*, and *resistant*. These resistance levels originate from bacterial resistance against antibiotics[256]. They are formally described according to the ISO 20776-1:2006 standard[257]:

> *Susceptible:* A bacterial strain is said to be susceptible to a given antibiotic when it is inhibited *in vitro* by a concentration of this drug that is associated with a high likelihood of therapeutic success.
>
> *Intermediate:* The sensitivity of a bacterial strain to a given antibiotic is said to be intermediate when it is inhibited *in vitro* by a concentration of this drug that is associated with an uncertain therapeutic effect.
>
> *Resistant:* A bacterial strain is said to be resistant to a given antibiotic when it is inhibited *in vitro* by a concentration of this drug that is associated with a high likelihood of therapeutic failure.

[254] Palmer et al. 2001

[255] Beerenwinkel et al. 2003

[256] Rodloff et al. 2008

[257] ISO 2006

These classes of resistance are used correspondingly in the realm of HIV. In clinical applications, where therapeutic drug dosing schemes are used, *susceptible* corresponds to full activity, *intermediate* to residual activity, and *resistant* to inactivity[258].

It is possible to map from *in vitro* measurements of resistance to levels of resistance by selecting cutoffs from the RF distribution. Initially, HIV resistance levels were determined by applying technically-motivated cutoffs to the RFs obtained from phenotypic tests[259]. However, these cutoffs were inaccurate because they did not take drug-specific effects into account, which are important for two reasons[260]. First, the susceptibility of HIV in treatment-naive persons differs across drugs[261,262]. Second, the *in vivo* effect of drugs is moderated by different metabolic pathways. PIs are a prime example for this. When prescribed, these drugs are usually combined with boosters that extend the half-life of these compounds[263]. As a consequence, boosted PIs exhibit superior potency *in vivo* than suggested by *in vitro* tests, which do not take boosting into account.

To circumvent these shortcomings, biologically-motivated cutoffs were developed. These cutoffs are selected based on the RF distribution of viral strains from therapy-naive patients. For example, Harrigan et al. (2001) defined a biological cutoff for differentiating susceptible and resistant strains at an RF defined by the distribution's mean plus two standard deviations. Although these cutoffs afforded an improvement over the technical cutoffs, they did not necessarily correlate well with treatment outcomes. Thus, clinically-motivated cutoffs were developed. These cutoffs are derived by determining the association between treatment outcomes and RF values. Initial clinical cutoffs relied on experts who would associate RFs, which were measured for specific drugs, with clinical outcomes. This approach was originally used by the geno2pheno[resistance] web server but was subsequently discontinued because it is time-intensive, not automatable, and subject to human biases. These problems were solved by statistical approaches for finding clinical cutoffs[264], which are described in more detail in Section 4.1.2.

2.5.3 *Sequencing of Viral Genomes*

The application of genotypic methods for the identification of viral drug resistance (Section 4.1) requires prior sequencing of viral genomes. Two approaches can be used for this purpose. Sanger sequencing has been the predominant sequencing method for the last three decades but the more recently established NGS is currently being widely adopted.

[258] Pironti et al. 2017b; Paredes et al. 2017

[259] Qari et al. 2002

[260] Perno and Bertoli 2006

[261] Harrigan et al. 2001; Parkin et al. 2004

[262] This variation is due to different wild-type IC50s and the differing impact of naturally occurring polymorphisms.

[263] Zeldin and Petruschke 2003

[264] Winters et al. 2008, 2009; Pironti et al. 2017b

Sanger Sequencing At the end of the 1970s, Sanger sequencing[265,266] heralded the era of first-generation sequencing technologies[267,268]. The approach by Sanger is based on the observation that DNA is formed by incremental linkage of individual deoxyribose nucleoside triphosphates (dNTPs), a reaction that is catalyzed by polymerase. In contrast to dNTPs, didexobyribose nucleoside triphosphates (ddNTPs) do not have the 3′ hydroxyl group that is required for bonding with other dNTPs. The incorporation of a ddNTP into a nascent nucleic acid terminates its synthesis by polymerase[269], which can be exploited for determining the sequence of a nucleic acid in the following way.

In the approach proposed by Sanger in 1977, ddNTPs are radiolabeled in order to differentiate the four types of nucleobases A, C, G, and T. Given a template DNA whose sequence is to be determined, four different reaction containers are set up. Each of the four preparation contains one of the four types of ddNTPs (either ddATP, ddCTP, ddGTP, or ddTTP) as well as all types of dNTPs, primers, and polymerase. The labeled ddNTPs are present in lower concentrations than the conventional dNTPs, which ensures that some synthesized fragments can reach full length. Due to the presence of ddNTPs, each of the four reactions generates thousands of nucleic acid fragments with varying lengths, for which the terminal nucleotide is known. It is possible to determine the sequence of nucleotides by performing gel electrophoresis, which separates the fragments according to size. By assigning each of the four reactions to a single lane in the gel, the nucleotide sequence can be obtained by moving from the shortest segment to the longest segment while noting the lane in which each segment was found (Figure 2.21).

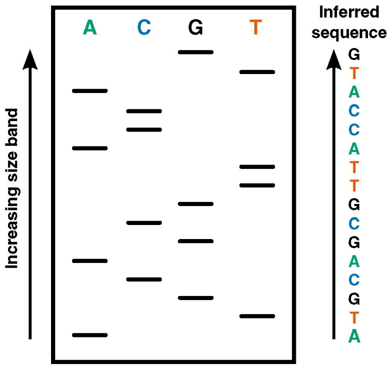


Figure 2.21: Sequencing via the Sanger method.
Adapted from Heather and Chain (2016) with permission from Elsevier.

The original Sanger sequencing approach was further refined throughout the years. One innovation was the substitution of radiolabeling with fluorescent dyes. The assignment of one of four fluorophores to each ddNTP gave rise to dye-terminator sequencing in which a single preparation is used instead of four preparations. This substantially decreased the extent of laboratory work and facilitated the automated computational analysis of gels[270]. Additionally, the use of fluorescent dyes removed the need of adhering to stringent safety protocols for handling radioactively labeled ddNTPs. By replacing gel electrophoresis with capillary electrophoresis in later years[271], the throughput of Sanger sequencing was further increased.

Current commercial Sanger sequencing systems[272] are extremely accurate with overall accuracies as high as 99.999%[273]. The errors that do occur are concentrated at the beginning and at the end of the reads. Errors are more prevalent in the first 50 bases of reads because smaller DNA fragments exhibit anomalies during gel elec-

[265] Sanger et al. 1977

[266] Sanger sequencing is also called the dideoxy chain-termination method whose characteristics are described later.

[267] Maxam and Gilbert 1977; Heather and Chain 2016

[268] The first-generation sequencing approach from Maxam and Gilbert is not discussed here for brevity's sake.

[269] Chidgeavadze et al. 1984

[270] Smith et al. 1986

[271] Luckey et al. 1990

[272] Karger and Guttman 2009

[273] Shendure and Ji 2008

trophoresis.[274] The end of the sequence is also subject to a higher error rate due to the reduced electrophoretic mobility of long reads and their relatively small number[275]. The maximal read length of Sanger sequence is at about 750 bp[276].

Sanger sequencing is suitable for the analysis of viral samples from patients that may contain a viral population consisting of different variants of the same viral species[277]. This is because individual variants of the same viral species can be detected in the form of overlapping peaks in the chromatogram. Since the level of detection is limited by noise in the fluorescence signal, current industrial sequencers can detect population variants at a prevalence of 10% to 20%[278].

Next-Generation Sequencing NGS[279] is characterized by the massive parallelization of the two steps of Sanger sequencing: synthesis of DNA and determination of the sequence. The greater depth of coverage afforded by NGS allows for the detection of variations at low population prevalence (e.g. at 1% in a population)[280]. This, however, comes at the cost of shorter reads[281] and higher error rates (at least 0.1%)[282]. In this section, I introduce the well-established sequencing approach that was developed by Solexa and later refined by Illumina, which is called the Solexa/Illumina approach in the following.

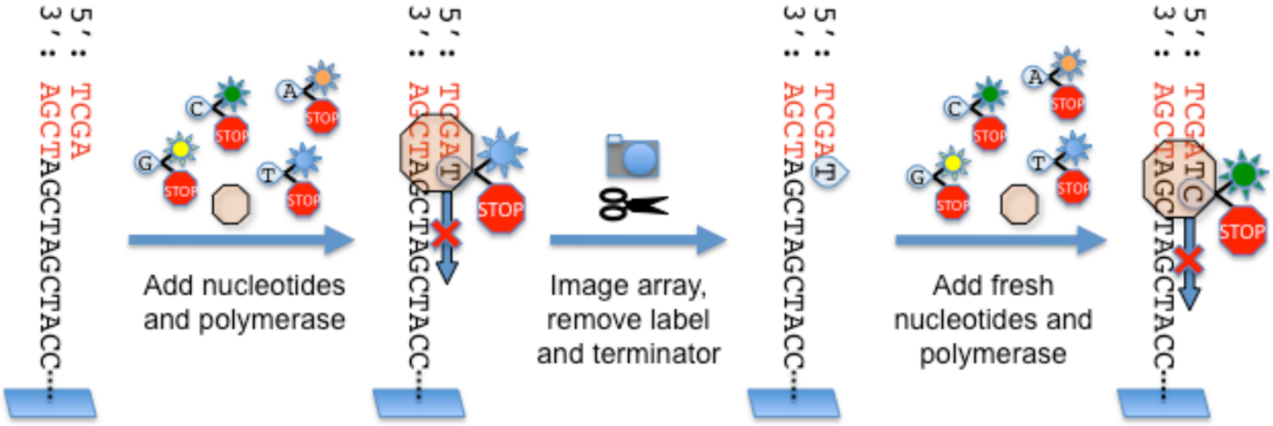

5':
3':
TCGA
AGCTAGCTAGCTACC...
C
A
G
T
STOP
Add nucleotides and polymerase
Image array, remove label and terminator
Add fresh nucleotides and polymerase

Figure adapted from Anderson and Schrijver (2010), licensed under CC BY-NC-SA 3.0.

Figure 2.22: Chemistry of Illumina/Solexa sequencing

In the approach by Solexa/Illumina, template DNA is amplified in a process termed bridge PCR. Single-stranded template DNA is added to a flow cell and immobilized by hybridization to anchor molecules such that templates arch over to adjacent anchor oligonucleotides during hybridization. By performing multiple rounds of PCR, arching clusters, that is, clusters containing thousands of clonal

[274] For short segments, the movement in the gel is substantially influenced by the dyes and the base composition.

[275] Ewing et al. 1998

[276] Anderson and Schrijver 2010

[277] Population variants that share specific mutations make up a viral quasispecies (Eigen, 1993).

[278] Tsiatis et al. 2010

[279] NGS is also called second-generation sequencing.

[280] Lin et al. 2014; Fox et al. 2014

[281] Shendure and Ji 2008

[282] Shendure and Ji 2008; Manley et al. 2016

nucleic acids are generated. About 50×10^6 of such clusters are generated[283]. After denaturing the clusters, sequencing can take place.

Sequencing based on the approach of Solexa/Illumina is similar to Sanger sequencing as it relies on the use of modified dNTPs[284] (Figure 2.22). Each type of dNTP carries a specific dye as well as a terminator for blocking the further polymerization of DNA. Once a nucleotide has been incorporated, further elongation is possible only after the terminator has been removed.

Sequencing is performed by iteratively cycling three steps. In the first step, all four types of dNTPs are added simultaneously such that the sequences are extended by a single nucleotide. In the second step, incorporated nucleotides are interrogated using an optical system that records the fluorescence signal. In the third step, the terminators are removed and the fluorescent labels are cleaved and degraded allowing for the next sequencing cycle to commence. Reads from the Solexa/Illumina approach suffer from increasing rates of errors at the end of the reads because nucleotides may be over- or underincorporated or terminator removal may fail[285].

The use of NGS poses two challenges. First NGS is computationally demanding because it requires the processing of thousands of short (e.g. 200 bp) reads. Second, NGS suffers from higher error rates than Sanger sequencing, which need to be taken into account when interpreting the results[286]. Despite these shortcomings, NGS is advantageous over Sanger sequencing as it affords a higher throughput at an improved resolution. A detailed account of the technical requirements for processing NGS data is provided in Section 4.1.4.

2.5.4 *Polymerase Chain Reaction*

PCR is the fundamental method that is used for the amplification and modification of DNA. PCR involves performing the following steps in multiple cycles (Figure 2.23): denaturation, annealing, and elongation. In the denaturation phase, double-stranded DNA is turned into single-stranded DNA by applying high temperatures (e.g. 95° C). Next, the temperature is reduced (e.g. to 50 °C) such that complimentary oligomers (primers) can attach to the single-stranded DNA. In the third step, the temperature is increased again (e.g. to 72 °C). Then, polymerase attaches to the region where the primer has bound and elongates the DNA segment to fully double-stranded DNA. Each PCR cycle roughly doubles the amount of DNA. For example, a single copy of DNA could be amplified to more than one million copies in only 20 PCR cycles.

[283] Voelkerding et al. 2009

[284] Voelkerding et al. 2009

[285] Dohm et al. 2008

[286] Loman et al. 2012

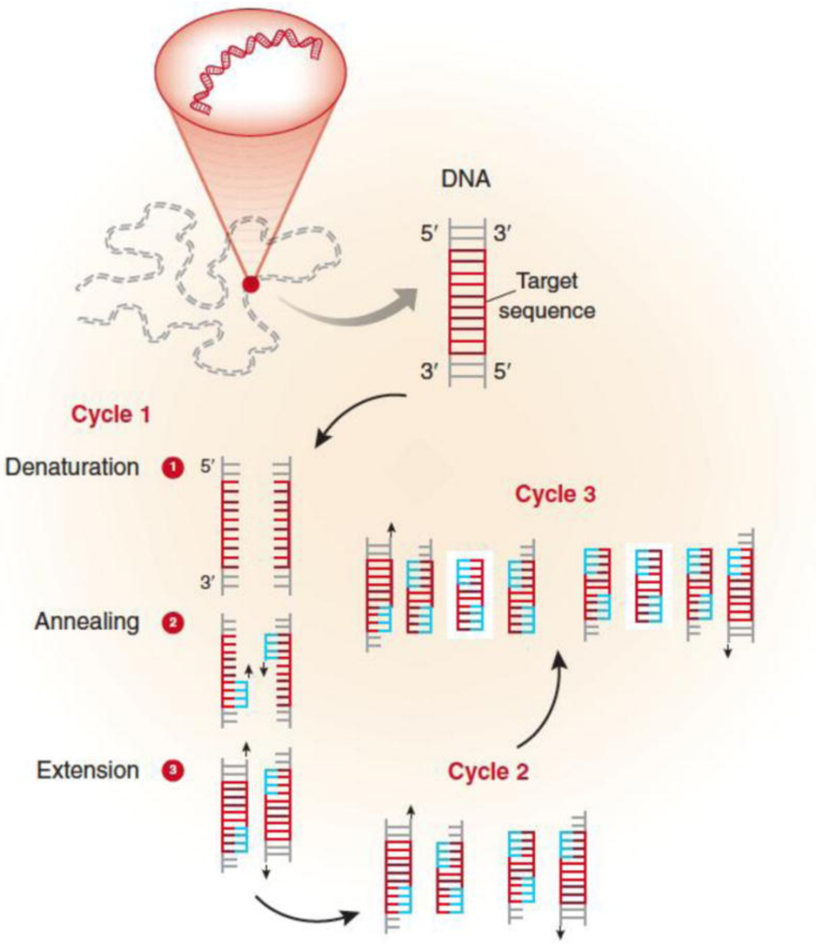


Figure 2.23: Steps involved in polymerase chain reaction.

Image reproduced with permission from Elsevier (Garibyan and Avashia, 2013).

Multiplex Polymerase Chain Reaction In contrast to conventional PCR, mPCR strives to amplify several template sequences simultaneously, which typically requires the use of multiple primers. Primer design for multiplex PCR is particularly challenging because the molecular characteristics of mPCR need to be considered (Section 6.1.1) and a combinatorial optimization problem has to be solved to find a suitable set of primers (Section 6.3.3).

Quantitative PCR Quantitative polymerase chain reaction (qPCR), which is also called real-time PCR, has the goal of measuring the concentration of amplified DNA in real time. The principle approach of qPCR is very similar to conventional PCR. In each cycle of PCR, amplified templates are tagged using specific fluorophores, which are excited with a beam of light. The emitted fluorescence is then detected. By calculating the increase in fluorescence between cycles, it is possible to quantify the template species[287]. While conventional PCR is used to amplify a genomic sequence for further experimental investigation, qPCR is often an experimental end point. Since qPCR is concerned with the quantification of nucleic acid sequences, primer amplification rates should be uniform to ensure that the estimates of qPCR are accurate.

To exemplify this point, let us consider the use of qPCR for studying gene expression levels. Assume we are studying three genes, A, B, and C, which have initial concentrations of $c_A = 1$ nM, $c_B = 2$ nM, and $c_C = 3$ nM. If we perform 20 cycles of qPCR, 1 nM of DNA could ideally be amplified to roughly 1 mM of DNA. If A, B, and C have maximum amplification rates of $\eta_A = \eta_B = \eta_C = 1$, the concentrations of the amplified templates would be $c'_A \approxeq 1$ mM, $c'_B \approxeq 2$ mM, and $c'_C \approxeq 3$ mM. Based on these results, we would correctly conclude that C is the most prevalent transcript, followed by B and C. Let us now assume that the amplification rates for A, B, and C are different, for example, $\eta_A = 1$, $\eta_B = 0.5$, and $\eta_C = 0.1$. In this case, each amplification cycle doubles the concentration of A but only increases the concentrations of B and C by 50% and 10%, respectively. Then, after performing qPCR, the concentrations would be $c'_A = 1 \times 10^{-9} \times 2^{20} \approxeq 1$ mM, $c'_B = 2 \times 10^{-9} \times 1.5^{20} \approxeq 6.6$ μM, and $c'_C = 3 \times 10^{-9} \times 1.1^{20} \approxeq 0.2$ nM. Based on these results, we would falsely conclude that A is the most prevalent transcript, followed by B and C, merely because the amplification rates were different.

There are two quantities that are essential for qPCR. The threshold cycle, Ct, measures the cycle at which the fluorescence signal becomes exponentially larger than the background noise. Based on the Ct of a reference sequence, Ct_{ref}, it is possible to define the cycle delay for the template of interest (denoted by OI), $\Delta Ct_{OI} = Ct_{\text{OI}} - Ct_{\text{ref}}$. The values of ΔCt are typically greater than zero. Templates that are amplified efficiently will have ΔCt close to 0, while those that are amplified poorly will have large values of ΔCt.

To exemplify the cycle delay, let us consider the previous example once again. Let us assume that $Ct = 1\mu$M such that $Ct_A = 10$, $Ct_B = 18$, and $Ct_C = 73$. If we define $Ct_{\text{ref}} := Ct_C$, then $\Delta Ct_A = Ct_A - Ct_A = 0$, $\Delta Ct_B = Ct_B - Ct_A = 8$, and $\Delta Ct_C = Ct_C - Ct_A = 63$. The cycle delays indicate that the amplification of B and C lag 8 and 63 cycles behind the amplification of A, respectively. The high value of ΔCt_C indicates that the amplification of C is exceedingly slow.

[287] Arya et al. 2005

3
Methodological Foundations

> [T]ruth ... is much too complicated to allow anything but approximations ...
>
> John von Neumann

The 2000s and 2010s signify periods in which machine learning has brought about technological innovations in various areas such as image recognition[1], face recognition[2], and gaming[3]. In the biomedical field, machine learning was successfully used to improve the treatment of diseases including diabetic retinopathy[4], cancer[5], and those caused by viral infections[6]. Since machine learning techniques form the methodological basis of this work, I first give an overview of machine learning (Section 3.1) and thereafter introduce the core concepts of supervised learning (Section 3.2). Measures for evaluating the predictive performance of classifiers are covered in Section 3.3. Since the supervised learning methods of support vector machines and logistic regression are especially relevant, these models are introduced in Section 3.4. The principles of clustering and significance tests are briefly covered in Section 3.5 and 3.6, respectively. Finally, an overview of optimization with linear programs is provided in Section 3.7.

3.1 Overview of Machine Learning

Machine learning is a subfield of artificial intelligence that is concerned with enabling computers to learn from data without explicit programming. There are three machine learning scenarios: supervised, unsupervised, and reinforcement learning[7]. Supervised learning requires the observation of input variables as well as corresponding outcomes. Its goal is to learn a generalized model that is capable of accurately estimating the outcomes for new inputs. In unsupervised learning, the outcomes are not available. Therefore, unsupervised learning is not concerned with prediction but rather with structuring related data according to their distribution. Applications of unsupervised learning include clustering and dimensionality reduction. In clustering, the goal is to assign observations into dis-

[1] Krizhevsky et al. 2012

[2] Taigman et al. 2014

[3] Chen 2016

[4] Gulshan et al. 2016

[5] Kourou et al. 2015

[6] Lengauer and Sing 2006

[7] There is also a combination of supervised and unsupervised learning, which is called semi-supervised learning. In semi-supervised learning, predictive models are fitted using a large amount of unlabeled data and a small amount of labeled data.

tinct groups, which represent their characteristics. Dimensionality reduction is concerned with projecting data into spaces with reduced dimensionality, while retaining as much information as possible.

Finally, reinforcement learning studies how an agent can learn from interactions with the environment. Reinforcement learning is different from both supervised and unsupervised learning because it does not rely on a fixed learning data set but on exploration of the environment, which produces non-deterministic outcomes. In particular, reinforcement learning is different from active learning, which refers to supervised learning problems in which the algorithm can retrieve additional labels. This is because reinforcement learning is not concerned with prediction[8]. For more information on reinforcement learning, I refer the interested reader to the excellent book by Sutton et al. (1998).

The following sections on machine learning are restricted to supervised and unsupervised learning. Their content is based on the books by Hastie et al. (2009) and Schölkopf and Smola (2001) if not stated otherwise.

3.2 *Supervised Learning*

Supervised learning is concerned with the identification of predictive models that are capable of accurately estimating the outcomes for new measurements. These models are fitted using a training data set containing measurements of features and their corresponding outcomes. Section 3.2.1 introduces the relevant notation and the two types of supervised learning tasks, regression and classification. Supervised learning can be seen as a form of function estimation in which we want to minimize the expected prediction error (Section 3.2.2). In order to reason about models for a specific prediction task, it is useful to think about the bias-variance tradeoff (Section 3.2.3). The errors of a model can be determined on the training data or on an independent test set (Section 3.2.4), giving rise to estimates of the in-sample or the extra-sample error, respectively (Section 3.2.5). To limit the complexity of a model, regularization (Section 3.2.6) or feature selection can be performed (Section 3.2.7).

3.2.1 *Preliminaries*

Input data for supervised learning consist of pairs of features X and outcomes Y. Features are represented by the feature matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, where N indicates the number of observations and p gives the number of features. The feature vector of the i-th sample is denoted by $x_i \in \mathbb{R}^p$. In case of a regression task, the vector of

[8] Reinforcement learning can be seen as an instance of a state space search problem.

outcomes is given by $Y \in \mathbb{R}^N$. In case of a classification task, the set of possible outcomes, G, is discrete and the vector of outcomes is denoted by $Y \in G^N$. The outcome, G, can indicate categorical or ordinal data.

The two types of learning scenarios can be illustrated by the following example. Assume that we would like to predict the level of precipitation, for example, given observations of temperature, atmospheric pressure, humidity, and so on. When we are interested in predicting the precipitation per square meter (i.e. a continuous variable), this gives rise to a regression problem.

The same problem can be formulated as a classification task by discretizing the outcome. For example, we could use the categorical outcomes $G = \{-1, 1\}$ where the two classes -1 and 1 represent the states *No Rain* and *Rain*, respectively. Classification problems that deal with two classes are particularly common and are called binary classification tasks. A classifier based on ordinal outcomes may consider the classes 0 (*No Precipitation*), 1 (*Little Precipitation*), 2 (*Medium Precipitation*), and 3 (*High Precipitation*). Note that there is an ordering associated with the classes such that $0 < 1 < 2 < 3$. For example, the ordering indicates that class 1 is associated with a smaller level of precipitation than class 3.

3.2.2 *Supervised Learning as Function Estimation*

Supervised learning is a form of function estimation, in which we would like to obtain a function $f(X)$ for predicting Y given X that minimizes the risk, where risk is defined by the expected prediction error. The expected prediction error can be estimated by defining a loss function $L(Y, f(X))$, which determines the penalties that are incurred when $f(X)$ deviates from Y. Let X and Y be random variables from the joint probability distribution, $\Pr(X, Y)$. Then, the risk, $R(f(X))$, can be formally defined as the expectation over the loss function:

$$\begin{aligned} R(f(X)) &= E(L(Y, f(X))) \\ &= \int L(Y, f(X)) \, \mathrm{d} \Pr(X, Y) \, . \end{aligned}$$

Since $\Pr(X, Y)$ is usually not known, the empirical risk needs to be determined instead. The empirical risk relies only on the available, discrete set of data and is defined as

$$R_{emp}\,(f(X)) = \frac{1}{n} \sum_{i=1}^{N} L(Y_i, f(X_i)) \, .$$

An optimal prediction function can be found by minimizing the empirical risk,

$$f(X)^* = \min_{f(X)} R_{\text{emp}}\left(f(X)\right) .$$

The selected loss function critically influences the result of empirical risk minimization. The set of eligible loss functions depends on the type of the prediction task. For example, squared-error loss, $L(f(X), Y) = (Y - f(X))^2$, is often used for regression, while classification often uses the zero-one loss,

$$L(f(X), Y) = \begin{cases} 1 & \text{if } Y \neq f(X) \\ 0 & \text{else} \end{cases} .$$

3.2.3 *The Bias-Variance Decomposition*

Minimizing the empirical risk is a useful tool for learning prediction functions. In order to select the best possible prediction function, however, the expected risk should be minimized over several choices of training sets $\mathcal{T}$. This idea gives rise to the notion of the expected prediction error, which can be decomposed into a term describing the squared bias and a term describing the variance of the prediction model. For example, the mean squared error (MSE) that results from the prediction $\hat{y}_0$ of the outcome y_0 at the test point x_0 can be decomposed as

$$\begin{aligned} MSE(x_0) &= E_{\mathcal{T}}[y_0 - \hat{y}_0]^2 \\ &= \sigma_\epsilon^2 + E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)]^2 + [E_{\mathcal{T}}(\hat{y}_0) - y_0]^2 \\ &= \sigma_\epsilon^2 + Var_{\mathcal{T}}(\hat{y}_0) + Bias^2(\hat{y}_0) . \end{aligned}$$

The first term in the final equation, σ_ϵ^2, is the irreducible error. This quantity describes the variance of the target around the mean. The remaining two terms depend on the prediction model, which should yield an appropriate trade-off between variance and bias. Variance describes the extent to which the predictions of a model would vary given different training data. Non-parametric models are prone to greater variance, while parametric models typically have low variance.[9] Bias represents the systematic deviation of model predictions from the true outcomes. Parametric models, which typically have strong assumptions, often have greater bias than non-parametric models, which typically have few assumptions. For example, using logistic regression on nonlinear data would incur greater bias than the use of support vector regression, which can handle nonlinearity, given an appropriate kernel function.

Minimizing both bias and variance simultaneously is impossible because these quantities have opposing trends. For example, flexible

[9] A parametric model (e.g. logistic regression) has a fixed number of parameters. A non-parametric model (e.g. nonlinear support vector machine), on the other hand, can fit a variable number of parameters.

models (e.g. nonlinear models) have low bias (due to weak assumptions) but high variance (due to many parameters), while stiff models (e.g. linear models) have high bias (due to strong assumptions) but low variance (due to few parameters).

The type of model that should be considered for a specific supervised learning task critically depends on the amount of data and its distribution. For example, models with low bias and high variance are often appropriate when $N \gg p$ because the large number of training samples N may reduce the variance of the model.[10] In a scenario where the distribution of the data follows a certain probabilistic distribution, a model with high bias and low variance may be suitable. For example, a linear model would make few systematic errors when the squared errors largely follow a normal distribution.

3.2.4 Training and Test Errors

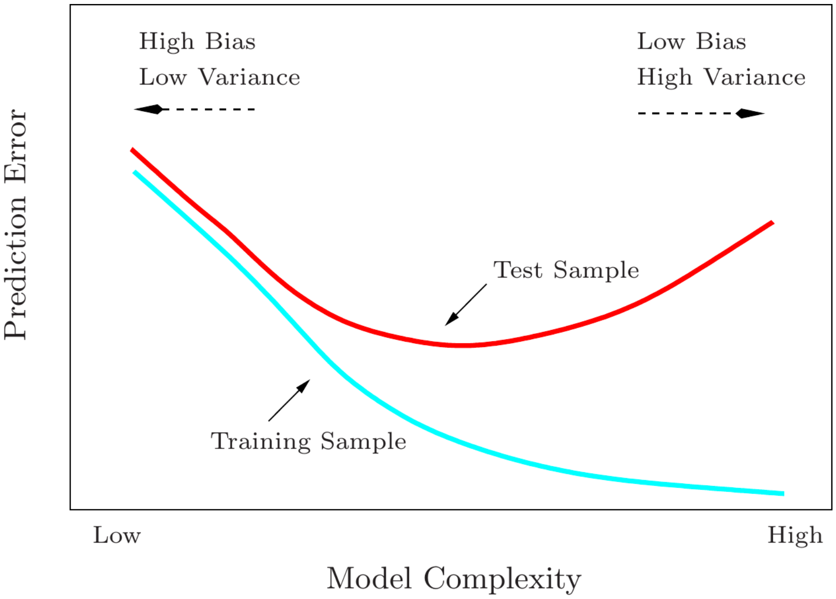


Figure reproduced with permission from Springer Series in Statistics (Hastie et al., 2009).

Figure 3.1: Relationship between training and test error. Models that are too complex allow for low training errors at the cost of high test errors. Models that are too simple exhibit similar, albeit relatively high test and training errors.

Predictive errors can occur either when training (fitting) a model (training error) or when the model is validated on an independent test set (test error). The training error is the mean loss over the training samples. It is defined by

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

where $\hat{f}(x_i) = \hat{y}_i$ indicates the prediction for the i-th sample. Models should never be selected based on the training error alone due to

[10] Note that the $\gg$ operator is an informal version of the $>$ operator. It models the notion that the left-hand side is substantially larger than the right-hand side.

the optimism of the training error. The training error is optimistic because it is typically lower than the error that would be observed if the model were applied on another data set (Figure 3.1). Instead, the generalization error (test error) should be used. Given a fixed training set T and samples of X and Y that are drawn from the population distribution, the test error is defined as the expected loss,

$$Err_T = E[L(Y, \hat{f}(X))|T] \, .$$

The generalization error is also called extra-sample error because X and Y do not necessarily coincide with the observations that are contained in the training set T. A related measure is the expected prediction error,

$$Err = E[L(Y, \hat{f}(x))] = E[Err_T] \, ,$$

which averages out the randomness of selecting a specific training set T. Ideally, model selection should be performed using the test error because supervised models are fitted using a single training set, T. If the test error cannot be determined, the expected prediction error can be used instead.

3.2.5 *Estimating Model Errors*

In order to select a model based on the performance on the training data alone, the in-sample error can be used. This measure penalizes model complexity in order to compensate for the optimism of the training error. The extra-sample error (test error), on the other hand, requires that the model is evaluated on a separate data set. In the following paragraphs, I first introduce methods for estimating the in-sample error and then describe methods for estimating the extra-sample error.

Estimating the In-Sample Error

In the following paragraphs, I present two approaches for estimating the in-sample prediction error, the C_p statistic and the Akaike information criterion (AIC). For a fixed training set T, the in-sample error is defined as

$$Err_{in} = \frac{1}{N} \sum_{i=1}^{N} E_{Y^0}[L(Y_i^0, \hat{f}(x_i))|T] \, ,$$

where Y^0 indicates the observation of new responses for each of the training points x_i with $i = 1, \ldots, N$. The optimism is defined as the difference between in-sample error and training error $\overline{err}$:

$$\text{op} = Err_{in} - \overline{err} \, .$$

The value of op is typically positive because $\overline{err}$ underestimates the error. The average optimism is the expected optimism over the training set outcomes,

$$\omega = E_y(\mathrm{op})\,.$$

The expected optimism can be expressed as

$$\omega = \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)\,,$$

where $Cov(\hat{y}_i, y_i) = E[(\hat{y}_i - E(\hat{y}_i))(y_i - E(y_i))]$ indicates the covariance between the estimates, $\hat{y}_i$, and the observed outcomes, y_i. Thus, the expected optimism decreases when the number of observations N increases or when y_i has little influence on the estimate $\hat{y}_i$. For linear fits with d parameters, the sum of covariances simplifies to

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = d\sigma_\epsilon^2\,,$$

where σ_ϵ is the standard deviation of the additive error in $Y = f(X) + \epsilon$. So, given a linear model, the fluctuation in the predictions is governed by the variance of the additive error σ_ϵ^2 and the number of model features d.

Since the estimate of the in-sample error can be decomposed into

$$\hat{Err}_{in} = \overline{err} + \hat{\omega}\,,$$

the in-sample error can be estimated by the C_p statistic,

$$C_p = \overline{err} + 2\frac{d}{N}\hat{\sigma}_\epsilon^2\,.$$

For log-likelihood loss functions, the AIC[11] can be used to estimate the in-sample error. It is based on the following relationship, which holds when $N \to \infty$:

$$-2E[\log \Pr_{\hat{\theta}}(Y)] \sim -\frac{2}{N}E[\mathrm{loglik}] + 2\frac{d}{N}\,.$$

Here, $\Pr_{\hat{\theta}}(Y)$ is a family of densities for Y and $\hat{\theta}$ indicates the maximum-likelihood estimate of the distribution parameters θ. The maximized log-likelihood is defined by

$$\mathrm{loglik} = \sum_{i=1}^{n} \log \Pr_{\hat{\theta}}(y_i)\,.$$

For the logistic regression model, the AIC is defined as

$$AIC = -\frac{2}{N}\mathrm{loglik} + 2\frac{d}{N}\,.$$

[11] Akaike 1998

Estimating the Extra-Sample Error

The extra-sample error is estimated by evaluating the model on data that was not used for training the model. In the following, I present two approaches for estimating the extra-sample error. The first approach directly estimates the generalization error, while the second approach estimates the expected prediction error.

Estimating the Generalization Error The generalization error can be estimated using hold-out validation. This entails splitting the data into three parts: a training set, a validation set, and a test set (Figure 3.2). The training set is used to fit the models. The validation set is used to select the parameters at which the models perform best. Finally, the test set is used to estimate the generalization error associated with the chosen models. The data are typically split such that 50% of the observations are used for training, while 25% of observations are used for validation and testing, respectively. The reason why models are selected on the validation set and not on the test set is to prevent optimistic estimates of the test error. By selecting model parameters on the independent validation set, the measured error on the test set resembles the generalization error more closely.

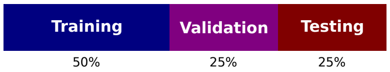


Figure 3.2: Hold-out validation.

Hold-out validation is easily implemented and requires a smaller runtime than cross validation, which is described later. However, hold-out validation is problematic for small data sets for which the random assignment of observations to training, validation, and test sets may critically influence the estimated predictive performance. This is particularly the case if specific subsets of data exhibit distinct feature distributions, which may require sampling strategies such as stratified sampling.

Estimating the Expected Prediction Error The expected prediction error can be estimated using methods such as the bootstrap or cross validation (CV). These methods are particularly useful when few data are available, in which case the generalization error cannot be estimated well. In the following, I limit myself to describing CV.

In k-fold cross validation the data set is split into k folds. The i-th of k rounds of CV entails training a model using the samples contained in all folds except for the i-th fold and then determining the predictive performance by applying the model on the observations of the i-th fold (Figure 3.3). Because the assignment of observations to folds is a random process, the estimates of the expected prediction error may vary across CV runs. To reduce the impact of randomness, several runs of CV are typically performed, after which the obtained errors are averaged.

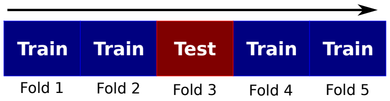


Figure 3.3: Cross validation for $k = 5$.

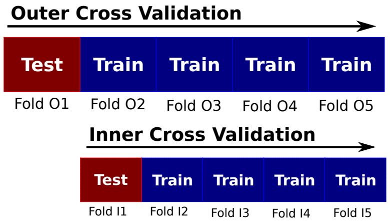


Figure 3.4: Nested cross validation for $k = 5$.

In order to incorporate model selection into CV, one can determine the in-sample error using the training data or perform nested cross validation (NCV). In NCV, two interlaced runs of CV are performed (Figure 3.4). One round of NCV entails the following steps. Having selected an outer CV fold for testing ($k = 1$ in Figure 3.4), an inner CV run is performed using the data from all other outer folds. Based on the results from this inner CV run, the parameters maximizing the predictive performance are selected. Subsequently, the prediction error for a single fold is determined by training a model with the selected parameters using the current outer training folds and evaluating it on the currently selected outer fold for testing. The expected prediction error can be obtained once the errors across all outer folds have been determined.

3.2.6 *Limiting Model Complexity via Regularization*

Models with many parameters can easily be overtrained (overfitted). A model is said to be overtrained if it has learned the peculiarities of the training data. Complex models are particularly prone to overfitting in which case they suffer from high variance and do not generalize well. Overtraining can be curbed through regularization. Regularization involves penalizing the model coefficients with a regularization parameter, λ, which controls model complexity. For least-squares loss, coefficients $\hat{\beta}$ can be regularized by setting

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} ,$$

where $\lambda \sum_{j=1}^{p} |\beta_j|^q$ is the regularization term. The choice of q defines which regularization method is used. For $q = 0$, the total number of coefficients is penalized. Therefore, $q = 0$ corresponds to best subset selection, which is described later. Setting $q = 1$ or $q = 2$, on the other hand, corresponds to the shrinkage methods of Lasso and ridge regression, respectively. The difference between the two methods is that Lasso regression uses an L1-norm, while ridge regression uses an L2-norm. The L1-norm that is used in Lasso regression ensures that the coefficients of some coefficients become zero.

3.2.7 *Feature Selection*

There are two situations in which it is useful to select a subset of the available features. On the one hand, features can be selected for model selection. In this case, feature selection mainly serves as a means of limiting model complexity. On the other hand, feature selection can be performed after a model has been trained. This allows for the interpretation of the most salient model features.

Feature Selection for Model Selection While regularization regulates model complexity at the time the model is fit, model complexity can also be reduced using feature selection. Feature selection determines a subset of features that yields the best tradeoff between model complexity and predictive performance. Suitable quantities for this purpose include the C_p statistic and the AIC. Best-subset selection trains a model for all 2^p subsets of features and then selects the model maximizing the chosen performance measure. This approach already becomes infeasible at small values of p. For example, for $p = 20$, one would have to fit $2^{20} = 1\,048\,576$ models. However, using the leaps-and-bounds algorithm[12], best subset selection can still be performed for p as large as 40.

Due to the infeasibility of best subset selection for data sets with many features, greedy subset selection is often performed. Greedy subset selection can be performed either in the form of forward-stepwise selection or in the form of backward-stepwise selection. Here, I limit myself to describing the backward-selection procedure because the procedure for forward-stepwise selection is analogous. In backward-stepwise selection, we start with a full model, that is, a model fitted using all p features. In each step of the procedure, the least predictive feature according to a specific criterion is eliminated. One way of selecting the least predictive feature is to consider the feature whose coefficient has the smallest absolute z-score. For a coefficient $\beta \in \mathbb{R}$, the z-score is defined as

$$z = \frac{\beta - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation of the normal distribution from which β was drawn. Once the least important feature has been removed, the model is refit and additional variables are eliminated until all coefficients have z-scores above a threshold.

Feature Selection for Model Interpretation In order to interpret the impact of features from a model that relies on many features, it is useful to limit the analysis to the most important features of the model. The importance of the i-th feature can be quantified using the value of the i-th model coefficient, β_i. Let $B = \sum_{i=1}^{p} |\beta_i|$ denote the sum of absolute feature weights. If we order the coefficients by decreasing absolute value, we obtain $\beta_{(1)}, \ldots, \beta_{(p)}$ with $|\beta_{(i-1)}| \geq |\beta_{(i)}|, \forall i \in \{2, \ldots, p\}$. We can select the ratio $r \in [0, 1]$ of features with the greatest absolute weight in the following way. Let $B_{\text{cut}} = rB$ indicate the cutoff on the total absolute model weight with respect to

[12] Furnival and Wilson 1974

Prediction/Reference	Class +1	Class -1
Class +1	TP	FP
Class -1	FN	TN

Table 3.1: Structure of the confusion matrix resulting from the predictions of a binary classifier.

r. Then, we set

$$k = \arg \max_{j \in \{1,\ldots,p\}} \min(0, (\sum_{i=1}^{j} \beta_i) - B_{\text{cut}})$$

and interpret only the features associated with the coefficients $\beta_{(1)}, \ldots, \beta_{(k)}$. Typically only those features making up at least 50% of the total absolute model weight are considered (i.e. $r \geq 0.5$). However, the specific choice of r depends on the distribution of the model weights, as this influences the number of selected features. The presented approach is particularly suitable for regularized models. This is because L_2 regularization leads to sparse models, which have many coefficients whose value is zero.

3.3 *Measures of Predictive Performance*

Regression and classification give rise to distinct measures of predictive performance. Since the machine learning models that were developed in this work perform binary classification, I only deal with performance measures for this task in the following. For binary classification, the confusion matrix (Section 3.3.1) can be used to derive performance measures for non-scoring (Section 3.3.2) or scoring classifiers (Section 3.3.3).

3.3.1 *The Confusion Matrix*

All quantities that describe the predictive performance of classification models can be derived from the confusion matrix. For a binary classification task with a positive class, $+1$, and a negative class, -1, the confusion matrix is a 2×2 matrix that indicates the number of times that a classifier correctly predicted the positive and the negative class, as well as the number of times that it confused the positive with the negative class, and vice versa (Table 3.1). The confusion matrix gives rise to four quantities: the number of true positives (TPs), which indicates how often the positive class was correctly predicted; the number of false positives (FPs), which indicates how often the positive class was falsely predicted; the number of false negatives (FNs), which indicates how often the negative class was falsely predicted; and the number of true negatives (TNs), which indicates how often the negative class was correctly predicted.

3.3.2 *Performance Measures for Non-Scoring Classifiers*

A non-scoring classifier computes estimates of the class labels via $\hat{y} \in G$. For example, for the prediction of precipitation, a non-scoring classifier such as k-nearest neighbors may output $\hat{y} \in \{\textit{No Rain}, \textit{Rain}\}$. For a non-scoring classifier, it is useful to summarize the entries in the confusion matrix that results from the application of the model. In the following, I introduce the most important quantities that can be obtained from the confusion matrix such as the accuracy or sensitivity and specificity.

Accuracy Accuracy captures the overall predictive performance of a classifier by calculating the fraction of correct predictions among all predictions via

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Accuracy is in the range $[0, 1]$. Well-performing classifiers have accuracies close to 100%. Note that accuracy is ill-suited when the class distribution is imbalanced. For example, imagine a scenario where 90% of the samples are from the negative class and only 10% are from the positive class. In this case, we can easily construct a well-performing classifier (accuracy of 90%) by assigning all samples to the negative class. This showcases why it is useful to consider performance measures that are appropriate when the class distribution is imbalanced.

Sensitivity and Specificity Sensitivity and specificity are well-suited when the distribution of class labels is imbalanced because sensitivity represents the performance on the samples from the positive class, while specificity represents the performance on the samples from the negative class. These two quantities are defined in the following way:

$$\begin{aligned}
\text{sensitivity} = \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR} = 1 - \frac{\text{FN}}{\text{FN} + \text{TP}} \\
\text{specificity} = \text{TNR} &= \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{FPR} = 1 - \frac{\text{FP}}{\text{FP} + \text{TN}}
\end{aligned}$$

Note that true positive rate (TPR) and true negative rate (TNR) are synonyms for sensitivity and specificity, respectively. Moreover, sensitivity and specificity correspond to one minus false negative rate (FNR) and one minus false positive rate (FPR), respectively.

There are several ways in which sensitivity and specificity can be summarized as a single quantity. One way is presented by the balanced accuracy, which is the arithmetic mean of sensitivity and specificity. Another way of summarizing sensitivity and specificity is Youden's index, which is defined as $Y =$ sensitivity + specificity $- 1$.

Additional Quantities Using the confusion matrix, we can define additional useful quantities. Precision gives the fraction of positively predicted samples that were correctly predicted. The false discovery rate (FDR) indicates the fraction of positively labeled samples that were falsely predicted. The negative predictive value (NPV) gives the fraction of negatively predicted samples that were correctly predicted. These quantities are formally defined as follows:

$$\begin{aligned} \text{precision} = \text{PPV} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{FDR} &= \frac{\text{FP}}{\text{FP} + \text{TP}} \\ \text{NPV} &= \frac{\text{TN}}{\text{TN} + \text{FN}} \end{aligned}$$

Note that precision and positive predictive value (PPV) are synonymous. When precision is used as a performance measure, sensitivity is typically referred to as recall.

3.3.3 *Performance Measures for Scoring Classifiers*

A scoring classifier outputs numeric values (scores) that are associated with the labels via $\hat{y} \in \mathbb{R}$. To obtain estimates of the class labels from a scoring classifier, it is necessary to define a cutoff that separates the estimates for observations from the two classes. For example, in support vector classifications (SVCs) *No Rain* would be encoded as -1 and *Rain* as 1. If the classifier performs well, it will predict negative values when it will not rain but positive values when it will rain. However, the question is where the decision boundary (cutoff) should lie. For example, is it better to classify observations with scores $\hat{y} > 0$ as *Rain* or would the predictive performance be higher if $\hat{y} > 0.5$ were required? To determine the quality of a scoring classifier, it is useful to consider its performance across multiple cutoffs.

To evaluate the predictive performance of a classifier across all relevant cutoffs, the area under the receiver operating characteristic curve (AUC) is frequently used[13]. The receiver operating characteristic (ROC) curve plots sensitivity as a function of the FPR; the AUC is simply the area under this curve. The AUC is in the range $[0, 1]$. Since a random classifier has an AUC of 0.5, the AUC is typically greater than 0.5. A classifier that allows for the perfect separation of the estimates for observations from two classes has an AUC of 1.

[13] Hanley and McNeil 1982

AUCs below 0.5 typically do not occur as they only arise when the class labels have been switched.

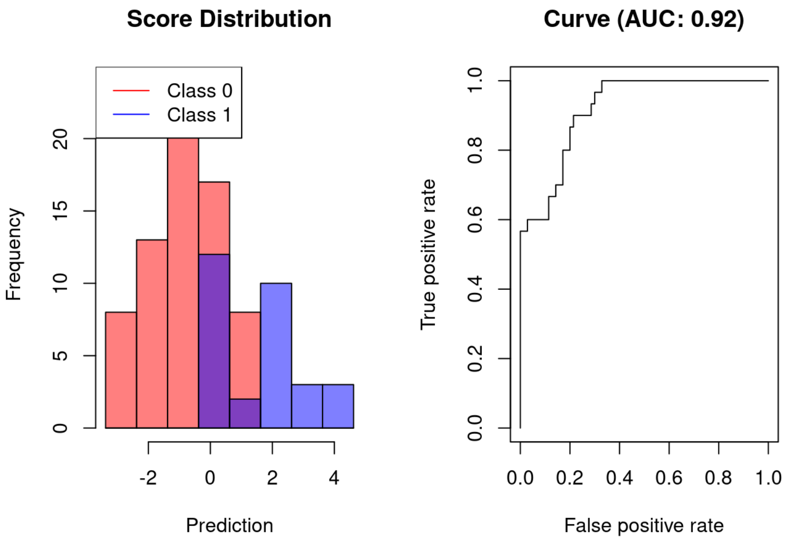


Figure 3.5: Example for the AUC. Scores for the positive class (70 samples) were drawn from $\mathcal{N}(-1, 1)$, while those from the negative class (30 samples) were drawn from $\mathcal{N}(1, 1.25)$.

Figure 3.5 exemplifies the AUC. The left-hand side of the plot shows the distribution of the classifier estimates. Evidently, the scores allow for separating the samples from the two classes well. However, since the separation is not perfect, the AUC, which is shown on the right-hand side of the figure, is slightly below 1.

3.4 *Models for Supervised Learning*

A multitude of machine learning models for supervised learning is available. In this work, two supervised models were used: logistic regression and support vector machines (SVMs). Logistic regression (Section 3.4.1) is a simple and interpretable approach for classification. SVMs (Section 3.4.2) allow for models with varying levels of complexity through the use of kernel functions and can be used for both, classification and regression.

3.4.1 *Logistic Regression*

Logistic regression models the posterior probabilities of K classes via linear functions in X. The model can be formulated in terms of $K - 1$

log-odds:

$$\begin{aligned}
\log \frac{\Pr(Y=1|X=x)}{\Pr(Y=K|X=x)} &= \beta_{10} + \beta_1^T x \\
\log \frac{\Pr(Y=2|X=x)}{\Pr(Y=K|X=x)} &= \beta_{20} + \beta_2^T x \\
&\vdots \\
\log \frac{\Pr(Y=K-1|X=x)}{\Pr(Y=K|X=x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x
\end{aligned}$$

The probabilities of the individual classes can be obtained via

$$\begin{aligned}
\Pr(Y=k|X=x) &= \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}, \forall k \in \{1\ldots, K-1\} \\
\Pr(Y=K|X=x) &= \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}
\end{aligned}$$

Logistic regression models are fit by maximizing the conditional likelihood of Y given X. Given model parameters θ, the log-likelihood is

$$l(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta)\,,$$

where $p_k(x_i;\theta) = \Pr(Y=k|x_i;\theta)$. Assume that there are two classes that are encoded by $y_i \in \{0,1\}$. Further, let $p_1(x;\theta) = p(x;\theta)$ and $p_2(x;\theta) = 1 - p(x;\theta)$. Then the log-likelihood can be written as

$$\begin{aligned}
l(\beta) &= \sum_{i=1}^{N} y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta)) \\
&= \sum_{i=1}^{N} y_i \beta^T - \log(1 + e^{\beta^T x_i})\,.
\end{aligned}$$

The parameters of the two-class logistic regression model, β, are defined by $\beta = \{\beta_{10}, \beta_1\}$. The feature vector x_i is assumed to contain the constant term 1 for the intercept. The log-likelihood can be maximized by setting its derivatives to zero via

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i(y_i - p(x_i;\beta)) = 0\,.$$

The model coefficients can be determined by solving these score equations using the Newton-Raphson algorithm.

3.4.2 *Support Vector Machines*

SVMs were introduced by Cortes and Vapnik (1995). They minimize a vector of coefficients β to find a hyperplane that maximizes the

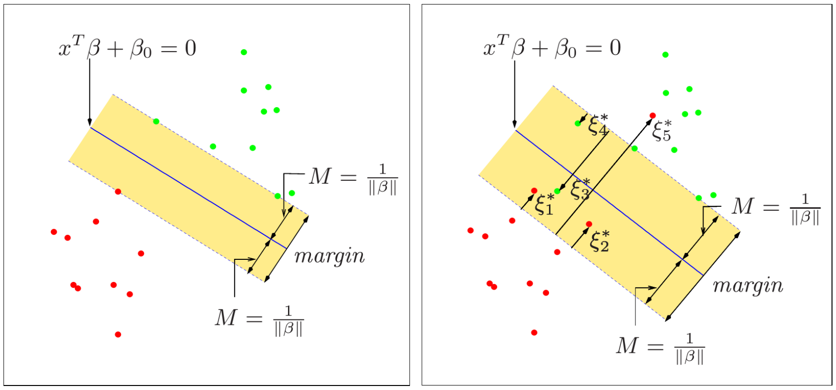


Figure 3.6: Hyperplanes of SVMs. The left panel shows the separable case. The right panel shows the inseparable case where the hyperplane was constructed using slack variables ξ^*.

Figure reproduced with permission from Springer Series in Statistics (Hastie et al., 2009).

margin between observations from two classes whose observations have $y_i \in \{-1, +1\}$. The SVM hyperplane is defined by the following optimization problem:

$$\begin{aligned}
&\min_{\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i \\
&\text{subject to} \\
&y_i(x_i^T\beta + \beta_0) \geq (1 - \xi_i) \\
&\xi_i \geq 0, \forall i\,.
\end{aligned}$$

Here, $||\beta||$ indicates the L2-norm of the vector of coefficients, which is defined as $||\beta|| = \sqrt{\sum_{i=1}^{n}\beta_i^2}$. The outputs of the SVM decision function are called decision values. A suitable function for the classification of a new measurement $x \in \mathbb{R}^p$ based on the decision function

$$\hat{f}(x) = x^T\hat{\beta} + \hat{\beta}_0$$

is given by

$$\hat{g}(x) = \begin{cases} 1 & \text{if } x^T\hat{\beta} + \hat{\beta}_0 > 0 \\ -1 & \text{else} \end{cases}.$$

In order to select a separating hyperplane for scenarios in which observations are inseparable, the SVM relies on slack variables $\xi_i \geq 0$ that indicate the extent to which the i-th observation lies within the margin or on the wrong side of the margin (Figure 3.6). The constraints guarantee that the estimate for x_i agrees with the outcome y_i up to a slack of ξ_i. The user-defined regularization parameter $C \geq 0$ regulates the extent to which misclassifications are allowed. For large C, the model is barely regularized as the hyperplane is adjusted with the goal of preventing any form of misclassification (little slack).

For low C, on the other hand, the model is more regularized because misclassifications are rather tolerated (much slack).

By solving the dual optimization problem using Lagrange multipliers, α_i, it is possible to estimate the coefficients of the model as

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \,.$$

SVM models can be specified based solely on their eponymous support vectors. If the observations from two classes are perfectly separable, the support vectors are those feature vectors x_i that lie inside the margin of the optimal hyperplane. More generally, the support vectors have $\alpha_i > 0$ and are sufficient for defining the optimal decision hyperplane. Let $\mathcal{S}$ be the set containing the indices of the support vectors. Then, the SVM decision function can be formulated solely based on the support vectors via

$$\hat{f}(x) = x^T \hat{\beta} + \hat{\beta}_0 = \sum_{i \in \mathcal{S}} \hat{\alpha}_i y_i x^T x_i + \hat{\beta}_0 \,.$$

The ν-SVM formulation, which was introduced by Schölkopf et al. (2000), uses $\nu \in [0, 1]$ rather than the regularization parameter C. The optimization problem of the ν-SVM can be formulated as[14]:

$$\begin{aligned} \min_{\beta,\xi \in \mathbb{R}^N, \rho, \beta_0 \in \mathbb{R}} \quad & \frac{1}{2}||\beta||^2 - \nu\rho + \frac{1}{2}\sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & y_i(x_i^T \beta + \beta_0) \geq \rho - \xi_i,\ \forall i = 1, \dots, N \\ \text{with} \quad & \xi_i \geq 0,\ \rho \geq 0 \end{aligned}$$

The parameter ν is an upper bound on the samples with $\xi_i > 0$ and a lower bound on the fraction of support vectors.[15] For example for $\nu = 0.3$, at most 30% of the training samples are either inside the margin or on the wrong side of the hyperplane, while at least 30% of samples are support vectors. Asymptotically, under certain conditions[16], ν equals both the fraction of support vectors and the fraction of samples with $\xi_i > 0$. In practice, it is often more convenient to use the ν-formulation of SVMs because it is easier to tune ν than C via grid search[17].

Support Vector Regression

The approach of SVMs can be applied to regression tasks through the use of support vector regression (SVR). Similarly to SVMs, SVR also uses the concept of a margin. Here, the margin is a consequence of Vapnik's ϵ-insensitive loss (Figure 3.7), which is defined by

$$|y - f(x)|_\epsilon := \max\{0, |y - f(x)| - \epsilon\} \,.$$

[14] Chen et al. 2005

[15] Not every sample with $\xi_i > 0$ is necessarily a support vector, that is, has $\alpha_i > 0$.

[16] Chen et al. 2005

[17] Grid search determines a suitable combination of model parameters by empirically evaluating predictive performance using a discrete set of parameter combinations.

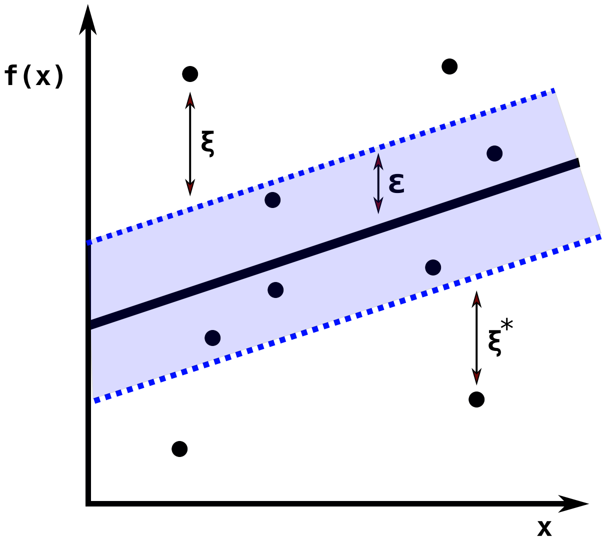


Figure 3.7: The epsilon-insensitive tube in support vector regression. Slack variables ξ are associated with observations that lie outside of the ϵ-tube. Observations that are located inside the ϵ-region do not appear in the target function.

The ϵ-insensitive loss curbs overfitting as it ensures that small absolute prediction errors are permissible. The decision function of linear SVR is $\hat{f}(x) = x^T\hat{\beta} + \hat{\beta}_0$ where $\hat{\beta} \in \mathbb{R}^p$ is the vector of estimated feature coefficients and $\hat{\beta}_0$ indicates the intercept. To estimate coefficients, β, the following function is minimized:

$$\frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N} |y_i - \hat{f}(x_i)|_{\epsilon} \, .$$

Here, C is a tuning parameter that determines the extent to which prediction errors are penalized and β is L2-regularized.

To ensure that predictions are sufficiently close to the observed outcomes a slack variable is introduced for each side of the ϵ-insensitive region: ξ for cases with $f(x_i) - y_i > \epsilon$ and ξ^* for those observations with $y_i - f(x_i) > \epsilon$. The collection of all slack variables is referred to as $\xi^{(*)}$. The introduction of slack variables leads to the following constrained optimization problem:

$$\begin{aligned}
&\text{minimize}_{\beta,\xi^{(*)}} && \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\
&\text{subject to} && f(x_i) - y_i \leq \epsilon + \xi_i \\
& && y_i - f(x_i) \leq \epsilon + \xi_i^* \\
& && \xi_i, \xi_i^* \geq 0, \quad \forall i = 1, \ldots, N
\end{aligned}$$

Only observations i with $\xi_i + \xi_i^* > 0$, that is, those observations lying outside or on the border of the ϵ-region are relevant for the objective function. The subset of these observations that is characterized by nonvanishing coefficients gives rise to the support vectors of SVR.

The SVR optimization problem is solved by introducing Lagrange multipliers, α_i^* and α_i:

$$\begin{aligned}
\underset{\alpha,\alpha^*\in\mathbb{R}^N}{\text{maximize}} \quad & -\epsilon\sum_{i=1}^{N}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)y_i \\
& -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)x_i^T x_j \\
\text{subject to} \quad & 0 \le \alpha_i, \alpha_i^* \le C \quad \forall i = 1,\dots,N \\
\text{and} \quad & \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0.
\end{aligned}$$

The regression estimate takes the form

$$\hat{f}(x) = \sum_{i\in\mathcal{S}}(\hat{\alpha}_i^* - \hat{\alpha}_i)x_i^T x + \hat{\beta}_0.$$

This formulation of SVR is called ϵ-SVR because the model is specified in terms of the hyperparameter ϵ. In ν-SVR, the following primal objective function is used:

$$\frac{1}{2}||\beta||^2 + C\left(\nu N\epsilon + \sum_{i=1}^{N}|y_i - f(x_i)|_\epsilon\right).$$

Here, $\epsilon \ge 0$ is considered as a parameter over which one minimizes and its value is implicitly determined by the choice of ν. The interpretation of the hyperparameter ν is the same as for SVMs for classification: ν is an upper bound on the fraction of points allowed outside the ϵ-tube and a lower bound on the fraction of support vectors[18]. For example, with $\nu = 0.3$, at least 30% of the training samples are support vectors and at most 30% of samples lie outside the ϵ region.

Until now, we have only dealt with SVMs that model linear relationships via dot products. In the following, I will introduce kernel functions, one of the main assets of SVMs.

Kernel Functions

In many prediction scenarios, there exist higher-order interactions between features, which cannot be captured by linear models. To deal with such prediction settings, feature vectors x_i, x_j can be implicitly mapped into another space through kernel functions $K(x_i, x_j)$. Kernel functions represent the inner product of two observations x_i and x_j that are mapped to reproducing kernel Hilbert space via a mapping function, Φ:

$$K(x_i, x_j) = \Phi(x_i)^T\Phi(x_j).$$

[18] Smola and Schölkopf 2004

Kernel functions must be continuous, symmetric, and must result in a positive-semidefinite matrix[19] K with entries $K_{ij} = K(x_i, x_j)$. The requirement that kernel functions are positive-semidefinite ensures that the SVM optimization problem is convex.

The technique of replacing the dot product in the SVM formulation with a kernel function is called the kernel trick. Applying the kernel trick to the SVM estimates yields

$$\hat{f}(x) = \sum_{i \in S} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0 \,.$$

Applying the kernel trick to the SVR estimate results in

$$\hat{f}(x) = \sum_{i \in S} (\hat{\alpha}_i^* - \hat{\alpha}_i) K(x, x_i) + \hat{\beta}_0.$$

Note that linear SVMs, which are based on dot products, have to fit merely $p + 1$ parameters. The use of kernel functions that act on pairs of observations means that nonlinear SVMs have inputs on the scale of N. Therefore, linear SVMs are considered as a parametric method, while nonlinear SVMs are considered as a nonparametric method. The SVM decision function can still be computed quickly because it only involves the evaluation of $|S|$ kernel functions.

In the following, I present several relevant kernel functions.

Linear Kernel The linear (*vanilla*) kernel,

$$K(x, x') = x^T x' \,,$$

is defined by a conventional dot product. The linear kernel is suitable for data that has a linear relationship with the outcome.

Polynomial Kernel The polynomial kernel,

$$K(x, x') = (s x^T x' + o)^d \,,$$

takes the d-th power of the dot product. In this way, interactions between up to d features can be modeled. Further adjustment of the kernel is possible by tuning the offset o and the scalar s.

Gaussian Radial Basis Function Kernel The Gaussian radial basis function (RBF) is defined by

$$K(x, x') = \exp(-\sigma ||x - x'||^2) \,.$$

It is dependent on a parameter σ that determines the width of the Gaussian. For small values of σ, the RBF kernel has a narrow but peaky distribution. For large values of σ, it has a wide but flat distribution.

Edit Kernel Edit kernels provide a means for determining the similarity between observations representing nucleotide sequences[20]. The edit kernel function is defined as

$$K(x, x') = \exp(-\gamma \times \text{edit}(x, x'))$$

where $\text{edit}(x, x')$ indicates the edit distance between x and x'. The positive parameter $\gamma \in \mathbb{R}$ scales the kernel value and ensures numerical stability. The edit function represents the genetic distance between two sequences. It can be defined by considering $\Pr(x'_i|x_i)$, the probability of observing mutation x'_i given x_i:

$$\text{edit}(x, x') = -\sum_i \log \Pr(x'_i|x_i)\,.$$

For modeling $\Pr(x'_i|x_i)$, matrices indicating the rate of amino-acid substitutions such as the PAM matrix[21] can be used.

3.5 Clustering

Clustering is a form of unsupervised learning that is concerned with the assignment of unlabeled data points into groups. In the following sections, I introduce two clustering methods that were employed in this work: K-means (Section 3.5.1) and hierarchical clustering (Section 3.5.2).

3.5.1 K-Means Clustering

K-means clustering groups observations into K clusters (Figure 3.8) by iteratively adjusting the K cluster centers with the goal of minimizing the total within-cluster variance. After fixing an initial set of clustering centers randomly, K-means alternates the following two steps:

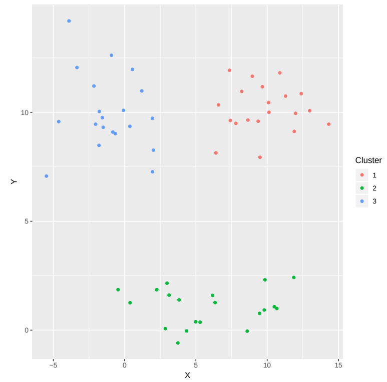


Figure 3.8: K-means clustering of artificial data with $K = 3$. The three clusters are indicated by distinct colors.

1. For each cluster center, determine which observations belong to the cluster. For this purpose, all points that are closer to this center than to any other center are selected.

2. For every cluster, determine a new clustering center by computing the mean of each feature from the observations belonging to the cluster.

These two steps are iterated until the procedure converges, that is, when the assignment of observations to clusters does not change anymore. The formal motivation for this procedure is provided in the following paragraphs.

The quality of a K-means cluster assignment, C, is described by the within-cluster point scatter, $W(C)$. This quantity describes the extent

[21] Dayhoff 1972

to which observations assigned to the same cluster tend to be close to one another. It is defined as

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \overline{x}_k||^2 .$$

The mean vector associated with the observations of the k-th cluster is indicated by $\overline{x}_k = (\overline{x}_{1k}, \ldots, \overline{x}_{pk})$ and $N_k = \sum_{i=1}^{N} I(C(i) = k)$ denotes the number of observations in cluster k.

In order to find a cluster assignment, C^*, minimizing the scatter, $W(C)$, it is useful to note that the mean of any set of observations S fulfills

$$\overline{x}_S = \arg\min_m \sum_{i \in S} ||x_i - m||^2 .$$

Thus, the optimization problem can be rewritten by fixing the clustering centers m_k:

$$\min_{C,\{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2 .$$

Using this formulation, it is possible to solve the problem using the previously described algorithm, which iteratively sets the current clustering means to m_k.

There are several drawbacks to K-means clustering. One drawback is the definition of cluster scatter according to variance, which enforces the assumption that clusters consist of observations that spherically scatter around a mean. Thus, K-means is unsuited for non-spherical clusters. Another problem is that the optimization procedure of K-means may become stuck in a local optimum, which can lead to misleading cluster assignments. The final shortcoming of K-means is that K has to be specified *a priori*. Quantitative methods for selecting K are based on analyzing the within-cluster dissimilarity W_k for several number of clusters $k = \{1, \ldots, K_{\max}\}$. Since W_k decreases with increasing k, it is not feasible to simply select the value of k minimizing W_k. Instead, the change of W_k for consecutive values of k needs to be considered. The intuition behind this approach becomes clear if we assume that the data is actually grouped into K^* clusters. Then, as long as $k < K^*$, we have $W_{k+1} \ll W_k$ because an increased number of observations will be assigned to their natural clusters. But when $k \geq K^*$, additional, artificial clusters are generated such that W_{k+1} will only be marginally smaller than W_k. A reasonable value for k can be obtained by plotting W_k against k and applying the elbow test, which selects the value of k above which W_k is merely being reduced marginally (Figure 3.9).

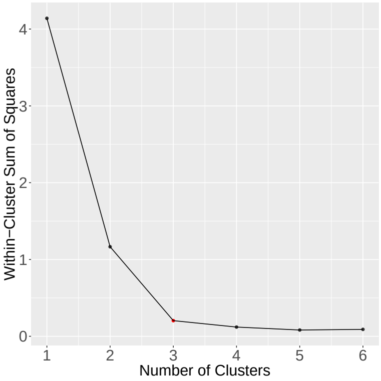


Figure 3.9: The elbow test for finding an appropriate number of clusters. Using the artificial data shown in Figure 3.8, the elbow test has identified the highlighted point at $K = 3$ as a suitable number of clusters.

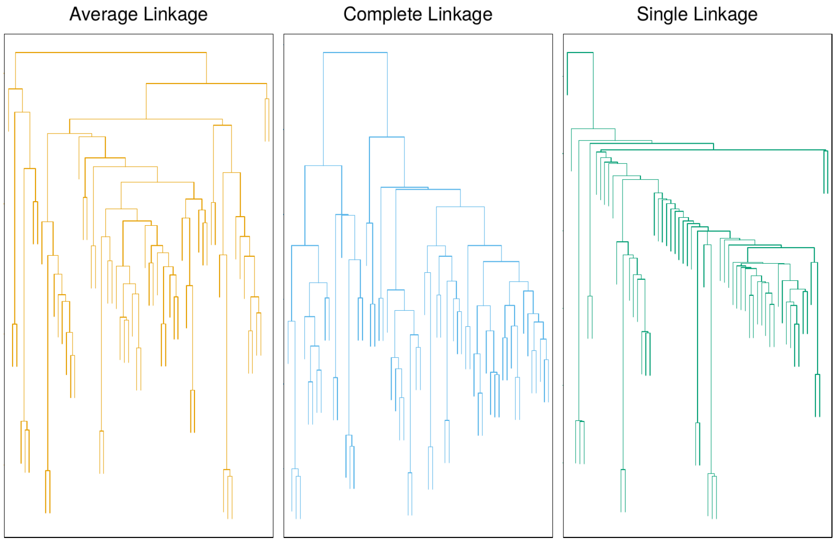


Figure 3.10: Hierarchical clustering. Comparison of dendrograms resulting from performing hierarchical clustering on the same data set using distinct intergroup dissimilarity measures.

Figure reproduced with permission from Springer Series in Statistics (Hastie et al., 2009).

3.5.2 *Hierarchical Clustering*

Hierarchical clustering does not require the prior definition of the number of clusters but requires a definition for the dissimilarity between observations in different groups instead. Hierarchical clustering provides a ranked representation of the data as clusters at one level of the hierarchy are constructed by merging clusters from the next lower level. At the lowest level, clusters are singletons consisting of individual observations. At the highest level, there is a single cluster comprising all data. Since hierarchical clustering enforces a hierarchical structure even if none is present in the data, it should be applied only if such a structure can be presumed.

There are two strategies for hierarchical clustering: agglomerative (bottom-up) and divisive (top-down) clustering. In the following, I describe the agglomerative strategy in more detail. Agglomerative strategies start with singletons and, at every level, merge the two clusters with the smallest intergroup dissimilarity into a single cluster. The resulting grouping at the next higher level of the hierarchy thus consists of one cluster less than before.

The popularity of hierarchical clustering is due to the fact that recursive binary agglomeration can be represented by a rooted binary tree. Since the dissimilarity between merged clusters is monotone increasing when an agglomerative approach is used, the binary tree can be visualized by setting the height of nodes according to the intergroup dissimilarity between its children. The leaves of the tree are plotted at height zero. This method of plotting a hierarchical clustering is called a dendrogram.

There are several variants of agglomerative clustering, which are characterized by their definition of dissimilarity between groups of observations. Given two groups of observations G and H, the dissimilarity $d(G, H)$ is determined from all pairwise dissimilarities $d_{ii'}$ where $i \in G$ and $i' \in H$. In contrast to K-means, hierarchical clustering also allows the use of dissimilarity metrics $d_{ii'}$ other than the Euclidean distance. There are three main approaches for determining intergroup dissimilarities, which are presented in the following.

Single linkage (SL) agglomerative clustering merely considers the dissimilarity of the least dissimilar pair via

$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'} \,.$$

The complete linkage (CL) approach is the opposite as it defines the intergroup dissimilarity according to the most dissimilar pair, by setting

$$d_{CL}(G, H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'} \,.$$

Another agglomerative clustering method is group average (GA) clustering, which defines the intergroup dissimilarity by taking the average dissimilarity between pairs through

$$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'} \,.$$

Exemplary results for the three agglomerative clustering strategies are shown in Figure 3.10. The figure demonstrates that SL clustering can lead to series of clusters in which a cluster is expanded by similar individual observations. This effect is called chaining. CL clustering does not suffer from this phenomenon as it ensures that the maximal dissimilarity between points in a cluster is small. At the same time, however, members assigned to a cluster may be closer to members of another cluster than to some members of their own cluster. GA clustering is a compromise between SL and CL clustering. Its disadvantage is that it is not invariant to monotone strictly increasing transformations on the dissimilarities.

3.6 *Statistical Significance Tests*

Significance tests are used to identify whether an observed effect was due to chance or whether the effect truly exists. Significance tests are crucial for the analysis of data because these tests take the sample size into account. Assume two classifiers are evaluated on a data set consisting of ten samples. One classifier obtains an accuracy of 70%, while the other classifier obtains an accuracy of 80%. Although the performance difference of 10% may suggest that the second classifier performs considerably better than the first, it actually classifies only a single additional sample correctly. Therefore, the performance difference is probably not significant.

In Section 3.6.1, I introduce the fundamentals of hypothesis testing. Since different tests are appropriate for different tasks, I discuss three statistical tests : McNemar's test (Section 3.6.2), Fisher's exact test (Section 3.6.3), and Wilcoxon rank-sum test (Section 3.6.4). Finally, I offer two techniques that correct for multiple hypothesis testing (Section 3.6.5).

3.6.1 *Introduction to Hypothesis Testing*

Statistical significance tests aim at quantifying the likelihood that a certain effect was merely observed by chance. This quantity, called the p-value, answers the following question: What is the probability, under the null hypothesis, that an observation at least as extreme as the current observation is made? The result of a hypothesis test is considered significant when the p-value is below a prespecified significance threshold α. Commonly, a significance threshold at $\alpha = 5\%$, is used.[22]

Each significance test considers two hypotheses. Typically, the null hypothesis, H_0, represents the absence of an effect (e.g. no difference in group means), while the alternative hypothesis, H_a, represents the presence of an effect (e.g. a difference in group means). A statistical hypothesis test identifies which of the two alternatives is more likely by calculating a decision value, x, on which basis the p-value is computed. For a one-tailed test, the p-value represents the probability of observing a decision value that is either larger or smaller than x. Let X indicate the random variable associated with the decision value x. Then, a right-tailed test computes the p-value as $\Pr(X \geq x|H_0)$, while a left-tailed test uses $\Pr(X \leq x|H_0)$. For a two-tailed test, both tails of the decision value distribution are considered. The p-value is therefore two times larger than the p-value of a single-tailed test, namely $2 \min\left(\Pr(X \leq x|H_0), \Pr(X \geq x|H)\right)$. When the p-value is smaller than the significance level, the result is significant and the null hypothesis can be rejected. This suggests that the alternative hypothesis is a more suitable hypothesis given the data at hand. If, however, the p-value is greater or equal to the significance level, the result is not significant and one has failed to reject the null hypothesis.

Different types of statistical tests are characterized by the use of distinct techniques for the calculation of decision values. There are two classes of significance tests: parametric and non-parametric tests.

[22] In case of exact tests, this setting limits the FDR to 5%.

	Classifier 2 Correct	Classifier 2 Incorrect	Row total
Classifier 1 Correct	a	b	$a+b$
Classifier 1 Incorrect	c	d	$c+d$
Column total	$a+c$	$b+d$	n

Table 3.2: Structure of a contingency table. The row totals as well as the column totals are also called the margin totals.

Parametric tests assume that the data follow a specific distribution (mostly the Gaussian distribution), while non-parametric tests do not make such assumptions. Although parametric tests allow for greater statistical power[23] when their assumptions are met, they can be rendered invalid if the test assumptions are not fulfilled[24]. In the following, I describe three non-parametric tests: McNemar's test, Fisher's exact test, and Wilcoxon rank-sum test.

3.6.2 McNemar's Test

McNemar's test[25] is a non-parametric method for testing the symmetry (marginal homogeneity) of entries in a 2 × 2 contingency table that was constructed from pairwise measurements pertaining to two groups. The test relies on identifying whether there exists a difference in the distribution of the marginal frequencies associated with the measurements from the two groups. In this work, I used McNemar's test to compare the predictive performance of two classifiers. Therefore, I will use the classification terminology in the following.

To construct a 2× 2 contingency table for two classifiers (Table 3.2), we simply calculate a, the number of cases in which both classifiers were correct; d, the number of cases in which both classifiers were incorrect; b, the number of cases where the first classifier was correct and the second classifier was incorrect; and c, the number of cases in which the first classifier was incorrect and the second classifier was correct. Let p_a, p_b, p_c, and p_d indicate the probabilities of the outcomes shown in Table 3.2. The test's null hypothesis assumes that both classifiers have the same ratios of correct and incorrect predictions, in which case we have $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$. Hence, the null hypothesis of the test is defined as $H_0 : p_b = p_c$, while the alternative hypothesis is $H_a : p_b \neq p_c$. The test statistic[26], $\chi^2 = \frac{(b-c)^2}{b+c}$, can be rejected when the test statistic is significantly extreme, which would suggest that the tested classifiers exhibit substantially different predictive performances.

3.6.3 Fisher's Exact Test

Fisher's exact test[27] is a non-parametric method for testing whether the frequencies of measurements from two groups are independent of each other. The test is typically applied on 2 × 2 contingency

[23] The statistical power of a significance test denotes its sensitivity, that is, whether the test will reject the null hypothesis if the alternative hypothesis is true. A study is said to be well powered if statistical testing will most likely be able to detect an effect, if one exists.

[24] Hoskin 2012

[25] McNemar 1947

[26] The test statistic of McNemar's test is denoted by χ^2 because it follows a chi-squared distribution.

[27] The test is exact because its false rejection rate exactly matches the specified significance level.

matrices. Given a contingency table with the structure indicated in Table 3.2, the test uses the hypergeometric distribution to compute

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}},$$

the probability of observing a certain contingency table. The p-value is determined as the sum of probabilities from all contingency tables with the same margin totals that are at least as extreme as the observed table.[28] Since the factorials can become very large when computing the p-value, Fisher's exact test may not be applicable for large sample sizes. The effect size associated with Fisher's exact test is determined by the odds-ratio,

$$OR = \frac{a}{b} / \frac{c}{d}.$$

Fisher's exact test is usually performed to identify whether the odds ratio differs from 1.

3.6.4 Wilcoxon Rank-Sum Test

Wilcoxon rank-sum test is a non-parametric method that evaluates whether the measurements from two groups have distinct orderings. The test can be performed in the following, procedural manner.

After pooling the measurements from both groups, we determine the rank of each observation[29]. For identical values, the arithmetic means of the unadjusted ranks are used. Next, we sum the ranks of observations from the first and the second group, obtaining R_1 and R_2, respectively. In the final step, the ordering of values from the first group is evaluated using

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2},$$

while the ordering of values from the second group is determined as

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}.$$

The variables n_1 and n_2 in these formulas correspond to the number of samples from the first and the second group, respectively.[30] Then, the test statistic is calculated as $U = \min(U_1, U_2)$.

3.6.5 Multiple Hypothesis Testing

The multiple hypothesis testing problem is concerned with the increased number of false positives that ensues when several hypotheses are tested on a single data set. To ensure that the number

[28] For a two-tailed test, a table is at least as extreme if its probability is less or equal to the observed probability of the table.

[29] The rank of an observation is the one-based index of that observation in a list of increasing order.

[30] The formulas are based on the famous rule by Gauss according to which $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.

of false positive test results is controlled, approaches that correct for multiple comparisons are used. Two frequently used methods for this task are Bonferroni correction and the Benjamini-Hochberg method. For discussing these methods, let us assume that we have performed m hypothesis tests. Then, we need to consider a family of hypothesis tests[31], $H_1, \dots, H_m$, with p-values $p_1, \dots, p_m$. Further, let m_0 indicate the number of null hypotheses that are correct (i.e. the number of null hypotheses that should not be rejected).

Bonferroni Correction Bonferroni correction adjusts the significance level, α, according to the number of tests, $m > 1$, that are performed. The adjusted significance level is defined as $\alpha_{\text{adj}} = \frac{\alpha}{m}$. The null hypothesis of the i-th test is rejected if its p-value fulfills $p_i \leq \alpha_{\text{adj}}$.

By setting α_{adj} in the specified manner, Bonferroni correction bounds the familywise error rate (FWER), which is defined as

$$\text{FWER} = \Pr\left(\bigcup_{i=1}^{m_0}(p_i \leq \frac{\alpha}{m})\right)$$

and indicates the probability that at least one null hypothesis is falsely rejected. Bonferroni correction can be derived from Boole's inequality. The inequality states that, given a countable set of events $E_1, E_2, \dots$, the probability that at least one event happens is less or equal to the sum of the individual event probabilities:

$$\Pr\left(\bigcup_i E_i\right) \leq \sum_i \Pr(E_i)\,.$$

By applying Boole's inequality to the definition of the FWER, we can show that α is an upper bound of the FWER:

$$\text{FWER} = \Pr\left(\bigcup_{i=1}^{m_0}(p_i \leq \frac{\alpha}{m})\right) \leq \sum_{i=1}^{m_0} \Pr(p_i \leq \frac{\alpha}{m}) = m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha\,.$$

Bonferroni correction is a conservative method to correct for multiple testing because it controls the FWER. On the one hand, this is an advantage, because it reduces the number of false alarms. On the other hand, this is a disadvantage because Bonferroni correction may also considerably reduce the statistical power of the tests that are performed[32].

The Benjamini-Hochberg Procedure The approach from Benjamini-Hochberg is a method for controlling the FDR at a significance level α. The approach can be implemented in two steps. First, we order all p-values in ascending fashion and denote them by $p_{(1)}, \dots, p_{(m)}$. Second, we determine $k = \arg\max_i p_{(i)} : p_{(i)} \leq \frac{i}{m}\alpha$ and define

[31] A family of hypothesis tests arises when several tests are performed on different subsets of the same data set. For example, to determine whether HIV viral loads are different for patients that are infected with different HIV subtypes, one might first compare patients infected with subtype A to those infected with subtype B and then compare patients infected with subtype C to those infected with subtype B. In this case, two hypothesis tests were performed.

[32] Nakagawa 2004

the adjusted p-value as $\alpha_{\text{adj}} = p_{(k)}$. The i-th null hypothesis is then rejected if $p_i \leq \alpha_{\text{adj}}$. The Benjamini-Hochberg procedure is valid when the m tests are independent. For a proof why the Benjamini-Hochberg procedure controls the FDR, I refer to Benjamini and Hochberg (1995).

Compared to Bonferroni correction, the use of the Benjamini-Hochberg procedure allows for greater statistical power at the price of a larger number of false discoveries. This is because the FDR denotes the fraction of false positive tests among all positive tests, while the FWER is the probability that any of the m tests yields a false positive result. To illustrate the two concepts, consider a situation where both the FWER and the FDR are at 5%. An FWER at 5% merely states that the probability of at least one false positive result is at 5%. On the other hand, when the FDR is at 5%, this means that 5% of all positive tests are expected to be false positives. When 100 statistical tests are performed, we would expect five false positives at an FDR of 5%. However, with an FWER at 5%, it is unlikely that any of the tests are false positives.

3.7 *Optimization with Linear Programs*

Mathematical optimization is concerned with the determination of solutions to maximization or minimization problems. Optimization is pervasive in machine learning. For example, quadratic programming is used to solve the SVM optimization problem that I introduced in Section 3.4.2, while logistic regression models (Section 3.4.1) are fitted using gradient descent. In this dissertation, I used integer linear programs (ILPs) to solve an instance of the set cover problem (SCP) for multiplex primer design. After an introduction to linear and integer linear programming (Section 3.7.1), I present the branch-and-bound algorithm to solving ILPs in Section 3.7.2. The set cover problem is introduced in Section 3.7.3.

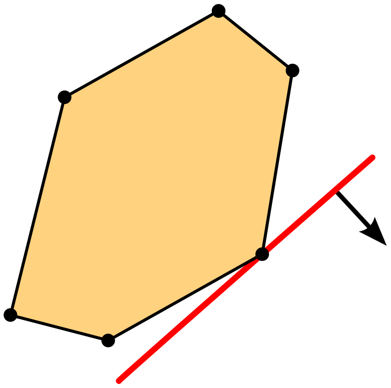


Figure 3.11: Representation of a linear program with two variables and six inequalities. The 2-dimensional polytope indicates the set of feasible solutions. The cost function is shown by the red line and the direction of optimization is shown by the black arrow.

Retrieved from Wikipedia, licensed under CC0 1.0 Universal Public Domain Dedication.

3.7.1 *Linear and Integer Linear Programming*

Linear programming involves the optimization of a linear objective function subject to linear constraints. The feasible region of a linear program is a convex polytope, which is determined by the intersection of finitely many half spaces that are each defined by a linear inequality. Linear programming algorithms find points within the polytope where the objective function assumes an extreme (minimum or maximum) value (Figure 3.11). Note that, in the context of this thesis, the terms maximum and minimum refer to the global maximum and minimum, respectively, while the terms maximal and minimal refer to local maximum and minimum, respectively.

In their canonical form, linear programs are defined as

$$\begin{aligned} \text{minimize } & c^T x \\ \text{subject to } & Ax \leq b \\ \text{with } & x \geq 0 \end{aligned}$$

where x is the vector of variables that is to be determined. The vector c provides the coefficients of the variables. The vector b determines the inequalities associated with the constraint matrix A. The inequalities $Ax \leq b$ and $x \geq 0$ are the constraints that define the set of feasible solutions.

Linear programs are often solved using the Simplex algorithm. The idea of the algorithm is to generate a feasible solution at a vertex of the polytope and to move along the edges to vertices with non-decreasing objective functions until the optimum is found. The Simplex algorithm does not guarantee an optimal solution for ILPs, in which all variables are required to be integer.[33]

3.7.2 *Branch and Bound*

Branch and bound is an algorithmic approach for solving combinatorial optimization problems. Branch-and-bound algorithms enumerate candidate solutions using a state space search. The set of candidate solutions forms a rooted tree where the root represents the full set of solutions, while branches represent solution subsets. Before entering a branch, estimates for the lower and upper bounds of the optimal solution are determined. Branches that cannot improve the currently best solution according to the estimates are discarded. The main challenge of implementing a branch and bound algorithm is the requirement of heuristics for computing lower and upper bounds. The worst-case runtime of branch-and-bound algorithms is exponential.

3.7.3 *The Set Cover Problem*

The SCP is an NP-complete combinatorial optimization problem that can be used for formulating the primer design problem that is solved in this work. Given a universe $\mathcal{U}$ and a family $\mathcal{S}$ of subsets of $\mathcal{U}$, the SCP is concerned with finding a minimal subfamily $\mathcal{C} \subseteq \mathcal{S}$ whose union is $\mathcal{U}$. The problem can be formulated as an ILP by defining decision variables $x_S \in \{0,1\} \, \forall S \in \mathcal{S}$, which indicate whether the set S is selected ($x_S = 1$) or not ($x_S = 0$). The optimal assignment of decision variables can be found by solving the following constrained

[33] Integer linear programming is NP-hard, while linear programming is in P.

optimization problem:

$$\begin{aligned} \text{minimize } & \sum_{S \in \mathcal{S}} x_S \\ \text{subject to } & \sum_{S: e \in S} x_S \geq 1 \, \forall e \in \mathcal{U} \end{aligned}$$

Minimizing the target function ensures that we find the smallest combination of sets among all combinations whose union equals the universe.

The SCP can also be approximated in polynomial time using a greedy algorithm that selects the set cover with the largest number of uncovered elements at every step. Given a set with n elements that are to be covered, this algorithm achieves an approximation ratio that equals the n-th harmonic number[34],

$$H(n) = \sum_{i=1}^{n} \frac{1}{i} \approx \ln(n) \, .$$

This means that the set cover obtained from a greedy algorithm is at most $H(n)$ times larger than the minimum set cover.

[34] Young 2008

Part II:

Contributions

The following chapters describe the original scientific contributions of this dissertation. geno2pheno[ngs-freq] performs drug resistance testing using next-generation sequencing data (Chapter 4), while geno2pheno[coreceptor-hiv] determines HIV-2 coreceptor usage (Chapter 5). Both systems support the personalization of antiviral treatments. The approach of openPrimeR, which is presented in Chapter 6, enables the design of primers for multiplex polymerase chain reaction. Using the approach, the isolation of broadly neutralizing antibodies against HIV-1 can be supported. The statistical model presented in Chapter 7 can assist primer design through the identification of polymerse chain reaction amplification events.

4
Interpreting Drug Resistance from Next-Generation Sequencing Data

We need to proactively address the rising levels of resistance to HIV drugs if we are to achieve the global target of ending AIDS by 2030.

WHO director general Tedros Ghebreyesus, 2017

In this chapter, I introduce geno2pheno[ngs-freq], a web server for identifying drug resistance from HIV-1 or HCV samples that were processed using next-generation sequencing. I have conceptualized, prototyped, and implemented the web server. Alejandro Pironti and Achim Büch have developed the framework on whose basis geno2pheno[ngs-freq] was developed. Additionally, Achim Büch provided technical assistance during the development. Georg Friedrich implemented the core components of the web frontend of geno2pheno[ngs-freq], which were further refined by me. Prabhav Kalaghatgi provided an initial, stand-alone implementation of the geno2pheno[hcv] interpretation system, which I integrated into the web service. Martin Däumer and Alexander Thielen developed the frequency file format that is used by the server and provided their technical expertise on next-generation sequencing of viral populations. Elena Knops, Eva Heger, Martin Obermeier, and Rolf Kaiser supported the development process with their expertise in viral diagnostics and performed extensive validations. The project was initiated and supervised by Thomas Lengauer and Nico Pfeifer. The content of this chapter expands upon the publication by Döring et al. (2018) in two regards. First, an overview of genotypic resistance testing is provided. Second, the description of the methodology was updated and extended.

Drug resistance mutations can emerge rapidly in individuals infected with pathogens such as HIV-1 or HCV. In order to select effective combinations of antiviral drugs against HIV-1 (Section 2.3.6) and HCV (Section 2.4.5), genotypic resistance tests can be used. These tests consist of two steps: sequencing the relevant segments of the viral genome followed by the interpretation of drug resistance based on the amino-acid sequence of the viral proteins that are drug targets[1]. Sanger sequencing (Section 2.5.3) is unable to detect minority variants because it can only detect viral variants that comprise at least 10%–20% of the total viral RNA in the sample[2]. The more recently developed NGS, on the other hand, can detect treatment-relevant variants that are present at low abundances[3]. Thus, NGS has recently become more commonly used for drug resistance testing.

To date, few web services for the interpretation of NGS data with respect to drug resistance are available[4]. These services typically rely on the raw sequencing data resulting from subjecting a viral sample to NGS, for example, in terms of a FASTQ file. After a sample has been uploaded, the processing pipelines of these services perform the following tasks: (1.) Reads are trimmed in order to remove low-quality positions; (2.) reads are mapped[5] to a reference sequence; (3.) the abundance of mutations is quantified independently for each aligned position; and (4.) resistance is inferred. While web services that interpret Sanger sequences provide near-instant responses to a query, web services relying on NGS data such as HyDRa or PASeq perform more time-intensive computations and notify users via email when the results are available. These services also support only rules-based interpretations.

The goal of the work presented in this chapter was the development of a web server for determining HIV-1 and HCV resistance based on NGS data. Initially, the server was intended to offer a new prediction model for HIV-1 on the basis of NGS reads. Due to a lack of clinical data this intention became infeasible. Therefore, the decision was made to rely on the well-established methods of geno2pheno[resistance][6] and geno2pheno[hcv][7]. Since these approaches are based on Sanger sequences, it was no longer necessary to require the raw NGS data. Thus, we chose to develop a web service that predicts drug resistance on the basis of already processed NGS data, which are converted to an appropriate representation for the use of established approaches for Sanger sequences.

The structure of this chapter is as follows. Section 4.1 introduces the genotypic methods that are used by geno2pheno[ngs-freq] and summarizes available interpretation systems based on NGS. In Section 4.2, the approach of geno2pheno[ngs-freq] is delineated. Section 4.3 exemplifies the usefulness of the web service in two case studies. The chapter concludes with Section 4.4, in which the approach of geno2pheno[ngs-freq] is critically discussed.

[1] Vercauteren and Vandamme 2006

[2] Tsiatis et al. 2010

[3] Lin et al. 2014; Fox et al. 2014; Vrancken et al. 2016; Cozzi-Lepri et al. 2015

[4] Noguera-Julian et al. 2017

[5] In the NGS setting, the term *mapping* is typically used in favor of the term *alignment* although both terms describe the same concept.

[6] Beerenwinkel et al. 2003; Pironti et al. 2017b; Lengauer and Sing 2006

[7] Kalaghatgi et al. 2016

4.1 Genotypic Resistance Testing

Genotypic resistance tests analyze viral genomic sequences with respect to drug resistance mutations. These approaches are cheap, can be performed quickly, and correlate well with results from phenotypic drug resistance tests. After introducing genotypic resistance tests in Section 4.1.1, I describe geno2pheno[resistance] (Section 4.1.2) and geno2pheno[hcv] (Section 4.1.3). This section concludes with an overview of the processing steps required for performing resistance tests based on NGS data (Section 4.1.4).

4.1.1 Overview

Genotypic resistance testing is based on sequencing the viral genomic regions of interest and subsequently analyzing their translated amino-acid sequences[8]. The critical advantage of genotypic approaches over phenotypic methods (Section 2.5.2) is that genotypic approaches do not require any wet lab work except for sequencing, which is fast and cheap. Once the sequencing results are available, the remainder of the work is done *in silico*. Since genotypic approaches are not only less intensive with regard to cost and time than phenotypic approaches but also offer advantages in clinical applications[9], they have become the standard of care for determining drug resistance for viruses such as HIV[10].

Clinical testing of HIV drug resistance is based on viral RNA extracted from patient plasma. Only if resistance tests based on plasma RNA cannot be performed due to low VLs (e.g. < 400 copies per mL), is proviral DNA used as a substitute. Results from tests based on proviral DNA may not agree with those from plasma RNA for three reasons: (1.) proviral DNA may reflect non-replicating proviruses; (2). the negative predictive value of proviral DNA is low because PBMCs do not reflect the entirety of immune cells infected by HIV[11]; (3). proviral DNA may be mutated by APOBEC[12].

Genotypic tests overcome two shortcomings of phenotypic tests in clinical applications. First, phenotypic tests are not capable of detecting emerging resistance. This means that phenotypic tests may classify an isolate as susceptible although it contains mutations that lead to the emergence of resistance in the near future[13]. These associations can be taken into account by genotypic tests. Second, phenotypic tests lack sensitivity when they are applied to clinical samples representing heterogenous viral populations[14]. For example, consider a viral population in which 80% of the population are susceptible to a drug and 20% of the population carries resistance mutations. Then, the experimentally determined IC50 may only be

[8] Tang and Shafer 2012

[9] Genotypic drug resistance tests can identify resistance-associated mutations that do not induce resistance but are associated with the emergence of drug-resistance mutations in the future. By learning these associations, genotypic tests are able to forecast future drug resistance, which cannot be done using phenotypic tests.

[10] Günthard et al. 2016; Behrens et al. 2017

[11] Wirden et al. 2011

[12] Kim et al. 2014

[13] Garcia-Lerma et al. 2004

[14] Mayer et al. 2001

marginally higher than the wild-type IC50, although this viral population carries resistant strains. Due to the low RF, a phenotypic test may fail to detect resistance in such a sample. Genotypic methods, on the other hand, can consider the impact of variants at any abundance.

Several studies have shown that genotypic methods and phenotypic methods exhibit high concordances[15]. A perfect correlation between genotypic and phenotypic methods, however, should not be expected for the following reasons[16]. First, phenotypic tests are subject to a certain degree of technical variation[17]. Second, most genotypic tests do not model phenotypic resistance (*in vitro* resistance) but rather the impact of amino-acid mutations on treatment outcomes (*in vivo* resistance). Since *in vivo* resistance implicitly considers the influence of viral fitness[18], drug metabolism[19], and likely evolutionary trajectories[20], apparent differences between genotypic and phenotypic tests may simply reflect the diverging orientations of the tests.

The main challenge of genotypic resistance testing lies in interpreting the viral sequence with respect to drug resistance. Rules-based interpretation systems rely on the identification of individual mutations in order to estimate the clinical impact of resistance mutations. Rules-based systems are easy to interpret and are nevertheless capable of modeling relationships with varying degrees of complexity. For example, a simple rule is given by *If M184V is present in RT of an HIV-1 strain, then it exhibits intermediate resistance to FTC*, while the following rule is more intricate: *If M46I, I54V, and V82A are present in PR of an HIV-1 strain, then it is resistant to IDV*[21].

The sets of rules that are applied by rules-based system are curated by panels of experts that scrutinize available clinical and phenotypic data. The spectrum of expert opinions has given rise to several sets of rules for interpreting HIV-1 drug resistance, for example, the rule sets from ANRS, HIVdb[22], HIV-GRADE, and the Rega institute.[23] Rules-based systems typically provide only categorical estimates of resistance according to the SIR scheme[24]. Among rules-based engines, HIVdb is the only system that also provides a quantitative output, which is determined by assigning resistance scores to individual mutations and summing them up[25].

Statistical interpretation engines are principally based on the same data that is available to the expert panels. They, however, interpret resistance using statistical models that are generated via machine learning algorithms. While human experts structure information on drug resistance in terms of rules, statistical models afforded by machine learning algorithms enable more intricate forms of knowledge representation. Statistical approaches differ among each other with respect to the applied machine learning algorithms and the data sets that are

[15] Pironti et al. 2017b; Beerenwinkel et al. 2003

[16] Sarmati et al. 2002

[17] Lengauer and Sing 2006

[18] Diallo et al. 2003

[19] Zeldin and Petruschke 2003

[20] Brenner and Coutsinos 2009

[21] Zhang et al. 2010

[22] Tang et al. 2012

[23] All of these rules-based approaches are available via the HIV-GRADE website (Obermeier et al., 2012).

[24] The SIR scheme is the standard system for classifying drug resistance. It is based on assigning viral strains into the classes *susceptible*, *intermediate*, or *resistant*, see Section 2.5.2.

[25] Paredes et al. 2017

used for training the models. For example, geno2pheno[resistance][26] uses support vector regression and classification, while the more recent SHIVA software[27] employs random forests. An advantage of statistical interpretation systems is that they are not influenced by human biases. For example, expert panels could put greater weight on studies that were published in journals that are considered more influential while disregarding articles published in less respectable journals. Although machine learning models are limited by the assumptions of the model and influenced by the possible biases in the composition of the training data, these problems can be avoided by judiciously selecting the model and the training data. Beyond this, there are even problem-specific methods for debiasing the learning process[28] and techniques for analyzing fitted models with respect to their biases[29]. Human biases, on the other hand, can neither be corrected for nor identified so easily. Thus, statistical methods afford rationally determined, quantitative estimates of resistance levels.

Albeit methodologically distinct, rules-based and statistical methods are not completely independent from one another. Statistical engines have generated novel insights that were later incorporated into rules-based systems. For example, the HIV-GRADE engine was not only informed by expert knowledge but also by information encoded in the model of geno2pheno[resistance][30], which is described in the next section.

4.1.2 geno2pheno[resistance]: A Statistical Resistance Interpretation System

geno2pheno[resistance] is a statistical genotypic resistance testing system for Sanger sequences from HIV-1. At the heart of the tool are two types of predictive models. While the original model of geno2pheno[resistance] estimates phenotypic RFs[31], the recently developed drug-exposure model uses clinical data to estimate a quantity that is correlated with the exposure to a drug[32]. Both models output estimates in terms of z-scores, which indicate the number of standard deviations that an estimate is above or below the mean of estimates for treatment-naive patients. In the following, I describe the two models in more detail and compare them with each other. The approach that is used for turning the quantitative predictions of these models to the discretized SIR representation concludes this section.

Prediction of Phenotypic Drug Resistance

The original approach of geno2pheno[resistance] relies on drug-specific support vector regression models using linear kernel functions. These models were trained on genotype-phenotype pairs

[26] Beerenwinkel et al. 2003; Pironti et al. 2017b; Lengauer and Sing 2006

[27] Riemenschneider et al. 2016

[28] Schnabel et al. 2016

[29] Adebayo 2016

[30] Obermeier et al. 2012

[31] Beerenwinkel et al. 2003

[32] Pironti et al. 2017a

consisting of Sanger sequences of HIV-1 PR and RT as well as corresponding measurements of phenotypically determined RFs[33]. Phenotypic resistance testing was performed for at least 300 samples for drugs from the three classes of NRTIs, NNRTI, and PIs[34].[35]

For the development of these models, amino-acid sequences were represented using a binary encoding where 1 indicates the presence of an amino acid and 0 its absence. For each drug, a SVR model (Section 3.4.2) based on an epsilon-insensitive loss function with $\epsilon = 0.1$ was fitted. The regularization parameter C was optimized via grid searching and CV. The models were validated using 10-fold CV. The mean squared correlation with phenotypically determined RFs was 0.6 ± 0.14.

Prediction of Drug Exposure from Clinical Data

The drug-exposure approach of geno2pheno[resistance] is based on drug-specific SVC models using linear kernel functions. These models were trained on clinical data consisting of Sanger sequences and corresponding binary labels indicating whether a sequence originates from a patient that was treated with a drug or not[36]. The decision value resulting from application of an SVM (Section 3.4.2) is interpreted in terms of the drug-exposure score (DES) whose value indicates the degree of drug exposure.

The approach was validated in the following way. SVC models with linear kernel functions were trained for predicting exposure to individual drugs of the three classes PIs, NRTIs, and NNRTIs. The training sets for most drugs consisted of thousands of observations. Viral sequences were represented using a binary encoding. Performance was evaluated on independent test sets where the performance of DESs was analyzed in three prediction scenarios: (1.) prediction of drug exposure; (2.) prediction of phenotypic drug resistance, and (3.) prediction of treatment success. The drug-exposure model outperformed the original, phenotypic model of geno2pheno[resistance] in two of three scenarios: prediction of drug exposure (AUC 0.78 vs AUC 0.71) and prediction of treatment outcomes (AUC 0.73 vs AUC 0.68). With a mean correlation of 0.51 to the results from the PhenoSense assay, the drug-exposure model exhibited poorer performance than geno2pheno[resistance] when estimating phenotypic resistance.

Comparison of the Phenotypic and Clinical Model

Since the original model of geno2pheno[resistance] estimates RFs resulting from phenotypic resistance tests, this model is best suited for the determination of *in vitro* drug resistance. As the drug-exposure

[33] Beerenwinkel et al. 2003; Lengauer and Sing 2006

[34] Walter et al. 1999

[35] The following drugs were considered for each class: ABC, ddI, 3TC, d4T, TDF, and ZDV for NRTIs, EFV, ETR, NVP, and RPV for NNRTIs, and APV, ATV, DRV, IDV, LPV, NFV, SQV, and TPV for PIs.

[36] Pironti et al. 2017a

model is based on clinical data, it is better suited for *in vivo* applications, for example taking treatment decisions. This is clearly demonstrated by the results from the validation of the drug-exposure model. Investigating the M184V mutation, can make this difference more tangible.

M184 is known for its association with FTC resistance *in vitro*[37] but not *in vivo*[38]. When the two geno2pheno[resistance] models are presented with the same viral sequence carrying M184V, the phenotypic model classifies the sequence as *resistant* to FTC, while the clinical model classifies the sequence as *intermediate*. Obviously, the prediction that is obtained with the phenotypic model agrees well with the *in vitro* evidence, while the prediction of the clinical model agrees well with the *in vivo* evidence. In this case, the prediction from the drug-exposure model can be directly integrated into clinical decision making, while the prediction from the phenotypic model requires expert interpretation.

The drug-exposure model has a critical advantage over the phenotypic model. In contrast to phenotypic data whose generation is costly and time-consuming, clinical data is generated as a by-product of clinical routine and more widely available. Although models based on clinical data cannot immediately take novel drugs into account[39], these models may still include novel drugs sooner than phenotypic models because they do not require experts that determine drug resistance experimentally. This can be exemplified by the geno2pheno[integrase] server that provides predictions of phenotypic resistance for INSTIs. Although DTG was FDA-approved already in 2013, geno2pheno[integrase] still (as of January 2019) provides merely rules-based predictions for DTG due to a lack of phenotypic data. Clinical data on DTG, on the other hand, are already numerous. Besides the point that models based on clinical data may be updated faster, there is another benefit of these models: Clinical databases are large. This means that, for many drugs, the natural variation of drug resistance mutations is represented well.[40].

Phenotypic methods, on the other hand, have two disadvantages. First, phenotypic methods that rely on measurements from multiple assays may suffer from batch effects. Even commercially available assays exhibit different RF distributions as evidenced by the requirement for different cutoffs when transforming RFs to levels of resistance[41]. Without correction, this problem can be circumvented by ensuring that the measurements for individual drugs were all performed with the same assay, under the same experimental conditions. Second, phenotypic models may also be limited by the diversity of viral sequences that can be observed *in vitro*. Since training data reflect the adaption of viruses subject to the selection pressure from a single

[37] Schinazi et al. 1993

[38] Campbell et al. 2005; Castagna et al. 2006

[39] This is because a suitable amount of data in which resistance mutations are observed needs to be generated first.

[40] Yet, the low coverage of drugs that are infrequently used may be unsuitable for statistical modeling.

[41] Wang et al. 2004b

antiretroviral drug, not all patterns of mutations that are observed *in vivo* may be represented.

A limitation of the drug-exposure model is that it derives DESs for individual drugs based on the properties of sequences that were exposed to combinations of drugs. If treatment selection were a random process, this would not be a problem. However, ART is strongly influenced by recommended first-line treatments, which consist of a handful of combinations[42]. Thus, it may not be possible to discriminate between features that are predictive of exposure to coprescribed drugs if the activity of these drugs is modulated by the same viral protein. As a result, the DESs for such drugs may be misleading because they may reflect the mutational profiles from multiple drugs rather than individual drugs. For example, this could be the case for the NRTIs TDF and FTC, which are often combined in the first-line regimen TDF + FTC + DRV. To alleviate this problem, two approaches could be considered. The first approach relies on changing the feature encoding for drug-specific models using prior knowledge. For example, when training and applying the TDF model, the features that are known to be associated with FTC drug resistance but not with TDF resistance could be encoded by zeros to limit their influence. The second approach is based on multi-task learning. By learning from data for all antiviral drugs, the resulting model may be able to discern between mutations that only arise for a specific drug and those mutations that arise due to coprescribed agents. geno2pheno[resistance] corrects for the coprescription of drugs only after DESs have been computed, namely at the time when DESs are discretized to resistance levels. This approach is described in the following section.

Discretization of Quantitative Estimates

The estimates of geno2pheno[resistance] (RFs or DESs) are transformed to three interpretable, clinically-motivated levels of resistance: *susceptible*, *intermediate*, or *resistant*. In the following, I outline how geno2pheno[resistance] determines cutoffs based on the approach by Pironti et al. (2017b), which relies on estimated RFs.[43]

In order to determine the natural variation of resistance among therapy-naive patients, RFs are determined by applying geno2pheno[resistance] to viral sequences from this population. To exclude RFs from samples that are subject to transmitted drug resistance (TDR), all strains exhibiting any major drug resistance mutation (DRM) are excluded. In the next step, the probability of susceptibility (POS) is computed for each clinical isolate in the EuResist database by fitting a two-component Gaussian mixture model to the RF distribution of every

[42] Günthard et al. 2016; Behrens et al. 2017

[43] The same approach is used to find cutoffs for DESs.

drug[44].

In the following step, treatment episodes (TEs) are collected. In short, TEs encompass the following measurements: the genomic regions of interest at baseline, the treatment regimen, a follow-up VL, and, optionally, a baseline VL. Each TE is labeled either as a success or a failure depending on the follow-up VL. All TEs with a VL below 400 HIV-1 RNA copies per mL or with a VL reduction of at least 100-fold with respect to the baseline are considered successful treatments and all others as failed. Treatments for which resistance testing was performed during treatment or close to treatment interruption are also considered as failed treatments.

Finally, cutoffs are determined in the following way. For each TE, RFs are predicted from baseline genotypes using geno2pheno[resistance]. Using weighted-kernel density estimation[45], RF-dependent densities are estimated for treatment failure and success, respectively. Susceptibility to the other drugs in the treatment regimen is taken into account by considering the POS of each backbone compound. Since combination treatments are considered, it is necessary to correct RFs estimated for individual drugs based on the potency of the other drugs in the regimen.

In order to correct the density representing successful treatments, the RF of a drug is weighted based on the estimated susceptibility to the other drugs in the regimen. For example, if a viral strain is susceptible to the considered drug but resistant to the other drugs in the combination, then the impact of the considered drug is increased because the success of the treatment seems to hinge on the susceptibility to the considered drug. However, if a viral strain is susceptible to all drugs in the regimen, then the impact of the considered drug is reduced since the success of the treatment cannot be attributed to the considered drug alone. The density associated with failed treatments also needs to be corrected because not all failures are due to drug resistance; other factors such patient non-compliance also play a role. Therefore, the impact of failed treatments for which no resistance was detected, is reduced.

For the analytic determination of suitable cutoffs, the following sigmoidal is fit to the probabilities of success for the i-th drug:

$$\hat{P}_i(x) = \frac{a-d}{1+\exp\left(-b(x-c)\right)} + d$$

where a, b, c, $d \in \mathbb{R}_0$ and non-negative. A lower and an upper cutoff can be found from $\hat{P}_i$ by determining the inflection points, which represent the points at which the probability of success changes considerably or is marginal. In case that the determined lower cutoff is below the 95th percentile of RFs from treatment-naive persons,

[44] Beerenwinkel et al. 2003

[45] Duong 2007

the lower cutoff is set to this value in order to curb overcalling of intermediate resistance.

The method from Pironti et al. (2017b) was compared to the approach from Stanford's HIVdb by computing genotypic susceptibility scores (GSSs). For this purpose, each drug was scored according to its resistance level (1 for *susceptible*, 0.5 for *intermediate*, 0 for *resistant*). Next, the GSS of a treatment was determined by summing up the scores from all constituent drugs. The GSSs resulting from the two methods achieved similar levels of performance when predicting treatment success on a test data set (AUC of 0.63 for the presented approach vs 0.65 for HIVdb). This result indicates that the presented approach affords a suitable discretization to resistance levels.

4.1.3 *geno2pheno[hcv]: A Rules-Based Resistance Interpretation System*

geno2pheno[hcv] relies on a set of rules that was chosen by an expert panel through extensive reviewing and weighting of literature related to HCV drug resistance[46]. These rules are drug- and genotype/subtype-specific. For example, given a virus with subtype 1b, the mutation 41R would not affect susceptibility to the NS3 inhibitor asunaprevir but susceptibility would be considered to be reduced if both 41R and 80R were present. To determine the level of drug resistance associated with associated with an input sequence, geno2pheno[hcv] scans the amino-acid mutations of NS3, NS5A, and NS5B for matches to any of the rules. When a rule matches an observed mutation, the resistance level is updated if the resistance level has increased compared to the prior resistance level. In this way, the highest level of resistance from all matched rules is reported for every drug. For example, when a sequence carries 41R and 80R (*reduced susceptibility*) but also 168A (*resistant*), then the resistance level for asunaprevir is *resistant*.

Since geno2pheno[hcv] performs geno- and subtype-specific predictions, it is necessary to obtain alignments to a large set of HCV reference sequences representing several geno- and subtypes. For every genomic region of HCV, just under 200 reference sequences are considered.[47] Computing alignments for such a large number of sequences would be infeasible, which is why geno2pheno[hcv] aligns input sequences only with respect to the HCV reference sequence H77[48], which has subtype 1a. The tool then relies on stored alignments of the subtype-specific reference sequences with respect to the H77 sequence. In this way, the similarity between query and reference can be determined by iterating along the aligned sequences and increasing the match score by 1 when the amino acids at a position

[46] Kalaghatgi et al. 2016

[47] Due to the high sequence diversity of HCV it is particularly important that local alignments with region-specific references are used. Global alignments against the whole HCV genome would be inaccurate for sub-genomic queries.

[48] Kolykhalov et al. 1997

in the alignment agree. The inferred subtype is that of the reference sequence maximizing the similarity to the aligned query sequence. For the determination of drug resistance, a subtype-specific reference that exhibits a minimum number of polymorphisms associated with resistance is selected.

4.1.4 Resistance Testing Based on Next-Generation Sequencing

While the result of Sanger sequencing is a nucleotide sequence, in which ambiguous bases indicate the presence of multiple peaks at the same position in the chromatogram, NGS (Section 2.5.3) produces millions of ultra-short reads that require further processing. The basic processing of NGS reads begins with quality control procedures. These procedures ensure that only high-quality positions (i.e. those that likely reflect the true nucleotide) are considered. After cleaning the reads, the fragmented data needs to be structured. For this purpose, one can either perform *de novo* read assembly or read mapping[49]. Assembly tries to determine scaffolds by finding sets of contiguous segments. Mapping, on the other hand, reconstructs the original, unsegmented sequence by aligning reads to a reference genome. *De novo* assembly is particularly useful when a reference sequence is not available (e.g. for less studied species) or when reads originate from mixtures of species (e.g. in metagenomics). In other cases, mapping is typically used because it is less time-intensive and provides a frame of reference, which is useful for downstream analyses. Since samples for genotypic resistance testing only contain genomic material from a single viral species and since a frame of reference is expedient, the data are processed with read mapping.

Once the reads have been mapped, variant calling is performed. Variant calling is particularly important for samples representing heterogenous populations. Here, the question is, whether a variant that was detected at a low frequency (e.g. at 1%) is actually present in the sample or whether it is the result of sequencing errors. This question can be answered using statistical tests that consider the technology-dependent error distributions in unison with the evidence for the variants.

Due to the time intensity of processing NGS data, there are many stand-alone tools for the interpretation of drug resistance from NGS data. Examples include the freely available tools ShoRah[50], V-Phaser[51], VirVarSeq[52], and MinVar[53] as well as the commercial DeepCheck software[54]. To my best knowledge, the only two freely available web services for interpreting HIV-1 drug resistance are HyDRa and PASeq[55]. In the following, I provide more details on the steps that are typically performed when processing NGS data.[56]

[49] Bao et al. 2011

[50] Zagordi et al. 2011

[51] Yang et al. 2013

[52] Verbist et al. 2014

[53] Huber et al. 2017

[54] Garcia-Diaz et al. 2014

[55] Noguera-Julian et al. 2017

[56] A more comprehensive overview on NGS data processing is provided by DePristo et al. (2011).

Quality Control

The most important quantity for quality control are Phred scores[57], Q, which measure the quality of each position according to

$$Q = -10\ \log_{10} P\,.$$

Here, P is the probability of a base calling error. High values of Q indicate low error rates. For example, Phred scores of 10, 20, and 30 correspond to base calling accuracies of 90%, 99%, and 99.9%, respectively. Phred qualities are determined during the sequencing procedure and are encoded together with the sequence using the FASTQ format. Since sequencing is subject to inherent, technology-dependent biases, quality scores should be recalibrated before they are interpreted[58]. Thereafter, complete reads or individual nucleotides can be excluded based on Phred quality scores. A simple strategy for improving the quality of NGS data is the exclusion of reads with median qualities below a cutoff (e.g. 20). Additionally, for data generated by the Illumina/Solexa approach, the ends of reads should be truncated when a drop in Phred quality scores is detected.

Once the reads have been mapped to a reference sequence, it is crucial to consider the depth of coverage, which describes the number of nucleotides that are observed at a genomic position. While human genome sequencing is typically performed with low coverages (often less than 30-fold)[59], sequencing of viral genomic regions can be performed with 10 000-fold coverages due to the smaller size of viral genes.[60] By considering the depth of coverage, it is possible to estimate the population prevalence of variants in the viral sequence.

Since HIV-1 plasma samples may be contaminated with proviral DNA, it is important to investigate the presence of APOBEC mutations.[61] A mutation is called an APOBEC mutation if it matches the mutational pattern of the APOBEC protein. APOBEC3G, the cytidine deaminase that targets the HIV provirus, causes G → A substitutions in the **GG** and **GA** dinucleotide context. If the activity of APOBEC3G is not inhibited by *vif*, the provirus is rendered replication-incompetent. Since APOBEC mutations can occur at resistance-associated positions, the presence of APOBEC mutations can distort the results of resistance analysis. Due to the action of APOBEC3G, the RT resistance mutations E138K (**GA**G → AAG) and M184I (AT**GG** → ATAA) can emerge even without prior drug exposure[62]. Since APOBEC activity typically results in extensive hypermutation[63], reads exhibiting extensive patterns of APOBEC activity can be excluded[64]. Note that there is evidence indicating that APOBEC may also induce sparse, non-lethal mutations whose detection may be difficult[65].

[57] Ewing et al. 1998

[58] McKenna et al. 2010

[59] Telenti et al. 2016

[60] The potential depth of coverage depends on the VL of the sample. Samples with higher VLs are more easily amplified than those with smaller VLs.

[61] APOBEC mutations are of no concern when Sanger sequencing is performed on samples from plasma because proviral contamination occurs at very small abundances that cannot be taken into account by the approach from Sanger.

[62] Fourati et al. 2012

[63] Armitage et al. 2012

[64] Reuman et al. 2010; Rhee et al. 2016

Read Mapping

There is a large number of methods for aligning reads to a reference genome. Algorithmically, this is either achieved by the use of Burrows-Wheeler transform, suffix trees, or hashing of the template sequence[66]. In practice, the selection of a suitable mapping algorithm depends on the application and the constraints on available time and memory space. Comparing the performance of different mappers is hindered by the unavailability of suitable reference data sets. Popular mappers for short reads from DNA include MOSAIK[67], Bowtie[68], and BWA[69].

Variant Calling

Calling variants at low population frequencies (e.g. at a prevalence small than 1%) can be difficult due to the presence of sequencing errors, particularly at low coverage positions. Thus, variant calling should consider sequencing-run specific error rates in order to improve the specificity of variant detection. Available variant calling approaches provide different sensitivity/specificity trade-offs. Similarly to read mapping, however, benchmarking of variant callers is difficult due to limited availability of benchmarking data sets[70]. Thus, available benchmarking studies have provided only inconclusive results[71]. Well-established variant callers include LoFreq[72], FreeBayes[73], and, more recently, DeepVariant[74].

Resistance Interpretation

Available tools interpret drug resistance from NGS samples by determining consensus sequences for individual prevalence cutoffs and applying established rules-based systems for Sanger sequences such as the approach from Stanford's HIVdb[75].

4.2 Approach of geno2pheno[ngs-freq]

In the following sections, I detail how geno2pheno[ngs-freq] predicts resistance for NGS samples from HIV-1 and HCV. For this purpose, I introduce the frequency file format in Section 4.2.1. The prevalence cutoffs that are used by geno2pheno[coreceptor] are discussed in Section 4.2.2. Section 4.2.3 describes how input samples are processed by geno2pheno[ngs-freq]. Thereafter, I outline the considered quality control measures (Section 4.2.4). A novel method for visualizing viral resistance is introduced in Section 4.2.5. The manner in which the web service was validated is presented in Section 4.2.6. This section concludes with the technical details about the implementation of the

[66] Schbath et al. 2012; Li and Homer 2010

[67] Lee et al. 2014

[68] Langmead and Salzberg 2012

[69] Li and Durbin 2009

[70] Xu 2018

[71] Krøigård et al. 2016; Cai et al. 2016; Wang et al. 2013; Roberts et al. 2013; Spencer et al. 2014; Sandmann et al. 2017; Xu et al. 2014

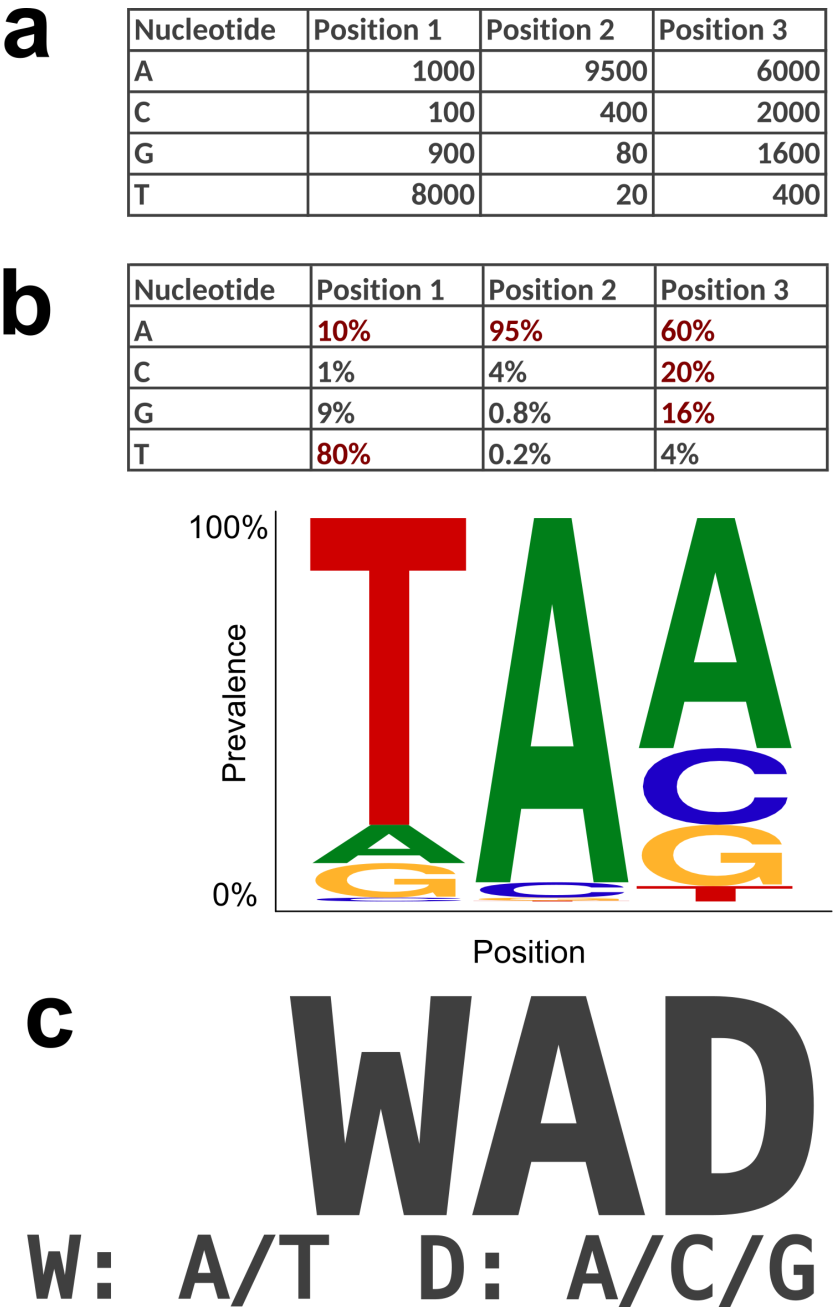
[72] Wilm et al. 2012

[73] Garrison and Marth 2012

[74] Poplin et al. 2018

[75] Paredes et al. 2017

web service (Section 4.2.7).



a

Nucleotide	Position 1	Position 2	Position 3
A	1000	9500	6000
C	100	400	2000
G	900	80	1600
T	8000	20	400

b

Nucleotide	Position 1	Position 2	Position 3
A	10%	95%	60%
C	1%	4%	20%
G	9%	0.8%	16%
T	80%	0.2%	4%

Figure 4.1: Transformation of a single-nucleotide frequency file to a consensus sequence at a prevalence cutoff of 10%. (a) Example of a nucleotide frequency file providing the nucleotide counts for the first three positions in a viral genome. (b) Table of prevalence ratios in which observations with ratios of at least 10% are shown in red. The corresponding motif logo in which the height of individual nucleotides reflects their prevalence is shown below. (c) Consensus sequence constructed for a prevalence cutoff of 10%. Ambiguous positions are encoded according to International Union of Pure and Applied Chemistry (IUPAC) nomenclature.

Figure 1 from Döring et al. (2018) is licensed under CC-BY 4.0

4.2.1 *Format of Input Files*

Frequency files (Figure 4.1) are comma-separated values (CSV) files containing either the counts of observed codons or nucleotides along a viral genome. These files can be generated as a result of applying NGS processing pipelines to raw NGS data (Section 4.1.4). In the following, I consider a frequency file as a matrix $F \in \mathbb{N}_0^{m \times n}$ whose number of rows, $m \in \mathbb{N}$, is determined by the number of genomic positions and whose number of columns, $n \in \mathbb{N}$, is defined either by the number of nucleotides or triplets. Let $\mathcal{A} = \{-, A, C, T, G, N\}$ be the nucleotide alphabet and let $\mathcal{A}_3 = (\mathcal{A} \setminus \{-, N\})^3 \cup \{-\}^3$ be the triplet alphabet. Nucleotide frequency files contain entries $f_{i,j}$, which denote the number of reads supporting the nucleotide $j \in \mathcal{A}$ at position i, while codon frequency files are defined by entries $f_{i,j}$ where $j \in \mathcal{A}_3$ relates to triplets instead.[76]

Note that single-nucleotide frequency files should not be used to indicate the presence of insertions as these could shift the translational frame.[77] Codon frequency files, on the other hand, respect the frame of translation and thus allow for the meaningful representation of insertions.

4.2.2 *Prevalence Cutoffs*

geno2pheno[ngs-freq] enables comparing the impact of mutations at low abundances with those at greater abundances through the selection of two prevalence cutoffs, the *personal cutoff* and the *reference cutoff*. By default, the personal cutoff is set to 2%. This cutoff serves two purposes. First, the cutoff is high enough to ensure tolerance with regard to sequencing errors[78]. Second, the cutoff is low enough to allow for the consideration of clinically relevant minority variants[79]. The default setting for the reference cutoff was chosen in order to obtain drug resistance interpretations that are comparable to the results that would be obtained with Sanger sequencing. The detection limit of Sanger sequencing[80] provides a reasonable frame of reference for choosing a default reference cutoff. The default reference cutoff for samples from HCV was set to 15% as this is the established cutoff for interpreting NGS data with respect to HCV drug resistance[81]. Data for HIV-1 suggest that Sanger sequencing is highly sensitive to NRTI mutations such that a lower NGS prevalence cutoff at 10% is necessary in order to recover all relevant NRTIs mutations with NGS[82]. Thus, the default reference cutoff for samples from HIV-1 was set to 10%.

4.2.3 *Workflow*

Once a user has supplied a set of prevalence cutoffs, a frequency file, and an optional sample identifier, geno2pheno[ngs-freq] performs the steps indicated in Figure 4.2: (1.) Generation of a consensus sequence for every prevalence cutoff; (2.) inference of the viral species; and

[76] Note that j represents a nucleotide or a triplet rather than an integer index. The column in F corresponding to j can be found using a bijective function that takes the observation j as an input.

[77] Frame shift mutations are typically not found when circulating viral RNA is sequenced because the presence of frame shifts leads to non-functional proteins, which prevent viral maturation or replication.

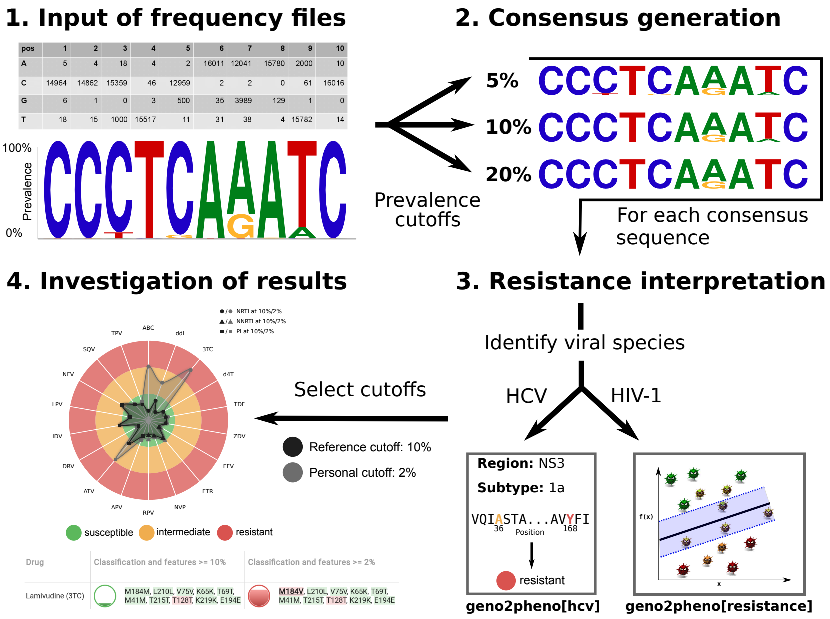
[78] Manley et al. 2016; Loman et al. 2012; Archer et al. 2012

[79] Goodman et al. 2009

[80] The detection limit of Sanger sequencing typically ranges between 10% and 20% (Tsiatis et al., 2010).

[81] Pawlotsky 2016; Zeuzem et al. 2017

[82] Thielen 2014



pos	1	2	3	4	5	6	7	8	9	10
A	5	4	18	4	2	16011	12041	15780	2000	10
C	14964	14862	15359	46	12959	2	2	0	61	16016
G	6	1	0	3	500	35	3989	129	1	0
T	18	15	1000	15517	11	31	38	4	15782	14

Drug	Classification and features >= 10%	Classification and features >= 2%
Lamivudine (3TC)	○ M184M, L210L, V75V, K65K, T69T, M41M, T215T, T128T, K219K, E194E	● M184V, L210L, V75V, K65K, T69T, M41M, T215T, T128T, K219K, E194E

Figure 4.2: Workflow of geno2pheno[ngs-freq].

Graphical abstract from Döring et al. (2018) is licensed under CC-BY 4.0.

(3.) identification of viral resistance for every consensus sequence. The following sections provide the computational details for the individual processing steps.

Generation of Consensus Sequences

For each prevalence cutoff $c_F \in [0,1]$ that is provided for a frequency file $F \in \mathbb{N}_0^{m \times n}$, the web service generates a consensus sequence in the following manner. Let $d_i = \sum_j f_{i,j}$ indicate the depth of coverage at position $i \in \{1, 2, \ldots, m\}$ in the frequency file. If F is a codon frequency file, the ratio at which the codon $z \in \mathcal{A}_3$ occurs at position i is determined by $\mathcal{P}_{i,z} = \frac{f_{i,z}}{d_i}$.

If F is a single-nucleotide file, the computations are slightly more involved because we need to determine the prevalence of codons in order to compute the prevalence of individual amino acids. Starting with the first nucleotide of the protein sequence of interest, we only consider positions i that form the first position of a coding triplet consisting of positions i, $i+1$, and $i+2$. Without loss of generality, assume that nucleotides $j \in \mathcal{A}$, $k \in \mathcal{A}$, and $l \in \mathcal{A}$ are observed at the three respective positions. Under the assumption of positional independence, single-nucleotide prevalence ratios

$$x_{i,j} = \frac{f_{i,j}}{d_i}$$
$$x_{i+1,k} = \frac{f_{i+1,k}}{d_{i+1}}$$
$$x_{i+2,l} = \frac{f_{i+2,l}}{d_{i+2}}$$

are computed to determine the codon prevalence ratio $x_{i,z} = x_{i,j} \times x_{i+1,k} \times x_{i+2,l}$ at which $z = (j, k, l)$ is expected to be observed. Given that

$$D_i = \min \{d_w | w \in \{i, i+1, i+2\}\}$$

indicates the maximal possible codon coverage, the expected number of reads supporting codon z is $D_i x_{i,z}$. Thus, the prevalence of codon z at position i is

$$\mathcal{P}_{i,z} = \frac{D_i x_{i,z}}{\sum_{z' \in \mathcal{A}^3} D_i x_{i,z'}} .$$

Having determined the codon frequencies, the consensus sequence is constructed in the following way. For the i-th position, only the set of observations (i.e. nucleotides or triplets) $\mathcal{A}_{i,c_F} = \{j | x_{i,j} \geq c_F\} \subseteq \mathcal{A}$, whose prevalence is at least c_F is considered. Then, position i of the consensus sequence is set to $s_{i,c_F} = \phi(\mathcal{A}_{i,c_F})$. The function ϕ converts nucleotides or codons into their corresponding IUPAC representation[83] (Table A.1). In case that $\mathcal{A}_{i,c_F}$ is empty (i.e. no observations were selected because all frequencies were less than the cutoff), the greedy criterion $\mathcal{A}_{i,c_f} = \arg\max_j x_{i,j}$ is used instead. The use of this criterion ensures that no artificial deletions are introduced into the consensus sequence. As an example, let us consider the first position shown in Figure 4.1. Given a prevalence cutoff of $c_F = 10\%$ and the observed prevalence ratios $x_{iA} = 10\%$, $x_{iC} = 1\%$, $x_{iG} = 9\%$, and $x_{iT} = 80\%$, we would set $s_{i,10\%} = \phi(\mathcal{A}_{i,10\%}) = \phi(A, T) = W$ because only the frequencies of A and T exceed the cutoff.

Due to the redundancy of the genetic code, there are amino acids that are encoded by multiple nucleotide triplets (Figure 4.3).[84] To compute the frequency of an amino acid at a certain position, it is therefore necessary to sum over the frequencies of all codons encoding that amino acid. Let Φ indicate a function that translates from codons to amino acids and let

$$\mathcal{B} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

indicate the alphabet for the twenty amino acids. The prevalence of

[83] Cornish-Bowden 1985

[84] For example, leucine is encoded by six different triplets: CTT, CTC, CTA, CTG, TTA, and TTG.

amino acid $a \in \mathcal{B}$ at position i is determined as

$$\mathcal{P}_{i,a} = \sum_{\text{codon } z \text{ with } \Phi(z)=a} \mathcal{P}_{i,z}\,.$$

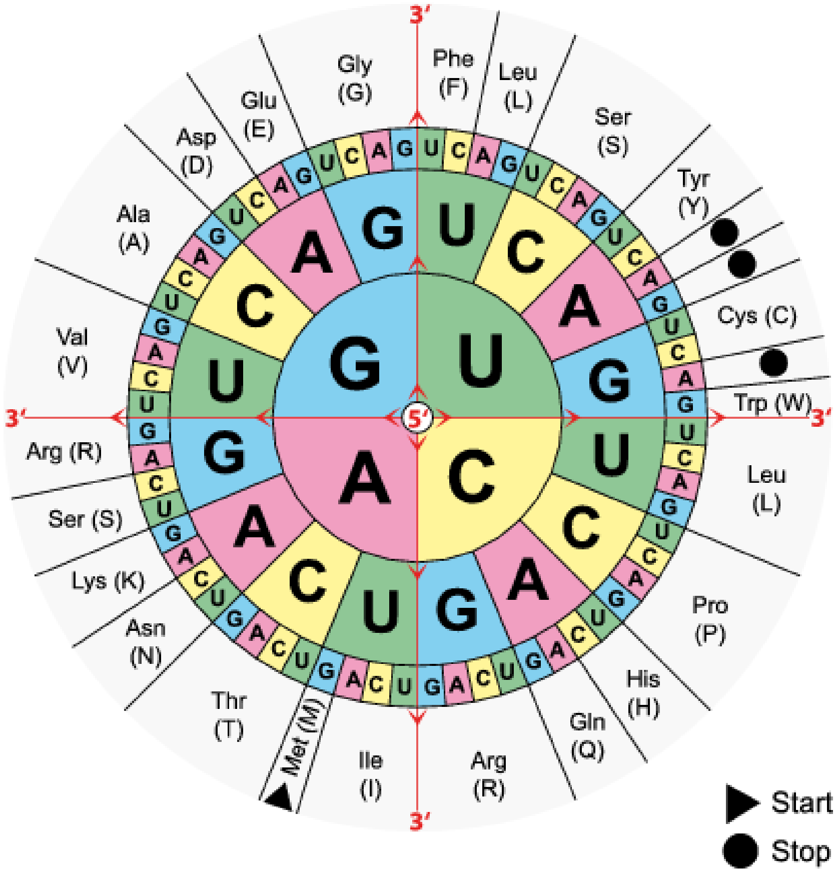


Figure 4.3: Translation of codons to amino acids. To obtain the translation of a nucleotide triplet, one needs to move from the innermost to the outermost circle. Note that uracil (U), which occurs in RNA, corresponds to thymine (T) in DNA.

Retrieved from Wikipedia, released into the public domain.

In order to prevent the consideration of positions with insufficient coverage, a coverage cutoff, which is set to $d_{\text{cut}} = 20$, was defined. Positions with $d_i < d_{\text{cut}}$ are assigned the ambiguity code N that represents any possible base. The benefit of this approach over the previously used, more stringent truncation strategy[85,86] is that it allows for the simultaneous interpretation of resistance for multiple genomic regions even if their amplicons are separated by stretches of low coverage (e.g. NS3, NS5A, and NS5B).

In the following, I use the term *default consensus sequence* to denote the consensus sequence of a sample that was constructed at the default reference cutoff for the corresponding virus. Let s_{c_F} indicate a consensus sequence at a specific cutoff $c_F \in [0, 1]$. Then, the default consensus sequences for HIV-1 and HCV are denoted by $s_{10\%}$ and $s_{15\%}$, respectively.

[85] Döring et al. 2018

[86] Previously, the cutoff was set to $d_{\text{cut}} = \max(20, \min(100, 0.1 d_{\text{med}}))$ where d_{med} is the median depth of coverage.

Inference of the Viral Species

The viral species from which an input sample originates is inferred by aligning the sample's default consensus sequence to the genomic regions of the reference sequences for HIV-1 and HCV, HXB2[87] and H77[88], respectively. To ensure that resistance is interpreted only for supported viral species, the existence of at least one high-quality alignment is required. An alignment is considered a high-quality alignment if it induces a high degree of similarity between query and reference sequence. More formally, a high-quality alignment must satisfy two similarity criteria, which are defined by dividing the number of matching amino acids in the alignment either by the length of the alignment (alignment similarity) or by the length of the reference sequence (reference similarity).

For sequences from HIV-1, a minimal alignment similarity of 60% and a minimal reference similarity of 50% are used for all regions, except for RT. Since all major drug resistance mutations are located within the first half of the gene[89], RT is frequently merely partially amplified. Thus, a reference similarity of only 20% is required for the RT. Due to the greater phylogenetic divergence of HCV, an alignment similarity of 40% and a reference similarity of 20% are required for all HCV regions. If no high-quality alignments are available, it is assumed that the sample derives from a species that is not supported and no further computations are performed. In all other cases, the viral species of the reference sequence with the greatest alignment score is used.

Identification of Viral Resistance

Viral resistance of HIV-1 and HCV samples is interpreted using the approach of geno2pheno[resistance][90] and a reimplemented version of geno2pheno[hcv][91], respectively. HIV-1 drug resistance is classified using the SIR scheme (Section 2.5.2).

For geno2pheno[ngs-freq], geno2pheno[hcv] was reimplemented in C++, with the following two changes. First, when the similarity to the reference sequence is calculated, ambiguous nucleotides are taken into account in the following way. Let n indicate the number of nucleotides that are represented by an (ambiguous) nucleotide in the reference sequence (e.g. $n = 1$ for T but $n = 2$ for W) and let m be the number of disambiguated nucleotides at the current position in the query that match the disambiguated nucleotides in the reference sequence. Then, the match score is $\frac{m}{n}$. Thus, when the query nucleotide is W and the reference nucleotide is T, then W is disambiguated to A and T. Since T is observed in both query and reference, we have $m = n = 1$ and the match score is 1 rather than

[87] Ratner et al. 1985

[88] Kolykhalov et al. 1997

[89] Shafer et al. 2007

[90] Beerenwinkel et al. 2003; Lengauer and Sing 2006; Pironti et al. 2017b,a

[91] Kalaghatgi et al. 2016

0 as before (W does not match T). Second, the original approach of geno2pheno[hcv] may ignore resistance-relevant positions because it considers only mutated positions. However, since some of the HCV reference sequences carry resistance-associated positions, this leads to an underestimation of resistance for sequences exhibiting these wild-type amino acids. Therefore, when evaluating the rule set, the reimplementation does not only scan mutated amino acids but all amino acids. The following resistance levels are used for samples from HCV: *susceptible*, *substitution on scored position* (rule matches the position but not the amino acid), *resistance-associated mutation in closest subtype* (for rare subtypes only[92]: matched a rule for the closest non-rare subtype), *reduced susceptibility*, *resistant*, and *unlicensed* (drug is not approved for the identified subtype).

The presence of ambiguities at more than a single codon position can lead to the inclusion of spurious codons into consensus sequences. Imagine, for example, that the codons ATA (Ile) and TTT (Phe) are observed at the same genomic position. In this case, the triplets ATT (Ile) and TTA (Leu) would be considered in addition to the observed codons when constructing the consensus sequence. Then, the resistance interpretation algorithm would incorrectly consider three (Ile, Phe, and Leu) rather than two (Ile and Phe) amino acids at this position. To prevent this, geno2pheno[ngs-freq] considers the prevalence of codons when translating consensus sequences in order to ignore codons that were not actually observed. Assume that the codons $\{z_1, z_2, \ldots, z_n\}$ were observed at the i-th position. Then, only codons $z' \in \{z_1, z_2, \ldots, z_n\}$ with $\mathcal{P}_{i,z'} \geq c_F$, where $c_F \in [0, 1]$ is the selected prevalence cutoff, are translated.

4.2.4 Quality Control Mechanisms

Besides the requirement of high-quality alignments, geno2pheno[ngs-freq] uses several additional measures for evaluating the quality of a sample. With regard to coverage, the tool outputs the extent to which individual genomic regions were covered as well as the depth of coverage. Moreover, individual positions are annotated with their depth of coverage and warnings are issued if positions associated with resistance exhibit a coverage less than 100.[93] The quality of consensus sequences constructed at different cutoffs can be analyzed with respect to the presence of stop codons, frame shifts, and ambiguous positions. For HIV-1, potential APOBEC mutations are annotated using the approach from Rhee et al. (2016). Here, APOBEC signature mutations are defined as G to A mutations at GG or GA dinucleotides that are conserved across 98% of Los Alamos National Laboratory (LANL) HIV sequences.[94] If more than one

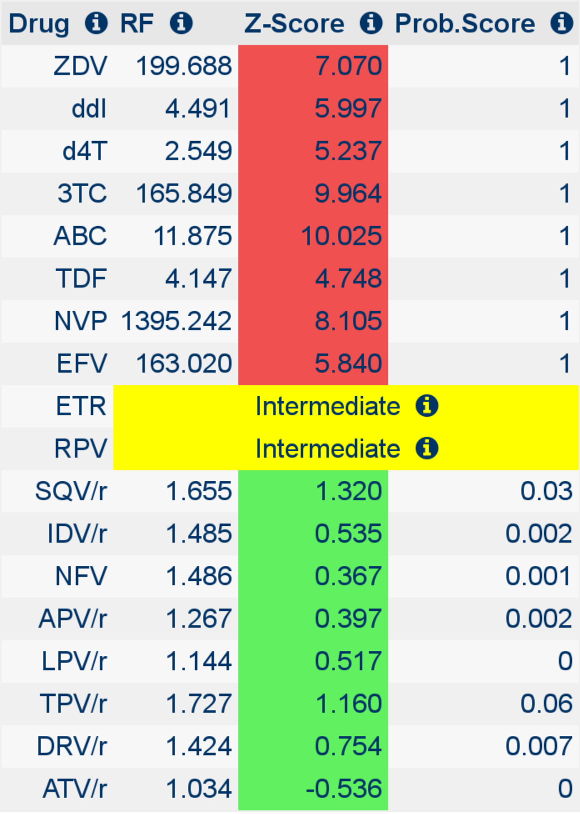
[92] All subtypes other than 1a, 1b, 2a, 3a, 4a, and 4d are considered rare subtypes. For these rare subtypes, drug resistance is not studied sufficiently to allow for resistance interpretation.

[93] For viral sequencing where the depth of coverage typically exceeds 10 000 reads, a coverage of 100 is relatively small because it represents less than 1% of the typical coverage.

[94] Due to technical reasons, geno2pheno[ngs-freq] currently cannot identify APOBEC mutations that span across codons.

APOBEC signature mutation is identified, a warning is issued.

4.2.5 *Visualization of Drug Resistance*



Drug ℹ	RF ℹ	Z-Score ℹ	Prob.Score ℹ
ZDV	199.688	7.070	1
ddI	4.491	5.997	1
d4T	2.549	5.237	1
3TC	165.849	9.964	1
ABC	11.875	10.025	1
TDF	4.147	4.748	1
NVP	1395.242	8.105	1
EFV	163.020	5.840	1
ETR	Intermediate ℹ		
RPV	Intermediate ℹ		
SQV/r	1.655	1.320	0.03
IDV/r	1.485	0.535	0.002
NFV	1.486	0.367	0.001
APV/r	1.267	0.397	0.002
LPV/r	1.144	0.517	0
TPV/r	1.727	1.160	0.06
DRV/r	1.424	0.754	0.007
ATV/r	1.034	-0.536	0

Figure 4.4: Resistance table for a sample sequence, as determined by the geno2pheno[resistance] web service.

A new approach for the visualization of quantitative predictions of drug resistance was developed in order to improve the interpretation of resistance estimates. geno2pheno[resistance] provides a table of drug resistance such as the one shown in Figure 4.4. Here, the level of resistance is provided in terms of the SIR system such that susceptibility is shown in green, intermediate resistance in yellow, and, resistance in red. Interpreting whether the z-score for a drug is relatively high or low given its SIR class requires knowledge of the lower and upper z-score cutoffs that were used for the classification, which are not publicly available.

Therefore, the following approach was used to arrive at a visualization for the relative level of resistance. Z-scores were estimated for all sequences contained in the geno2pheno[resistance] training data set. The minimum and maximum z-score, denoted by z_{min} and z_{max}, respectively, were determined for each drug. Let z_{low} and z_{high} denote the corresponding lower and upper cutoffs of the SIR classification as determined by the approach presented in Section 4.1.2.

Based on these values, z-score vectors $z^l \in \mathbb{R}^2$ representing each level of resistance $l \in \{S, I, R\}$ were constructed:

$$\begin{aligned} z^S &= (z_{\text{min}}, z_{\text{low}}) \\ z^I &= (z_{\text{low}}, z_{\text{high}}) \\ z^R &= (z_{\text{high}}, z_{\text{max}}) \end{aligned}$$

Given a prediction consisting of a z-score, $z \in \mathbb{R}$, and a resistance level, l, the relative level of resistance, $z_l \in [0, 1]$, is computed as

$$z_l = \begin{cases} 0 & \text{if } z < z_1^l \\ 1 & \text{if } z > z_2^l \\ \frac{z - z_1^l}{z_2^l - z_1^l} & \text{otherwise} \end{cases} .$$

Thus, z_l of 1 indicates the highest possible level of resistance for a resistance class and 0 the lowest. In order to obtain an overall measure of resistance, z_l is adjusted in dependence on the level of resistance l:

$$\phi(z_l) = \begin{cases} 0 + \frac{1}{3} z_l & \text{if } l = S \\ \frac{1}{3} + \frac{1}{3} z_l & \text{if } l = I \\ \frac{2}{3} + \frac{1}{3} z_l & \text{if } l = R \end{cases} .$$

The values plotted in the radar visualization of drug resistance for HIV-1 (Figure 4.6) and the gauges shown in Figure 4.7 are based on the mapping provided by ϕ. Besides improving the interpretability of z-scores, the transformation has the benefit of enabling the comparison of z-scores resulting from different prediction models. This is important since the default HIV-1 prediction model of geno2pheno[ngs-freq] relies on a combination of prediction models: The phenotypic model of geno2pheno[resistance] is applied for all drugs except for ETR and RPV, for which the drug-exposure model is used.[95]

4.2.6 Validation

geno2pheno[ngs-freq] was validated by analyzing a total of 3 844 frequency files of which 926 files represented samples from HIV-1 (24.1%) and 2 918 files represented samples from HCV (75.9%). Resistance interpretations were obtained for 922 of 926 HIV-1 samples (99.6%) and 2 898 of 2 918 HCV samples (99.3%). For the remaining samples, geno2pheno[ngs-freq] did not provide a result due to low depth of coverage.

To validate the reimplemented version of geno2pheno[hcv], I investigated the concordance between the predictions of geno2pheno[ngs-

[95] The phenotypic predictions for ETR and RPV are substituted by default because they are based on insufficient training data.

freq] and the geno2pheno[hcv] web service using the default consensus sequences constructed from the 2 866 successfully analyzed HCV frequency files. The concordance between geno2pheno[ngs-freq] and geno2pheno[hcv] was 99.7%. The differences are a result of two modifications of the original approach from geno2pheno[hcv], which are described in Section 4.1.3. Since the reimplementation takes ambiguous nucleotides into account when computing sequence similarities, there are instances in which different reference sequences (i.e. different genotypes) were selected such that different resistance levels ensued. Moreover, the reimplementation scans all of the rules rather than evaluating merely those rules matching the mutations in the sequence. Therefore, the resistance levels of sequences whose wild-type amino acids are associated with resistance were different in the reimplementation.

An analogous validation was not performed for the HIV-1 samples because predictions for HIV-1 samples are based on the established version of geno2pheno[resistance], which has already been extensively validated[96]. The median runtimes required for analyzing HIV-1 and HCV samples were 6 seconds and 4 seconds, respectively.

4.2.7 *Technical Details*

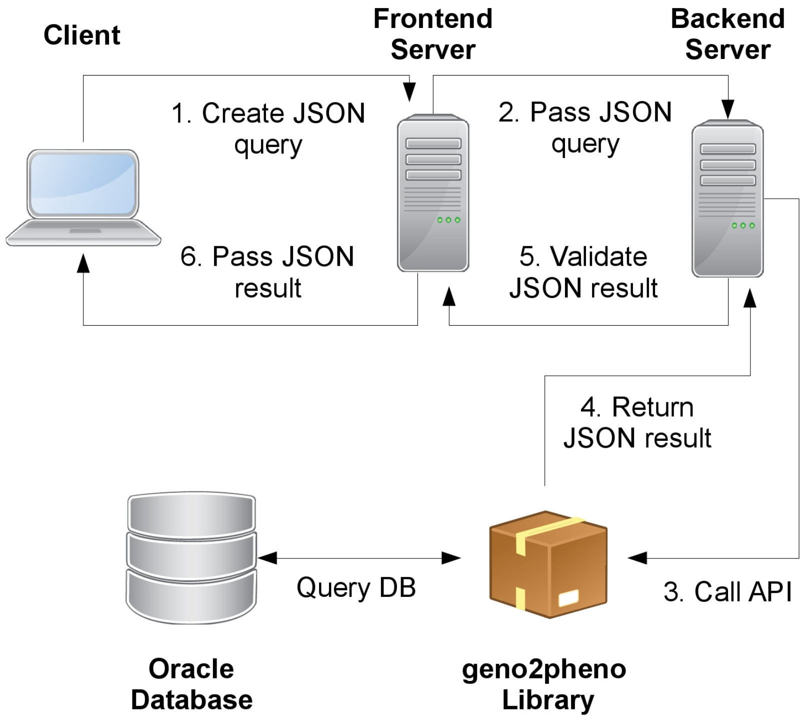


Figure 4.5: Computational architecture of the geno2pheno[ngs-freq] web service.

[96] Beerenwinkel et al. 2003; Pironti et al. 2017b

Figure 4.5 depicts the computational architecture of the geno2pheno[ngs-freq] web service. The frontend has been implemented in Typescript and relies on the React library. Once a user has requested the geno2pheno[ngs-freq] website, the JavaScript application that provides the user interface is retrieved from the backend server. The subsequent communication between client and server is facilitated through a representational state transfer (REST) service. When a user performs an action in the web interface, a javasript object notation (JSON) query is created. This query is passed from the frontend web server (Apache) to the backend server, which runs an instance of the Play framework (Java). The backend server queries the application programming interface (API) that is defined by the geno2pheno[ngs-freq] C++ library, which interprets the request and computes the result of the query. The library uses an Oracle database for data retrieval. Once the result has been determined, it is returned to the client via JSON and the results are displayed by the client's web browser. The web interface affords the analysis of batches containing at most 20 files.

4.3 Case Studies

In the following, I provide two case studies that illustrate how geno2pheno[ngs-freq] can offer insights that may impact clinical decision making. The case studies can be replicated by accessing the geno2pheno[ngs-freq] web service, ensuring that the default prevalence cutoffs (2%, 10%, and 15%) are selected, and loading the exemplary frequency files. The HIV-1 case study (Section 4.3.1) was performed on the basis of the phenotypic geno2pheno[resistance] model, while the HCV case study (Section 4.3.2) was performed using the approach of geno2pheno[hcv].

4.3.1 HIV-1 Resistance Interpretation

This case study is based on a plasma isolate from an HIV-1 infected patient with a VL of 102,000 copies per mL. The plot of viral drug resistance (Figure 4.6) reveals that the major viral populations, at the reference cutoff of 10%, seem to be susceptible to nearly all drugs. When minor viral populations are included in the analysis by considering the consensus sequence at the personal cutoff of 2%, considerably increased levels of resistance against the NRTIs ABC, ddI, and 3TC are found. Using the resistance table shown in Figure 4.7, it is possible to see that the increased levels of resistance are caused by the well-studied resistance mutation M184V, which was not considered in the prediction for the reference cutoff because the

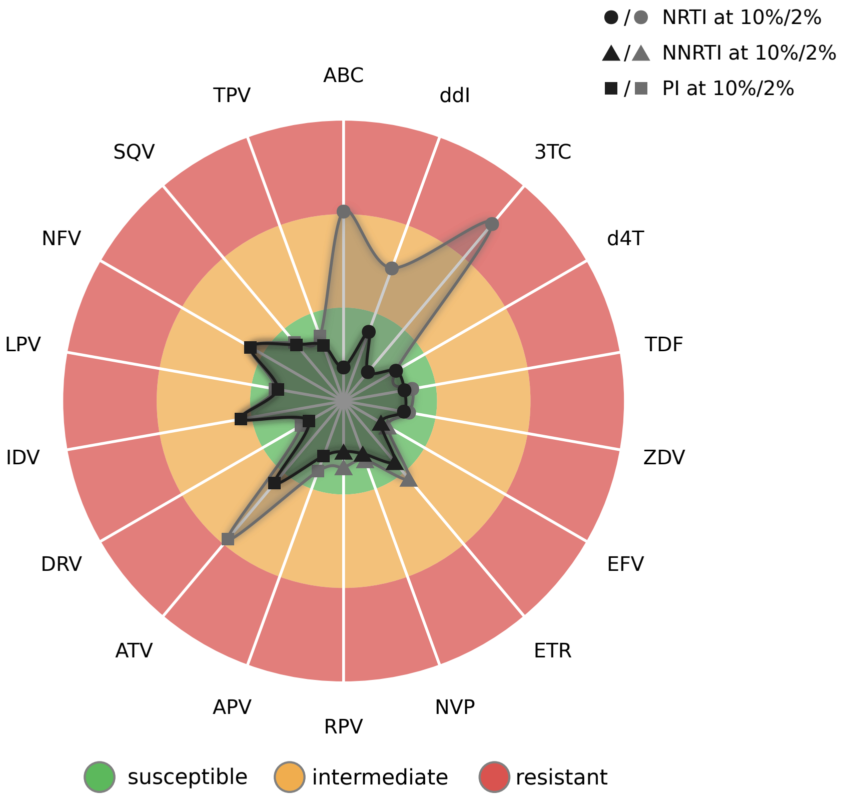
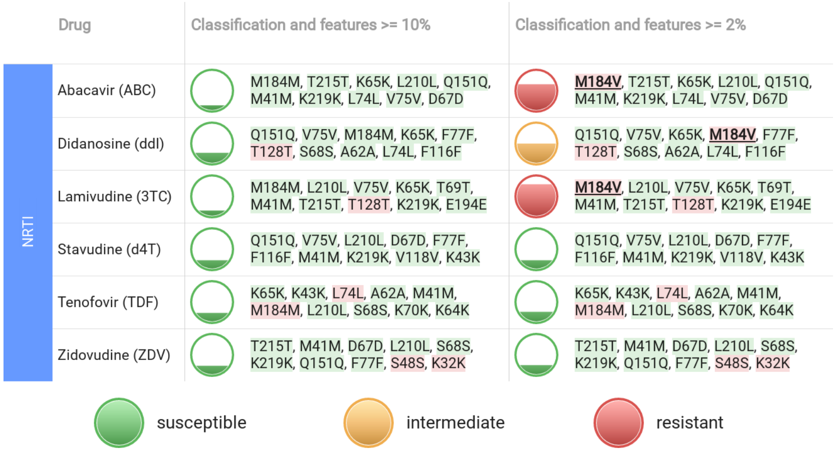


Figure 2 from Döring et al. (2018) is licensed under CC-BY 4.0

Figure 4.6: Radar plot of predicted drug resistance for the sample from the HIV-1 case study. Each spoke in the plot relates to an antiretroviral drug. The points defining the surfaces are determined through the standardized RF that are predicted by geno2pheno[resistance]. The three colored circle sectors indicate the estimated levels of drug resistance. The inner surface shows the estimated level of resistance for the consensus sequence based on the reference prevalence cutoff at 10%, while the outer surface indicates the level of resistance for the consensus sequence based on the personal cutoff at 2%.



	Drug	Classification and features >= 10%	Classification and features >= 2%
NRTI	Abacavir (ABC)	🟢 M184M, T215T, K65K, L210L, Q151Q, M41M, K219K, L74L, V75V, D67D	🔴 **M184V**, T215T, K65K, L210L, Q151Q, M41M, K219K, L74L, V75V, D67D
	Didanosine (ddl)	🟢 Q151Q, V75V, M184M, K65K, F77F, T128T, S68S, A62A, L74L, F116F	🟠 Q151Q, V75V, K65K, **M184V**, F77F, T128T, S68S, A62A, L74L, F116F
	Lamivudine (3TC)	🟢 M184M, L210L, V75V, K65K, T69T, M41M, T215T, T128T, K219K, E194E	🔴 **M184V**, L210L, V75V, K65K, T69T, M41M, T215T, T128T, K219K, E194E
	Stavudine (d4T)	🟢 Q151Q, V75V, L210L, D67D, F77F, F116F, M41M, K219K, V118V, K43K	🟢 Q151Q, V75V, L210L, D67D, F77F, F116F, M41M, K219K, V118V, K43K
	Tenofovir (TDF)	🟢 K65K, K43K, L74L, A62A, M41M, M184M, L210L, S68S, K70K, K64K	🟢 K65K, K43K, L74L, A62A, M41M, M184M, L210L, S68S, K70K, K64K
	Zidovudine (ZDV)	🟢 T215T, M41M, D67D, L210L, S68S, K219K, Q151Q, F77F, S48S, K32K	🟢 T215T, M41M, D67D, L210L, S68S, K219K, Q151Q, F77F, S48S, K32K

🟢 susceptible 🟠 intermediate 🔴 resistant

Figure 4.7: Resistance table for the sample from the HIV-1 case study. The color of the gauges indicates the predicted level of resistance; their fill level reflects the extent of resistance. The ten features with the highest impact on the prediction are shown, ordered by decreasing weight. The features are shown using color (green/red indicates decreased/increased resistance), underlining (mutations), and bold font (only reported at the personal cutoff).

mutation occurs at a population prevalence of 2.36%. M184V is not only known for enhancing the susceptibility to the NRTIs ZDV, d4T, and TDF, but also for delaying the emergence of resistance to these drugs.[97] Therefore, a combination therapy consisting of two such NRTIs and one PI such as TDF + ZDV + DRV would be a reasonable choice.

An alternative treatment with fewer side effects could consist of TDF + FTC + DRV. The idea behind this treatment is that FTC could stabilize M184V such that susceptibility to TDF is ensured[98]. Moreover, although M184V is associated with a more than 100-fold reduction in susceptibility to FTC *in vitro*[99], FTC exhibits residual activity in the presence of M184V *in vivo*[100]. Therefore, even if the minority population characterized by M184V were to become more prevalent over time, FTC would still be residually active.

4.3.2 *HCV Resistance Interpretation*

In this HCV case study, resistance to NS5A inhibitors is investigated. The visualization of resistance for the considered sample (Figure 4.8) shows that the viral population at the 15% cutoff seems to be susceptible to all DAAs targeting NS5A, while the population at the 2% cutoff seems to be resistant to most NS5A inhibitors due to the presence of the resistance mutation 30R (Figure 4.9), which was found at a prevalence of 6.1%. Using this information, the treating clinician may decide to avoid the use of the NS5A inhibitors DCV, EBR, LDV, and OBV to which the minor viral population seems to be resistant and instead use VEL, for which no resistance was reported.

[97] Gallant 2006

[98] Gallant 2006

[99] Schinazi et al. 1993

[100] Campbell et al. 2005; Castagna et al. 2006

4.4 Discussion

This chapter presented a new approach for genotypic resistance testing based on NGS data. The result of this work is geno2pheno[ngs-freq], a web server for the genotypic interpretation of drug resistance for HIV-1 or HCV samples, which is freely available via the internet. geno2pheno[ngs-freq] facilitates the analysis of viral drug resistance in minority populations. While other web services offering NGS resistance interpretation for HIV-1 are based on rules-based approaches, geno2pheno[ngs-freq] enables the use of the statistical models from geno2pheno[resistance] in the NGS setting. Moreover, to the best of my knowledge, the server provides the first freely available interpretation engine for NGS samples from HCV.

In contrast to other web services for interpreting viral resistance for NGS samples, geno2pheno[ngs-freq] uses frequency files instead of raw NGS data. The use of frequency files over raw sequencing data offers three benefits. First, due to their small size (kilobytes vs megabytes), frequency files can be uploaded quickly even in settings with limited bandwidth. Second, skipping the time-intensive step of processing the raw NGS data allows for rapid analyses (a few seconds vs several minutes or hours). Third, the use of frequency files decouples the step of NGS data processing from drug resistance analysis. Services requiring the input of raw sequencing data are limited by the application of largely pre-determined processing pipelines. geno2pheno[ngs-freq], on the other hand, does not impose limitations on the manner in which NGS samples are processed. This enables the use of customized NGS processing pipelines. The fact that geno2pheno[ngs-freq] requires local data preprocessing is not a barrier to its application because I have developed a reference implementation for transforming binary alignment map (BAM) files to frequency files, which is publicly available via bamToFreq.

The use of frequency files naturally also entails a loss of information. For nucleotide frequency files, amino-acid frequencies need to be estimated and spurious amino acids may be generated. These problems can be avoided by using codon rather than single-nucleotide frequency files because this file format accurately models amino-acid frequencies and allows for the correct translation of codons containing multiple ambiguous positions. This is also an advantage of geno2pheno[ngs-freq] over the direct input of consensus sequences to geno2pheno[resistance] or geno2pheno[hcv]: since these services do not have prevalence information available, they cannot correct for spurious observations. The frequency file format does not allow for the consideration of co-occurrence patterns beyond the codon level. However, this is not a practical disadvantage, because such patterns are not yet taken into account by any genotypic drug resistance test. Frequency files also do not allow for quasispecies reconstruction, which is possible with raw sequencing data. However, the low precision/recall trade-off of methods for quasispecies reconstruction poses the question whether these methods are mature enough for routine use[101], although the improved performance of more recent approaches inspires hope[102].

Recently, the Winnipeg consortium has proposed the amino-acid variant format (AAVF) as a means of representing the NGS results for viral samples.[103] This format is similar to the variant call format (VCF)[104], which is a general format for denoting variants detected by NGS. However, the AAVF is tailored towards viral sequencing because it provides the prevalence of observed amino-acid variants. In contrast to the frequency file format, which considers all possible nucleotides or codons for every position, the AAVF is more succinct as it considers only the observed amino acids. Time will tell whether the AAVF will establish itself as the standard for summarizing the results of viral sequencing and whether it will form a basis for interpreting viral drug resistance.

An important contribution of geno2pheno[ngs-freq] is that the service eases the interpretation of viral drug resistance compared to previous approaches. The visualization of viral drug resistance as a radar plot has several advantages over other representations (compare Figure 4.6 with Figure 4.4). First, drug resistance for all antiviral drugs is intuitively represented in a single diagram. Second, drug resistance for minor and major viral populations can be contrasted. Third, for predictions of HIV-1 drug resistance, relative differences in resistance become evident. To illustrate the third point, consider the predictions for APV shown in Figure 4.6. Although the discretized level of resistance is identical for the populations at the 2% and the 10% prevalence cutoff, it becomes apparent that resistance is markedly increased in the minor population.

A general limitation of genotypic resistance testing based on NGS data is that the clinical impact of minor populations carrying resistance mutations is still largely unclear. While NNRTI resistance mutations at low abundances are associated with virological failure[105], the general impact of minority resistant variants on the treatment outcomes of HIV-1 infected patients is still unclear[106]. The impact of HCV minority resistant variants is less studied than for HIV-1 but the presence of minority resistant variants has recently been shown to deteriorate the outcomes in subtype-1 patients being treated with NS5A inhibitors[107]. Treatment choices that are informed by minority variants should be taken with care[108] for two reasons. First, the presence of minority resistance mutations seem to have different ramifications

[101] Prosperi et al. 2013

[102] Töpfer et al. 2014; Posada-Cespedes et al. 2017

[103] Ji et al. 2018

[104] Danecek et al. 2011

[105] Cozzi-Lepri et al. 2015; Van Laethem et al. 2007; Delobel et al. 2011; Simen et al. 2009

[106] Johnson et al. 2008; Lataillade et al. 2010; Stekler et al. 2011; Peuchant et al. 2008; Moscona et al. 2017; Callegaro et al. 2014

[107] Zeuzem et al. 2017

[108] Johnson and Geretti 2010

on treatment success for different drugs. Second, the detected level of resistance for a minor viral population may be inaccurate due to biological and technical factors. On the one hand, minority resistance mutations are not always biologically meaningful because they may be found on reads that belong to viral variants with low replicative capacity (due to low fitness) or without any replicative capacity (due to the action of APOBEC). On the other hand, applying prediction models that were developed for Sanger sequences to NGS data may incur inaccuracies. This is because currently available approaches consider only the amino acid with the greatest impact on drug resistance when multiple amino acids are observed at a single position. Selecting only the amino acids most strongly associated with drug resistance is particularly problematic for consensus sequences at low abundances (e.g. at 1%) as it is likely that constructed sequences contain resistance-associated mutations that do not occur on the same viral strand *in vivo*. This can lead to overestimated levels of drug resistance because all of these mutations are taken into account in unison. Models that do not rely on consensus generation would allow for more accurate estimates of biological reality.

The following work could be done in the future. For HIV, prediction models for other genomic regions could be added to geno2pheno[ngs-freq]. Most importantly, the server should be able to identify viral susceptibility towards INSTIs, which have become the pillars of antiretroviral therapy. Although coreceptor prediction[109] could also be integrated into geno2pheno[ngs-freq], this would not be a priority because the geno2pheno[454] web server[110] already predicts HIV-1 coreceptor usage under consideration of co-occurring mutations on individual reads, which is not possible using the frequency file format. At a later point in time, support for samples from other viral species such as hepatitis B virus (HBV), for which the emergence of resistance is relevant[111], could be added. The usefulness of geno2pheno[ngs-freq] could be enhanced through a publicly available API with which the service could be queried programmatically. Moreover, the prediction models underlying geno2pheno[ngs-freq] could be further improved. For example, clinical models that determine drug exposure[112] could benefit from learning from the combinations in which drugs are prescribed through the use of multi-task learning[113].

In the future, a greater amount of NGS data with corresponding clinical outcomes should allow for the development of novel models trained on NGS data. By directly encoding either the frequencies of amino-acids (e.g. using frequency files) or incorporating information on the level of individual reads (e.g. using the FASTQ format), these models could learn the specific circumstances under

[109] Lengauer et al. 2007; Döring et al. 2016

[110] Thielen and Lengauer 2012

[111] Shaw et al. 2006; Beggel et al. 2012

[112] Pironti et al. 2017a

[113] Bickel et al. 2008

which minority variants influence therapeutic outcomes. For example, read-based models could learn the co-occurrence patterns of mutations in individual reads. These models could be formulated in terms of multiple-instance learning problems, for example, using set kernels[114]. An alternative would be to form prediction models based on reconstructed quasispecies[115], under consideration of their prevalence and fitness[116]. These new models could simplify the interpretation of drug resistance from NGS data because they would provide a single prediction for every drug instead of multiple predictions for populations at different abundances, which is currently the case due to consensus generation. By learning from the frequencies at which mutations are observed, these models would also have an edge over currently used rules-based approaches, which do not consider the abundance of mutations yet.

In summary, geno2pheno[ngs-freq] is a web service that allows for the detection of minor viral populations carrying drug resistance mutations, which is important for two reasons. First, although the impact of minority variants is still unclear, it is never detrimental to exclude drugs from consideration when their effectiveness could be impaired by resistant minorities, as long as suitable alternative treatment options are available. Second, by improving the surveillance of drug-resistant minorities, it should be possible to further elucidate the impact of resistant minor variants. geno2pheno[ngs-freq] is freely available via ngs.geno2pheno.org.

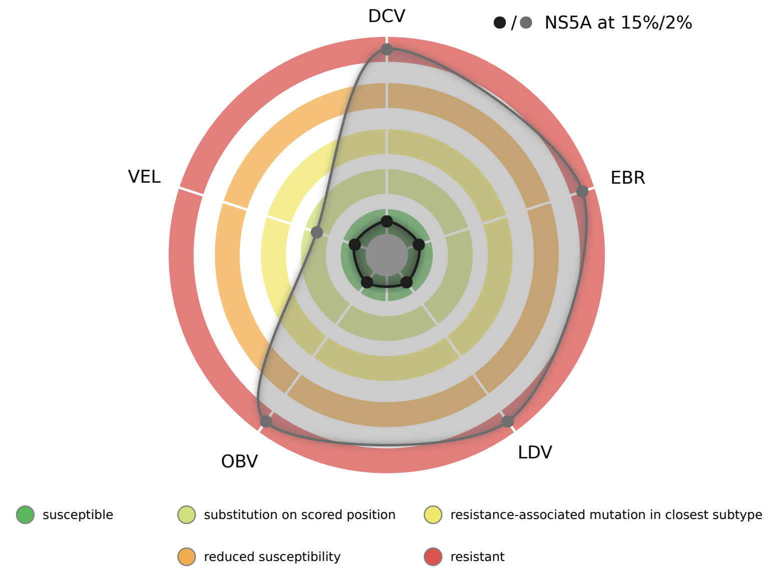
Having shown how genotypic drug resistance testing based on NGS data can improve the selection of antiviral therapies, it is important to note that drug resistance testing by itself cannot guide the selection of all ARVs. Let me provide two examples. First, the antiretroviral drug ABC should be prescribed only to persons that do not carry the HLA variant HLA-B*5701 as this variant is linked to hypersensitivity to the drug, which can be life-threatening[117]. Second, the coreceptor antagonist maraviroc is an HIV entry inhibitor that blocks one of the human coreceptors that is required for HIV cell entry. The next chapter (Chapter 5) presents a method for the genotypic testing of HIV-2 coreceptor usage, which can guide the prescription of maraviroc.

[114] Gärtner et al. 2002; Pfeifer and Lengauer 2012

[115] Posada-Cespedes et al. 2017

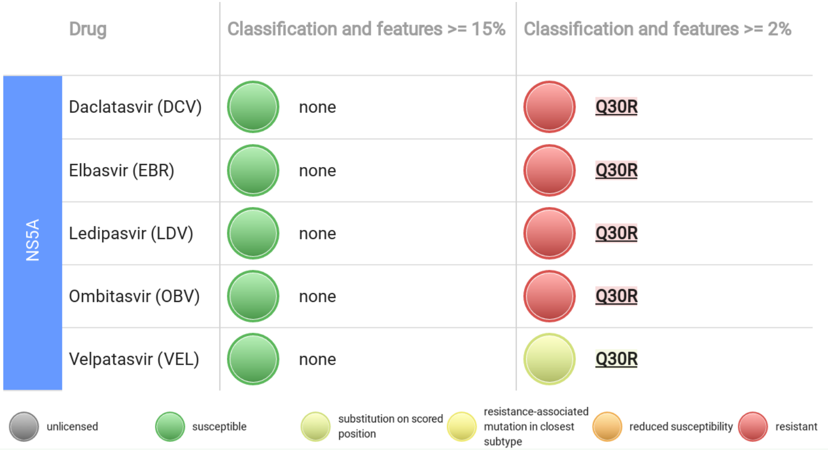
[116] Seifert et al. 2015

[117] Ma et al. 2010



Supplementary Figure S4 from Döring et al. (2018) is licensed under CC-BY 4.0

Figure 4.8: Radar plot of predicted viral drug resistance for the HCV sample from the case study. Each spoke of the plot relates to a DAA and the colored circle sectors indicate different levels of drug resistance. Predicted drug resistance is indicated by two surfaces: one surface showing the drug resistance estimate for the consensus sequence based on a prevalence cutoff at 2% and the other for a prevalence cutoff at 15%.



	Drug	Classification and features >= 15%	Classification and features >= 2%
NS5A	Daclatasvir (DCV)	🟢 none	🔴 Q30R
	Elbasvir (EBR)	🟢 none	🔴 Q30R
	Ledipasvir (LDV)	🟢 none	🔴 Q30R
	Ombitasvir (OBV)	🟢 none	🔴 Q30R
	Velpatasvir (VEL)	🟢 none	🟡 Q30R

● unlicensed 🟢 susceptible 🟡 substitution on scored position 🟡 resistance-associated mutation in closest subtype 🟠 reduced susceptibility 🔴 resistant

Figure 4.9: Table of resistance mutations for the sample from the HCV case study.

5
Predicting HIV-2 Coreceptor Usage

In this chapter, I introduce geno2pheno[coreceptor-hiv2], a freely available web server for identifying the human coreceptor that is used by HIV-2 during cell entry. I developed, validated, and interpreted the model that is used for coreceptor identification. The web server was implemented with support from Achim Büch and Georg Friedrich. Pedro Borrego, Andreia Martins, and Nuno Taveira made experimental measurements that were used for training and validating the model. Ricardo Camacho, Josef Eberle, and Rolf Kaiser provided expert insights. Nico Pfeifer and Thomas Lengauer supervised the project. This chapter augments the publication by Döring et al. (2016) with an overview of established genotypic approaches for coreceptor identification.

> For all these infectious diseases, the goal is to eventually get rid of them. And to do that we need to invent new tools, but nobody was doing that because there was no money to buy on behalf of the poorest, even the existing tools.
>
> Bill Gates, 2016

Studying HIV coreceptor usage provides insights into disease progression and allows for improved clinical decision making. Coreceptor usage (Section 2.3.4) determines viral tropism, which describes the types of immune cells that can be infected by HIV (e.g. macrophages and T cells). Upon initial infection with HIV, the virus typically binds to the CCR5 coreceptor to infect macrophages. In the course of infection, HIV can evolve the ability to use the CXCR4 coreceptor[1], an event referred to as a *coreceptor switch*. Viruses using the CXCR4 coreceptor can infect naive T cells[2] and are associated with increased severity of immunosuppression[3]. Thus, coreceptor testing can be used for monitoring HIV disease progression.

Coreceptor usage influences available treatment options. Coreceptor antagonists such as maraviroc, a drug that blocks the CCR5 coreceptor, prevent viral cell entry by inhibiting the interaction of HIV with its coreceptors[4]. Since maraviroc is only effective against viruses using CCR5, coreceptor usage should be determined before the start of treatment. Besides modulating the effectiveness of coreceptor antagonists, coreceptor usage also modulates the effectiveness

[1] Connor et al. 1997

[2] A naive T cell has not has not yet encountered its cognate antigen.

[3] Connor et al. 1997

[4] Dorr et al. 2005; Baba et al. 1999; Strizki et al. 2005; Nakata et al. 2005; Donzella et al. 1998

of bNAbs. For example, the bNAbs PG9 and PG 16 bind mostly within V2 and V3[5]. Considering that these regions are important determinants of HIV coreceptor usage, it could be shown that both antibodies neutralize R5 variants more effectively[6].

HIV-1 coreceptor usage has been intensively studied[7] and many methods for identifying its coreceptor usage are available[8]. In contrast, much less is known about HIV-2 coreceptor usage[9]; only a single rules-based method for identifying HIV-2 coreceptor usage has been developed[10]. However, this rules-based approach is not available as an online tool. Thus, prescription of maraviroc to HIV-2-infected patients was based on phenotypic coreceptor testing (Section 2.5.1) for a long time. Phenotypic testing is generally more cost- and time-consuming than genotypic testing. This is particularly true for HIV-2 for which no standardized phenotypic assay[11] exists. Therefore, it is desirable to have a web service that accurately determines HIV-2 coreceptor usage based on the composition of the viral sequence.

The goal of this study was to develop a genotypic tool for predicting whether a population of HIV-2 viruses uses only the CCR5 coreceptor or whether the population can use the CXCR4 coreceptor. Since the interaction with cellular coreceptors is facilitated by viral surface glycoproteins, only genomic sequences from Env were considered. Because nucleotide sequences of full-length Env and corresponding phenotypic coreceptor usage are uncommon, I focused on the more prevalent V3[12]. The use of coreceptors other than CCR5 and CXCR4 was not considered because the use of other coreceptors is correlated with CXCR4 usage[13] and cell entry independent of CCR5 and CXCR4 is rarely observed[14].

The significance of this work lies in the contribution of a publicly available tool for identifying HIV-2 coreceptor usage based on the V3 amino-acid sequence alone. Such a tool is useful for three groups of users. Clinicians can use such a tool in order to take informed treatment decisions. Epidemiologists can use the tool to perform large-scale studies that investigate the association between tropism and clinical markers of infection. Last, virologists can use the tool to study the influence of individual V3 amino acids on coreceptor usage.

An overview of established methods for genotypic coreceptor determination is provided in Section 5.1. Section 5.2 explains the steps that are required for the implementation of a statistical model for the prediction of HIV-2 coreceptor usage. The resulting prediction model is validated and interpreted in Section 5.3, in which I also introduce the geno2pheno[coreceptor-hiv2] web server. I conclude this chapter with a discussion of the results in Section 5.4.

[5] Walker et al. 2009

[6] Pfeifer et al. 2014

[7] Clapham and McKnight 2002; Doms and Trono 2000

[8] Pillai et al. 2003; Jensen et al. 2003; Lengauer et al. 2007; Pfeifer and Lengauer 2012; Thielen and Lengauer 2012

[9] Mörner et al. 2002; Blaak et al. 2005

[10] Visseaux et al. 2011

[11] Low et al. 2009

[12] Visseaux et al. 2011

[13] Blaak et al. 2005

[14] Azevedo-Pereira et al. 2003

5.1 Genotypic Methods for Coreceptor Identification

While phenotypic assays can accurately detect a broad range of coreceptor usage patterns, using such assays is expensive and time-consuming. Genotypic testing of coreceptor usage, on the other hand, requires only the genomic sequence of the virus. The advantage of these approaches is that they are fast, inexpensive, and at the same time agree well with the results from phenotypic tests[15]. Similarly to HIV resistance testing, genotypic approaches for determining coreceptor usage can be differentiated into rules-based and statistical methods. In the following, I first give an overview of the genotypic approaches that are used for identifying HIV-1 coreceptor usage (Section 5.1.1) and then come to HIV-2 coreceptor identification (Section 5.1.2).

When reporting the performances of genotypic approaches for coreceptor prediction, it is important to differentiate whether these approaches were evaluated on clonal or bulk data (population-based data). While clonal data reflect the genetic makeup of individual viruses, bulk data result from applying Sanger sequencing to viral populations, which is typically done in clinical applications. Therefore, population-based sequences contain ambiguities at genomic positions for which different variants exist in the viral population.

Determining coreceptor usage on population-based Sanger sequencing data is more challenging than on clonal sequences[16] for two reasons. First, Sanger sequencing does not allow for the detection of minority variants (e.g. variants with an abundance smaller than 10%). This means that approaches based on Sanger sequences cannot consider the coreceptors that are used by minor variants although they may be relevant. Second, genotypic methods need to devise suitable ways for modeling the presence of ambiguities. For example, if multiple amino acids are observed at a single position, *worst-case encodings*[17] consider only those amino acids that are most associated with X4-capability. While such an encoding improves sensitivity, it also potentially increases the rate at which false positives are reported. Thus, in the following, I will refrain from comparing the predictive performance of methods that were evaluated on data of disparate origin (i.e. clonal vs bulk).

5.1.1 Identification of HIV-1 Coreceptor Usage

The first available genotypic approach for the identification of HIV-1 coreceptor usage was established by the 11/25 rule[18]. Under this rule, an HIV-1 strain is considered to be X4-capable if it exhibits a positively charged amino acid at the 11th or 25th position in the V3.

[15] Prosperi et al. 2010; Skrabal et al. 2007; Sing et al. 2007

[16] Sing et al. 2007

[17] The term *worst-case* indicates that the constructed sequence does not necessarily exist *in vivo* but that a sequence most associated with CXCR4 usage is constructed. This is the worst case with regard to the potential usage of maraviroc.

[18] Lengauer et al. 2007

Otherwise, the sequence is considered to be from an R5-tropic virus. Although evaluations on clinical isolates have shown that the 11/25 rule is highly specific (93%), it severely lacks sensitivity (30%)[19]. Therefore, several new approaches for determining coreceptor usage were developed in the 2000s, with one branch of research investigating extended sets of rules and the other studying quantitative models brought forth by statistical learning.

Rules-based Approaches The predictive performance of rules-based approaches was improved considerably by extensions of the 11/25 rule. Examples for such extensions are the 11/24/25 rule and Garrido's rule. The 11/24/25 rule classifies a sequence as X4-capable if a positively charged amino acid is found at the 11th, 24th, or 25th V3 position (sensitivity of 89% at a specificity of 96% on a small data set of mostly clonal sequences)[20]. Garrido's rule classifies a sequence as X4-capable if it fulfills the 11/25 rule or if it exhibits a net charge $\geq +5$ (sensitivity of 80% at a specificity of 79% for population-based sequences)[21].

Statistical Approaches The most frequently used bioinformatic approaches for the identification of HIV-1 coreceptor usage are WetCat[22], WebPSSM[23], and geno2pheno[coreceptor][24], all of which are available as web servers processing the V3 amino-acid sequence. WetCat offers a SVM and two types of decision trees that were trained on 292 V3 amino-acid sequences. WebPSSM, on the other hand, uses position-specific scoring matrices, which are defined according to the observed frequencies of amino acids in the V3 for X4-capable and R5 variants. Based on the scoring matrices, the odds of a position being rather associated with X4-capable or R5-using variants are computed and integrated into a final score.

geno2pheno[coreceptor] relies on a linear support vector machine whose decision values are transformed to FPRs, which provide a measure of confidence for the predictions. Due to its high predictive accuracy and interpretability, geno2pheno[coreceptor] has achieved considerable popularity in the field[25]. The initial version of geno2pheno[coreceptor], which was released in 2004, has been further developed throughout the years. The geno2pheno[454] service, which relies on geno2pheno[coreceptor] and was released in 2010, provides an interface for identifying HIV-1 coreceptor usage based on data from 454 next-generation sequencing[26]. In 2012, geno2pheno[coreceptor] was extended by a model trained on next-generation sequencing data that is also applicable to Sanger sequencing data[27].

[19] Low et al. 2007

[20] Cardozo et al. 2007

[21] Seclén et al. 2010

[22] Pillai et al. 2003

[23] Jensen et al. 2003

[24] Lengauer et al. 2007

[25] Vandekerckhove et al. 2011; Harrigan et al. 2009; Eberle et al. 2014

[26] Thielen and Lengauer 2012

[27] Pfeifer and Lengauer 2012

Performance Comparison Both SVMs and position-specific scoring matrices were shown to constitute a considerable improvement over the established 11/25 rule (sensitivity 59.5%, specificity 92.5%), achieving a sensitivity of 71.9% and 76.4% at a specificity of 92.5%, respectively[28]. Despite their different methodological approaches, current rules-based and statistical approaches for the genotypic identification of HIV-1 coreceptor usage seem to be similarly accurate[29].

5.1.2 *Identification of HIV-2 Coreceptor Usage*

The first genotypic approach for predicting the coreceptor usage of HIV-2 was developed by Visseaux et al. (2011). In their work, they studied the amino-acid sequence of the HIV-2 V3 in order to find associations between coreceptor usage and specific features in the V3 region. They generated a new data set of V3 amino-acid sequences from 53 HIV-2 infected persons. Using these data, they identified nine markers in V3 exhibiting significant associations with coreceptor usage according to Fisher's exact test or the χ^2 test. The four markers that performed best at identifying X4-capable variants (sensitivities greater than 70% and specificities of 100%) were classified as the major determinants of X4-capable HIV-2: any substitution at position 18, V19K/R[30], any insertion after position 24, and a V3 net charge exceeding 6. The remaining five markers of X4-capability were considered to be minor markers of X4-capability: S22A/F/Y, Q23R, I25L/Y, R28K, and R30K.

Using the four major determinants of X4-capable variants, they defined the following rules-based approach. If an HIV-2 V3 exhibits any of the four major markers, it is identified as X4-capable and, otherwise, as R5. Applying this approach on an independent data set consisting of 51 V3 sequences from the literature and the LANL HIV database yielded a sensitivity of 65% at a specificity of 100% for the detection of X4-capable variants.

5.2 *Model Development*

In this section, I describe the data and methods that were used for formulating a new model for HIV-2 coreceptor usage prediction. In Section 5.2.1, I describe the phenotypic measurements that were performed in order to generate an independent validation set. The steps that were necessary for obtaining the learning data set are described in Section 5.2.2. The characteristics of the data set are presented in Section 5.2.3. Section 5.2.4 describes the framework for model selection and validation. Finally, Section 5.2.5 summarizes the work that was done for implementing the geno2pheno[coreceptor-

[28] Sing et al. 2007

[29] Raymond et al. 2008; Seclén et al. 2010

[30] V19K/R denotes that the wild-type amino acid valine (V) at position 19 of the V3 is mutated either to lysine (K) or arginine (R). In general, mutations are indicated by a triplet consisting of the wild-type amino acid (e.g. V), the position (e.g. 19), and the observed substitutions (e.g. K/R).

hiv2] web server.

5.2.1 *Phenotypic Measurements*

The phenotypic measurements that are described in the following paragraphs were carried out by Pedro Borrego, Andreia Martins, and Nuno Taveira.

Reagents HEK293T cells were purchased from American Type Culture Collection (Rockville, MD). The following reagents were provided by the AIDS Research and Reference Reagent Program, National Institutes of Health: TZM-bl cells[31], TAK-779[32], and bicyclam JM-2987, a hydrobromide salt of AMD-3100[33]. The wild-type pROD10 plasmid was a gift from Keith Peden[34]. HEK293T and TZM-bl cells were cultured in complete growth medium consisting of Dulbecco's modified eagle medium (DMEM) supplemented with 10% of fetal bovine serum, 100 U/ml of penicillin-streptomycin, 2 mM of L-glutamine, 1 mM sodium pyruvate, and 1× of MEM non-essential amino acids (Gibco/Invitrogen, USA). All cell cultures were maintained at 37 °C in 5% of CO_2.

Viral Isolates Two new primary isolates, *15PTHSJIG* and *15PTHCEC*, were obtained from HIV-2-infected Portuguese patients by cocultivation with peripheral blood mononuclear cells from seronegative subjects, as described previously[35]. In addition, six new HIV-2 ROD10 mutants were analyzed that contained the following mutations in the V3: H18L, H23Δ + Y24Δ, K29T, H18L + H23Δ + Y24Δ, H18L + K29T, and H18L + H23Δ + Y24Δ + K29T[36]. HIV-2 ROD10 mutants were obtained by transient transfection of HEK293T cells. Transfections were performed with 10 μg of DNA in a 100 mm tissue culture dish, using the jetPrime transfection reagent (Polyplus) according to the instructions of the manufacturer. Cell culture supernatants were collected 48 h post-transfection, filtered, and stored at −80 °C.

The half maximum tissue culture infectious dose (TCID50) of each isolate was determined in a single-round viral infectivity assay using a luciferase reporter assay with TZM-bl cells. First, 10 000 TZM-bl reporter cells were seeded in 96-well tissue culture plates and incubated overnight. On the next day, the growth medium was removed and replaced by 200 μL of fresh growth medium supplemented with 19.7 μg/ml of DEAE-dextran. A total of 100 μL of virus supernatant was added to the first well, from which serial threefold dilutions were prepared in the next wells. The assay was performed in quadruplets. After 48 h, luciferase expression was quantified by measuring luminescence with the Pierce Firefly Luciferase Glow

[31] Platt et al. 1998; Wei et al. 2002; Takeuchi et al. 2008; Derdeyn et al. 2000; Platt et al. 2009

[32] Baba et al. 1999; Dragic et al. 2000

[33] Hendrix et al. 2000; Bridger et al. 1995; De Clercq et al. 1994

[34] Ryan-Graham and Peden 1995

[35] Cavaco-Silva et al. 1998

[36] Martins et al. 2016

Assay Kit (Thermo Fisher, USA) and the Infinite M200 luminometer (TECAN), according to manufacturer's instructions. Control wells containing only target cells and growth medium were used to measure background luminescence. The TCID50 was calculated using the statistical method of Reed and Muench[37].

Phenotypic Determination of Coreceptor Usage CCR5 and CXCR4 coreceptor usage was determined in a single-round viral infectivity assay with TZM-bl cells[38]. First, 10 000 TZM-bl reporter cells were seeded in 96-well tissue culture plates and incubated overnight. On the next day, the growth medium was removed and the cells were incubated for 1 h (at 37 °C in 5% CO_2) with growth medium either in the presence or in the absence of excessive amounts of the CCR5 antagonist TAK-779 (10 μM) and/or of the CXCR4 antagonist AMD3100 (1.2 μM). A fixed amount of virus supernatant, corresponding to 200 TCID50 was added to each well and cells were cultured with a total volume of up to 200 μL of growth medium in the presence of 19.7 μg/mL of DEAE-dextran. After 48 h, luciferase expression was quantified by measuring luminescence with the Pierce Firefly Luciferase Glow Assay Kit (Thermo Fisher, USA) and the Infinite M200 luminometer (TECAN), according to manufacturer's instructions. Control wells containing only target cells and medium were used to measure background luminescence.

A viral population was classified as R5-tropic when viral infectivity was inhibited in the presence of TAK-779 but unaltered in the presence of AMD3100, and, as X4-tropic when infectivity was inhibited in the presence of AMD3100 but unaltered in the presence of TAK-779. When infectivity was completely inhibited only by the simultaneous presence of TAK-779 and AMD3100, the virus population was classified as dual/mixed (D/M) for viral isolates or as R5/X4 tropic for ROD10 mutants.

5.2.2 Data Collection and Processing

Due to the established association between V3 and HIV-2 coreceptor usage[39], I aimed at gathering all available amino-acid sequences of V3 with corresponding measurements of phenotypic coreceptor usage. The majority of samples[40] were retrieved from the LANL HIV database. Further observations were acquired from the literature[41] and my collaborators as described earlier. In total, 314 genotype-phenotype pairs were included in the training data set. Since a fraction of the samples represented full-length Env, a small number of V1 and V2, which have also been reported to influence coreceptor usage[42], were obtained additionally.

[37] Reed and Muench 1938

[38] Borrego et al. 2012; Davis et al. 2008

[39] Shi et al. 2005; Isaka et al. 1999; Kulkarni et al. 2005; Visseaux et al. 2011

[40] Visseaux et al. 2011; Skar et al. 2010; Jadhav et al. 2009; Borrego et al. 2008; de Silva et al. 2012; Barnett et al. 1993; Franchini et al. 1989; Clavel et al. 1986; Breuer et al. 1995; Barnett et al. 1996

[41] Isaka et al. 1999; Kulkarni et al. 2005; Owen et al. 1998

[42] Santos-Costa et al. 2014; Skar et al. 2010

Identifier	#R5	#X4	V3	Decision
DQ870430	21	1	`CKRPGNKTVVPITLMSGLVFHSQPINKRPRQAWC`	R5
NARI-12	5	1	`CKRPGNKTVLPITLMSGLVFHSQPINTRPRQAWC`	R5
GU204945	3	1	`CKRPGNKTVRPITLLSGRRFHSQVYTVNPKQAWC`	Exclude
310248	1	1	`CRRPGNKTVVPITLMSGLVFHSQPINKRPRQAWC`	X4-capable

Table 5.1: Overview of observations with identical V3 amino-acid sequence but discordant annotation of phenotypic coreceptor usage. *#R5* and *#X4* indicate the number of observations that were annotated as *R5* and *X4-capable*, respectively. The identifier and V3 sequence of the X4-capable isolate is shown.

Data Processing Each observation in the data set was labeled either as *R5* or *X4-capable*. Isolates for which CXCR4 usage was reported (X4, R5X4, or dual/mixed) were annotated as *X4-capable* and isolates for which only usage of the CCR5 coreceptor was reported were annotated as *R5*. All of the samples using additional coreceptors (e.g. GPR15 or CXCR6) were capable of using the CXCR4 coreceptor, which is in line with the literature[43]. Observations representing clonal genotype-phenotype pairs from the same patient were merged by forming consensus sequences if their phenotypes agreed.

Duplicate Removal With the intent of constructing a representative training data set, I removed duplicated observations from the V3 data set (i.e. observations sharing the same V3 amino-acid sequence). One of two operations was performed for every set of duplicate observations with discordant phenotypes: either one of the discordant observations was included in the data set or all of the sequences were excluded. The decisions were made by first considering the frequency at which a duplicate observation was phenotyped in a certain way and then, if necessary, considering further evidence from the literature. In the following, I discuss the decisions taken for the four sets of discordant observations that are summarized in Table 5.1.

Each of the samples sharing the same V3 amino-acid sequence as *DQ870430*[44] and NARI-12[45] was phenotyped as an X4-capable variant only once, while a decidedly larger number of samples was phenotyped as R5 (21 and 5 sequences, respectively). Hence, I regarded the X4-capable measurements as outliers and the respective sequences were included with the R5 label. The V3 sequence with the accession GU204945[46] was identified as X4-capable once and as R5 thrice. Due to lacking evidence of actual coreceptor usage, this sequence was removed from the data set.

For the V3 sequence with the identifier *310248*, one study reported the use of CCR5 and another the use of CXCR4. The sequence had been identified in an X4-capable isolate by Owen et al. (1998), but also in an R5 isolate with the same V3 sequence except for an R/K ambiguity at position 27[47]. Interestingly, the R5 isolate showed a marginal signal for the CXCR4 coreceptor, which was discarded because the signal was smaller than 5% of the signal for CCR5 us-

[43] Blaak et al. 2005; van Der Ende et al. 2000; Owen et al. 1998

[44] Visseaux et al. 2011; Kulkarni et al. 2005; Skar et al. 2010; Jadhav et al. 2009; Borrego et al. 2008; de Silva et al. 2012

[45] Visseaux et al. 2011; Kulkarni et al. 2005; Owen et al. 1998; Jadhav et al. 2009

[46] Skar et al. 2010

[47] Visseaux et al. 2011

age. Further evidence pointing towards the usage of CXCR4 was presented by Owen et al. (1998) who reported a minor induction of syncytia for their isolate. Additionally, applying a CXCR4 antagonist to cells lacking the CCR5 coreceptor revealed a reduction in infectivity between 40% and 90% for this strain, which was enough evidence to include this sequence in the data set with the *X4-capable* label.

Profile Alignments To allow for the alignment of a single V3 amino-acid sequence with respect to the composition of the overall HIV-2 V3, I used profile alignments. While standard pairwise alignment algorithms[48] are based on a single reference sequence, profile alignments are better suited for divergent sequences such as the V3 from HIV-2. Here, alignment scores were computed under consideration of both, the frequency of amino-acid substitutions given by the alignment profile and the value from an amino-acid substitution matrix[49].

In order to construct an amino-acid profile for the HIV-2 V3, I retrieved all available amino-acid sequences of the HIV-2 envelope region from the LANL HIV database and selected the V3 region through pattern matching in the following manner. If a sequence exhibited the highly conserved V3 start motif (`CKRP` or `CRRP`) as well as the end motif (`QAWC`), the corresponding subsequence was selected. In cases where either only the start or end motif could be found, a search for a substring of the missing motif was conducted and the corresponding subsequence was selected if a substring of the missing motif could be found. The 1 979 retrieved V3 amino-acid sequences were aligned using the well-established tool ClustalW (version 2.1 with the *accurate* switch and otherwise default parameters)[50]. The V3 profile was then constructed by determining the frequency of each amino acid for every alignment position.

Feature Encoding Let $\mathcal{B}$ indicate the set of 20 amino acids augmented with the gap character, –. The feature vector $x_i \in \mathbb{R}^p$ of observation $i \in \mathbb{N}$ is determined from the amino-acid sequences of the considered HIV-2 regions in the following way. Let s_i indicate the i-th position of an aligned sequence and let s_{ij} indicate the set of disambiguated amino acids (Table A.1) occurring at position j in the i-th input sequence. We use $x_{ij}[c]$ to denote whether the character $c \in \mathcal{B}$ appears at position j in the sequence of the i-th observation. For each position $j \in \{1, 2, \ldots, |s_i|\}$ in an aligned sequence, s_i, we set

$$x_{ij}[c] = \begin{cases} \frac{1}{|s_{ij}|}, & \text{for } c \in s_{ij} \\ 0, & \text{for } c \notin s_{ij} \end{cases} .$$

Data set labels were encoded by setting $y_i = -1$ for observations labeled as *X4-capable* and $y_i = 1$ for samples labeled as *R5*.[51]

[48] Smith and Waterman 1981

[49] Gonnet et al. 1992

[50] Higgins et al. 1996

Phenotype	Phylogenetic Group	Frequency
R5	A	61 (48.4%)
X4-capable	A	46 (36.5%)
R5	B	12 (9.5%)
X4-capable	B	5 (3.9%)
X4-capable	D	1 (0.08%)
R5	Unknown	1 (0.08%)

Table 5.2: Distribution of phenotypically determined coreceptor usage and phylogenetic groups of HIV-2 in the V3 data set.

5.2.3 *Characteristics of the V3 Data Set*

The distribution of phenotypic coreceptor annotations and HIV-2 phylogenetic groups for the V3 data set are shown in Table 5.2. After filtering the V3 data set, 126 genotype-phenotype pairs remained of which 74 (58.7%) were labeled as R5 and 52 (41.3%) as X4-capable. The samples in the data set originate from diverse geographic regions. In total, 87 (69%) samples were collected in Europe, of which 42 (48.3%) come from France, 33 (37.9%) from Portugal, and 12 (13.8%) from Sweden. All of the 10 (10.3%) Asian samples originate from India. Of the 24 (19%) West African samples, 15 (60%) were collected in Guinea-Bissau, 5 (20.8%) in Ivory Coast, 2 (8.3%) in Gambia, and 2 (8.3%) in Senegal.

Most isolates in the data set (84.9%) had been genotyped as HIV-2 group A. Only a minority of samples (13.5%) had been identified as group B. The remaining samples (1.6%) either had been identified as group D or had not been genotyped. The group distribution of the samples in the data set reflects the global distribution of HIV-2 groups[52].

5.2.4 *Model Selection and Validation*

Since SVMs (Section 3.4.2) that use the amino-acid sequence of the V3 region as an input have already been successfully applied for identifying the coreceptor usage of HIV-1 (Section 5.1), I also considered the use of SVMs in the HIV-2 setting. For the coreceptor prediction problem, SVMs find a vector of coefficients and an intercept that define a hyperplane maximizing the margin between observations from the two classes *X4-capable* and *R5*.

As kernel functions, I considered linear, RBF, polynomial, and edit kernels[53] (Section 3.4.2). To simplify model selection via grid search, I opted for the ν-SVM formulation. This formulation has the benefit that its regularization parameter ν is bounded by the interval $[0, 1]$, while the original formulation uses the soft-margin parameter $C \in \mathbb{R}$. The grid search was conducted with ν between 0.1 and 0.4. Higher values were not considered as these led to infeasible optimization

[52] Groups A and B are most prevalent phylogenetic groups. Group A strains cause the majority of infections (Chen et al., 1997; Gao et al., 1994; Marlink, 1996).

[53] Li and Jiang 2005

problems. Additionally, the following hyperparameters were considered for the individual kernels: $\sigma \in \{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$ for the RBF kernel, $d \in \{2, 3, 4\}$ for the polynomial kernel, and

$\gamma \in \{1 \times 10^{-3}, 2.5 \times 10^{-3}, 5 \times 10^{-3}, 7.5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}\}$ as well as PAM $\in \{30, 70, 250\}$ for the edit kernel.

The model parameters were tuned by maximizing the AUC (Section 3.3.3) in ten runs of tenfold CV (Section 3.2.5). To determine the expected performance of the models taking into account the model selection procedure, I performed tenfold NCV. To compare the performance of the rules-based approach by Visseaux et al. (2011) with the statistical models, it was necessary to define a data set containing only those observations that had not been used for identifying discriminatory features by Visseaux et al. in order to prevent overestimating the performance of their approach. Thus, I selected an appropriate subset of the available data ($N = 84$), which is referred to as the *comparison data set* in the following.

To evaluate whether there exists a significant difference between the rules-based approach and SVMs, I applied McNemar's test (Section 3.6.2). The result of the test indicates whether the predictions from the two approaches have differential rates of agreement with phenotypically determined coreceptor usage.

5.2.5 Development of geno2pheno[coreceptor-hiv2]

The following paragraphs describe how the geno2pheno[coreceptor-hiv2] web server was developed. I first explain how the interpretability of the SVM can be improved by transforming decision values to FPRs and visualizing the SVM model weights. Thereafter, I describe the feature encoding that is used by the web service, which allows for increased sensitivities for low-quality sequences and clinical isolates.

Transforming Decision Values

SVM decision values were transformed to probabilities (Wu et al., 2004) indicating the likelihood that a V3 amino-acid sequence originates from an X4-capable virus. These probabilities are called *X4-probabilities* in the following. Since these probabilities do not offer insights into the specificity of coreceptor prediction, X4-probabilities were transformed to FPRs, which have already been established for the quantification of HIV-1 coreceptor usage[54]. Here, the FPR indicates the estimated rate at which *R5* samples would be falsely predicted as *X4-capable* if the SVM estimate were used as a cutoff for classifying observations into the two classes.

To transform X4-probabilities to FPRs, I constructed a mapping from predicted X4-probabilities to FPRs during the training stage.

[54] Lengauer et al. 2007

Each predicted X4-probability was used as a cutoff for classifying the samples once: All samples with X4-probabilities below the cutoff were assigned the label *R5* and all samples with X4-probabilities greater or equal to the cutoff were assigned the label *X4-capable*. This cutoff-dependent class assignment in combination with the phenotypic labels for each observation yielded a 2 × 2 contingency table indicating FPs and TNs, from which I could compute the FPR as $FPR = \frac{FP}{FP+TN}$. Low FPRs indicate confident predictions of X4-capable variants, while high FPRs designate R5-tropic viruses.

Choice of Reference Sequence

To display alignments of query sequences in the user interface of the web service, the well-described HIV-2 reference strain M33262 (Mac239)[55] is used.

Visualizing the Impact of Individual Amino Acids

Although SVMs are often considered as black boxes, their interpretability depends on the used kernel functions. Linear SVMs are particularly easy to interpret because a fitted linear SVM can be compactly described by the vector $\alpha^* \in \mathbb{R}^{|S|}$ whose length, $|S|$, is defined by the number of support vectors. The vector's entries, $\alpha_i^* = \hat{\alpha}_i y_i$, are defined as the product of the weight, $\hat{\alpha}_i$, and the corresponding outcome y_i. Let $X^* \in \mathbb{R}^{|S| \times p}$ be the feature matrix containing only the support vectors x^*. Then the impact of individual amino acids according to the model can be determined via $\beta = ({\alpha^*}^T X^*)^T \in \mathbb{R}^p$. Given a new input sequence with the feature vector $x \in \mathbb{R}^p$, the contribution of feature $j \in \{1, \ldots, p\}$ is given by $b(j) = x_j \times \beta_j \in \mathbb{R}$. This quantity can be used for visualizing how individual amino acids in a query sequence impact coreceptor usage.

Modified Feature Encoding

When applying the SVM model to input sequences, the query sequence is modified in two ways in order to improve predictive performance. The first modification concerns gaps in the sequence and the second relates to ambiguous positions. Note that since the labels for training the SVM were encoded by -1 for *X4-capable* and 1 for *R5*, positive coefficients designate features associated with R5 variants and negative coefficients designate features associated with X4-capable variants.

Gap Replacement Errors during sequencing or problems with the alignment can lead to the introduction of gaps in the V3 sequence

[55] Regier and Desrosiers 1990; Chen et al. 1997; Kestler et al. 1990

that have no functional meaning. Although functionally irrelevant, gaps with absolute weights close to 0 can still influence the outcome because their absence reduces the decision value, thereby pulling the prediction towards *X4-capable*. Therefore, the following approach is used to deal with this problem. Let $\beta_j(c)$ be the coefficient that corresponds to the character c at sequence position j and let $\epsilon = 0.01$. For every position j where c represents a gap, the model weight associated with the gap, $\beta_j(c)$, is considered. If $|\beta_j(c)| < \epsilon$, then the gap is replaced with the encoded consensus amino acid c' from position i of the V3 alignment profile by setting $x_j[c] = c'$ before predicting coreceptor usage. Otherwise, if $|\beta_j(c)| \geq \epsilon$, the gap is deemed to be functionally relevant and is retained.

Worst-Case Encoding Ambiguous positions typically indicate the presence of multiple viral variants within the same host. These variants may use different coreceptors for cell entry. Therefore ambiguous positions can encompass amino acids representative of both, R5 and X4-capable viruses. To be more sensitive towards X4-capable variants, the encoding is changed in the following way. For every ambiguous sequence position j with observed amino acids s_j, the position is adjusted to $s_j = \arg\min_{c \in s_j} \beta_j(c)$.

The fact that this worst-case encoding may give rise to sequences that might not exist *in vivo* is only a minor concern due to the following reasoning. Assume that a viral population consists of an R5- and an X4-capable quasispecies, which means that the prediction should be *X4-capable*. In this case, every ambiguous position should contain an amino acid representing the X4-capable variant such that for every ambiguous position j, there exists $\beta_j(c) \leq 0$. Selecting the observed amino acid whose weight contributes most to X4-capability means choosing the character c obtaining the most negative weight, $\beta_j(c)$. Consequently, the decreased decision value enhances the prediction of X4-capable variants. The same logic can be applied to two distinct X4-capable quasispecies.

Assume now that there exist two quasispecies that use only the CCR5-coreceptor. In this case, the prediction should be R5 and the weights of ambiguous positions should be positive, because no amino acids associated with X4-capability should be observable. Hence, the worst-case choice would result in $\min \beta_j(c) \geq 0$ for all characters c at every ambiguous position j. This, however, does not enhance the prediction of X4-capable variants and thus does not degrade the prediction of R5 when the decision boundary is set to 0. Even for decision boundaries at values above zero, the prediction would not be influenced much because it is likely that there is a sufficient number of non-ambiguous positions with positive weights for sequences from R5 viruses.

5.3 Results

To generate statistical models capable of predicting HIV-2 coreceptor usage, a data set of 126 pairs of HIV-2 genomic amino-acid sequences and phenotypic coreceptor usage annotations (either *R5* or *X4-capable*) was gathered. Based on this data set, SVMs with various kernel functions were trained and validated using the amino-acid sequences of either the V1, V2, V3, or all three regions.

Section 5.3.1 describes the selection of the linear SVM trained on V3 amino-acid sequences that was used for all subsequent analyses. Section 5.3.2 analyzes the predictive performance of the approach from Visseaux et al. (2011) and Section 5.3.3 compares the rules-based approach with the linear SVM. The features in the V3 imparting the X4-capable phenotype are investigated in Section 5.3.4. A cutoff for separating *R5* and *X4-capable* observations is presented in Section 5.3.5. Finally, the geno2pheno[coreceptor-hiv2] web service and its validation are presented in Sections 5.3.6 and 5.3.7, respectively.

5.3.1 Model Selection and Performance

Model Selection In the following, I report the best performing models as determined by 10-fold CV for every considered genomic region of HIV-2. Linear SVMs based on V1 and V2 (N = 62) achieved AUCs of 0.84 and 0.65, respectively. For the V3 region (N = 126), an SVM with a linear kernel performed best (AUC of 0.95). SVMs based on other kernel functions achieved similarly high performances except for the SVMs based on the edit kernel, which had distinctly smaller AUCs (Table 5.3).

The performance of SVMs trained on a concatenation of the amino-acid sequences of all three variable regions V1/V2/V3 (N = 62 samples) was also investigated. The best model resulting from the use of all three regions performed worse (AUC of 0.89) than the best model based on the V3 alone (AUC of 0.95). Models based on V1/V2 were not further investigated for the following reasons. First, models incorporating information from the V1/V2 region performed worse than models based on V3 alone. Second, V1/V2 are more variable than V3, which may cause problems with the alignment. Third, data for V1/V2 is limited, while genomic data from the V3 is more commonly available. Thus, the linear ν-SVM ($\nu = 0.3$) trained on 126 V3 amino-acid sequences was selected for all further applications. In the following, this SVM is referred to as *the linear SVM*.

CV Run	RBF	Linear	Polynomial	Edit Kernel
1	0.9475	0.9459	0.941	0.8629
2	0.9509	0.9506	0.9452	0.851
3	0.9504	0.9579	0.9444	0.8655
4	0.9449	0.947	0.9379	0.8634
5	0.9472	0.9467	0.9413	0.8744
6	0.9467	0.9467	0.9457	0.8689
7	0.9532	0.9535	0.9475	0.8377
8	0.9522	0.9532	0.9306	0.8623
9	0.9524	0.9524	0.9478	0.9012
10	0.9441	0.9431	0.9384	0.8672
μ	0.949	0.9497	0.942	0.8654
σ	0.0033	0.0045	0.0053	0.0162

Table 5.3: AUCs per run of tenfold cross validation for SVMs trained on V3 amino-acid sequences. For every kernel function, only the best performing parameter combination is shown. All classifiers performed best with the SVM parameter $\nu = 0.3$. The hyperparameters for the kernel functions resulting in the best performance were $\sigma = 1 \times 10^{-3}$ for the RBF kernel, a degree of 2 for the polynomial kernel, and $\gamma = 5 \times 10^{-3}$ with a PAM 70 matrix for the edit kernel.

Model Performance The AUCs reported in Table 5.3 are optimistic because the best performing models were chosen *a posteriori*. In order to obtain a more accurate estimate of the prediction error, NCV was used. In the 10 inner runs of NCV, SVMs using a linear kernel were chosen seven times and SVMs using an RBF kernel were chosen three times according to their AUCs. The AUC of tenfold NCV was 0.88 (sensitivity of 76.9% and specificity of 97.3%).

5.3.2 *Validation of the Rules-Based Approach*

In order to validate the major markers of HIV-2 coreceptor usage that were established by Visseaux et al. (2011), profile alignments were computed for all V3 amino-acid sequences in the comparison data set. The alignment positions were enumerated according to the HIV-2 reference M33262. Of the 34 X4-capable sequences in the comparison data set, only 5 (14.7%) samples did not have any major marker, 2 (5.9%) had a single marker, 2 (5.9%) had two markers, 4 (11.8%) had three markers, and 21 (61.8%) had four markers. Interestingly, the five X4-capable sequences without any markers for CXCR4 usage (accession numbers/isolate identifiers: DQ213035[56], GU204944[57], consensus V3 from clones JX219591-JX219598, GB87[58], and 310248[59]) could be identified as X4-capable neither by the rules-based method nor by the linear SVM.

The performance of the rules-based approach from Visseaux et al. (2011) was evaluated on the comparison set by requiring different numbers of major markers of X4-capability (either 1, 2, 3, or 4). The balanced accuracy decreased when the required number of major markers was increased (balanced accuracies 0.89, 0.88, 0.85, and 0.81, respectively). This confirms that requiring one major marker for

[56] Shi et al. 2005

[57] Skar et al. 2010

[58] Owen et al. 1998

[59] Owen et al. 1998

Number of Rules	Sensitivity	Specificity	Balanced Accuracy
1	0.85	0.94	0.89
2	0.79	0.96	0.88
3	0.74	0.96	0.85
4	0.62	1	0.81

Table 5.4: Performance of the approach from Visseaux et al. on the comparison data set. The column *Number of Rules* indicates the number of required major rules for calling X4-capability.

Rule	Sensitivity	Specificity	Accuracy	P-value
L18X	0.79	0.96	0.88	2.3×10^{-13} (∗)
Insertion after position 24	0.74	1	0.87	3.4×10^{-14} (∗)
Net charge $> +6$	0.77	0.96	0.86	6.8×10^{-11} (∗)
V19K/R	0.74	0.96	0.85	8.7×10^{-11} (∗)
R28K	0.5	0.96	0.73	8.9×10^{-7} (∗)
Q23R	0.29	1	0.65	4.7×10^{-5} (∗)
R30K	0.47	0.7	0.57	1.7×10^{-1}
S22A/F/Y	0.15	1	0.59	9×10^{-3} (∗)
I25L/Y	0.08	0.97	0.53	4.7×10^{-1}

Table 5.5: Performance of individual rules from Visseaux et al. on the comparison data set. The column *Accuracy* provides the balanced accuracy. The major discriminatory features are highlighted in bold. Significant p-values as identified by Fisher's exact test are shown by asterisks.

X4-capability provides the best trade-off between sensitivity and specificity. Note, however, that the presence of additional markers could be used to corroborate predictions since the requirement of a larger number of major markers increases specificity (Table 5.4). For example, requiring four major markers results in a specificity of 100% at a sensitivity of 62%.

To determine the predictive performance of individual markers of X4-capability, I applied a two-sided Fisher's exact test on the confusion matrices resulting from the application of individual rules. P-values were corrected for multiple testing using the Benjamini-Hochberg procedure (Section 3.6.5), for which the FDR was set to 5%. All established discriminatory features except for R30K and I25L/Y were significant predictors of X4-capability on the comparison data set (Table 5.5). The four major discriminatory markers of X4-capability (accuracies of 85%–88%) outperformed the other markers considerably with respect to predictive performance (accuracies 53%–73%).

5.3.3 *Comparison of Predictions from SVMs and Rules*

For comparing the predictive performances of SVMs and the rules-based method, both approaches were evaluated on the comparison data set. The rules-based method from Visseaux et al. (2011), which requires just a single major determinant to predict X4-capability, achieved a sensitivity of 85.3% at a specificity of 94% (balanced accuracy 89.6%). Tenfold NCV of SVMs resulted in a sensitivity of 73.5% at a specificity of 96% (balanced accuracy 84.7%). Despite the slightly higher performance of the rules-based method on this data set, the predictive performances were not found to be significantly different according to McNemar's test (p-value 0.37).

5.3.4 Learning Discriminatory Features from SVMs

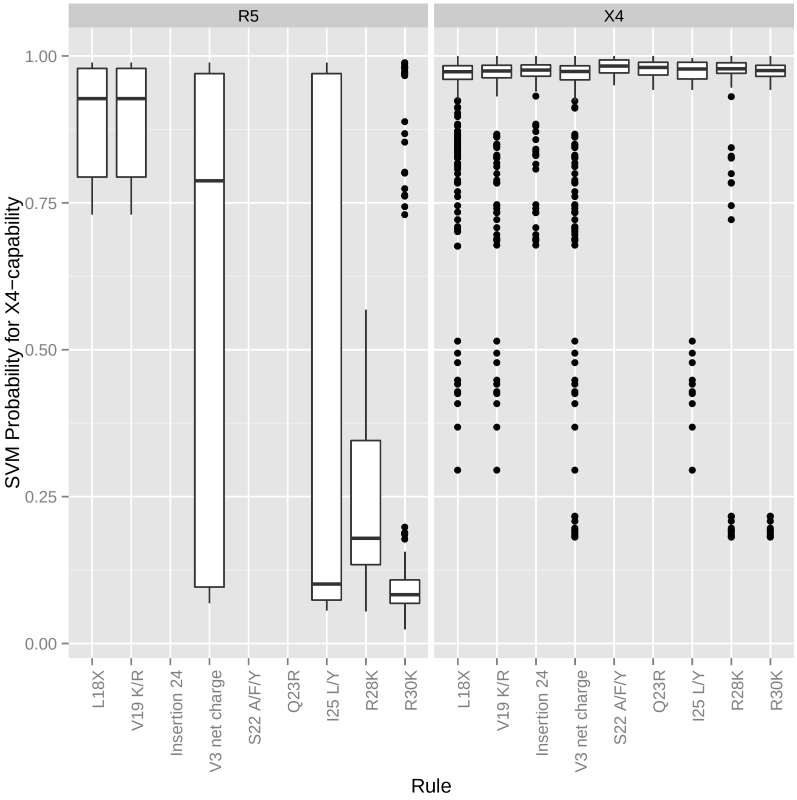


Figure 1 from Döring et al. (2016) is licensed under CC-BY 4.0

Figure 5.1: X4-probabilities predicted by the linear SVM for V3 amino-acid sequences exhibiting the established discriminatory features indicative of X4-capability listed on the x-axis. The left-hand panel shows the predicted X4-probabilities for sequences labeled as R5, while the right-hand panel shows the predicted X4-probabilities for sequences labeled as X4-capable. The bottom line of a box indicates the 1st quartile (Q1) of predicted X4-probabilities, the bar inside the box indicates the median, and the top line indicates the 3rd quartile (Q3). The whiskers extending from a box indicate predicted X4-probabilities that lie within 1.5 × IQR (interquartile range, IQR = Q3 - Q1). Outlier values that are not within the whisker region are shown as dots. Segments on the x-axis without a box reflect the absence of the corresponding feature and phenotype.

I investigated how well the linear SVM reproduces the established markers for X4-capability by analyzing the SVM estimates for sequences exhibiting specific markers of X4-capability. Figure 5.1 shows that the SVM assigns high X4-probabilities to sequences from X4-capable viruses exhibiting the established X4-markers. However, the SVM also assigns high X4-probabilities to the rare *R5* samples carrying markers of X4-capability.

Table 5.6 shows the SVM features with the greatest impact on the prediction. All major markers of X4-capability are among the features contributing 75% of the total absolute model weight. The model also considers additional features that have not been described previously. To determine whether these features are significant, I applied Fisher's exact test on the 2 × 2 matrix resulting from using these features as rules for predicting X4-capability. The multiple-testing corrected

V3 Position	R5 feature	X4 feature	R5 weights	X4 weights
18	L	H, Q, F, M	0.69	-0.23, -0.15, -0.12, -0.1
Ins 24	-	I, V	0.45	-0.22, -0.21
19	I	R, K, V	0.19	-0.25, -0.23, -0.19
Ins 22	-	H, Y	0.36	-0.18, -0.18
24	P	NA	0.17	NA
23	Q	R	0.14	-0.14
27	Q	K	0.09	-0.12
13	T	R	0.11	-0.07
26	NA	N	NA	-0.09
10	A	K	0.09	-0.07
14	I	L	0.08	-0.08
22	S	NA	0.08	NA
15	A	G	0.08	-0.07
8	K	S	0.07	-0.07

Table 5.6: Features of the linear SVM contributing 75% of the total absolute model weight. Positions of discriminatory features that were not described previously are shown in bold. Entries prefixed with *Ins* indicate the presence of any insertion after a certain V3 position. Entries annotated with *NA* indicate that the corresponding feature did not contribute to 75% of the model weight.

p-values (as above, FDR 5%) indicate that the previously undescribed substitutions 8S, 15G, and 27K are indeed predictive of X4-capability.

5.3.5 *Separation of R5 and X4-Capable Variants*

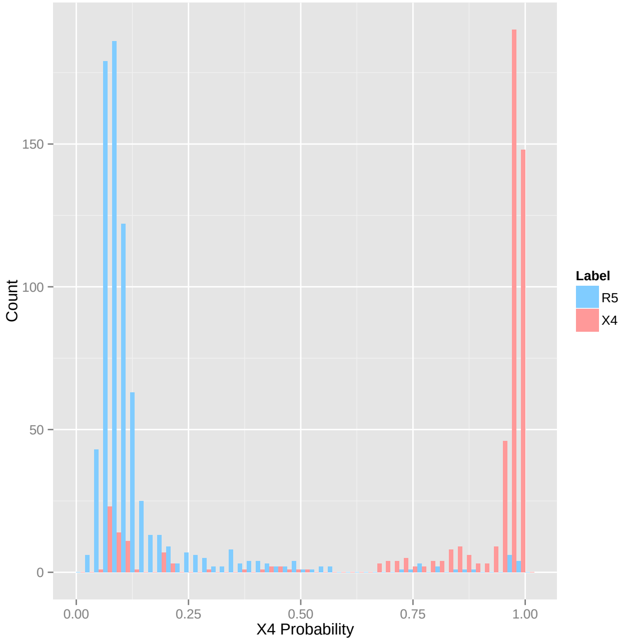


Figure 5.2: Distribution of X4-probabilities estimated by the linear SVM. Blue bars indicate sequences labeled as R5, while red bars indicate sequences labeled as X4-capable.

The distribution of predicted X4-probabilities was determined by applying the linear SVM on the V3 data set using ten runs of tenfold CV. Figure 5.2 shows that X4-probabilities separate the V3 amino-acid sequences from R5 and X4-capable viruses well. Note that the region of low X4-probabilities is interspersed with a greater number of observations labeled as *X4-capable* than there are observations labeled *R5* in the high X4-probability region. This indicates that the SVM more easily obtains a high specificity than a high sensitivity.

In order to dichotomize the SVM estimates for classifying samples into one of the two classes *R5* and *X4-capable*, I performed *K*-means clustering (Section 3.5.1) on the X4-probabilities. A suitable number of clusters was selected using the elbow test on the within sum of squares[60] and X4-probabilities were subsequently clustered using the *K*-means algorithm. From the cluster representing X4-capable viruses, the minimal X4-probability (53.4%) was selected and the corresponding FPR was determined (3.4%). The recommended cutoff for HIV-2 coreceptor prediction was set to an FPR of 5% for two reasons. First, in case that the model's TPR (as determined from the population sample) overestimates the TPR that would be observed for the population, it would be worthwhile to increase the TPR at the cost of an increased FPR. This is because incorrectly classifying an *R5* sample as *X4-Capable* has little consequences (e.g. maraviroc will not be prescribed), while incorrectly classifying an *X4-Capable* sample as *R5* may have severe consequences (e.g. treatment failure because maraviroc is not effective). Second, a value of 5% is more memorable than a value of 3.4%.

5.3.6 The geno2pheno[coreceptor-hiv2] Web Server

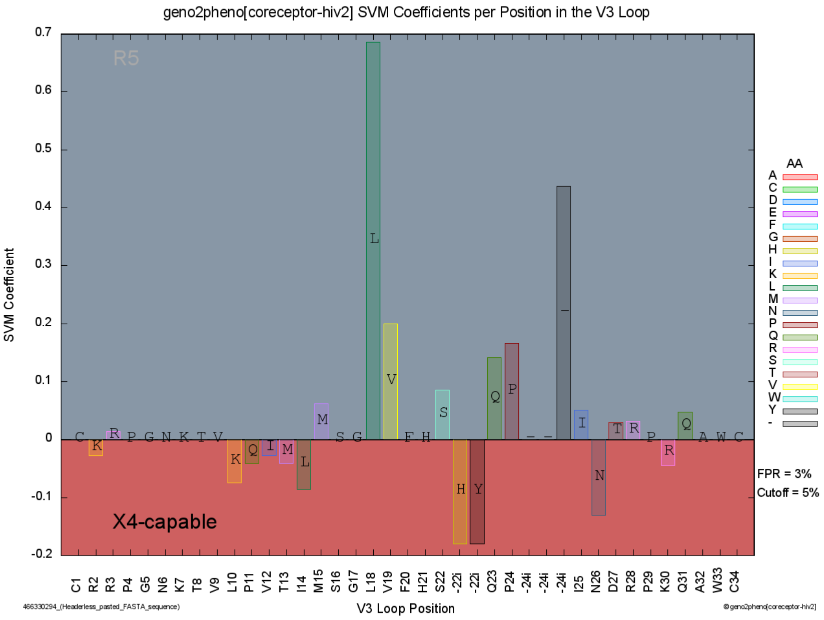


Figure 2 from Döring et al. (2016) is licensed under CC-BY 4.0

Figure 5.3: Visualization of the model coefficients for the V3 amino-acid sequence of the mutant ROD10 isolate (H18L + K29T). Amino acids with positive coefficients are associated with R5-tropic viruses, while negative coefficients are associated with X4-capable variants. The legend on the right indicates the color-coded amino acids and gives the FPR of the prediction. Because the predicted FPR is below the selected cutoff at 5%, the sequence is predicted to be X4-capable, which is indicated by the dark color of the X4-capable label in the bottom left corner. The labels of the x-axis refer to the positions and amino acids of the HIV-2 reference strain M33262.

[60] Tibshirani et al. 2001

Implementation The linear SVM for the prediction of HIV-2 coreceptor usage is available via the geno2pheno[coreceptor-hiv2] web service . The backend of the service was implemented in C++ with LibSVM[61] and the frontend was developed in PHP. Based on the approach presented in Section 5.2.5, a visualization for the feature weights associated with a query sequence was implemented with gnuplot.

Usage The geno2pheno[coreceptor-hiv2] web service requires the selection of a cutoff on the FPR. The selected cutoff specifies the trade-off between sensitivity and specificity. The user can either simply use the recommended cutoff at 5%, which maximizes the overall predictive performance, or customize the cutoff depending on the application scenario. Finally, one or multiple nucleotide/amino-acid sequences in FASTA format (at most 500) containing the V3 of an HIV-2 sample can be uploaded. Once the input data have been provided, each input sequence is aligned to a profile of HIV-2 V3 amino-acid sequences (Section 5.2.2) and the FPR is estimated using the linear SVM (Section 5.3.1). Once the predictions are available, the user can obtain an alignment of the input V3 sequence relative to the reference M33262 in order to investigate relevant mutations.

To facilitate the interpretation of the results, the web service does not only provide the classification (either *X4-capable* or *R5*) according to the selected FPR cutoff, but also the FPR itself. Additionally, the web service generates a visualization of the weights assigned to individual amino acids in the input sequences for a more intuitive way of interpreting the results (Figure 5.3). For the application of geno2pheno[coreceptor-hiv2] in clinical settings, a PDF report summarizing the results of an analysis is available. To facilitate the application of geno2pheno[coreceptor-hiv2] in research settings, the results of batch computations can be retrieved in CSV format.

5.3.7 *Validation of the Web Service*

The implementation of the geno2pheno[coreceptor-hiv2] web service was validated using an independent test set containing eight additional V3 samples (Section 5.2.1). These samples had not been used for training the predictive model. Predictions from geno2pheno[coreceptor-hiv2] were compared to the phenotypically measured coreceptor usages for these samples. Using the recommended FPR cutoff at 5%, all samples were classified correctly by geno2pheno[coreceptor-hiv2], while the approach from Visseaux et al. (2011) misclassified two of the samples (Table 5.7). The R5-isolate ROD10 (H18L + H23Δ + Y24Δ) was incorrectly classified as

[61] Chang and Lin 2011

Isolate	FPR	Major markers	Rules	SVM	Phenotype
ROD10 (wild type)	0.01	L18X, V3 net charge > 6	X4	X4	X4
ROD10 (K29T)	0.01	L18X	X4	X4	X4
ROD10 (H18L)	0.03	V3 net charge > 6	X4	X4	X4
ROD10 (H23Δ + Y24Δ)	0.01	L18X	X4	X4	X4
ROD10 (H18L + K29T)	0.03	NA	R5	X4	X4
ROD10 (H18L + H23Δ + Y24Δ)	0.11	V3 net charge > 6	X4	R5	R5
ROD10 (H18L + H23Δ + Y24Δ + K29T)	0.15	NA	R5	R5	R5
15PTHSJIG	0.36	NA	R5	R5	R5
15PTHCEC	0.01	L18X, V19K/R, Ins 24, V3 net charge > 6	X4	X4	X4

Table 5.7: Results from validating geno2pheno[coreceptor-hiv2] using additional V3 sequences. The *FPR* column indicates the FPR that is estimated by geno2pheno[coreceptor-hiv2] and the *SVM* column indicates the corresponding prediction according to the 5% FPR cutoff. The *Major markers* column shows the major markers of X4-capability that were detected according to the approach from Visseaux et al. and the *Rules* column shows the corresponding prediction. The *Phenotype* column provides the gold standard according to phenotypic measurements of coreceptor usage. Note that table entries annotated as *X4* indicate X4-capable variants. Correct classifications are indicated by blue font, while incorrect classifications are shown in red font. ROD10 refers to the HIV-2 group A reference strain, which uses both CCR5 and CXCR4. Mutations from the ROD10 wild-type sequence are indicated in brackets, where Δ indicates deletions.

X4-capable due to its net charge of +7 and the X4-capable sequence ROD10 (H18L + K29T) was misclassified as *R5* because it did not exhibit any of the four major markers for X4-capability.

5.4 *Discussion*

This chapter investigated the prediction of HIV-2 coreceptor usage from amino-acid segments of the viral surface glycoprotein and corresponding phenotypic measurements. The results suggest that specific V3 amino acids are the key markers of HIV-2 coreceptor usage. In fact, three novel markers that are significantly associated with X4-capability, namely 8S, 15G, and 27K were identified (Table 5.6). It could also be shown that geno2pheno[coreceptor-hiv2], which uses a linear SVM to predict coreceptor usage from V3 amino-acid sequences, is interpretable and accurate.

Analyzing the predictive performance of SVMs based on various kernel functions revealed that linear kernel functions are well suited for the prediction of HIV-2 coreceptor usage. Kernel functions capturing higher-order interactions do not seem to offer additional benefits in this prediction scenario. This finding confirms that HIV-2 coreceptor is largely based on individual amino-acid mutations in V3 rather than on interdependent substitutions of amino acids as in HIV-1[62]. The hypothesized open structure of the HIV-2 V3, which might reduce the role of interactions among the amino acids in the V3[63], would support this result. However, determining and analyzing the structure of gp125 with intact and ordered V3 would be crucial for confirming the independence of positions by elucidating the accessibility of the V3[64].

Since geno2pheno[coreceptor-hiv2] is based on an SVM, it considers all V3 positions when predicting coreceptor usage. Rules-based systems, on the other hand, use only a pre-selected set of discriminatory features from the V3 to identify coreceptor usage. Investigating the model coefficients of the X4-isolate ROD10 (H18L + K29T), which are shown in Figure 5.3, highlights the advantage of statistical over rules-based methods. In this case, the combination of multiple negative weights associated with the features R2K, P11K, V12K, T13M, I14L, insertions after position 22, and N26N resulted in the prediction of X4-capability rather than fulfilling any of the established rules. The superior performance of the approach from geno2pheno[coreceptor-hiv2] over the rules-based approach from Visseaux et al. (2011) was shown on a small test set (Table 5.7). Recently, further phenotypic measurements by Cerejo et al. (2018) showed that the identification of HIV-2 coreceptor usage based on a limited number of positions can be problematic. The predictive performance of geno2pheno[coreceptor-hiv2] is at least as high as the predictive performance of geno2pheno[coreceptor] for HIV-1[65].

This study confirmed that the V3 is the major determinant for the usage of the CCR5 and the CXCR4 coreceptors by HIV-2. However, there is also further evidence[66] indicating that other envelope regions besides V3 seem to contribute to HIV-2 coreceptor usage. First, SVMs based on the V1 and V2 regions achieved substantial predictive accuracies. Second, the V3 sequences of some X4-capable viruses did not exhibit any known features indicative of CXCR4 usage and some V3 sequences of R5-tropic isolates exhibited markers of X4-capability (Figure 5.1). Third, several samples sharing the same V3 amino-acid sequence but exhibiting discordant measurements of phenotypic coreceptor usage were observed (Table 5.1).[67]

R5-tropic HIV-2 exhibiting X4-markers could also be explained by a switch from CXCR4 to CCR5 usage (X4-R5 reversion). X4-R5 reversions have already been reported in HIV-1-infected patients after immune reconstitution[68]. Because recent findings indicate that X4-capable HIV-1 viruses are less susceptible to neutralization by autologous antibodies than R5 viruses from the same host[69], X4-

[62] Pfeifer and Lengauer 2012

[63] Shi et al. 2005

[64] Davenport et al. 2016

[65] The established cutoffs for geno2pheno[coreceptor] (EU: 10%/20%, UK: 5.75%, Germany/Austria: 5-15%) (Vandekerckhove et al., 2011; Harrigan et al., 2009; Eberle et al., 2014) exceed the recommended 5% cutoff for geno2pheno[coreceptor-hiv2].

[66] Santos-Costa et al. 2014

[67] This finding may also be the result of varying sensitivities among the different phenotypic assays (e.g. GHOST (3) cells, PBMCs with the Δ32 mutation, U87 cells) as well as experimental conditions.

[68] Ribeiro et al. 2006; Ercoli et al. 1997; Philpott et al. 2001; Skrabal et al. 2003

[69] Lin et al. 2016

R5 reversions could result from the normalization of naive T-cell turnover following immunological recovery[70], after which the infection of naive T-cells by X4-capable variants may not be productive enough[71]. Since X4-capable HIV-2 also seem to be less susceptible to neutralization than CCR5-using strains[72], X4-R5 reversions in HIV-2 could be explained by the same mechanism.

Besides these interpretations, discrepancies between the measured phenotypic coreceptor usage and features in the V3 amino-acid sequence could also be a by-product of the qualitative interpretation of phenotypic assays. *In vivo*, coreceptor usage is on a continuous scale and several, consecutive structural changes within the surface glycoprotein occurring along the viral evolutionary trajectory enable an increasingly effective coreceptor usage. However, this fact is neglected when the results of phenotypic assays are reported. Although the assays produce quantitative measurements (e.g. fluorescence, luminescence, or antigen levels), these measurements are typically converted to a qualitative scale for the sake of convenience regarding further analyses[73]. For the sake of accuracy, however, it would be paramount to work on the raw data (e.g. fluorescence, luminescence, or antigen levels). Using these quantitative measurements, it would be possible to develop models capable of placing a virus onto the evolutionary continuum stretching from viruses using only CCR5 to dual-tropic viruses, and finally viruses using only CXCR4. Moreover, working on raw data from phenotypic assays would facilitate the application of established statistical techniques for the normalization of data subject to batch effects (e.g. due to different types of phenotypic assays), which could improve the accuracy of large-scale studies on coreceptor usage considerably. In the meantime, the FPRs provided by geno2pheno[coreceptor-hiv2] could serve as a useful quantity for placing a virus on the coreceptor continuum.

To shed more light on why identical V3 amino-acid sequences with discordant phenotypic measurements of coreceptor usage exist, three aspects should be investigated. First, the agreement between different phenotypic assays should be validated or, even better, a standardized phenotypic assay for identifying HIV-2 coreceptor usage should be developed. Second, further research investigating the intra-host evolution of HIV-2 with respect to coreceptor usage and its impact on viral fitness seems necessary to determine whether X4-R5 reversions do occur. Third and most importantly, it should be investigated whether amino acid substitutions in the V1/V2 region can impart the X4-phenotype independently of substitutions in the V3, a question for whose resolution more data is required[74].

Based on the conducted analyses, it was possible to identify four characteristics that differentiate the V3 of HIV-2 and HIV-1 with

[70] Vrisekoop et al. 2015

[71] Ribeiro et al. 2006

[72] Marcelino et al. 2012

[73] Typical qualitative scales are the annotation of coreceptor usage (e.g. R5/X4-capable) or the efficiency of coreceptor usage (e.g. -/+/++/+++).

[74] Shi et al. 2005

respect to coreceptor usage. While individual mutations in the V3 of HIV-2 are highly predictive of coreceptor usage (e.g. 18X has a sensitivity of 79% and a specificity of 96%), there is no discriminatory signal in the HIV-1 V3 that allows for the accurate identification of coreceptor usage by itself. For example, the 11/25 rule, which classifies HIV-1 as X4-capable if its V3 contains a positively charged amino acid at the 11th or 25th position only obtains a sensitivity of 30% albeit at a specificity of 93%[75]. Second, while the major discriminatory markers for CXCR4 usage of HIV-2 (18X, 19K/R, insertions after position 24) appear at the V3 C-terminus, discriminatory features of HIV-1 coreceptor usage occur along the full extent of the V3 region. Third, while a V3 net charge exceeding six is significantly associated with the usage of CXCR4 by HIV-2, there is no significant association between the overall charge of the HIV-1 V3 and coreceptor usage[76], although CCR5 and CXCR4 exhibit contrasting electrostatic potential surfaces[77]. Fourth, a comparison of the predictions for X4-capable variants from HIV-1 (Figure A.1) and HIV-2 (Figure 5.2) suggests that the V3s of X4-capable HIV are genetically more distant to R5 variants in HIV-2 than in HIV-1.

In the future, geno2pheno[coreceptor-hiv2] could be improved by ensuring that input sequences originate from HIV-2, for example, by implementing a cutoff on the alignment score. In this way, misuse of the service could be prevented, for example, when a V3 of HIV-1 is unintentionally provided.

To conclude, geno2pheno[coreceptor-hiv2] is the first web service for the prediction of HIV-2 coreceptor usage. By using geno2pheno[coreceptor-hiv2], clinicians can decide whether they can prescribe the CCR5 coreceptor antagonist maraviroc, while epidemiologists can use the tool to investigate the association between HIV-2 coreceptor usage and disease progression. geno2pheno[coreceptor-hiv2] is freely available via coreceptor-hiv2.geno2pheno.org.

This chapter and the previous chapter (Chapter 4) described genotypic methods that enable the personalization of antiviral therapies. Persons that are under antiviral therapy often have to ingest multiple drugs daily, which can result in side effects including fatigue, lipodystrophy, nausea, or diarrhea[78]. This is one of the reasons why it is not only necessary to support therapy selection but also to support the development of novel antiviral agents. The following chapter (Chapter 6) is concerned with an approach that can advance the development of antibody-based antiretroviral treatments.

[75] Low et al. 2007

[76] Kalinina et al. 2013

[77] Tan et al. 2013

[78] Carr and Cooper 2000b

6
Designing Multiplex PCR Primers for Human Immunoglobulins

In this chapter, I present a new primer design approach that improves the amplification of cDNA from highly mutated antibodies via PCR. I devised and implemented the approach that resulted in the primer design tool openPrimeR. The primer design approach was developed in collaboration with Christoph Kreer, Nico Pfeifer, and Florian Klein. In close collaboration with Christoph Kreer, I used openPrimeR to design and computationally validate primer sets for IGHV, IGKV, and IGLV. The designed primer sets were experimentally validated by Nathalie Lehnen, Philipp Schommers, Meryem Seda Ercanoglu, and Christoph Kreer. The project was initialized by Florian Klein. The work was supervised by him and Nico Pfeifer, who contributed to the development of openPrimeR. This chapter expands upon a manuscript entitled "openPrimeR for Multiplex Amplification of Highly Diverse Templates", which has been submitted to Nature Biotechnology.

> Few scientists acquainted with the chemistry of biological systems at the molecular level can avoid being inspired. Evolution has produced chemical compounds exquisitely organized to accomplish the most complicated and delicate of tasks.
>
> Donald J. Cram, 1987

Elite neutralizers, which make up a small fraction of HIV-1 infected persons (≈ 1%), develop potent bNAbs against HIV-1 that are capable of neutralizing HIV-1 virions from several clades even at low concentrations[1]. Therefore, bNAbs are currently being investigated as a novel strategy for treating and preventing HIV-1 infection[2].

The *in vitro* investigation of human antibodies requires the extraction of single B-cell transcripts from immunoglobulin heavy chain genes (IGHs) and corresponding immunoglobulin lambda genes (IGLs) or immunoglobulin kappa genes (IGKs), which encode the immunoglobulin light chains[3]. After amplification of these transcripts via multiplex RT-PCR[4], the antibody cDNA is cloned into eukaryotic expression vectors that are inserted into cells for antibody secretion[5]. Designing primers for mPCR is challenging because the smallest

[1] Simek et al. 2009; Scheid et al. 2011; Wu et al. 2010a; West et al. 2014

[2] Klein et al. 2013; Moldt et al. 2012

[3] Tiller et al. 2008; Ippolito et al. 2012

[4] In contrast to conventional PCR, RT-PCR is performed on RNA instead of DNA as it involves the transformation of RNA to cDNA via reverse transcriptase.

[5] These expression vectors are plasmids that encode the constant regions of human IGH, IGK, and IGL. They provide a framework for inserting variable segments of interest into the vectors.

possible combination of primers covering (amplifying) all templates has to be found. bNAbs introduce an additional burden because they are highly mutated[6] and may prevent the annealing of primers that have been optimized for germline immunoglobulins.

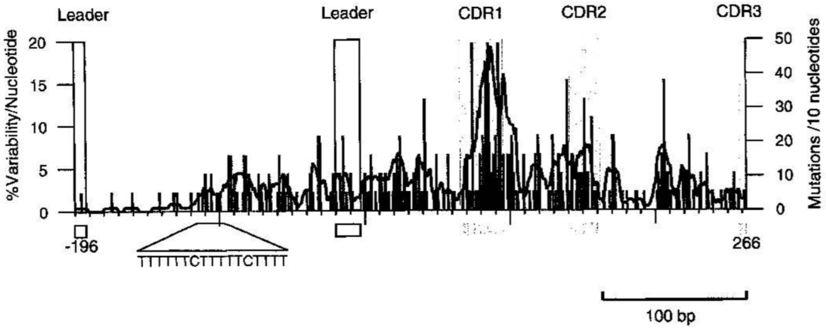


Figure 6.1: Variability in different antibody regions. The leader region at the 5′ end consists of two parts (shown as rectangles) that are separated by an intron.

Permission for reproduction was kindly granted by John Wiley and Sons (Rada et al., 1994).

While conventional immunoglobulin G (IgG) antibodies (Section 2.2.4) carry 18.0 ± 8.1[7] somatic mutations in their heavy chain variable (V_H) genes[8], HIV-1-specific IgG antibodies exhibit roughly twice as many V_H mutations[9].[10] Second-generation bNAbs (Section 2.3.7) even frequently carry in excess of 60 V_H mutations[11,12]. However, immunoglobulin heavy chain variable region genes (IGHVs) show a decrease in mutation frequencies towards the leader region[13] (Figure 6.1). The leader is a short 5′ untranslated region of immunoglobulin mRNA upstream of the translation start site. Designing primers in the conserved immunoglobulin leader region has the potential for improving the amplification of IGHVs from bNAbs[14]. Therefore, the goal of this work was to develop a rational approach for designing primers targeting the leaders of IGHV, immunoglobulin kappa variable region gene (IGKV), and immunoglobulin lambda variable region gene (IGLV) in order to improve the amplification of highly mutated human immunoglobulin sequences.

This chapter is structured as follows. Section 6.1 provides an overview of the considerations for mPCR primer design and introduces related work. The approach of openPrimeR is delineated in two sections. In the first methodological section (Section 6.2), I introduce the notation that is used in this chapter and describe techniques for the evaluation of primers and primer sets. The second methodological section (Section 6.3) presents algorithms for the design and selection of mPCR primer sets. Section 6.4 presents the primer sets that were designed for IGHV, IGKV, and IGLV. This chapter concludes with Section 6.5 in which I discuss the results of this work.

[6] Klein et al. 2013

[7] IgG from memory B cells of several patients were pooled.

[8] Tiller et al. 2007

[9] Scheid et al. 2009; Mouquet et al. 2011

[10] Scheid et al. (2009) analyzed 502 antibodies from HIV envelope-binding memory B cells from six HIV-infected patients with broadly neutralizing antibodies. Mouquet et al. (2011) studied 51 B cell clones from the IgG memory B cells of three patients infected with clade A and B HIV-1 that elicit broadly neutralizing antibodies.

[11] Wu et al. 2011

[12] Nine VRC01-like antibodies were constituted from B cells of three HIV-infected donors.

[13] Rada et al. 1994; Yaari et al. 2013

[14] Scheid et al. 2011

6.1 Multiplex Primer Design

The design of primers for mPCR requires a deep understanding of the mechanism by which mPCR allows for the successful amplification of nucleic acids (Section 6.1.1). Approaches for mPCR primer design have to consider the physicochemical properties of the primers in order to find suitable primer sets (Section 6.1.2). Due to the unavailability of a suitable primer design tool for human immunoglobulin sequences, openPrimeR was developed (Section 6.1.3).

6.1.1 Biophysical Characteristics of PCR

For the design of mPCR (Section 2.5.4) primers, the following aspects should be taken into account. A template can only be amplified if both primer annealing and elongation by polymerase are successful[15]. Efficient primer annealing is determined by the overall complementarity of primer and template. Elongation, on the other hand, critically depends on the structure of the primer 3′ hexamer, which forms the binding region of the polymerase[16].[17]

Primers should fulfill a multitude of physicochemical properties that are relevant for the success of mPCR, which are discussed in the following.

Melting Temperature The melting temperature, T_m, of an oligonucleotide is defined as the temperature at which half of the DNA strands are denatured, while the other half is still double-stranded. The melting temperature provides insights into the stability of the primer-template duplex: primers with high melting temperatures bind tightly to complementary templates, while primers with low melting temperatures bind more loosely.

The PCR annealing temperature depends on the melting temperatures of the primers. A simple rule of thumb is to choose the annealing temperature to be about 5°C lower than the smallest T_m. While low annealing temperatures may lead to unspecific annealing, high annealing temperatures may prevent annealing.

The two factors that influence melting temperature the most are primer length and GC content. This is because longer primers exhibit a larger number of nucleotide interactions and high GC contents facilitate a greater number of stable G-C pairings. Hence, a simple formula based on the nucleotide distribution was developed for calculating the melting temperature for sequences of length n exceeding 13 nucleotides[18]:

$$T_m = 64.9 + 41 \frac{|G| + |C| - 16.4}{n}$$

[15] Pan et al. 2014

[16] Kwok et al. 1990; Stadhouders et al. 2010; Wright et al. 2014

[17] See Chapter 7 for more information on the molecular determinants of successful PCR amplification events.

[18] Wallace et al. 1979

Here, $|G|$ and $|C|$ indicate the number of guanines and cytosines in the sequence. A typical primer has a length of 18 nucleotides and a typical GC ratio is 50%. If we assume that $|G| = 4$ and $|C| = 5$, we would have

$$T_m = 64.9 + 41\frac{4+5-16.4}{18} = 64.9 + 41\frac{-7.4}{18} = 48°\text{C}$$

according to the formula.

Current approaches for determining the melting temperature correct for the PCR ion concentrations[19] and rely on nearest-neighbor thermodynamics, which considers the thermodynamic contributions of neighboring base pairs[20].

GC Clamp Primers are often designed to exhibit a so-called GC clamp[21], which typically consists of one to three Gs or Cs at the 3′ end. Having one or multiple GCs at the 3′ end of primers is desirable because GC pairs, which are sustained by three hydrogen bonds, are more stable than AT pairs, which are stabilized by two hydrogen bonds. Since the 3′ region is crucial for polymerase binding, high affinity in this region ensures stable binding of polymerase. A large number of GCs (e.g. more than 3) among the last 3′ bases should be avoided, as this could facilitate primer dimer formation and lead to mispriming.

GC Content The GC content of a primer is defined by the ratio of Gs or Cs among its constituent nucleotides. Primers should have balanced GC contents (e.g. ranging from 40%–60%) because high GC contents are associated with secondary structure formation, while primers exhibiting low GC contents may not stably hybridize to templates.

Primer Length The length of primers should be chosen under consideration of priming specificity, efficiency of hybridization, and experimental costs. While short primers may save costs, they also lack specificity as their use may result in off-target amplifications. Longer primers ensure higher specificities due to their increased sequence complexity. The higher specificity of longer primers can be detrimental to multiplex PCR where higher specificity also implies lower coverage. The length of typical primers ranges from 18 to 22 nucleotides[22]. Since longer primers have higher melting temperatures than shorter primers, primer sets typically consist of primers with similar lengths.

Nucleotide Runs and Repeats A run refers to the consecutive repetition of a single nucleotide in a primer. Large number of repetitions

[19] Panjkovich and Melo 2005

[20] SantaLucia 1998

[21] No et al. 2014; Lorenz 2012; Thornton and Basu 2011

[22] Burpo 2001

(e.g. more than 4) should be avoided because they can lead to mispriming. The same holds for repeats, that is, consecutive repetitions of the same dinucleotide.

Secondary Structures Primers can form secondary structures due to interactions between complementary nucleotides within the primer that lead to spatial orientations other than simple coils. Primers exhibiting secondary structures should be avoided because secondary structures can prevent annealing to the template. Experimentally, secondary structures can be avoided by using dimethyl sulfoxide (DMSO).

Primer Dimerization Primer dimerization refers to the hybridization of primers to other primers rather than hybridization to the template sequence. Preventing the formation of primer dimers is one of the main concerns when designing primers because primer dimers reduce product yields and can lead to unintended amplicons when dimers are formed at the 3′ ends. A primer that binds to a copy of itself is called a self-dimer, while a primer that binds to another oligomer is called a cross-dimer.

6.1.2 Related Work

In the following, established approaches are evaluated with respect to the following demands on a suitable tool for the design of immunoglobulin primers:

- The tool should estimate the coverage of the primers. This is a necessary requirement in order to find a minimum set of primers maximizing the coverage by solving the SCP.

- Designed primers should fulfill stringent quality criteria regarding their physicochemical properties. Low-quality primers may fail to amplify their target templates.

- The tool should be easily usable and provide a graphical user interface (GUI). The end users of primer design tools typically have life science backgrounds. Thus, they may shy away from unintuitive software, for example software that is only usable through a command line interface.

- The tool should be able to design degenerate primers. A degenerate primer is an oligonucleotide sequence that is not only made up of the four standard nucleobases but also contains IUPAC ambiguity codes, which indicate the presence of multiple bases.

Degenerate primers are economically desirable because the corresponding mixtures of primers can be ordered at the same cost as individual primers.

- The tool should facilitate the comparison of primer sets. By comparing primer sets *in silico*, it is possible to identify which set performs favorably without any experimental expenses.

These requirements for multiplex primer design give rise to the notion of *rational primer design*. In rational primer design, the minimum number of high-quality primers maximizing the coverage is selected. To find high-quality primers, the properties of the primers should be comprehensively analyzed. Since rational primer design is computationally intensive, such approaches are only feasible using specialized software and cannot be performed by hand.

Table 6.1 provides an overview of published primer design approaches. Unfortunately, for the majority of published primer design approaches, no corresponding tool is available. Nearly all available tools provide GUIs, which affords a high usability. However, the GUIs of most tools such as GeneFisher[23], PrimerStation[24], or PRIMEGENS[25] are rather rudimentary in that they do not allow for the interactive investigation of the results. In this regard, the commercial softwares OLIGO[26] and FastPCR[27] are considerably more advanced. Still, their GUIs neither allow for investigating primer sets with regard to coverage nor for comparing the properties of sets.

The main challenge of primer design for multiplex PCR lies in finding the smallest possible set of primers that ensures the amplification of all template sequences[28]. This optimization problem can be formulated in terms of the NP-complete SCP (Section 3.7.3). However, only few approaches actually solve the SCP. This is because one branch of primer design methods solves the degenerate primer design problem, which is discussed in the next paragraph, and another branch aims at high throughput (HT) applications such as tiling whole genomes[29]. Most approaches that solve the SCP either do so approximately using a greedy algorithm[30] or exactly using an ILP formulation[31]. Genetic algorithms[32], which can find approximate solutions, are less frequently used. Approaches that do not solve the SCP such as PRIMEGENS typically output a list with suggested primers for each template. This, however, leaves the user with the burden of finding an optimal combination of primers — a task that typically cannot be completed by hand.

Relatively few approaches are able to find degenerate primers. Degenerate primers are typically constructed via hashing[33], beam search[34], or alignment[35]. There are two groups of methods for computing degenerate primers. The first group consists of approaches

[23] Giegerich et al. 1996

[24] Jabado et al. 2006

[25] Srivastava and Xu 2007

[26] Rychlik 2007

[27] Kalendar et al. 2014

[28] Pearson et al. 1996

[29] Gardner et al. 2014

[30] Jabado et al. 2006; Gardner et al. 2009

[31] Hsieh et al. 2003; Bashir et al. 2007

[32] Huang et al. 2005

[33] Linhart and Shamir 2002

[34] Souvenir et al. 2003

[35] Jabado et al. 2006

Year	Reference	Tool	Availability	GUI	SCP	HT	Degeneracy
1989	(Rychlik, 2007)	OLIGO	✓	✓	✗	✗	✗
1996	(Giegerich et al., 1996)	GeneFisher	✓	✓	✗	✗	✓
1998	(Pesole et al., 1998)	GeneUp	✗	✗	✓	✗	✗
2003	(Rose et al., 1998, 2003)	j-CODEHOP	✓	✓	✗	✓	✓
2001	(Kämpke et al., 2001)	DoPrimer	✗	✗	✗	✓	✗
2002	(Linhart and Shamir, 2002)	HYDEN	✓	✗	✗	✗	✓
2003	(Emrich et al., 2003)	PROBEMER	✗	✓	✗	✗	✗
2003	(Souvenir et al., 2003)	MIPS	✗	✗	✗	✗	✓
2004	(Jarman, 2004)	Amplicon	✓	✓	✗	✗	✓
2004	(Wang et al., 2004a)	G-PRIMER	✗	✓	✓	✗	✗
2005	(Huang et al., 2005)	PDA-MS/UniQ	✗	✗	✓	✗	✓
2005	(Rachlin et al., 2005)	MuPlex	✗	✓	✗	✓	✗
2006	(Jabado et al., 2006)	Greene SCPrimer	✗	✓	✓	✗	✗
2006	(Yamada et al., 2006)	PrimerStation	✓	✓	✗	✓	✗
2006	(Lee et al., 2006)	MultiPrimer	✗	✗	✗	✗	✗
2007	(Srivastava and Xu, 2007)	PRIMEGENS	✓	✓	✗	✓	✗
2007	(Bashir et al., 2007)	*NA*	✗	✗	✓	✓	✗
2009	(Gardner et al., 2009)	MPP	✗	✗	✓	✗	✗
2009	(Kalendar et al., 2009, 2014)	FastPCR	✓	✓	✗	✓	✓
2010	(Shen et al., 2010)	MPPrimer	✓	✓	✗	✗	✗
2012	(Chuang et al., 2012)	URPD	✗	✓	✗	✗	✗
2012	(Hysom et al., 2012; Gardner et al., 2014)	PriMux	✓	✗	✓	✓	✓
2016	(O'Halloran, 2016)	PrimerMapper	✓	✓	✗	✓	✗

Table 6.1: Overview of multiplex primer design tools. *SCP* indicates that a set cover problem is solved in order to minimize the primer set. *Availability* indicates whether an implementation of the tool is publicly available (as of June 2018). *HT* indicates whether the tool is high-throughput, that is, whether it is suitable for large-scale applications. *Degeneracy* indicates whether primers containing degeneracies can be designed.

that solve the degenerate primer design problem in which a single primer with minimum degeneracy[36] and maximum coverage is sought. Tools solving the degenerate primer design problem such as HYDEN[37] and MIPS[38] can provide elegant solutions since they output only a single primer. However, these tools are also subject to severe limitations. For example, it may not be possible to obtain a single degenerate, high-quality primer that covers all of the templates. In this case, the degenerate primer design problem needs to be iteratively solved for several positions in the templates (similar to a greedy algorithm), which may lead to suboptimal solutions.

The second group consists of approaches that determine several degenerate primer candidates and then solve the SCP. An example for such an approach is the primer design tool GreeneSCPrimer[39]. Since the approach of openPrimeR also implements this strategy, I refer the reader to Section 6.3, in which I provide more details.

To my best knowledge, the only published primer set for immunoglobulin sequences that was designed using a computational tool was described by Sun et al. (2012) who designed primer sets for individual IGHV gene groups using CODEHOP[40] and manually combined them. However, the designed forward primers did not find acceptance in the immunological community, presumably due to two reasons. First, the designed primer set was quite large because the seven primers were heavily degenerate, representing a total of 112 individual oligomers. Second, the combination of primers was not selected with stringent quality criteria in mind.

Popular primer sets such as the one from Tiller et al. (2008) or the one from Lim et al. (2010) consist of fewer primers[41] that were manually chosen by studying multiple sequence alignments. In another

[36] The degeneracy of a primer is the number of non-degenerate sequences it represents. For example, the oligonucleotide *acrgacgtgacr* has degeneracy $2 \times 2 = 4$ since **r** occurs two times and represents A or G.

[37] Linhart and Shamir 2002

[38] Souvenir et al. 2003

[39] Jabado et al. 2006

[40] Rose et al. 2003

[41] The set from Tiller et al. (2008) has 4 (5) primers, while the set from Lim et al. (2010) has 8 (23) primers. The number of disambiguated primers are shown in parentheses.

study, Scheid et al. (2011) pioneered a new primer design strategy in which primers were designed for the leader region of IGHV. However, these primers were not rationally designed. Therefore, the resulting primer set is quite large (21 primers) and evaluations with openPrimeR have shown that the set seems to suffer from cross dimerization (22 pairs with $\Delta G < -6$ kcal/mol) and high melting temperature differences (melting temperature differences are between 10°C and 20°C).

6.1.3 *Technical Details of openPrimeR*

The approaches and data presented in the following sections have been made available through openPrimeR, an open-source tool for designing, evaluating, and comparing primer sets for mPCR. To design primers, the tool solves a SCP using either a greedy algorithm or an ILP. openPrimeR was implemented in R. The graphical user interface was developed using the Shiny framework. The tool is available in the form of two Bioconductor packages. While *openPrimeR* provides a programmatic interface, *openPrimeRui* provides a GUI in terms of a Shiny application.

To use all functionalities of openPrimeR, the following programs should be installed[42]:

MELTING Melting temperature calculations[43]

ViennaRNA Secondary structure detection[44]

OligoArrayAux Thermodynamic evaluation of primer binding events[45]

MAFFT Determination of multiple sequence alignments[46] for initializing degenerate primers

Pandoc Generation of PDF reports

A Docker container satisfying all dependencies of openPrimeR is available at Docker Hub.

6.2 *Evaluation of Primers*

This section introduces the methods for the evaluation of mPCR primer sets that are used by openPrimeR. Section 6.2.1 defines the notation that is used for the remainder of this chapter. Section 6.2.2 introduces an approach for the estimation of primer coverage. The physicochemical properties and quality metrics that are considered by openPrimeR are presented in Section 6.2.3. These metrics are important for the primer design algorithms that are presented in Section 6.3.

[42] If software dependencies are not fulfilled, the corresponding features are not available. For example, if ViennaRNA is not installed, constraints on the secondary structure cannot be considered. Similarly, if MAFFT is not installed, it is not possible to design degenerate primers.

[43] Dumousseau et al. 2012

[44] Tafer et al. 2011

[45] Markham and Zuker 2008

[46] Kuraku et al. 2013

6.2.1 Preliminaries

Templates $t \in T = \{t_1, \dots, t_{|T|}\}$ and primers $p \in P = \{p_1, \dots, p_{|P|}\}$ are sequences of length $n, m \in \mathbb{N}$ such that $t \in \mathcal{A}^n$ and $p \in \mathcal{A}^m$. The alphabet, $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, consists of conventional nucleotides,

$$\mathcal{A}_1 = \{A, C, G, T\},$$

and ambiguous IUPAC nucleotides (Table A.1),

$$\mathcal{A}_2 = \{M, R, W, S, Y, K, V, H, D, B, X, N\}.$$

When primer $p \in P$ is expected to cover template $t \in T$, we write $p \triangleright t$. Equivalently, we can write $t \triangleleft p$ to express that t is expected to be covered by p. Note that $p \triangleright t$ does not necessarily guarantee that t is actually amplified by p during PCR. This is because the theoretical coverage may deviate from the actual coverage; the accuracy with which the coverage can be approximated depends on which notion of coverage is employed (Section 6.2.2). The set of expected amplification events (EAEs) of primer p with respect to the set of templates T is denoted by

$$p_{\triangleright T} = \{t_i \in T | p \triangleright t_i\}.$$

The number of templates in T that are covered by p are determined via $|p_{\triangleright T}|$, which is also referred to as *primer coverage*. The set of primers that cover a template $t \in T$ is given by

$$t_{\triangleleft P} = \{p_i \in P | t \triangleleft p_i\}.$$

A template $t \in T$ is considered to be covered if and only if $|t_{\triangleleft P}| \geq 1$, that is, if there is at least one primer covering the template.

The set of templates that is covered by a set of primers P is defined by

$$P_{\triangleright T} = \bigcup_{t \in T} \bigcup_{p \in t_{\triangleleft P}} t.$$

In the following, the coverage of P with respect to T, which is given by $|P_{\triangleright T}|$, is called the (overall) coverage.

Primers with Multiple Orientations Since PCR requires the amplification of sense- and anti-sense strands of DNA, every template should be covered by a forward primer, $p^{\rightarrow}$, which allows for synthesizing the sense strand, and a reverse primer, $p^{\leftarrow}$, which allows for synthesizing the anti-sense strand. In case that only primers of a single orientation, that is, either only forward or only reverse primers are present, coverage is defined as above. Otherwise, if primers of both orientations are present, the set of primers covering a template $t \in T$

is defined by considering primers of both orientations via

$$t_{\triangleleft P} = \begin{cases} \emptyset & \text{if } \nexists p^{\rightarrow}, p^{\leftarrow} : p^{\rightarrow} \triangleright t, p^{\leftarrow} \triangleright t \\ \{p^{\rightarrow} \in P | t \triangleleft p^{\rightarrow}\} \cup \{p^{\leftarrow} \in P | t \triangleleft p^{\leftarrow}\} & \text{else}\,. \end{cases}$$

This means that the number of primers that cover a template is zero if there is no pair of forward and reverse primer that covers the template. Otherwise, the number of primers that cover the template is defined by the union of forward and reverse primers covering the template. The coverage of a primer pair $(p^{\rightarrow}, p^{\leftarrow})$ is defined as the intersection of EAEs from its constituent primers:

$$(p^{\rightarrow}, p^{\leftarrow})_{\triangleright T} = {p^{\rightarrow}}_{\triangleright T} \cap {p^{\leftarrow}}_{\triangleright T}\,.$$

Note that when primers of multiple orientations are considered, they are not represented as pairs before the optimization has completed since it would be unclear how pairing should be performed. This is because the primers constituting a pair should have overlapping EAEs and similar physicochemical properties (Section 6.3.3).

6.2.2 *Detection of Amplification Events*

openPrimeR supports several notions of coverage because different coverage conditions are suitable in dependence on the intended use of the primers. For example, primers for qPCR should be designed using a more conservative coverage definition than those for conventional PCR. Since complementarity between primer and template is a fundamental requirement for the amplification of a template, it is necessary to determine the most likely binding region of the primer in the template. Since primers may also bind with mismatches, I introduce the parameter n_{mm}, which defines the maximally allowed number of mismatches between a primer and a template. For every primer, the most likely binding region in every template is determined by scanning the template for matches between the primer and template sequence with at most n_{mm} mismatches and returning the match with the smallest number of mismatches, if available.

Determining primer coverage events in this way gives rise to the notion of *basic coverage* according to which a primer covers a template if a match with at most n_{mm} mismatches is found. Including further requirements (Table 6.2) leads to the notion of *constrained coverage*.[47] In the following, I discuss the parameters that influence the notion of primer coverage.

Allowed Mismatches The choice of a suitable value for n_{mm} depends on two factors. The first factor is the fidelity of amplification. Since

[47] The term *coverage* refers to the constrained coverage if not stated otherwise.

Property	Enabled	Setting
Mismatches	✓	≤ 7
Prevention 3′ terminal mismatches	✗	≤ 0
Prevention of stop codons	✗	NA
Prevention of amino-acid substitutions	✗	NA
Free energy of annealing	✗	≤ -5 kcal/mol
Amplification efficiency	✗	$\geq 0.1\%$
TMM model	✓	FPR $\leq 6\%$

Table 6.2: Default settings for the evaluation of primer coverage. If the prevention of stop codons or substitutions is active, then the EAEs of primers that would induce stop codons or amino-acid substitutions into the amplicons are discarded.

mismatch binding events induce changes in the nucleotide sequence of amplicons, large values of n_{mm} should be avoided in order to preserve the original template sequences. The second factor is the sensitivity and specificity at which amplification events should be called. While high values of n_{mm} (e.g. $n_{\text{mm}} = 5$) allow for increased sensitivities, they also reduce the specificity of calling amplification events. Low values of n_{mm} (e.g. $n_{\text{mm}} = 1$), on the other hand, ensure higher specificities at reduced sensitivity.

For primer design, it is crucial to achieve high specificities in order to ensure that all templates can be covered. By selecting a small value for n_{mm}, it is possible to reduce false positive coverage calls. This ensures that designed primers actually allow for amplifying all of the templates. However, a conservative (i.e. small) choice of n_{mm} may also lead to prohibitively large primer sets since mismatch amplification events are discarded. In these scenarios, a larger value for n_{mm} may be chosen. Then, however, a model that estimates whether EAEs correspond to actual amplification events, should be employed. Such models are introduced in the next paragraph.

Additional Criteria There are further criteria besides the maximal number of allowed mismatches according to which proposed amplification events can be limited. These approaches are particularly useful when n_{mm} is large because they can increase the specificity of the proposed amplification events. For example, a simple criterion may involve setting a minimal value for the free energy of annealing, ΔG. However, there are also more intricate models for the determination of amplification events. Chapter 7 provides a comprehensive discussion of models for this task and introduces the logistic regression model *TMM* that is available via openPrimeR.

openPrimeR provides the following approaches for detecting amplification events:

- The thermodynamic model from DECIPHER[48] that estimates amplification efficiency.

[48] Wright et al. 2014

Constraint	Denomination	Target range	Limit range
Length	$\|p\|, \forall p \in P$	[18, 22]	[18, 22]
Specificity	spec	[1, 1]	[0.8, 1]
GC clamp	|GC|	[1, 3]	[0, 4]
GC ratio	GC%	[40%, 60%]	[30%, 70%]
Runs	Runs	[0, 4]	[0, 6]
Repeats	Reps	[0, 4]	[0, 6]
Self dimers	ΔG_{sd} [kcal/mol]	[-5, ∞]	[-7, ∞]
Melting temperature	T_m [°C]	[50, 70]	[50, 70]
Folding	ΔG_f [kcal/mol]	[-1, ∞]	[-2, ∞]
T_m deviation*	ΔT_m [°C]	[0, 5]	[0, 7.5]
Cross dimers*	ΔG_{cd} [kcal/mol]	[-5, ∞]	[-7, ∞]

Table 6.3: Permissible properties of PCR primers as defined in the default settings of openPrimeR. The column *Target range* indicates the permissible values for each constraint, which are used for filtering primers. The column *Limit range* indicates the value range that is used when constraints are relaxed during the selection procedure. Constraints annotated with asterisks are not included in the set of filtering constraints and instead considered only during the optimization phase.

- The logistic regression model *TMM* for estimating the likelihood of amplification.

- The free energy of annealing.

- The positions in the 3′ hexamer in which mismatches are forbidden.

There are also other criteria for ensuring the fidelity of amplification for designed primers. These criteria influence the selection of EAEs. Using openPrimeR, it is possible to disregard EAEs that are associated with stop codons or amino-acid substitutions.

6.2.3 *Physicochemical Constraints*

In order to evaluate the quality of a primer, it is useful to define constraints on its physicochemical properties (Section 6.1.1). The constraint on the *i*-th physicochemical property is defined as the pair $f_i = (f_i^{\text{min}}, f_i^{\text{max}})$ where f_i^{min} and f_i^{max} indicate the minimum and maximum desired value, respectively. Constraints define desirable ranges on the physicochemical properties of primers. For example, the GC ratio of primers should be within the interval $[40\%, 60\%]$. Thus, if the *i*-th constraint relates to the GC ratio, we would include the constraint $f_{=}(0.4, 0.6)$. Based on the set of active constraints $\mathcal{F} = \{f_1, \ldots, f_{|\mathcal{F}|}\}$, the quality of a primer can be evaluated. An overview of the default constraint settings is shown in Table 6.3. Table 6.4 shows the PCR parameters that are applied when these properties are computed. In the following, I provide details on the computation of the constraints for melting temperatures, primer dimerization, and secondary structures.

Melting Temperature The melting temperature can be constrained in two ways. First, a desired range of primer melting temperatures $[T_1, T_2]$ can be specified. Let $T_m(p)$ indicate the melting temperature of primer p. The constraint $[T_1, T_2]$ ensures that only primers $p \in P$ with $T_1 \leq T_m(p) \leq T_2$ are retained. Second, the maximum allowed difference between the melting temperatures of selected primers can be constrained via ΔT_m^{max}. This constraint contributes to the compatibility of primers[49] by requiring that primers in the same set exhibit similar melting temperatures via

$$\max_{p_i, p_j \in P} |T_m(p_i) - T_m(p_j)| \leq \Delta T_m^{\text{max}} \,.$$

Since the maximum melting temperature difference depends on the set of selected primers, the constraint on ΔT_m is applied when primers are optimized, while the allowed melting temperature range is used when primers are filtered.

openPrimeR computes melting temperatures via the MELTING software[50]. The program uses a nearest-neighbor approach based on the parameters from SantaLucia (1998) and corrects for the salt concentration.

Dimerization There are two types of dimerization that are considered: self dimerization and cross dimerization. Self-dimerization refers to the tendency of a primer to bind to a copy of itself. Since self dimerization does not depend on the set of selected primers, self-dimerizing primers can be excluded during the filtering procedure. Cross-dimerization, on the other hand, describes the association of a primer with another primer. Since cross dimerization requires the consideration of the selected set of primers, this constraint is only considered by the primer design optimization procedure.

The selection of cross dimers is prevented through symmetric dimerization matrix $\mathbf{D} \in \{0, 1\}^{m \times n}$, which is defined by its entries

$$d_{p_i, p_j} = \begin{cases} 1 & \text{if } \Delta G(p_i, p_j) < \Delta G_{\text{cd}}^{\text{min}} \\ 0 & \text{else} \end{cases} \,.$$

Table 6.4: PCR settings that enter the computation of constraints. T_m refers to the melting temperature. ΔG_{cd} and ΔG_{sd} refer to free energy of cross dimerization and self dimerization, respectively. ΔG_f indicates secondary structure formation.

Parameter	Value	Affected quantities
[Na+]	Concentration	T_m, ΔG_{cd}, ΔG_{sd}
[Mg2+]	Concentration	T_m, ΔG_{cd}, ΔG_{sd}
[K+]	Concentration	T_m, ΔG_{cd}, ΔG_{sd}
[Tris]	Concentration	T_m, ΔG_{cd}, ΔG_{sd}
Polymerase type	Taq/Non-Taq	$p_{\triangleright T}, \forall p \in P$
Annealing temperature	Manual/Automatic	ΔG_f, ΔG_{cd}, ΔG_{sd}

[49] Two primers are considered *compatible* if combining them in a multiplex reaction does not negatively affect their amplification efficiency.

[50] Dumousseau et al. 2012

The entry $\mathbf{D}_{p_i,p_j}$ indicates whether primers p_i and p_j dimerize based on $\Delta G_{\text{cd}}^{\text{min}}$, a threshold on the smallest allowed free energy of cross dimerization. Therefore, entries d_{p_i,p_j} with $p_i = p_j$ indicate the presence of self dimers, while entries d_{p_i,p_j} with $p_i \neq p_j$ indicate the existence of cross dimers. To exemplify the use of $\mathbf{D}$, let us consider three primers p_1, p_2, and p_3 with the following dimerization matrix:

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} .$$

In this case, p_1 forms cross dimers with primers p_2 and p_3. Thus, p_1 can be combined with neither p_2 nor p_3. Since p_2 and p_3 do not hybridize, they can be included in the same primer set. More formally, this means that a combination of primers that is compatible with regard to cross dimerization induce a submatrix whose entries are zero.

Dimerizing primers are detected by computing the free energy using OligoArrayAux[51] and applying the user-definable cutoffs ΔG_{sd} (for self dimers) and ΔG_{cd} (for cross dimers).

Secondary Structure Secondary structure refers to the folding of a primer onto itself. In order to identify possible primer secondary structures, openPrimeR employs ViennaRNA[52] to compute ΔG_f based on the thermodynamic parameters from Turner and Mathews (2010). Since secondary structures should be prevented during the primer annealing phase, the temperature for computing ΔG_f is set to the annealing temperature.

6.2.4 *Metrics for Physicochemical Constraints*

To determine the properties of a primer p, I define the evaluation vector $\text{eval}(p, \mathcal{F}, T) \in \mathbb{R}^{|\mathcal{F}|}$ with respect to the template set T and the constraint set $\mathcal{F}$. Its i-th entry, $\text{eval}(p, \mathcal{F}, T)_i$, contains the i-th property of primer $p \in P$.

In the following paragraphs, I introduce several useful measures for evaluating the quality of primers. For this purpose, I first define the conditions under which a constraint is fulfilled and then introduce a way to test whether a set of primers exhibits significant constraint fulfillment. Next, I define the rate of constraint fulfillment and introduce the deviation from the target constraints.

Constraint Fulfillment A primer p is said to fulfill the i-th constraint, $f_i \in \mathcal{F}$, that is, $p \models f_i$ if and only if

$$f_i^{\text{max}} \geq \text{eval}(p, \mathcal{F}, T)_i \geq f_i^{\text{min}} .$$

[51] Markham and Zuker 2008

[52] Tafer et al. 2011

Number	$\lvert p \rvert$	$\lvert \mathrm{GC} \rvert$	GC%	Runs	Reps	ΔG_{sd}	$[T_m]$	ΔG_f	ΔT_m	ΔG_{cd}
Fulfilled	212	285	245	407	427	232	235	301	88	351
Failed	215	142	182	20	0	195	192	126	339	76

Table 6.5: Reference distribution of fulfilled and failed constraints.

Data Set	Satisfied Constraints	Failed Constraints
Current Set	n_{CS}	n_{CF}
Reference	n_{RS}	n_{RF}

Table 6.6: Structure of the constraint fulfillment matrix for testing whether the fulfillment of constraints is significant. The second and third column show the total number of fulfilled and failed constraints, respectively. The first row contains the values for the primer set whose significance is to be tested, while the second column contains the tallies from the reference distribution.

In case that, $p \not\models f_i$, we say that p does not fulfill (violates) f_i.

Testing the Significance of Primer Set Quality The theoretical significance of the quality of a primer set can be computed by determining the number of primers that either fulfill or violate individual constraints. Then, these counts are compared with the reference distribution shown in Table 6.5. This reference distribution was obtained from tabulating the constraint fulfillment of the established primer sets listed in Tables A.3, A.4, and A.5, which were evaluated using the default constraint settings shown in Table 6.3.

The total number of fulfilled and failed constraints for the input and the reference primer sets, respectively, gives rise to a two by two constraint fulfillment matrix whose structure is shown in Table 6.6. Using a one-tailed Fisher's exact test (Section 3.6.3), it is possible to determine whether the odds ratio $OR = \frac{n_{CS}}{n_{CF}} / \frac{n_{RS}}{n_{RF}}$ is significantly greater than 1. If this is the case, the tested primer set fulfills a significantly greater number of constraints than the sets from the reference distribution. Thus, primer sets with significant p-values exhibit favorable properties in comparison to the reference primer sets.

Rate of Constraint Fulfillment The constraint fulfillment vector $\mathrm{Ful}(P, \mathcal{F}, T) \in \mathbb{R}^{|\mathcal{F}|}$ indicates the fraction of primers that fulfill individual constraints. The rate of constraint fulfillment is formulated using an indicator function,

$$\mathbb{1}_{f_i}(p) = \begin{cases} 1 & \text{if } p \models f_i \\ 0 & \text{else} \end{cases},$$

which determines whether the constraint $f_i \in \mathcal{F}$ is fulfilled by primer p. Given $j \in \{1, 2, \ldots, |\mathcal{F}|\}$, the j-th entry of the constraint fulfillment vector is defined by the fraction of primers fulfilling constraint $f_j \in \mathcal{F}$:

$$\mathrm{Ful}(P, \mathcal{F}, T)_j = \frac{\sum_{p_i \in P} \mathbb{1}_{f_j}(p_i)}{|P|}.$$

The rate of constraint fulfillment is a useful quantity for evaluating the quality of a primer set. A disadvantage of this measure is that it does not consider the extent to which a constraint is broken. For example, consider a primer that fulfills 9 of 10 constraints. Based on the high fulfillment rate, one might conclude that this primer is of high-quality even though the primer may substantially deviate from one of the constraints. For example, with $\Delta G_{\text{cd}}^{\text{min}} = -15$ kcal/mol, the primer would excessively deviate from the default minimum free energy of cross dimerization of $\Delta G_{\text{cd}}^{\text{min}} = -5$ kcal/mol, which may jeopardize the success of PCR. Thus, primer sets exhibiting high rates of constraint fulfillment should still be investigated with respect to their deviation from the individual constraints.

Constraint Deviation The extent to which the physicochemical properties of a primer, p, deviate from the constraints is represented by the constraint deviation vector, $\text{Dev}(p, \mathcal{F}, T) \in \mathbb{R}^{|\mathcal{F}|}$. For $i \in \{1, 2, \ldots, |\mathcal{F}|\}$, its entries are defined by

$$\text{Dev}(p, \mathcal{F}, T)_i = \begin{cases} \frac{\text{eval}(p,\mathcal{F},T)_i}{f_i^{\text{max}}} - 1 & \text{if eval}(p, \mathcal{F}, T)_i > f_i^{\text{max}} \\ \frac{\text{eval}(p,\mathcal{F},T)_i}{f_i^{\text{min}}} - 1 & \text{if eval}(p, \mathcal{F}, T)_i < f_i^{\text{min}} \\ 0 & \text{else} \end{cases} .$$

In the deviation vector, positive values indicate that a property exceeded the maximal allowed value, negative values indicate that a property was below the minimal required value, and a value of zero indicates constraint fulfillment.

Quality Penalty Based on the total absolute and maximum deviation of a primer from the constraints, the quality penalty associated with selecting the primer p can be defined using a tuning parameter $0 \leq \alpha \leq 1$:

$$\text{Pen}(p, \mathcal{F}, T, \alpha) = \alpha ||\text{Dev}(p, \mathcal{F}, T)||_\infty + (1 - \alpha) ||\text{Dev}(p, \mathcal{F}, T)||_1 .$$

Here, $||x||_\infty = \max_{i=\{1,\ldots,n\}} |x_i|$ is the maximum norm and $||x||_1 = \sum_{i=1}^n |x_i|$ is the L1-norm. The maximum norm ensures that extreme deviations from individual constraints are penalized, while the L1-norm ensures that overall deviations across all constraints are considered. The parameter α defines the trade-off between the maximum norm and the L1-norm. For $\alpha \to 1$, the maximum deviation dominates giving rise to a local penalty reflecting the largest absolute deviation for a single constraint. For $\alpha \to 0$, the total deviation dominates giving rise to a global penalty reflecting the sum of all absolute constraint deviations. By default α is set to 0.5. This setting ensures that among two primers with the same total constraint deviation, the primer with the smaller maximum deviation is less penalized. The quality penalty that is associated with the selection of a primer is a useful quantity for performing global primer design (Section 6.3.5).

6.3 Primer Design and Selection Algorithms

This section introduces algorithms for designing and selecting mPCR primer sets. The local primer design procedure (Section 6.3.4) is based on the subsequent steps of initialization (Section 6.3.1), filtering (Section 6.3.2), and optimization of primers (Section 6.3.3). The global primer design algorithm (Section 6.3.5) goes through the same steps but replaces the filtering procedure with a scoring procedure. An approach for checking the feasibility of a primer design problem is discussed in Section 6.3.6. An algorithm for subsetting already optimized primer sets is provided in Section 6.3.7.

6.3.1 Primer Initialization

The goal of primer initialization is the allocation of a candidate set of primers that fulfills the following three properties. First, the primers should be sufficiently complementary to the target binding region in the templates. Second, the primers should have appropriate lengths. Third, the primers should carry ambiguous nucleotides at positions that allow for increasing the number of EAEs of the primer.

Since initialization procedures for degenerate primers have different requirements, I implemented two strategies for primer initialization: nondegenerate and degenerate primer initialization. While the first strategy is simple and fast, the second strategy is more intricate and time-consuming. Both procedures require the input of a set of templates T with annotated binding regions as well as the permissible lengths $(l_{\text{min}}, l_{\text{max}})$ of the primers.

Nondegenerate Primer Initialization The initialization of nondegenerate candidate primers is achieved by enumerating all possible substrings with length $l \in [l_{\text{min}}, l_{\text{max}}]$ for each template binding region as illustrated in Algorithm 1.

Degenerate Primer Initialization The initialization of degenerate primers is afforded by a clustering-based strategy that consists of the five steps that are performed by the initPrimersDegenerate procedure, which is described in Algorithm 2. In the first step, the template binding regions are aligned using MAFFT[53] (line 2), which is a fast and accurate tool for the determination of multiple sequence alignments[54,55]. Next, for each subalignment of a given primer length l, the following steps are performed. The dissimilarities of all sequences in a subalignment are computed based on their Hamming distance (line 12). Given two strings $x = (x_1, \ldots, x_n) \in \Sigma^n$ and

[53] Kuraku et al. 2013

[54] Thompson et al. 2011

[55] The generated degenerate primers depend on the quality of the alignment.

Algorithm 1: Nondegenerate primer initialization.

```
Input:
    Set of templates $T$, range of allowed primer lengths $(l_{\text{min}}, l_{\text{max}})$
Output:
    An initial set of primer strings $P$
 1: procedure INITPRIMERSNAIVE($T$, $(l_{\text{min}}, l_{\text{max}})$)
 2:     $P \leftarrow \emptyset$
 3:     for $t$ in $T$ do
 4:         $t \leftarrow \text{getBindingRegion}(t)$
 5:         for $s \in \{1, 2, \ldots, |t| - l_{\text{max}} + 1\}$ do           ▷ Start position
 6:             for $l \in \{l_{\text{min}}, l_{\text{min}} + 1, \ldots, l_{\text{max}}\}$ do   ▷ Primer length
 7:                 $e \leftarrow s + l - 1$                                  ▷ End position
 8:                 if $e > |t|$ then
 9:                     break
10:                 else
11:                     $p \leftarrow t.\text{substring}(s, e)$
12:                     $P \leftarrow P \cup p$
13:     return $P$
```

$y = (y_1, \ldots, y_n) \in \Sigma^n$, their Hamming distance is defined as

$$\Delta(x, y) = |\{i \in \{1, \ldots, n\} | x_i \neq y_i\}|$$

and the corresponding dissimilarity is $d(x, y) = \frac{\Delta(x,y)}{n}$. Based on the dissimilarities, complete-linkage hierarchical clustering (Section 3.5.2) is performed in order to construct a dendrogram representing the relationship of the subsequences (line 13). Then, the GENERATEDEGENERATEPRIMERS procedure is called, which processes the leaves of the dendrogram (line 21).

Starting from the leaves l, the procedure iterates towards the root of the tree in the following way. Line 22 determines the degeneracy of the consensus sequences of the ungapped sequences associated with l and its descendants as

$$\text{degen}(s) = \prod_{i=1}^{|s|} |\text{disambig}(s_i)|\,,$$

where $\text{disambig}(s_i)$ disambiguates the IUPAC nucleotide s_i to a set of conventional nucleotides. Only consensus sequences with $\text{degen}(s)$ below a cutoff, n_{degen} (default: 16), are considered as primer candidates in order to exclude highly degenerate primers (e.g. primers with many amibiguous positions). The iteration for a single leaf terminates either when the degeneracy of the consensus sequence resulting from the current node, l, exceeds n_{degen} (line 22) or when the root of the dendrogram is reached (line 26).

Since ungapped sequences of length smaller than l_{min} have to be discarded, it is possible that the generated primers do not cover all target regions. To ensure that the alignment-based initialization strategy can obtain the same coverage as the nondegenerate initialization approach, line 16 augments the set of degenerate primers by the set of nondegenerate primers.

6.3.2 Primer Filtering

In order to select only primers with favorable physicochemical properties, a filtering procedure is applied to the initial set of primer candidates. The filtering procedure iterates over each filtering constraint[56] and eliminates a primer from the set as soon as a constraint is broken. The selection of high-quality primers is not the only purpose of the filtering procedure. Since the filtering procedure removes a large number of candidate primers, it is critical for rendering the optimization procedure, which is performed subsequently, computationally feasible.

If the target coverage (typically 100%) cannot be reached because too many primer candidates were excluded during the filtering phase, the constraints are relaxed in order to retain primers that cover missing templates. For this purpose, Algorithm 3 couples filtering of primers with a procedure that relaxes the constraints until the target coverage can be reached. The FILTERPRIMERS procedure executes a while loop that performs the following steps. First, the primers are filtered according the current constraints by calling the FILTERPRIMERSNAIVE procedure (line 5). Note that line 16 of this procedure implicitly evaluates the property of the primer associated with a given constraint if it has not been determined yet and otherwise retrieves the previously computed value. Second, the constraints are relaxed using the RELAXCONSTRAINTS procedure (line 23), which relaxes a constraint only if there exists a primer that provides additional coverage and violates the current constraint.

The extent to which a constraint is relaxed in the RELAXCONSTRAINTS procedure is determined by the target constraints and the constraint limits, which are included in the set of filtering constraints, F. While target constraints indicate the desired properties of primers, constraint limits indicate properties that would not be ideal but still acceptable. The relaxation is performed by the RELAX procedure (line 28) in which the current constraint is adjusted according to the difference between the initially defined constraint limit and the target value. To showcase a relaxation, consider the entry for GC ratio in Table 6.3, which defines a target range of 40%–60% and a limit range of 30%–70%. Thus, if required for reaching the target coverage, the GC

[56] The constraints that can be considered in the filtering procedure via the set of filtering constraints F are those entries in Table 6.3 without an asterisk.

Algorithm 2: Degenerate primer initialization.

```
Input:
    Set of templates $T$, range of allowed primer lengths $(l_{\text{min}}, l_{\text{max}})$,
    maximum degeneracy per primer $n_{\text{degen}}$
Output:
    An initial set of primers $P$
procedure initPrimersDegenerate($T$, $(l_{\text{min}}, l_{\text{max}}), n_{\text{degen}}$)
    $A \leftarrow$ alignBindingRegions($T$)
    $|A| \leftarrow$ lengthOfAlignment($A$)
    $P \leftarrow \emptyset$
    for $s \in \{1, 2, \ldots, |A| - l_{\text{max}} + 1\}$ do        ▷ Start of subalignment
        for $l \in \{l_{\text{min}}, l_{\text{min}} + 1, \ldots, l_{\text{max}}\}$ do                 ▷ Length
            $e \leftarrow s + l - 1$                                ▷ End of subalignment
            if $e > |A|$ then
                break
            else
                $A_{[s,e]} \leftarrow A$.getSubAlignment($s, e$)
                $d \leftarrow$ computeHammingDissimilarities($A_{[s,e]}$)
                $H \leftarrow$ computeHierarchicalClustering($d$)
                $p \leftarrow$ generateDegeneratePrimers($H, n_{\text{degen}}$)
                $P \leftarrow P \cup p$
    $P \leftarrow P \cup$ initPrimersNaive($T, (l_{\text{min}}, l_{\text{max}})$)
    return $P$

procedure generateDegeneratePrimers($H$, $n_{\text{degen}}$)
    $P \leftarrow \emptyset$
    for each $l \in$ leaves($H$) do
        while degen(computeConsensus($l$)) $\leq n_{\text{degen}}$ do
            $p \leftarrow$ computeConsensus($l$)
            $P \leftarrow P \cup p$
            $l \leftarrow$ getAncestor($l$)
            if $l$ is NONE then
                break
    return $P$
```

ratio constraint would be relaxed from 40%–60% to 30%–70% in the first relaxation, from 30%–70% to 20%–80% in the second relaxation, and so on. Since reaching the target coverage has a higher priority than primer quality, breaches of the constraint limits are allowed by default.

6.3.3 *Primer Set Optimization*

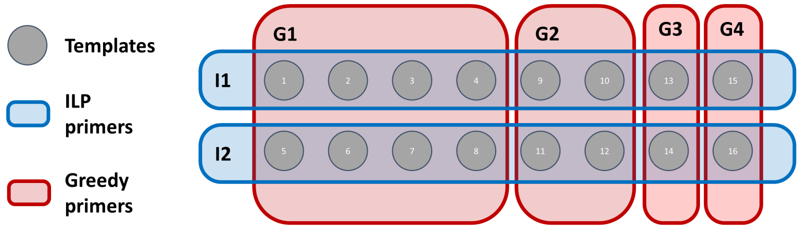


Figure 6.2: The set cover problem. Circles represent templates and primers are indicated by rectangles. Templates that are enclosed by a primer are considered to be covered by that primer. The minimum primer set covering all templates, which can be determined by an integer linear program, is $\{I1, I2\}$. An approximate solution consisting of $\{G1, G2, G3, G4\}$ is obtainable with a greedy algorithm.

This figure is based on an example provided by Pearson et al. (1996).

openPrimeR provides a greedy algorithm and an ILP formulation for solving the SCP. The practical difference between these two approaches is that an ILP can find the optimal solution, while a greedy algorithm may only find an approximate solution. Figure 6.2 illustrates the set cover problem for mPCR primer design using a toy example with 16 templates, that is, $T = \{t_1, t_2, \ldots, t_{16}\}$ and 6 primers, that is, $P = \{G1, G2, G3, G4, I1, I2\}$. Assume the primers have the following amplification events:

$$\begin{aligned}
G1_{\triangleright T} &= \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\} \\
G2_{\triangleright T} &= \{t_9, t_{10}, t_{11}, t_{12}\} \\
G3_{\triangleright T} &= \{t_{13}, t_{14}\} \\
G4_{\triangleright T} &= \{t_{15}, t_{16}\} \\
I1_{\triangleright T} &= \{t_1, t_2, t_3, t_4, t_9, t_{10}, t_{13}, t_{15}\} \\
I2_{\triangleright T} &= \{t_5, t_6, t_7, t_8, t_{11}, t_{12}, t_{14}, t_{16}\}
\end{aligned}$$

An ILP would find the optimal solution consisting of the two primers $I1 \cup I2$, while a greedy algorithm would only find the approximate solution consisting of the four primers, $G1 \cup G2 \cup G3 \cup G4$. This is because the greedy approach, at every iteration, selects the primer providing the maximal gain in coverage. Hence, the algorithm can get stuck in a local minimum. To illustrate this, consider the following iterations of a greedy algorithm:

Algorithm 3: Filtering of primers.

Input:

Set of templates T, set of primers P, filtering constraints F, target coverage $c \geq 0$

Output:

Reduced set of primers fulfilling the (relaxed) filtering constraints F

```
procedure filterPrimers($T$, $P$, $F$, $c$)
    $c \leftarrow \min(c, |P_{\triangleright T}|)$                ▷ Number of templates to cover
    $P_F \leftarrow \emptyset$
    while $|P_{F \triangleright T}| < c$ do
        $P_F \leftarrow$ filterPrimersNaive$(T, P, F)$
        $P_E \leftarrow P \setminus P_F$                         ▷ Excluded primers
        $T_M \leftarrow P_{\triangleright T} \setminus P_{F \triangleright T}$                   ▷ Missing templates
        $P_C \leftarrow T_{M \triangleleft P_E}$   ▷ Excluded primers with additional coverage
        $F \leftarrow$ relaxConstraints$(F, P_C)$
    return $P_F$

procedure filterPrimersNaive($T$, $P$, $F$)
    $P_F \leftarrow P$
    for $f$ in $F$ do
        for $p$ in $P_F$ do
            if $p \not\models f$ then
                $P_F \leftarrow P_F \setminus p$
    return $P_F$

procedure relaxConstraints($F$, $P$)
    for $f$ in $F$ do
        if $\exists p \in P \not\models f$ then
            $f \leftarrow$ relax$(f)$
    return $F$

procedure relax($f$)
    $(\Delta_{\min}, \Delta_{\max}) \leftarrow \text{initialLimit}(f) - \text{initialSetting}(f)$     ▷ Step size
    $f \leftarrow f + (\Delta_{\min}, \Delta_{\max})$
    return $f$
```

1. Initially (Figure 6.3), the selection of either $G1$, $I1$, or $I2$ would lead to the maximal coverage gain of 8. Thus, the algorithm may arbitrarily select $G1$, which has $G1_{\triangleright T} = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}$. Due to the selection of $G1$, it is no longer necessary to consider coverage events relating to the templates t_1, t_2, t_3, t_4, t_5, t_6, t_7, or t_8.

2. Based on the updated coverage (Figure 6.4), the selection of either $G2$, $I1$, or $I3$ would lead to the maximal coverage gain of 4. Thus, the algorithm may select primer $G2$. Since $G2$ provides coverage of t_9, t_{10}, t_{11}, t_{12}, these templates do no longer have to be considered.

3. Based on the updated coverage (Figure 6.5), the selection of either $G3$, $I1$, or $I2$ would lead to the maximal coverage gain of 2. Hence, the algorithm may select primer $G3$. As $G3$ covers t_{13} and t_{14}, these templates no longer need to be considered.

4. Based on the updated coverage (Figure 6.6), $G4$ has the maximal coverage gain of 2 and is therefore selected. As $G2$ covers t_{15} and t_{16}, these templates no longer have to be considered (Figure 6.7). The greedy procedure terminates after this step since no additional coverage could be gained by considering additional primers.

$$\begin{aligned}
G1_{\triangleright T} &= \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}\\
G2_{\triangleright T} &= \{t_9, t_{10}, t_{11}, t_{12}\}\\
G3_{\triangleright T} &= \{t_{13}, t_{14}\}\\
G4_{\triangleright T} &= \{t_{15}, t_{16}\}\\
I1_{\triangleright T} &= \{t_1, t_2, t_3, t_4, t_9, t_{10}, t_{13}, t_{15}\}\\
I2_{\triangleright T} &= \{t_5, t_6, t_7, t_8, t_{11}, t_{12}, t_{14}, t_{16}\}
\end{aligned}$$

Figure 6.3: Greedy state 0.

$$\begin{aligned}
G1_{\triangleright T} &= \{\}\\
G2_{\triangleright T} &= \{t_9, t_{10}, t_{11}, t_{12}\}\\
G3_{\triangleright T} &= \{t_{13}, t_{14}\}\\
G4_{\triangleright T} &= \{t_{15}, t_{16}\}\\
I1_{\triangleright T} &= \{t_9, t_{10}, t_{13}, t_{15}\}\\
I2_{\triangleright T} &= \{t_{11}, t_{12}, t_{14}, t_{16}\}
\end{aligned}$$

Figure 6.4: Greedy state 1.

$$\begin{aligned}
G1_{\triangleright T} &= \{\}\\
G2_{\triangleright T} &= \{\}\\
G3_{\triangleright T} &= \{t_{13}, t_{14}\}\\
G4_{\triangleright T} &= \{t_{15}, t_{16}\}\\
I1_{\triangleright T} &= \{t_{13}, t_{15}\}\\
I2_{\triangleright T} &= \{t_{14}, t_{16}\}
\end{aligned}$$

Figure 6.5: Greedy state 2.

$$\begin{aligned}
G1_{\triangleright T} &= \{\}\\
G2_{\triangleright T} &= \{\}\\
G3_{\triangleright T} &= \{\}\\
G4_{\triangleright T} &= \{t_{15}, t_{16}\}\\
I1_{\triangleright T} &= \{t_{15}\}\\
I2_{\triangleright T} &= \{t_{16}\}
\end{aligned}$$

Figure 6.6: Greedy state 3.

$$\begin{aligned}
G1_{\triangleright T} &= \{\}\\
G2_{\triangleright T} &= \{\}\\
G3_{\triangleright T} &= \{\}\\
G4_{\triangleright T} &= \{\}\\
I1_{\triangleright T} &= \{\}\\
I2_{\triangleright T} &= \{\}
\end{aligned}$$

Figure 6.7: Greedy state 4.

Note that, primer optimization necessitates that two additional constraints are included in the SCP: the maximum allowed melting temperature difference, $\Delta T_m^{\text{max}} \geq 0$, and the minimum allowed free energy of cross dimerization, $\Delta G_{\text{cd}}^{\text{min}} \leq 0$. In the following, the greedy and the ILP-based optimization strategies are presented.

Formulation as a Greedy Algorithm Algorithm 4 illustrates the greedy procedure, optimizeGreedy, for the design of multiplex PCR primers. The algorithm constructs a small set of primers according to the following criterion: *In every iteration, select a compatible primer that provides the greatest gain in coverage until the maximal possible coverage has been reached*. A primer is considered compatible if it fulfills the constraints on the permissible melting temperature difference and on the free energy of cross dimerization. These constraints ensure that only non-dimerizing primers with suitable melting temperatures are combined into the same set. To ensure that primer combinations exhibit similar melting temperatures, primer subsets within a melting temperature range $[t_i, t_j]$, that is, $P_{(t_i, t_j)}$ with $t_j - t_i \leq \Delta T_m$, are constructed by the createTemperatureRanges procedure (line 2). Cross-dimerizing primers are prevented by the greedyChoice procedure, which selects only non-dimerizing primers (line 55).

The createTemperatureRanges procedure selects suitable melting temperatures using the following approach. First, the createTemperatureRangesNaive procedure determines evenlyspaced temperatures whose subsequent differences are below ΔT_m (line 35). Second, the isCoverageSufficient procedure determines whether any of the primer sets that are induced by the selected melting temperatures reach the target coverage. While this is not the case (line 16), the maximal allowed melting temperature difference, ΔT_m, is relaxed (line 17) and a new array of melting temperatures is constructed via createTemperatureRangesNaive (line 18). Finally, the constructed array of melting temperatures is returned.

The optimizeGreedy procedure performs the following computations until at least one primer set whose coverage exceeds the target coverage has been found (line 4). First, the cross-dimerization matrix, D, is computed (line 5). Second (line 6), for each melting temperature, t_i, greedy primer selection is performed via greedySet. If a primer set with sufficient coverage is found, it is stored in the list of primer sets, R (line 9). Third, once all melting temperatures have been considered, the cross-dimerization constraint is relaxed (line 10).

Since the optimizeGreedy procedure generates small primer sets for several melting temperature ranges, line 11 determines the smallest set with the greatest coverage among all constructed primer sets. To exemplify this point, assume that the procedure constructs three primer sets with optimal annealing temperatures at 50° C, 55° C, and 60° C, denoted by P^{50°, P^{55°, and P^{60°, respectively. Further assume $|T| = 150$ and let the sets have coverages of $|P^{50^\circ}_{\triangleright T}| = 120$, $|P^{55^\circ}_{\triangleright T}| = 148$, and $|P^{60^\circ}_{\triangleright T}| = 130$. Then, primer set P^{55° is selected because it has the largest coverage (148 of 150 templates) among all constructed sets.

The greedySet procedure works as follows. Only primers whose melting temperature is in the current melting temperature range, $[t_1, t_2]$, are considered (line 41). Then, primers are selected according to the greedy criterion until the maximal possible coverage is obtained (line 42). When the greedyChoice procedure has selected a primer according the greedy criterion, the primer is removed from the candidate set of primers, added to the selected set of primers, and the coverage of the remaining primers is updated using the updateCoverage procedure (line 48), which discounts the templates that are covered by the selected primer. If no primer could be selected (i.e. no compatible primer could be found), the procedure terminates (line 44).

The greedyChoice procedure, which selects a primer according to the greedy criterion, is implemented as follows. After sorting the primers according to decreasing coverage (line 52), the non-dimerizing primer providing the greatest gain in coverage is selected by considering the entries of the dimerization matrix **D** (line 55). If no primer fulfilling the dimerization constraint could be found, the procedure returns NONE.

Theoretically, the size of a greedy set cover is at most approximately $\ln(|T|)$ times larger than the size of the minimum cover[57]. This upper bound also holds for the presented greedy primer design strategy. However, in this case, the size of the minimum cover relates to the minimum primer set that fulfills the (relaxed) filtering constraints as well as the (relaxed) optimization constraints.

Formulation as an Integer Linear Program Multiplex PCR primer design can be formulated as an ILP in the following manner. First, we define the indicator vector $x \in \{0,1\}^{|P|}$ according to its entries

$$x_i = \begin{cases} 1 & \text{if primer } p_i \in P \text{ is selected} \\ 0 & \text{if primer } p_i \in P \text{ is not selected} \end{cases}$$

such that entry x_i indicates whether the candidate primer p_i is part of the optimal primer set. Further, the coverage information is summarized in the coverage matrix $\mathbf{C} \in \{0,1\}^{|T| \times |P|}$, which is defined by its entries

$$c_{ij} = \begin{cases} 1 & \text{if } p_j \triangleright t_i \\ 0 & \text{else} \end{cases} .$$

Based on these definitions, the primer selection ILP can be written as

$$\begin{aligned} \min \sum_{i=1}^{N} x_i & \quad \text{Minimize number of primers} && (6.1) \\ \text{subject to} & && (6.2) \\ \mathbf{C}x \geq 1 & \quad \text{Cover each template} && (6.3) \\ (x_i + x_j)\mathbf{D}_{p_i,p_j} \leq 1 \, \forall p_i, p_j \in P & \quad \text{Prevent dimers} && (6.4) \end{aligned}$$

In this formulation, dimers are prevented using side constraint 6.4, which ensures that never both of two dimerizing primers are selected[58]. In order to keep the number of side constraints manageable, the melting temperature constraint is not explicitly modeled. Instead, similarly to the greedy approach, the ILP is solved for primer subsets at various melting temperature ranges, as depicted in Algorithm 5. To select an optimal set of primers through an ILP, the OPTIMIZEILP procedure is used. Similarly to the greedy approach, the ILP procedure uses CREATETEMPERATURERANGES to select appropriate melting temperatures (line 2). The procedure then performs the following steps until an optimal primer set has been found.

After the dimerization matrix, $\mathbf{D}$, has been calculated (line 5), for each selected melting temperature (line 6), a solution to the primer design ILP is determined by calling the *solveILP* function (line 7), which uses the exact branch-and-bound implementation of

[57] Young 2008

[58] Bashir et al. 2007

Algorithm 4: Greedy primer optimization.

Input:

Set of templates T, set of primers P, melting temperature constraint ΔT_m, dimerization constraint ΔG_{cd}, target coverage $c \geq 0$

Output:

Approximation of a minimum primer set $\hat{P}^*$ with maximum possible coverage of template sequences T subject to the temperature constraint ΔT_m and the dimerization constraint ΔG_{cd}

```
procedure OPTIMIZEGREEDY($T$, $P$, $\Delta T_m$, $\Delta G_{\text{cd}}$, $c$)
    $[t_1, \dots, t_N] \leftarrow$ CREATETEMPERATURERANGES$(P, T, \Delta T_m, c)$
    $R \leftarrow \emptyset$                ▷ Optimal primer sets for temperature ranges
    while $R = \emptyset$ do
        $\mathbf{D} \leftarrow (d_{p_i,p_j}) = \begin{cases} 1 & \text{if } \Delta G(p_i, p_j) < \Delta G_{\text{cd}}^{\text{min}} \\ 0 & \text{else} \end{cases}$
        for $i \in [1, \dots, N-1]$ do
            $\hat{P}^* \leftarrow$ GREEDYSET$(T, P, [t_i, t_{i+1}], \mathbf{D})$        ▷ Best set for [$t_i$, $t_{i+1}$]
            if $|\hat{P}^*_{\triangleright T}| \geq c$ then
                $R \leftarrow R \cup \{\hat{P}^*\}$
        $\Delta G_{\text{cd}} \leftarrow$ RELAX$(\Delta G_{\text{cd}})$
    $\hat{P}^* \leftarrow$ SELECTBESTSET$(R, T)$ ▷ Best set across all temperatures
    return $\hat{P}^*$

procedure CREATETEMPERATURERANGES($P$, $T$, $\Delta T_m$, $c$)
    $[t_1, \dots, t_N] \leftarrow$ CREATETEMPERATURERANGESNAIVE$(P, \Delta T_m)$
    while $\neg$ISCOVERAGESUFFICIENT$(P, T, [t_1, \dots, t_N], c)$ do
        $\Delta T_m \leftarrow$ RELAX$(\Delta T_m)$
        $[t_1, \dots, t_N] \leftarrow$ CREATETEMPERATURERANGESNAIVE$(P, \Delta T_m)$
    return $[t_1, \dots, t_N]$

procedure ISCOVERAGESUFFICIENT($P$, $T$, $[t_1, \dots, t_N]$, $c$)
    $c_{\text{OK}} \leftarrow$ FALSE
    for $i \in [1, \dots, N-1]$ do
        $P_{(t_i, t_{i+1})} \leftarrow \{p | p \in P, t_i \leq T_m(p) \leq t_{i+1}\}$
        if $|P_{(t_i, t_{i+1})_{\triangleright T}}| \geq c$ then ▷ Set with sufficient coverage found
            $c_{\text{OK}} \leftarrow$ TRUE
            break
    return $c_{\text{OK}}$
```

Algorithm 4: Greedy primer optimization (continued).

```
procedure createTemperatureRangesNaive($P$, $\Delta T_m$)
    $T = [\,]$                                            ▷ Array with temperatures
    $(t_{\min}, t_{\max}) \leftarrow (\min_{p_i \in P} T_m(p_i), \max_{p_i \in P} T_m(p_i))$
    $t_i \leftarrow t_{\min}$
    while $t_i < t_{\max}$ do
        $T$.append($t_i$)
        $t_i \leftarrow t_i + \Delta T_m^{\max}$
    $T$.append($t_{\max}$)
    return $T$

procedure greedySet($T$, $P$, $(t_1, t_2)$, $\mathbf{D}$)
    $\hat{P}^* \leftarrow \emptyset$                                  ▷ Selected primer set
    $P_{(t_1,t_2)} \leftarrow \{p \in P | t_2 \geq T_m(p) \geq t_1\}$
    while $|\hat{P}^*_{\triangleright T}| \neq |P_{\triangleright T}|$ do
        $p \leftarrow$ greedyChoice($T, P_{(t_1,t_2)}, \mathbf{D}, \hat{P}^*$)
        if $p =$ none then
            return $\hat{P}^*$
        $P_{(t_1,t_2)} \leftarrow P_{(t_1,t_2)} \setminus \{p\}$
        $\hat{P}^* \leftarrow \hat{P}^* \cup \{p\}$
        $P_{(t_1,t_2)} \leftarrow$ updateCoverage($P_{(t_1,t_2)}, T, p$)
    return $\hat{P}^*$

procedure greedyChoice($T$, $P$, $\mathbf{D}$, $\hat{P}^*$)
    $P \leftarrow$ sortByDecreasingCoverage($P$)
    $p \leftarrow$ none                                         ▷ Selected primer
    for each $p_i \in P$ do
        $n_{\text{dimers}} \leftarrow \sum_{j=1}^{|\hat{P}^*|} \mathbf{D}_{p_i,p_j}$
        if $n_{\text{dimers}} = 0$ then
            $p \leftarrow p_i$
            break
    return $p$
```

Algorithm 4: Greedy primer optimization (continued).

```
61: procedure SELECTBESTSET($R$, $T$)
62:     $\hat{P}^* \leftarrow$ NONE                                  ▷ The best primer set
63:     $C_{\max} = 0$                                             ▷ The highest coverage
64:     for each $R_i \in R$ do
65:         $C_i \leftarrow R_{i \triangleright T}$
66:         if $(C_i > C_{\max})$ or $(C_i = C_{\max}$ and $|R_i| < |\hat{P}^*|)$ then
67:             $\hat{P}^* \leftarrow R_i$
68:             $C_{\max} \leftarrow C_i$
69:     return $\hat{P}^*$
70:
71: procedure UPDATECOVERAGE($P$, $T$, $p$)
72:     for each $p_i \in P$ do
73:         $p_{i \triangleright T} \leftarrow p_{i \triangleright T} \setminus p_{\triangleright T}$
74:     return $P$
```

lpsolve[59]. If the constructed set reaches the target coverage (line 8), it is included in the set of optimal primer sets, R. Once all melting temperatures have been considered, the cross-dimerization constraint is relaxed (line 10). Finally, in line 11, the best set across all temperature ranges is returned, in the same way as described for the greedy algorithm. Note that Lagrangian-based heuristics for solving the set cover problem[60], which could considerably improve the runtime, are not applicable to the formulated ILP due to the dimerization constraint.

Reporting Results for Forward and Reverse Primers When both forward and reverse primers are designed, pairs of forward and reverse primers are formed for clarifying the results. Pairing primers is possible only after the optimization procedure because designed primers have passed stringent quality control criteria and should therefore be compatible with each other. To pair primers of contrary orientation, the following procedure is used. First, all combinations of forward and reverse primers are generated. Second, the EAEs of primers constituting a pair are intersected. Pairs whose EAE intersection is empty are removed. Third, since the remaining pairs may exhibit redundancies (i.e. they may cover only templates that are already covered by the other primer pairs), a set cover ILP (Problem 6.1 without the dimerization constraint) is solved to find the minimum combination of primer pairs that retains the overall coverage.

[60] Caprara et al. 1999

Algorithm 5: ILP primer optimization.

```
Input:
    Set of templates $T$, set of primers $P$, melting temperature constraint $\Delta T_m$, dimerization constraint $\Delta G_{\text{cd}}$, target coverage $c \geq 0$
Output:
    A minimal primer set $\hat{P}^*$ fulfilling the filtering constraints $F$ with maximal possible coverage of template sequences $T$ subject to the temperature constraint $\Delta T_m$ and the dimerization constraint $\Delta G_{\text{cd}}$
procedure OPTIMIZEILP($T$, $P$, $\Delta T_m$, $\Delta G_{\text{cd}}$, $c$)
    $[t_1, \ldots, t_N] \leftarrow$ CREATETEMPERATURERANGES$(P, T, \Delta T_m, c)$
    $R \leftarrow \emptyset$          ▷ Optimized primer sets for temperature ranges
    while $R = \emptyset$ do
        $\mathbf{D} \leftarrow (d_{p_i, p_j}) = \begin{cases} 1 & \text{if } \Delta G(p_i, p_j) < \Delta G_{\text{cd}}^{\text{min}} \\ 0 & \text{else} \end{cases}$
        for $i \in [1, \ldots, N-1]$ do
            $\hat{P}^* \leftarrow \text{solveILP}(T, P_{[t_i, t_{i+1}]}, \mathbf{D})$        ▷ Best set for $[t_i, t_{i+1}]$
            if $|\hat{P}^*_{\triangleright T}| \geq c$ then
                $R \leftarrow R \cup \{\hat{P}^*\}$
        $\Delta G_{\text{cd}} \leftarrow$ RELAX$(\Delta G_{\text{cd}})$
    $\hat{P}^* \leftarrow$ SELECTBESTSET$(R, T)$ ▷ Best set across all temperatures
    return $\hat{P}^*$
```

6.3.4 *Local Primer Design*

Based on the algorithms for initializing (Algorithms 1 and 2), filtering (Algorithm 3), and optimizing (Algorithm 4 and 5) primers, the local primer design procedure (Algorithm 6) determines an optimum set of primers for a single DNA strand via DESIGNPRIMERSSINGLE. This procedure first initializes a set of primers using INITPRIMERS, filters the primers using FILTERPRIMERS, and then optimizes the primers using OPTIMIZEPRIMERS. The INITPRIMERS procedure provides an option for initializing either nondegenerate primers via INITPRIMERSNAIVE (line 12) or degenerate primers via INITPRIMERSDEGENERATE (line 10). The procedure OPTIMIZEPRIMERS solves the optimization problem either using the greedy formulation provided by the OPTIMIZEGREEDY procedure (line 20) or the ILP formulation provided by the OPTIMIZEILP procedure (line 18).

The local primer design strategy for a single DNA strand is sufficient for designing primers targeting immunoglobulin variable regions because this task requires only the design of forward primers for the variable region; reverse primers for the constant region do not need to be optimized due to the high conservation of this region. In general, primer design is concerned with determining both forward and reverse primers. Algorithm 7, which presents the DESIGNPRIMERS procedure, is concerned with this problem. The procedure requires an argument *d* that specifies whether only forward primers, only reverse primers, or primers of both orientations should be designed. If primers of a single orientation are designed, the DESIGNPRIMERSSINGLE procedure is called by providing the templates in the appropriate orientation according to *d* (line 4). Otherwise, an optimal set of forward primers is constructed first (line 7) and then an optimal set of reverse primers is determined (line 9). Finally, the sets of forward and reverse primers are combined (line 10).

6.3.5 Global Primer Design

For small instances of the primer design problem (e.g. for few templates and short binding regions), it is not necessary to filter the primers before the optimization procedure commences. Algorithm 8 demonstrates such a global primer design strategy via the DESIGNPRIMERSGLOBALSINGLE procedure.[61]

The global primer design approach replaces the filtering procedure of the local primer design approach with SCOREPRIMERS, a procedure that determines the quality penalty associated with every primer (line 4, as described in Section 6.2.4). Primer sets with varying deviations from the constraints are generated by adjusting the maximum allowed quality penalty ξ (line 6). For each ξ, a subset of primers whose penalties do not exceed ξ is constructed and optimized (line 8). Finally, the procedure returns a list containing small primer sets fulfilling the constraints to varying degrees. The choice of a suitable primer set can be facilitated by considering the primer quality penalty with respect to coverage and set size. Because the global primer design approach can be adapted to multiple strands in the same way as the local strategy (Algorithm 7), the analogous implementation for the global strategy is not described in this dissertation.

The benefit of the global primer design approach is that it does not merely return a single primer set fulfilling the constraints but allows for designing primer sets fulfilling the constraints to differing degrees. Based on the results, it is possible to find a suitable tradeoff between relatively large primer sets with small deviations from the target constraints and relatively small primer sets with large deviations from the target constraints.

6.3.6 Feasibility of Primer Design

Since computational primer design is time-consuming, it is useful to have an approach for estimating whether it is possible to design a set

[61] This approach is called global because the primer search space is not limited by the filtering procedure.

Algorithm 6: Primer design for a single strand.

Input:

Set of templates T, filtering constraints F, melting temperature constraint ΔT_m, dimerization constraint ΔG_{cd}, $o_O \in \{\textsc{Greedy}, \text{ILP}\}$, $o_I \in \{\textsc{Naive}, \textsc{Degenerate}\}$, desired primer lengths $(l_{\min}, l_{\max})$, target coverage $c \geq 0$, maximal primer degeneracy n_{degen}

Output:

A minimum primer set $\hat{P}^*$ fulfilling the (relaxed) filtering constraints F with maximum possible coverage of template sequences T subject to the temperature constraint ΔT_m and the dimerization constraint ΔG_{cd}

```
procedure designPrimersSingle($T$, $F$, $\Delta T_m$, $\Delta G_{\text{cd}}$, $o_O$, $o_I$, $(l_{\min}, l_{\max})$, $c$, $n_{\text{degen}}$)
    $P \leftarrow$ initPrimers$(T, o_I, (l_{\min}, l_{\max}), n_{\text{degen}})$
    $P_F \leftarrow$ filterPrimers$(T, P, F, c)$
    $\hat{P}^* \leftarrow$ optimizePrimers$(T, P_F, \Delta T_m, \Delta G_{\text{cd}}, o_O, c)$
    return $\hat{P}^*$

procedure initPrimers($T$, $o_I$, $(l_{\min}, l_{\max})$, $n_{\text{degen}}$)
    $T_b \leftarrow$ getBindingRegions$(T)$
    if $o_I$ = Degenerate then
        $P \leftarrow$ initPrimersDegenerate$(T, (l_{\min}, l_{\max}), n_{\text{degen}})$
    else
        $P \leftarrow$ initPrimersNaive$(T, (l_{\min}, l_{\max}))$

procedure optimizePrimers($T$, $P$, $\Delta T_m$, $\Delta G_{\text{cd}}$, $o_O$, c)
    $c \leftarrow \min(c, |P_{\triangleright T}|)$                ▷ Number of templates to cover
    $\hat{P}^* \leftarrow$ none
    if $o_O = \text{ILP}$ then
        $\hat{P}^* \leftarrow$ optimizeILP$(T, P, \Delta T_m, \Delta G_{\text{cd}}, c)$
    else
        $\hat{P}^* \leftarrow$ optimizeGreedy$(T, P, \Delta T_m, \Delta G_{\text{cd}}, c)$
    return $\hat{P}^*$
```

Algorithm 7: Generalized primer design procedure.

```
Input:
    Set of templates $T$ with forward ($T_{\text{fw}}$) and reverse binding regions ($T_{\text{rev}}$), filtering constraints $F$, melting temperature constraint $\Delta T_m$, dimerization constraint $\Delta G_{\text{cd}}$, $o_O \in \{\textsc{Greedy}, \text{ILP}\}$, $o_I \in \{\textsc{Naive}, \textsc{Degenerate}\}$, desired primer lengths $(l_{\min}, l_{\max})$, target coverage ratio $c \in [0,1]$, maximal primer degeneracy $n_{\text{degen}}$, orientation of primers $d = \{\textsc{fw}, \textsc{rev}, \textsc{both}\}$
Output:
    A minimal primer set $\hat{P}^*$ with primers of orientation $d$ fulfilling the filtering constraints $F$ with maximal possible coverage of template sequences $T$ subject to the temperature constraint $\Delta T_m$ and the dimerization constraint $\Delta G_{\text{cd}}$
procedure designPrimers($T$, $F$, $\Delta T_m$, $\Delta G_{\text{cd}}$, $o_O$, $o_I$, $(l_{\min}, l_{\max})$, $c$, $n_{\text{degen}}$, $d$)
    $c \leftarrow c|T|$                                  ▷ Number of templates to cover
    if $d \in \{\textsc{fw}, \textsc{rev}\}$ then
        $\hat{P}^* \leftarrow$ designPrimersSingle($T_d, F, \Delta T_m, \Delta G_{\text{cd}}, o_O, o_I, (l_{\min}, l_{\max}), c, n_{\text{degen}}$)
        return $\hat{P}^*$
    else
        $\hat{P}^*_{\text{fw}} \leftarrow$ designPrimersSingle($T_{\text{fw}}, F, \Delta T_m, \Delta G_{\text{cd}}, o_O, o_I, (l_{\min}, l_{\max}), c, n_{\text{degen}}$)
        $T \leftarrow P^{\hat{*}}_{\text{fw}\triangleright T}$     ▷ Cover the appropriate templates of the other strand
        $\hat{P}^*_{\text{rev}} \leftarrow$ designPrimersSingle($T_{\text{rev}}, F, \Delta T_m, \Delta G_{\text{cd}}, o_O, o_I, (l_{\min}, l_{\max}), c, n_{\text{degen}}$)
        $\hat{P}^* \leftarrow P^{\hat{*}}_{\text{fw}} \cup \hat{P}^*_{\text{rev}}$
        return $\hat{P}^*$
```

Algorithm 8: Global primer design.

Input:

Set of templates T, filtering constraints F, melting temperature constraint ΔT_m, dimerization constraint ΔG_{cd}, $o_O \in \{\textsc{Greedy}, \text{ILP}\}$, $o_I \in \{\textsc{Naive}, \textsc{Degenerate}\}$, desired primer lengths $(l_{\min}, l_{\max})$, primer penalty parameter $\alpha \in [0, 1]$, maximum allowed primer penalty $\xi_{\max} \geq 0$, step size for primer penalties $\xi_\epsilon > 0$

Output:

A collection $\hat{\mathcal{P}}^*_\xi$ of optimal primer sets whose members exhibit differing deviations from the constraints F as defined by $\xi_{\max}$ and ξ_ϵ

```
procedure designPrimersGlobalSingle($T$, $F$, $\Delta T_m$, $\Delta G_{\text{cd}}$, $o_O$, $o_I$, $(l_{\min}, l_{\max})$, $\alpha = 0.5$, $\xi_{\max}$, $\xi_\epsilon$)
    $\hat{\mathcal{P}}^*_\xi \leftarrow \emptyset$
    $P \leftarrow$ initPrimers$(T, o_I, (l_{\min}, l_{\max}))$
    $P_E \leftarrow$ scorePrimers$(T, P, F, \alpha)$
    $\xi \leftarrow 0$                              ▷ Current maximal allowed penalty
    while $\xi \leq \xi_{\max}$ do
        $P_\xi \leftarrow \{p \in P_E | \text{Pen}(p, F, T, \alpha) \leq \xi\}$
        $\hat{P}^*_\xi \leftarrow$ optimizePrimers$(T, P_\xi, \Delta T_m, \Delta G_{\text{cd}}, o_O)$
        $\hat{\mathcal{P}}^*_\xi \leftarrow \hat{\mathcal{P}}^*_\xi \cup \hat{P}^*_\xi$
        $\xi \leftarrow \xi + \xi_\epsilon$
    return $\hat{\mathcal{P}}^*_\xi$

procedure scorePrimers($T$, $P$, $F$, $\alpha$)
    $P_E \leftarrow P$
    for $p$ in $P_E$ do
        $\text{Pen}(p, F, T, \alpha) \leftarrow \alpha ||\text{Dev}(p, F, T)||_\infty + (1 - \alpha)||\text{Dev}(p, F, T)||_1$
    return $P_E$
```

of primers of appropriate size before beginning with the design. The key idea for developing such an approach is that the time-intensive steps of evaluating and optimizing the primers should be avoided. In fact, it is possible to estimate the feasibility of a primer design problem solely based on the distribution of the EAEs because there is a relationship between the coverage distribution and the size of the optimized primer set. Templates that are well-conserved are associated with sets of primers containing multiple primers exhibiting high coverages, which lead to small optimized primer sets. Less conserved templates, on the other hand, are associated with low-coverage primers and, correspondingly, large primer sets. Based on this observation, the feasibility of a primer design problem can be evaluated by comparing the primer coverage distribution with reference distributions.

Each reference distribution represents a distinct amplification scenario exhibiting a certain level of feasibility with respect to designing primers. For feasible primer design problems, primer candidates comprise a considerable number of high-coverage primers (e.g. covering 10%–20% of templates), while infeasible primer design problems exhibit few such primers. To illustrate this, imagine a set of 100 templates. If there were 10 primers, each covering a distinct set containing 10% of the templates, it would be possible to obtain a primer set of size 10. However, if there were 10 complementary primers, each with a coverage of only 1%, the size of the primer set would be 100, which is too large.

By empirically investigating the coverage distribution for several sets of templates, I found that the coverage distribution can be modeled well using a beta distribution. The beta distribution is a continuous probability distribution that is controlled by two positive shape parameters α and β. Its probability density function for $0 \leq x \leq 1$ is defined by

$$\beta(\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}\,.$$

The beta function $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha)+\Gamma(\beta)}$ serves as a normalization constant, which ensures that the probability integrates to 1. For positive integers, the gamma function $\Gamma(n) = (n-1)!$ is simply an extension of the factorial function.

Algorithm 9 gives the CLASSIFYDESIGNPROBLEM procedure, which classifies the difficulty of a primer design problem. Using the *defineReferenceDistributions* function (line 2), it defines reference distributions representing primer design instances with distinct levels of feasibility (Figure 6.8): *Easily Feasible* with $\beta(1, 10)$, *Feasible* with $\beta(0.8, 20)$, *Hardly Feasible* with $\beta(0.6, 40)$, and *Unfeasible* with

$\beta(0.3, 200)$.[62] Since the computation of the coverage distribution would be too time-consuming, a lower bound on the coverage of each primer is computed through the *estimateCoverageDistribution* function (line 3) in the following way.

For a specific primer length k, all template k-mers are generated. Next, the fractional coverage is determined by counting the occurrences of each k-mer and dividing by the number of templates.[63] Based on the coverage distribution of the primers, the parameters of a beta distribution are determined using a maximum likelihood approach[64] via the function *fitBetaDistribution* (line 4). The fitted beta distribution is then compared to each of the reference distributions based on the total variation distance (line 8), which is introduced in the following.

Beta distributions representing coverage ratios induce an infinite probability space with $\Omega = [0, 1]$. Given two probability distributions P and Q on a sigma-algebra $\mathcal{C}$ of subsets of the sample space Ω, the total variation distance is defined as

$$\delta(P, Q) = \sup_{C \in \mathcal{C}} |P(C) - Q(C)| \,.$$

The total variation distance $\delta(P, Q)$ can be interpreted as the largest difference in probabilities that P and Q assign to the same event. By determining $\delta(P, Q)$ for the fitted beta distribution, P, and all reference distributions, Q, the feasibility of the primer design task is obtained according to the reference distribution minimizing $\delta(P, Q)$ (line 9). In the openPrimeR frontend, the difficulty of a primer design task is represented by a traffic light where *Easily Feasible* and *Feasible* problems are indicated by a green light, *Hardly Feasible* problems by an orange light, and *Unfeasible* problems by a red light (Figure 6.9).

Although the CLASSIFYDESIGNPROBLEM procedure makes several assumptions, its inaccuracies may theoretically balance out. First, since a lower bound on primer coverage the determined feasibility is rather pessimistic with respect to coverage. Second, since primers are not filtered and optimized, a set of primers exhibiting redundant coverage or low-quality primers would require a larger number of primers. Thus, regarding quality and redundancy of coverage, the estimated feasibility may be too optimistic.

6.3.7 *Optimal Subset Selection*

Determining subsets of an optimized primer set can be desirable in two scenarios: when a primer set is too large or when specific primers amplifying a subset of templates should be selected. In the following, I provide examples for both scenarios. For the first scenario, consider a set of 20 primers of which 5 primers provide

[62] The parameters of the beta distributions were selected by hand so as to obtain distributions that are representative of the corresponding levels of primer design feasibility.

[63] By using the fractional coverage rather than the coverage, the approach is independent of the number of templates.

[64] Delignette-Muller and Dutang 2015

Algorithm 9: Classifying the feasibility of a primer design task.

Input:

Set of templates T, primer length k

Output:

Primer design problem difficulty class

```
procedure CLASSIFYDESIGNPROBLEM($T$, $k$)
    $\beta \leftarrow$ defineReferenceDistributions()
    $x \leftarrow$ estimateCoverageDistribution($T, k$)
    $\beta_x \leftarrow$ fitBetaDistribution($x$)
    class $\leftarrow$ NONE
    $d_{\text{min}} \leftarrow$ NONE
    for each $\beta_i \in \beta$ do
        $d \leftarrow$ totalVariationDistance($\beta_x, \beta_i$)
        if $d < d_{\text{min}}$ then
            $d_{\text{min}} \leftarrow d$
            class $\leftarrow$ classOf($\beta_i$)
    return class
```

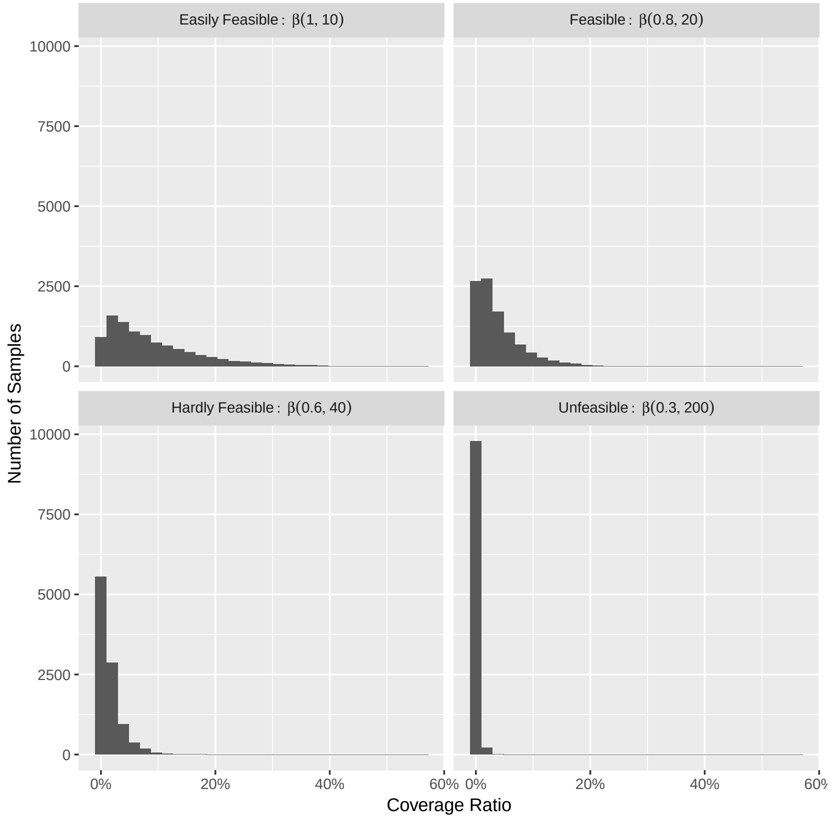


Figure 6.8: Reference beta distributions for estimating the feasibility of primer design tasks. The plots show the coverage distributions that were obtained by drawing 100 000 samples from the indicated beta distributions.

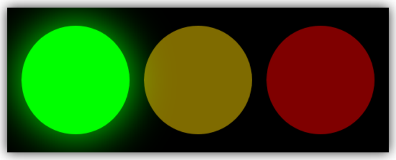


Figure 6.9: Traffic light representation for primer design difficulty. In this case, the light is green, so it should be possible to design a suitably small set of primers for the given templates.

90% coverage. If coverage of all templates is not required, it may be beneficial to remove the 15 primers that provide little additional coverage and work with the subset of 5 primers that already achieves 90% coverage. For the second scenario, imagine that there is a set of primers suitable for all subtypes of HIV-1 group M but one would like to isolate only sequences from subtype B. Then, one would need to determine the subset of primers maximizing the coverage of subtype B sequences.

Since best subset selection targets well-designed primer sets, the physicochemical properties of the primers do not have to be considered. The problem can be formulated in terms of the NP-hard maximum coverage problem. Since optimized primer sets are small, the problem can be solved exactly using the following ILP. As previously, let $x_i \in \{0,1\}$ refer to the decision variables and let $\mathbf{C} \in \{0,1\}^{|T| \times |P|}$ indicate the coverage matrix. Let P_x denote the primer set that is induced by x, that is, $P_x = \{p_i \in P | x_i = 1\}$. Additionally, let $y_j \in \{0,1\}$ refer to decision variables indicating whether template t_j is covered by any member of P_x, that is,

$$y_j = \begin{cases} 1 & \text{if } |t_{j \triangleleft P_x}| \geq 1 \\ 0 & \text{else} \end{cases} .$$

Further, let $k \in \mathbb{N}$ indicate the size of the primer subset. Then, the ILP for best subset selection is defined as follows:

$$\max \sum_{j=1}^{|T|} y_j \qquad \text{Maximize template coverage} \tag{6.5}$$

$$\text{subject to} \tag{6.6}$$

$$\sum_{i}^{|P|} x_i = k \qquad \text{Select a subset of size } k \tag{6.7}$$

$$\mathbf{C}x \geq y \qquad \text{Definition of } y \tag{6.8}$$

In this ILP, the critical part is the last constraint, $\mathbf{C}x \geq y$, which ensures that y_j is only set to 1 if template t_j is covered by at least one selected primer. The trade-off between the size of a primer set P and its coverage can be analyzed by solving the ILP for $k \in \{1, \ldots, |P|\}$ (Figure A.2).

6.4 Primer Design for Human Immunoglobulin Sequences

In this section, I present the application of openPrimeR on primers targeting human immunoglobulin variable sequences from IGH, IGK, and IGL (Section 6.4.1). Primers targeting the leaders of IGH (Table 6.7), IGK (Table 6.8), and IGL (Table 6.9) were designed according to the settings shown in Table A.2. These settings were obtained by iteratively adjusting the constraints according to experimentally obtained PCR yields. Since the number of overall mismatches was limited to either zero (for IGHV) or three (for IGKV and IGLV) and 3′ mismatches were not allowed at all, it was not necessary to apply a model for filtering the EAEs. The three designed primer sets were determined through the local primer design procedure; the SCP was solved using the ILP formulation.

The designed primer sets for IGH, IGK, and IGL were validated both *in silico* and *in vitro*. For the sake of brevity, the following sections provide only the validation results for IGHV.[65] The designed IGHV primers (*openPrimeR*, 15 primers) were compared to the well-established primer sets from Tiller et al. (2008) (*Tiller*, 4 primers) and Ippolito et al. (2012) (*Ippolito*, 8 primers) both *in silico* (Section 6.4.2) and *in vitro* (Section 6.4.3).

6.4.1 Data for Immunoglobulin Primer Design

For the purpose of designing primers for human immunoglobulin sequences, germline sequences for IGH, IGK, and IGL were retrieved from the international immunogenetics information system (IMGT) database[66]. Only those sequences that were classified as functional according to IMGT[67] were selected in order to limit primer design to sequences from expressed genes. Additionally, template sequences with partial leader sequences were either removed (for IGK and IGL) or augmented by our own NGS measurements (for IGH). Based on this procedure, 152 IGH, 62 IGK, and 35 IGL templates were selected (Table 6.10). For the purpose of comparing the properties of designed primer sets with existing sets, established primer sets for IGH (Table A.3), IGK (Table A.4), and IGL (Table A.5) were collected from IMGT and the literature.

6.4.2 In Silico Validation

Figure 6.10 shows that all of the *openPrimeR* primers are expected to exclusively bind in the conserved leader region, which is neither the case for the primers from *Tiller* nor those from *Ippolito*. Panel A of Figure 6.11 demonstrates that the designed primers fulfill most of the required physicochemical constraints and have a higher constraint

[65] For an overview of the *in silico* results for IGKV and IGLV, I refer the interested reader to the Figures A.4, A.5, A.6, and A.7 in the appendix.

[66] Ruiz et al. 2000; Lefranc 2004; Giudicelli et al. 2004; Lefranc et al. 2015

[67] A sequence is classified as functional by IMGT if it has coding regions with open reading frames without stop codons and there are no defects in splicing sites, recombination signals and/or regulatory elements.

Primer	Sequence	Main target	Coverage	T_m [°C]
1-IGHV	atggactggacctggagcatcc	IGHV1	8.6%	60.68
2-IGHV	atggactggacctggaggatcctc	IGHV1	11.2%	61.59
3-IGHV	atggactggacctggagggtcttc	IGHV1	2%	61.95
4-IGHV	atggactggatttggagggtcctcttc	IGHV1	1.3%	61.30
5-IGHV	atggacacactttgctacacactcctgc	IGHV2	0.7%	61.99
6-IGHV	actttgctccacgctcctgc	IGHV2	13.2%	60.07
7-IGHV	ggctgagctgggttttccttgttg	IGHV3	27%	60.33
8-IGHV	ggctccgctgggttttccttgttg	IGHV3	1.3%	62.80
9-IGHV	cacctgtggttcttcctcctgctg	IGHV4	28.3%	61.51
10-IGHV	atgaaacacctgtggttcttcctcctcc	IGHV4	27.6%	61.43
11-IGHV	acatctgtggttcttccttctcctggtg	IGHV4	1.3%	61.19
12-IGHV	gcctctccacttaaacccaggctc	IGHV5	0.7%	61.41
13-IGHV	atgtctgtctccttcctcatcttcctgc	IGHV6	1.3%	60.88
14-IGHV	atggagttggggctgagctgg	IGHV3	26.3%	61.74
15-IGHV	atggggtcaaccgccatcctc	IGHV5	2.6%	61.74

Table 6.7: Overview of designed IGHV primers. *Sequence* provides the oligomer sequence for the forward primer. *Main target* indicates the IGHV gene group that is predominantly targeted by the primer. *Coverage* refers to the overall coverage of the template sequences.

Primer	Sequence	Main target	Coverage	T_m [°C]
1-IGKV	atgaggctccttgctcagcttctgg	IGKV2	1.6%	62.59
2-IGKV	atggaagccccagctcagcttc	IGKV3D	14.5%	61.59
3-IGKV	cccagctcagcttctcttcctcctg	IGKV3D	16.1%	62.88
4-IGKV	tggtgttgcagacccaggtcttcatttc	IGKV4	1.6%	62.09
5-IGKV	gtcccaggttcacctcctcagcttc	IGKV5	1.6%	63.05
6-IGKV	gccatcacaactcattgggtttctgctg	IGKV6	6.5%	61.54
7-IGKV	tccctgctcagctcctggg	IGKV1	64.5%	61.68
8-IGKV	cctgggactcctgctgctctg	IGKV1	38.7%	62.22

Table 6.8: Overview of designed IGKV primers. *Sequence* provides the oligomer sequence for the forward primer. *Main target* indicates the IGKV gene group that is predominantly targeted by the primer. *Coverage* refers to the overall coverage of the template sequences.

Primer	Sequence	Main target	Coverage	T_m [°C]
1-IGLV	ccctgggtcatgctcctcctgaaatc	IGLV10	2.1%	62.91
2-IGLV	ctctgctgctcctcactctcctcac	IGLV2	10.6%	62.51
3-IGLV	atggcatggatccctctcttcctcg	IGLV3	2.1%	61.89
4-IGLV	cctctctggctcactctcctcactc	IGLV3	2.1%	61.94
5-IGLV	acactcctgctcccactcctcaac	IGLV3	4.3%	62.40
6-IGLV	atggcctggatccctctacttctcc	IGLV3	4.3%	61.34
7-IGLV	atggcctgggtctccttctacc	IGLV4	2.1%	60.21
8-IGLV	atggcctggactcctctctttctgttc	IGLV7	6.4%	61.34
9-IGLV	atggcctggatgatgcttctcctc	IGLV8	2.1%	60.39
10-IGLV	gtcccctctcttcctcaccctcatc	IGLV1	2.1%	62.15
11-IGLV	ctcctcgctcactgcacagg	IGLV3	21.3%	60.21
12-IGLV	cctctcctcctcaccctcctc	IGLV1	19.1%	60.56
13-IGLV	ctcctcctcaccctcctcactc	IGLV2	19.1%	60.41
14-IGLV	atggcctggacccctctcc	IGLV3	21.3%	61.59
15-IGLV	atggcctggaccccactcc	IGLV3	8.5%	62.12

Table 6.9: Overview of designed IGLV primers. *Sequence* provides the oligomer sequence for the forward primer. *Main target* indicates the IGLV gene group that is predominantly targeted by the primer. *Coverage* refers to the overall coverage of the template sequences.

Locus	Total	Functional	Functional with leader
IGH	243	156	152
IGK	132	64	62
IGL	61	36	35

Table 6.10: Overview of immunoglobulin templates.

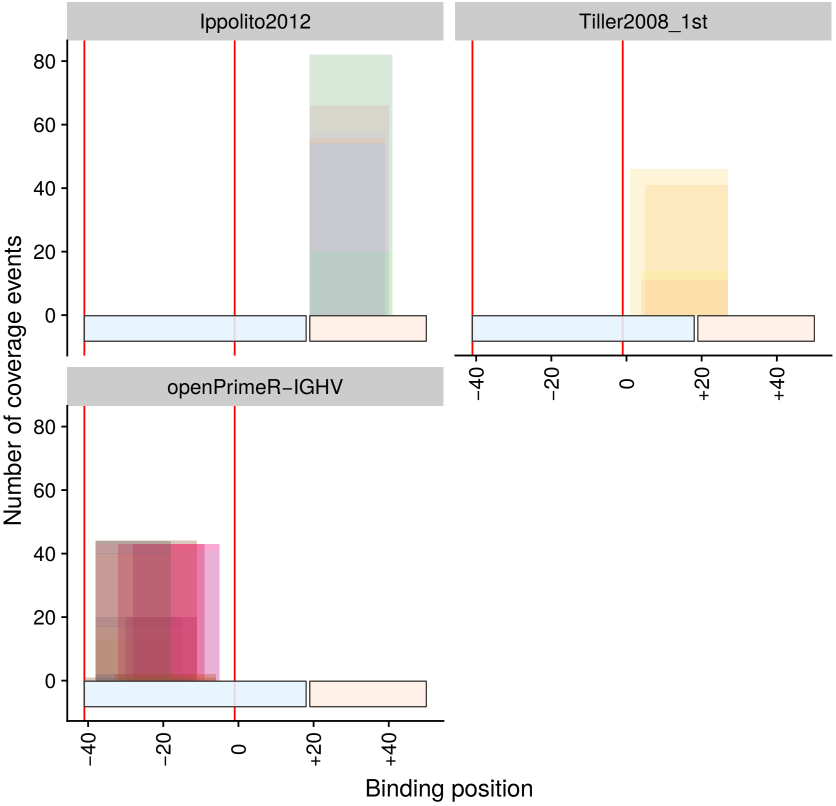


Figure 6.10: IGHV primer binding regions. The leader region is indicated by the horizontal blue bar, while the horizontal red bar indicates the variable region. The region for which the new primers were designed is indicated by vertical red lines. The vertical bars indicate the number of expected amplification events for individual primers.

fulfillment rate than the primers from *Tiller* and *Ippolito*. The only constraint that is minimally broken by two primers from *openPrimeR* (absolute deviation slightly above 3%) relates to the GC ratio (Figure A.3). In fact, the constraint fulfillment rate of the newly designed primers is significant at the 5% significance level according to Fisher's exact test (p-value of 1.6e-36). Panel B of Figure 6.11 shows that the newly designed primers are estimated to cover 100% of germline IGHVs. The coverage of the new IGHV primers seems to be particularly superior to the coverage of the primers from *Tiller*.

6.4.3 *In Vitro Validation*

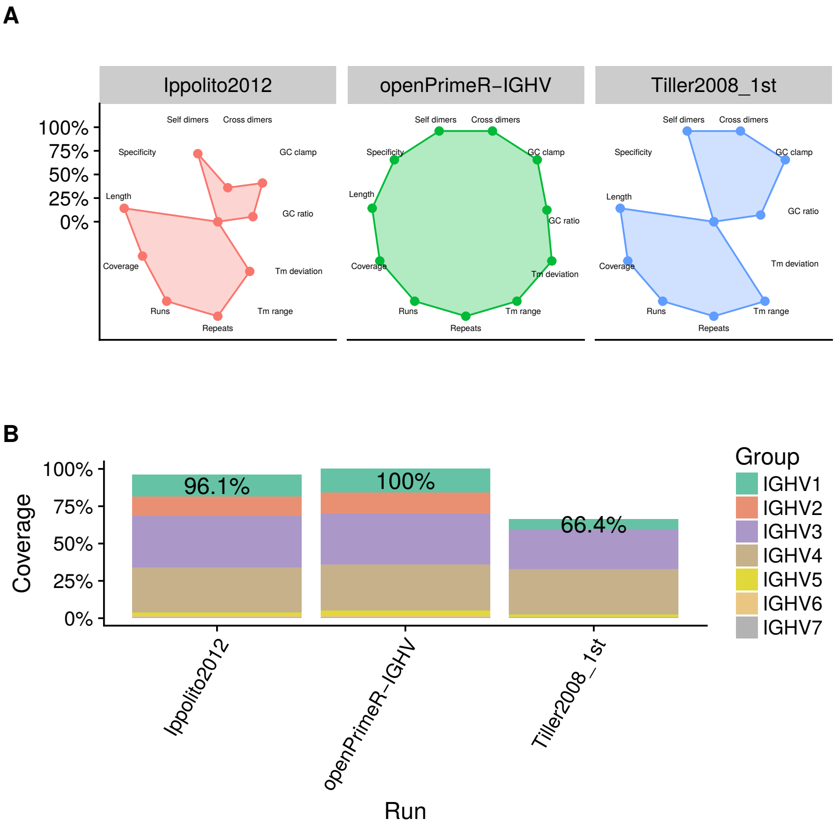
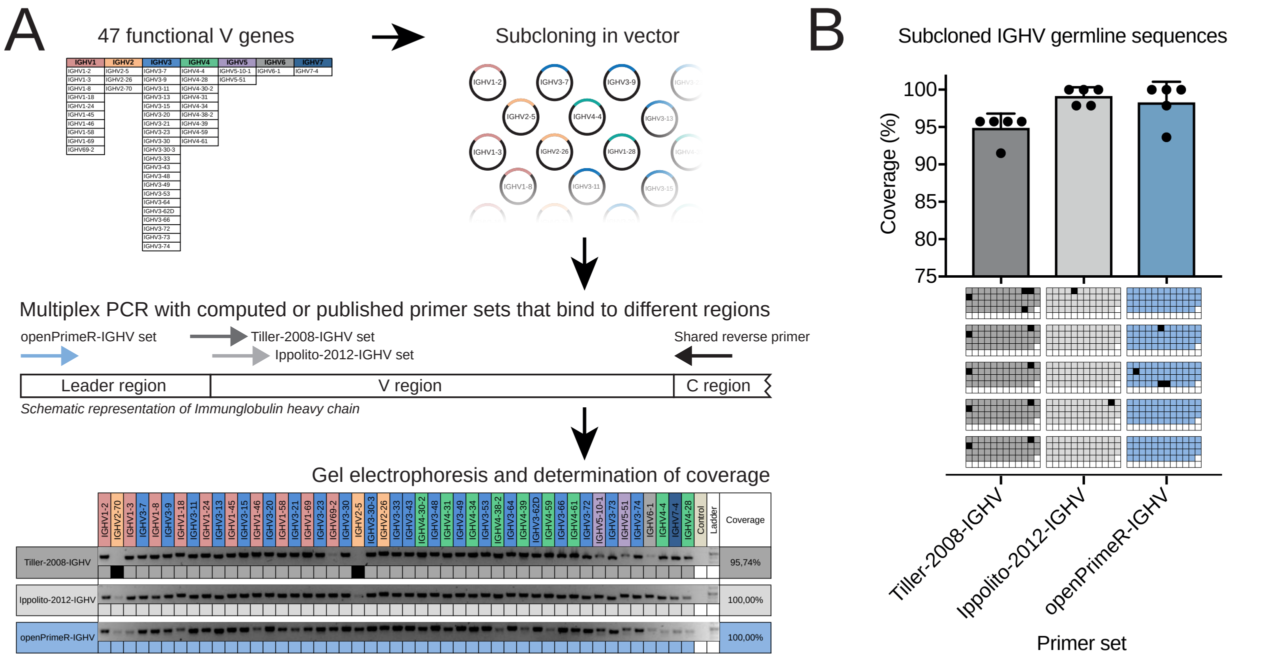


Figure 6.11: Properties of IGHV primer sets as determined by openPrimeR. (A) Rate of constraint fulfillment for each constraint as determined by $\mathrm{Ful}(P, \mathcal{F}, T)$. A value of 100% for a given constraint indicates that all primers in the set fulfill the constraint. Thus, the greater the surface area in the radar chart, the greater the quality of the primer set. (B) Percentage of covered templates per IGHV gene group. A cumulative value of 100% indicates that a set of primers is estimated to cover all templates.



IGHV1	IGHV2	IGHV3	IGHV4	IGHV5	IGHV6	IGHV7
IGHV1-2	IGHV2-5	IGHV3-7	IGHV4-4	IGHV5-10-1	IGHV6-1	IGHV7-4
IGHV1-3	IGHV2-26	IGHV3-9	IGHV4-28	IGHV5-51		
IGHV1-8	IGHV2-70	IGHV3-11	IGHV4-30-2			
IGHV1-18		IGHV3-13	IGHV4-31			
IGHV1-24		IGHV3-15	IGHV4-34			
IGHV1-45		IGHV3-20	IGHV4-38-2			
IGHV1-46		IGHV3-21	IGHV4-39			
IGHV1-58		IGHV3-23	IGHV4-59			
IGHV1-69		IGHV3-30	IGHV4-61			
IGHV69-2		IGHV3-30-3				
		IGHV3-33				
		IGHV3-43				
		IGHV3-48				
		IGHV3-49				
		IGHV3-53				
		IGHV3-64				
		IGHV3-62D				
		IGHV3-66				
		IGHV3-72				
		IGHV3-73				
		IGHV3-74				

Figure courtesy of Christoph Kreer.

Figure 6.12: Validation of primer sets on germline IGHVs. (A) Experimental procedure. 47 functional IGHV genes were cloned into expression vectors in order to obtain corresponding antibody transcripts. These transcripts were amplified with three different primer sets and the differential amplification of individual IGHV genes was analyzed using gel electrophoresis. (B) Amplified IGHV genes for five PCRs per primer set. The bar plot summarizes the percentage of amplified templates. The diagram below shows a digitized version of the five observed gel electrophoresis results. Here, black squares indicate failed amplifications, white squares indicate negative controls, and the remaining squares indicate successful amplifications.

For the *in vitro* validation of the IGHV primers, three wet lab experiments were performed by Nathalie Lehnen, Philipp Schommers, Meryem Seda Ercanoglu, and Christoph Kreer. The first experiment was done in order to determine whether the estimated coverage agreed with the experimentally observed coverage. In this experiment, a set of 47 functional IGHV genes representing all of the 7 heavy chain gene groups was selected. These sequences were cloned into eukaryotic expression vectors and subsequently expressed in order to amplify them via RT-PCR using the *openPrimeR*, *Tiller*, and *Ippolito* primer sets (Panel A of Figure 6.12). PCRs were performed in five replicates each and results were visualized using gel electrophoresis. The evaluation of the coverage revealed that all primer sets achieved close to 100% coverage (Panel B of Figure 6.12), although the coverage of *Tiller* was slightly lower than that of the other sets. The further investigation of the PCR results revealed that the lack of coverage of *Tiller* was attributable to the fact that not all IGHV2 templates were amplified, as predicted *in silico* (Figure 6.11).

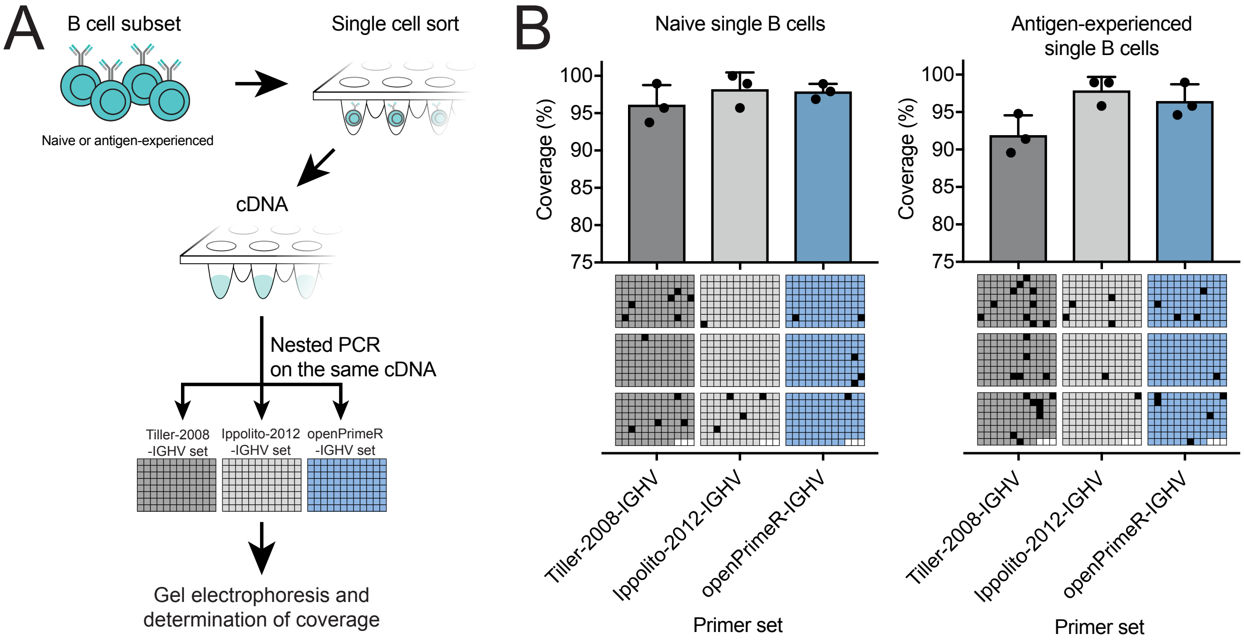


Figure courtesy of Christoph Kreer.

Figure 6.13: Validation of primer sets on IGHVs from single B cells. (A) Experimental procedure. Naive or antigen-experienced B-cells were sorted and the cDNA of their transcripts was determined. The cDNA was then amplified through triplicate PCRs using three different primer sets whose performance was evaluated using gel electrophoresis. (B) Amplified IGHV genes. The left-hand panel shows the coverage of IGHVs from naive B cells, while the right-hand panel shows the coverage of IGHVs from antigen-experienced B cells.

The goal of the second experiment was to investigate whether the results from the first experiment could be reproduced for immunoglobulin sequences obtained from single B cells. In order to differentiate how the presence of mutations influences the amplification rate, naive (unmutated) and antigen-experienced single B cells (mutated) were separated via cell sorting (Panel A of Figure 6.13). After transforming B-cell receptor transcripts to cDNA, PCRs were performed in triplicates for each of the three primer sets. The gel electrophoresis results for naive B cells closely mirrored the results that were obtained for germline sequences in the first experiment. All primer sets achieved close to 100% coverage but *Tiller* obtained a slightly smaller coverage (Panel B of Figure 6.13) . On antigen-experienced B cells, this effect was enhanced as the coverages for *Ippolito* and *openPrimeR* remained high but the coverage of *Tiller* dropped to slightly above 90%.

The final experiment was conducted in order to evaluate the performance of the primer sets on highly mutated antibody cDNA. For this purpose, HIV-1 reactive B cells were retrieved from an elite neutralizer and corresponding cDNA was amplified. Panel A of Figure 6.14 demonstrates that, on these sequences, *openPrimeR* achieved the highest coverage ($>$ 90% coverage), followed by *Tiller* and *Ippolito* (both with approximately 80% coverage). Note that *Tiller* slightly outperformed *Ippolito* with respect to both, coverage (Panel A of Figure 6.14) and amplification of heavily mutated sequences (Panel B of Figure 6.14).

The amplicons generated by *openPrimeR* were sequenced and mapped to their germline correspondents using IgBLAST[68]. This revealed that *openPrimeR* particularly improved the amplification of heavy chains with reference similarities less than 70% (Panel B of Figure 6.14). In fact, *openPrimeR* amplified roughly three times as many highly mutated sequences as each of the other primer sets. Most importantly, further experimental work demonstrated that these highly mutated IGHVs indeed constitute bNAbs.

[68] Ye et al. 2013b

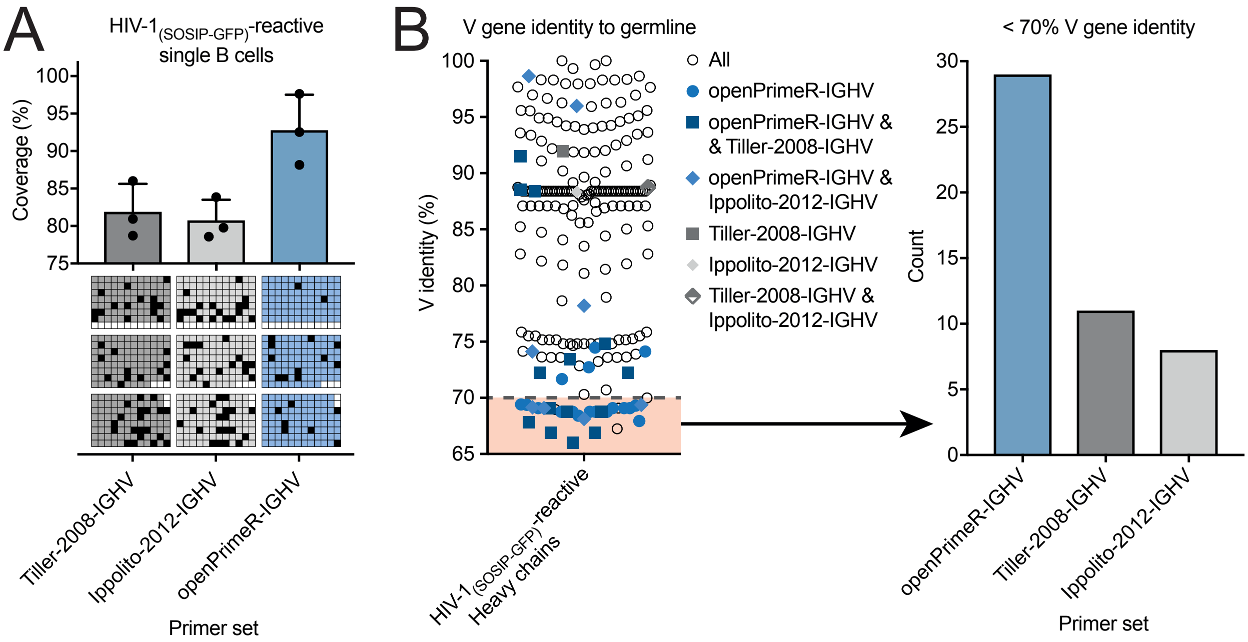


Figure courtesy of Christoph Kreer.

Figure 6.14: Validation on IGHVs from a HIV-1-infected person. (A) Amplification of IGHV transcripts from B cells of an HIV-1 positive person for three primer sets, each of which was measured in triplicates. (B) Amplification of IGHVs as a function of sequence similarity to corresponding germline sequences. Blue samples indicate sequences that were amplified by the primers designed by openPrimeR or at most one other primer set. The bar chart on the right shows the number of IGHVs with a germline similarity less than 70% that were amplified by each of the primer sets.

6.5 Discussion

This chapter introduced openPrimeR, a computational tool for the evaluation and design of primer sets for mPCR. Based on the hypothesis that primers binding to the immunoglobulin leader region should allow for the improved amplification of highly mutated antibody sequences, primer sets targeting the leaders of IGHV, IGKV, and IGLV were developed. The reliability of openPrimeR and the high performance of the newly designed primers targeting immunoglobulin heavy chain sequences was shown in three experiments. The first experiment revealed that the designed primers achieve coverage of all germline immunoglobulin variants (Figure 6.12). The second experimented showed that the primers also perform well on cDNA from single B cells (Figure 6.13), which suggests their suitability for repertoire analyses.

The third experiment revealed that *Tiller* slightly outperformed *Ippolito* on highly mutated sequences (Figure 6.14), which was interesting because this was the case neither on germline nor on naive nor on antigen-experienced sequences. An explanation for this is presented by the binding regions of *Tiller* and *Ippolito*. While *Tiller* overlaps with both leader and variable region, *Ippolito* solely binds in the variable region (Figure 6.10). Consequentially, the high mutational load of the variable region[69] may prevent the annealing of a larger number of primers from *Ippolito* than from *Tiller*. The superior performance of the primers designed by openPrimeR suggest that primers binding to the early leader region are more suitable for the amplification of highly mutated sequences than primers that bind to the late leader region (*Tiller*) or the variable region (*Ippolito*). This finding suggests that the first positions in the immunoglobulin leader are the most conserved stretch of the 5′ end.

Most importantly, it was possible to show that the newly designed primers facilitate the amplification of cDNA sequences from bNAbs. In the future, the newly designed primer sets could allow for the isolation of previously unknown bNAbs, which might be used in novel antibody-mediated treatment and prevention strategies against HIV-1 infection. To determine whether this is the case, previously collected samples from elite neutralizers such as the data from (Simek et al., 2009) could be re-analyzed using the novel primer sets.

openPrimeR fulfills all of the requirements for designing suitable primers for human immunoglobulins. Most importantly, the tool can be used to estimate the template amplification status with sufficient specificity for finding the smallest set of primers covering all templates. This is possible by enforcing a stringent notion of coverage and requiring tight constraints on the physicochemical properties of primers. The high coverage obtained in the experimental validation of openPrimeR suggests that the tool can identify coverage events with sufficient accuracy. Note that the experimental coverage was generally higher than the estimated coverage. This is attributable to the conservative definition of coverage that was used for the *in silico* evaluations, for which only a single mismatch was allowed. Since the tool allows for specifying allowed binding ranges in the templates, it is possible to design primers for specific regions such as the leader. Through its GUI, openPrimeR is intuitively usable. Additionally, openPrimeR contains a programmatic interface that provides access to a rich set of functions that can be used for various tasks such as performing batch runs or implementing custom primer design approaches. Although openPrimeR was developed with immunological applications in mind, the tool is versatile and could be a useful resource for other primer design tasks.

There are several ways in which openPrimeR could be improved in the future. One aspect for improvement regards the programming language. Currently, openPrimeR is implemented solely in R. Replacing some aspects with function calls to a C++ library would be beneficial. First, the runtime of the code could be improved. Second, the robustness of the tool could be improved because in contrast to C++, which is strictly typed, R is typed dynamically. Using a strictly

[69] Rada et al. 1994

typed language for critical functions could improve maintainability by reducing the potential for runtime errors.

The following paragraphs deal with algorithmic aspects. The procedure for the determination of primer coverage is a bottleneck when designing primers because all templates have to be scanned for thousands of primer candidates. It would be possible to considerably improve the runtime of the algorithm using hashing[70], for example, by hashing all hexamers in the template binding regions, matching all primer 3′ hexamers to the hash map, and then extending the hits. A limitation of this approach, however, is that it would not allow for the generation of primers exhibiting mismatches in the 3′ hexamers. Another possible improvement pertains to the selection of the expected region in the template to which a primer binds. This approach may be refined by not only searching for the primer-template conformation with the smallest number of mismatches but by considering several conformations and selecting the most likely one by using models that estimate the likelihood of amplification.

openPrimeR could also be further automated. For example, the approach for classifying the difficulty of a primer design task (Section 6.3.6) could be used to construct multi-tube solutions. Given an infeasible primer design task, the software could automatically iteratively split the set of templates into smaller subsets (e.g. using clustering) until each subset yields a feasible primer design problem and then design an individual set of primers for each subset. Further potential for optimization lies in the automatic selection of the parameters for primer design. For example, the tightness of the constraints could be based on the observed coverage distribution and the selection of the optimization algorithm could be based on the number of candidate primers.[71]

With regard to the optimization problem, the primer design problem was formulated in terms of a SCP. In the local primer design procedure, I introduced a relaxation procedure to ensure that a primer set obtaining full coverage can be found. In the global version, several solutions are computed in order to explore the trade-off between the three objectives of primer design (from highest to lowest): coverage, set size, and primer quality. An alternative way of formulating the global variant would have been via lexicographic goal programming[72] in which the priority of competing objectives is preemptively determined. Using goal programming, several linear programs are solved iteratively such that each one fulfills the constraints of the previous programs. In this case, the first solution would just satisfy the coverage criterion, the second would minimize the set size (subject to maximum coverage), and the third would maximize the quality (subject to maximum coverage and minimum set size). Although goal

[70] Huang et al. 2005

[71] Solving the primer design ILP may be infeasible for large sets of templates.

[72] Widhelm 1981; Romero 2001

programming represents a rational way of selecting a single, optimal set, it is also more of a black box than when users are provided with an overview of several possible results.

Another limitation concerns the computation of primer melting temperatures. Currently, only the melting temperature of complementary primer-template pairs are considered. The melting temperatures of primers binding with mismatches, however, are actually lower than those of complementary primers. Thus, primer sets that are designed with many allowed mismatches may not perform well because the primers likely do not amplify the templates at the same annealing temperature. A solution to this problem could be obtained by considering the mismatch melting temperatures for individual EAEs. This would, however, incur increased runtime.

In summary, openPrimeR uses techniques from mathematical optimization for the design of mPCR primer sets. In this work, the approach was used to improve the isolation of bNAbs targeting HIV-1. In the future, openPrimeR could be applied to improve the amplification of other variable sequences, for example, those of viral origin (e.g. HIV-1). openPrimeR is freely available via openprimer.mpi-inf.mpg.de.

This chapter has shown that methods for the estimation of amplification events are key for the design of mPCR primers. The subsequent chapter (Chapter 7) investigates the molecular characteristics of successful amplification events and presents *TMM*, a logistic regression model that estimates the likelihood of amplification.

7
Predicting PCR Amplification Events

This chapter deals with approaches for the identification of PCR amplification events. I developed a new statistical model and statistically analyzed the PCR data set. Christoph Kreer planned the IGHV template generation and the PCR experiments. Nathalie Lehnen performed the PCR reactions. Florian Klein initiated the project and guided the work. Nico Pfeifer supervised the analysis. This chapter largely corresponds to a manuscript entitled "Modeling the Amplification of Immunoglobulins through Machine Learning on Sequence-Specific Features", which has been submitted to Nature Scientific Reports.

> I was working for Cetus, making oligonucleotides. They were heady times. Biotechnology was in flower and one spring night while the California buckeyes were also in flower I came across the polymerase chain reaction. . . . It was the first day of the rest of my life.
>
> Kary B. Mullis on the discovery of polymerase chain reaction, 1994

PCR (Section 2.5.4) forms the foundation for a multitude of molecular methods. Typical applications in virological research involve the determination of viral drug resistance[1] and viral loads[2]. Primers — short nucleotide oligomers complementary to template DNA — are critical for the success of PCR because primer binding properties determine whether amplification is successful or not. Therefore, successful primer design hinges on the ability to model whether a primer allows for template amplification. Predictive models can be categorized into two groups. Models that estimate PCR efficiencies can guide primer design for qPCR[3] (Section 2.5.4), while models estimating the likelihood of amplification are suitable for conventional PCR[4]. These models need to consider the two consecutive molecular interactions that determine whether a primer allows for the amplification of a PCR template. In the first reaction, the primer anneals to the template, which leads to the formation of the primer-template heteroduplex. In the second reaction, polymerase attaches to the duplex region and extends the primer to a full-length sequence[5].

Efficient primer annealing is largely determined by the complementarity of primer and template[6]. More specifically, mismatches between the nucleotide sequences of primers and templates should

[1] Petropoulos et al. 2000; Hertogs et al. 1998

[2] Foulongne et al. 2006

[3] Klein et al. 2001; Whiley and Sloots 2005; Gibson 2006; Wright et al. 2014

[4] Yuryev et al. 2002

[5] Pan et al. 2014

[6] Sipos et al. 2007

be avoided as mismatches increase the free energy of annealing. Mismatches within the 3′ hexamer of the primer-template duplex (i.e. within the terminal six nucleotides) are especially detrimental as they can disrupt polymerase binding[7]. Since the disruptive effect of 3′ mismatches increases with growing proximity to the 3′ terminus, a mismatch at the 3′ terminus is more severe than a mismatch at the penultimate position[8]. The extent to which mismatches at the 3′ terminus reduce the efficiency of PCR depends on the type of mismatch.[9] For example, a terminal A/G misatch is more detrimental than a terminal A/C mismatch. To stabilize the 3′ region, primers are often designed to exhibit a so-called GC clamp[10], which typically consists of one to three Gs or Cs at the 3′ end.

Until March 2018, the thermodynamic model from DECIPHER[11] was the only publicly available model for identifying whether a primer allows for amplification. The model of DECIPHER incorporates empiric evidence about the impact of position- and nucleotide-specific mismatches within the last seven positions of the 3′ region. These data were gathered by measuring the elongation efficiency of Taq polymerase in PCRs performed with 171 primers exhibiting different binding properties. The model considers three reactions: the interaction between primer and template, unimolecular folding of the primer, and unimolecular folding of the template. Based on the underlying kinetic differential equations for these reactions, the concentrations of the considered molecular states are mechanistically computed so as to estimate the efficiency of PCR.

The work presented in this chapter investigates the molecular determinants of PCR amplification status. For this task, a novel Taq PCR data set providing the amplification status for 47 IGHV genes was generated.[12] Triplicate measurements were performed with primers from two sets. *Set1* consists of 16 forward primers that have been recently designed using openPrimeR[13], while *Set2* is a well-established set of four forward primer[14]. PCR was performed for each combination of the 20 primers and 47 templates giving rise to a total of 940 triplicate measurements. In contrast to other studies investigating PCR amplification, which are largely based on qPCR, the data that was generated in our work provides the amplification status according to gel electrophoresis.

The structure of this chapter is as follows. Section 7.1 describes the collected IGHV data and introduces the feature encodings and statistical methods that were used for analyzing the data. The identified associations between the molecular properties of primer-template pairs and their amplification status are investigated in Section 7.2. In the same section, a new model for estimating the likelihood of amplification events is presented and compared with other approaches.

[7] Klein et al. 2001; Whiley and Sloots 2005; Stadhouders et al. 2010; Kwok et al. 1990; Bru et al. 2008; Ghedira et al. 2009

[8] Kwok et al. 1990; Ghedira et al. 2009

[9] Stadhouders et al. 2010; Kwok et al. 1990; Ayyadevara et al. 2000; Day et al. 1999; Huang et al. 1992; Li et al. 2004; Wu et al. 2009

[10] No et al. 2014; Lorenz 2012; Thornton and Basu 2011

[11] Wright et al. 2014

[12] We used immunoglobulin sequences as templates because we were researching the amplification of antibody sequences, as described in Chapter 6.

[13] Döring and Pfeifer 2017

[14] Tiller et al. 2008

Primer ID	Sequence	GC Ratio	ΔG_{sd}	ΔG_{f}
Set1.1	cacctgtggttcttcctcct**cc**	59.1%	-0.8	0
Set1.2	cacctgtggttcttcctcct**gc**	59.1%	-0.8	0
Set1.3	atggagtttgggctgagct**gg**	57.1%	-2.3	0
Set1.4	atggagttggggctgagct**g**	60%	-2.3	0
Set1.5	tggagttttggctgagct**ggg**	57.1%	-2.3	-0.1
Set1.6	actttgctccacgctcct**gc**	60%	-0.3	0
Set1.7	atggactggacctggagcat**c**	57.1%	-1.9	0
Set1.8	atggactggacctggaggtt**cc**	59.1%	-2.1	-1.9
Set1.9	atggactgcacctggaggat**c**	57.1%	-1.9	0
Set1.10	atggactggacctggagggtctt**c**	58.3%	-1.9	-3.6
Set1.11	tctgtctccttcctcatcttcct**gc**	52%	0.4	0
Set1.12	ggactggatttggagggtcctctt**c**	56%	-2.2	-3.2
Set1.13	gctccgctgggttttcctt**g**	60%	0.4	0
Set1.14	tggggtcaaccgccat**cc**	66.7%	-0.7	-1.6
Set1.15	ggcctctccacttaaaccca**gg**	59.1%	-1.9	0
Set1.16	tggacacactttgctacacact**cc**	50%	0	0
Set2.1	acaggtgcccactcccaggtgca**g**	66.7%	-0.8	-1.2
Set2.2	aaggtgtccagtgtgargtgca**g**	54.3%	-1.2	0
Set2.3	cccagatgggtcctgtcccaggtgca**g**	66.7%	-1.3	-2.6
Set2.4	caaggagtctgttccgaggtgca**g**	58.3%	-0.8	-0.3

Table 7.1: Primers that were used in IGHV PCRs. The extent of the primer 3′ GC clamp is indicated in bold. Primers prefixed with *Set1* indicate primers from *Set1*, while those prefixed with *Set2* refer to primers from *Set2*. The free energies of self-dimerization and folding are indicated by ΔG_{sd} and ΔG_{f}, respectively.

Finally, Section 7.3 discusses the results and puts them into context with available knowledge.

7.1 PCR Data, Features, and Statistical Models

This section deals with the generation of IGHV PCR data and the manner in which these data were analyzed. Section 7.1.1 describes how IGHV data were generated and Section 7.1.2 describes how these data were transformed to a structured data set. The employed feature encodings are provided in Section 7.1.3. The use of logistic regression for estimating the likelihood of amplification is described in Section 7.1.4. Thereafter, additional approaches that can be used for identifying amplification events are introduced (Section 7.1.5).

7.1.1 Template Design and PCR Measurements

The experimental work described in the following paragraphs was carried out by Nathalie Lehnen and Christoph Kreer.

47 heavy chain fragments from naive B cells were cloned into pCR4-TOPO-vector backbones. Each fragment comprised a different functional IGHV gene and contained the complete leader region,

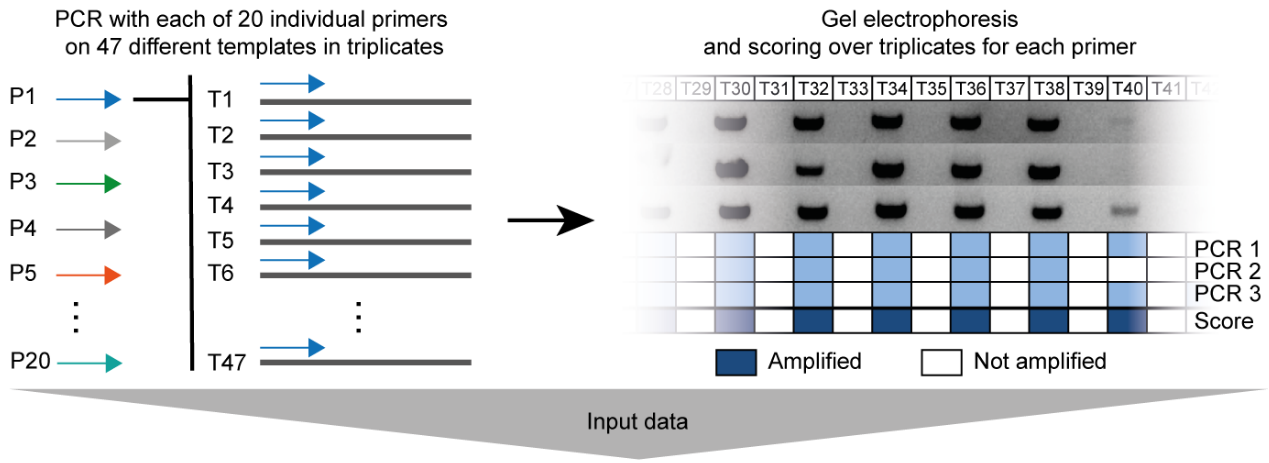


Figure courtesy of Christoph Kreer.

Figure 7.1: Construction of PCR data set. The right panel illustrates the classification of triplicates by a single independent evaluator (light blue colors). The overall classification is defined as the majority label (dark blue) from the triplicate labels.

the complete variable region, and a short part of the constant region. The individual variable genes served as representative templates for two different IGHV-specific primer sets. *Set1* is a set of 16 forward primers that was recently designed using openPrimeR, while *Set2* consists of four forward primers that were designed by Tiller et al. (2008). We performed three independent PCR reactions for each of the 20 primers on all 47 templates with the same IgM constant region-specific reverse primer from Ippolito et al. (2012) (GGTTGGGGCG-GATGCACTCC). All primers used in the experiments are listed in Table 7.1.

PCRs were performed in 25 μL reactions with 2U/rxn Platinum Taq (Thermofisher), 0.2 μM forward and reverse primer, 0.2 mM dNTPs, 1.5 mM MgCl2, and 6% kb extender under the following cycling conditions: 2 min initial denaturation at 94 °C followed by 25 cycles of 30 s at 94 °C, 30 s at 57 °C (*Set2*) or 55 °C (*Set1*), and 55 s at 72 °C. The expected 600–700 bp fragments were visualized on a 2% agarose gel supplemented with SYBR Safe (Thermofisher) and documented with the BioRAD Gel DocTM XR+ Imaging system.

7.1.2 *Data Set Construction*

The 47 IGHV fragments, which were described in the previous section, were sequenced using the approach from Sanger and annotated with IgBLAST[15]. Every experimentally evaluated primer-template pair (PTP) was assigned a label, $y_i \in \{Amplified, Unamplified\}$, based on the results of gel electrophoresis. If a band was visible in the gel, the corresponding PTP was labeled as *Amplified* and otherwise as *Unamplified* (Figure 7.1). The data set was independently labeled by five experts. The labels provided by each expert were integrated by

[15] Ye et al. 2013a

Data set	N	$N(y_i = \mathit{Amplified})$	$N(y_i = \mathit{Unamplified})$
Full	908 (100%)	382 (42.1%)	526 (57.9%)
Validation	227 (25%)	96 (42.3%)	131 (57.7%)
Training	454 (50%)	197 (43.4%)	256 (56.6%)
Testing	227 (25%)	92 (40.5%)	135 (59.5%)

Table 7.2: Distribution of data set labels. The number of observations are shown for the full data set and the constructed subsets for validation, training, and testing.

merging triplicate PCR measurements in the following way. If at least two of three measurements were labeled as *Amplified*, the corresponding PTP was included with the label *Amplified* and otherwise with the label *Unamplified*. Finally, for every PTP, the label that was provided by the majority of experts was selected.

I used openPrimeR to enrich the PCR data with several physicochemical properties relating to primers and PTPs. Using openPrimeR, the most likely binding mode for every PTP was identified by selecting the local alignment of primer and template subsequence minimizing the number of mismatches, as described in Chapter 6. Such a pairing is called *aligned PTP* in the following. In order to limit the analysis to PTPs that can be detected with a certain level of confidence, aligned PTPs with more than 12 mismatches were discarded. This reduced the size of the data set from 940 to 908 observations. Based on the aligned PTPs, further properties such as the positions of primer-template mismatches were derived. The free energy of annealing, ΔG, was computed with OligoArrayAux[16], a software for thermodynamic calculations on oligonucleotides. The calculations were performed using the optimal annealing temperatures for aligned PTPs from *Set1* and *Set2*, 55 °C and 57 °C, respectively. Additionally, the following primer-specific properties were computed via openPrimeR (see Chapter 6): primer length, extent of GC clamp, GC ratio, melting temperature, number of repeats/runs, free energy of folding ΔG_{f}, and free energy of self-dimerization ΔG_{sd}.

For model development purposes, I split the full data set into three distinct parts: validation set, training set, and test set (Table 7.2). For selecting classifier cutoffs, 25% of the observations were randomly sampled for inclusion in the validation set. Of the remaining observations, 50% were randomly sampled for inclusion in the training set, which was used for fitting a logistic regression model. The remainder of observations was included in the test set, which was used for evaluating model performance.

7.1.3 *Feature Encoding*

In order to investigate the impact of 3′ terminal mismatches, I implemented several encodings for 3′ mismatches (Figure 7.2). The mismatch feature vector $z \in \{0,1\}^6$ relies on a binary encoding to

[16] Markham and Zuker 2008

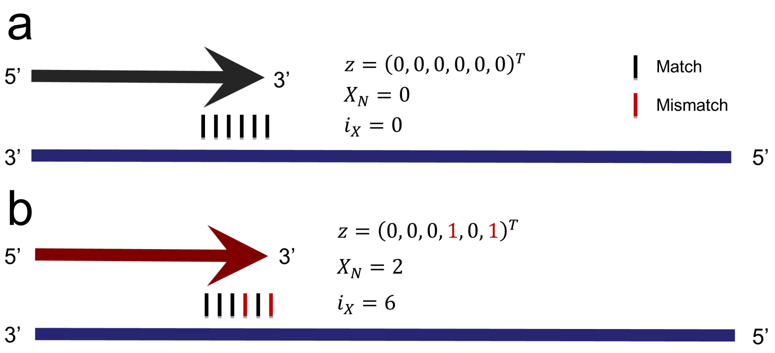


Figure 7.2: Encodings for primer-template 3′ mismatches. Primers are indicated by arrows and templates by rectangles. Black bars indicate complementary bases, while red bars indicate mismatches. (**a**) A primer annealing without any 3′ mismatches. (**b**) A primer annealing with two 3′ mismatches.

indicate whether a mismatch was identified at the j-th position in the 3′ hexamer via

$$z_j = \begin{cases} 1 & \text{if there was a mismatch at position } j \text{ in the } 3' \text{ hexamer} \\ 0 & \text{otherwise} \end{cases}.$$

Here, $j \in \{1, \ldots, 6\}$ identifies the 3′ hexamer position such that $z_j = 1$ indicates the first position in the 3′ hexamer and $z_j = 6$ indicates the 3′ terminal position. To explicitly model the augmenting effect of co-occurring mismatches in the 3′ hexamer[17], the total number of 3′ hexamer mismatches was encoded as $X_N = \sum_{j=1}^{6} z_j$.

Due to the small number of PTPs, it should be challenging to learn the association of z_j, $\forall j \in \{1, \ldots, 6\}$, with the outcome. Since positions closer to the 3′ terminus deteriorate PCR efficiency to a greater degree[18], a reasonable alternative strategy is to concentrate all the information on 3′ mismatches into a single quantity,

$$i_X = \begin{cases} \max_{j \in \{1,\ldots,6\}} \{j | z_j = 1\} & \text{if } X_N \neq 0 \\ 0 & \text{else} \end{cases},$$

the 3′ hexamer mismatch closest to the 3′ terminus.

For an example of the 3′ hexamer feature encodings consider Figure 7.2. A primer without 3′ mismatches has $z = (0,0,0,0,0,0)^T$, $X_N = 0$, and $i_X = 0$ (Panel a), while a primer exhibiting mismatches at positions 4 and 6 in the 3′ hexamer has $z = (0,0,0,1,0,1)^T$, $X_N = 2$, and $i_X = 6$ (Panel b).

7.1.4 Logistic Regression Models

I used multivariate logistic regression models (Section 3.4.1) in order to investigate the influence of individual features on the template amplification status. Logistic regression is a commonly used approach

[17] Wright et al. 2014

[18] Klein et al. 2001; Whiley and Sloots 2005; Stadhouders et al. 2010; Kwok et al. 1990; Bru et al. 2008; Ghedira et al. 2009

for problems with categorical outcomes. In this case, we would like to estimate the amplification status $y_i \in \{\mathit{Amplified}, \mathit{Unamplified}\}$. Let $\Pr(y_i = \mathit{Amplified})$ denote the probability that the i-th PTP is successfully amplified and let $\hat{p}$ indicate the corresponding estimated likelihood. Further, let β_0 indicate the model intercept and let β_i with $i \in \{1, \ldots, p\}$ indicate the weight associated with the i-th feature. Given a feature vector, $x \in \mathbb{R}^p$, the logistic regression model can be formulated as

$$\ln \frac{\hat{p}}{1 - \hat{p}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \,.$$

For the definition of a logistic regression model for PCR amplification, only features relating to PTPs were considered as terms of the logistic regression models[19]. This choice was motivated by the characteristics of the primers (Table 7.1). Since all of the selected primers were designed for the effective amplification of templates, they have similar physicochemical properties. Therefore, features based on the physicochemical properties of the primers would not be helpful for differentiating the amplification status.

Two logistic regression models relying on different sets of features were constructed for studying feature importance. The logistic regression model LR_1 was defined using the mismatch feature vector, $z \in \{0, 1\}^6$, the number of mismatches in the 3′ hexamer, X_N, and the free energy of primer-template annealing, ΔG. In order to study whether 3′ terminal mismatches can be summarized by a single feature, LR_2 was defined. This model additionally includes a term for the 3′ hexamer mismatch closest to the 3′ terminus, i_X, and the cross-term $\Delta G i_X$, which corrects for the association between ΔG and i_X.

For the definition of a logistic regression model estimating the probability of amplification, I performed feature selection using the best subset selection approach from Morgan-Tatar[20] using the features from LR_2. By minimizing the AIC (Section 3.2.5), a model with a reduced number of features, the thermodynamic mismatch model (TMM), was identified.

7.1.5 Validation of Models and Classifiers

In addition to *TMM*, I considered two other approaches for predicting template amplification status: a model based solely on the free energy of annealing (*FE*) and the thermodynamic model from DECIPHER[21] (*DE*), which considers the impact of mismatches on the efficiency of polymerase elongation. All of these models provide quantitative outputs. In order to evaluate their corresponding

[19] Examples of features relating to PTPs are i_X and ΔG. Examples of features that are not a consequence of primer-template binding are the free energy of self-dimerization or the free energy of folding.

[20] Morgan and Tatar 1972

[21] Wright et al. 2014

classifiers, cutoffs were used. For *FE*, I used the classification rule

$$f(x)_{\Delta G} = \begin{cases} \textit{Amplified} & \text{if } \Delta G(x) < \Delta G_c \\ \textit{Unamplified} & \text{else} \end{cases}$$

where $\Delta G(x)$ is the free energy of annealing of PTP x and ΔG_c is a cutoff on the free energy of annealing. For *DE*, I performed classification by applying a cutoff, η_c, on the PCR efficiency, $\eta(x)$, that is computed by DECIPHER:

$$f(x)_{\eta} = \begin{cases} \textit{Amplified} & \text{if } \eta(x) > \eta_c \\ \textit{Unamplified} & \text{else} \end{cases}$$

For *TMM*, I applied a cutoff, $\hat{p}_c$, on the estimated likelihood of amplification, $\hat{p}$:

$$f(x)_{\hat{p}} = \begin{cases} \textit{Amplified} & \text{if } \hat{p}(x) > \hat{p}_c \\ \textit{Unamplified} & \text{else} \end{cases}$$

To exemplify these classifiers, let us assume the following cutoffs: $\Delta G_c = -5$, $\eta_c = 0.01$, and $\hat{p}_c = 0.75$. Further, let us consider the PTP, x, with the following associated quantities: $\Delta G(x) = -10$ kcal/mol, $\eta(x) = 0.2$, $\hat{p}(x) = 0.6$. Then, $f(x)_{\Delta G} = \textit{Amplified}$ because -10 kcal/mol < -5 kcal/mol, $f(x)_{\eta} = \textit{Amplified}$ because $0.2 > 0.01$, and $f(x)_{\hat{p}} = \textit{Unamplified}$ because $0.6 \leq 0.75$.

Two cutoffs were selected for each of the three approaches: one cutoff ensuring an empiric specificity of at least 99% (denoted by *s*) and another cutoff maximizing Youden's index. For *FE* and *DE*, which did not require model training, I selected optimal cutoffs by maximizing the two criteria on a data set containing training and validation observations. For *TMM*, on the other hand, cutoffs were chosen by performing ten runs of fivefold CV on the validation data set, again maximizing either specificity or Youden's index. Finally, all model and classifier performances were determined on the independent test set.

In the following, classifiers optimized for overall performance and classifiers optimized for high specificity are denoted by subscription of Y and s, respectively. For example, $\textit{TMM}_s$ denotes the high-specificity *TMM* classifier and $\textit{TMM}_Y$ denotes the *TMM* classifier that was optimized for overall performance.

7.2 Results

This section presents the results from analyzing the IGHV data set. First, the results from the descriptive analysis of the IGHV data set are provided (Section 7.2.1). Next, the properties of the fitted logistic regression models, LR_1, LR_2, and *TMM* are presented (Section 7.2.2). Section 7.2.3 compares the predictive performance of *TMM*, *DE*, and *FE*. The properties of *TMM* are presented in Section 7.2.4.

7.2.1 *Properties of the Data Set*

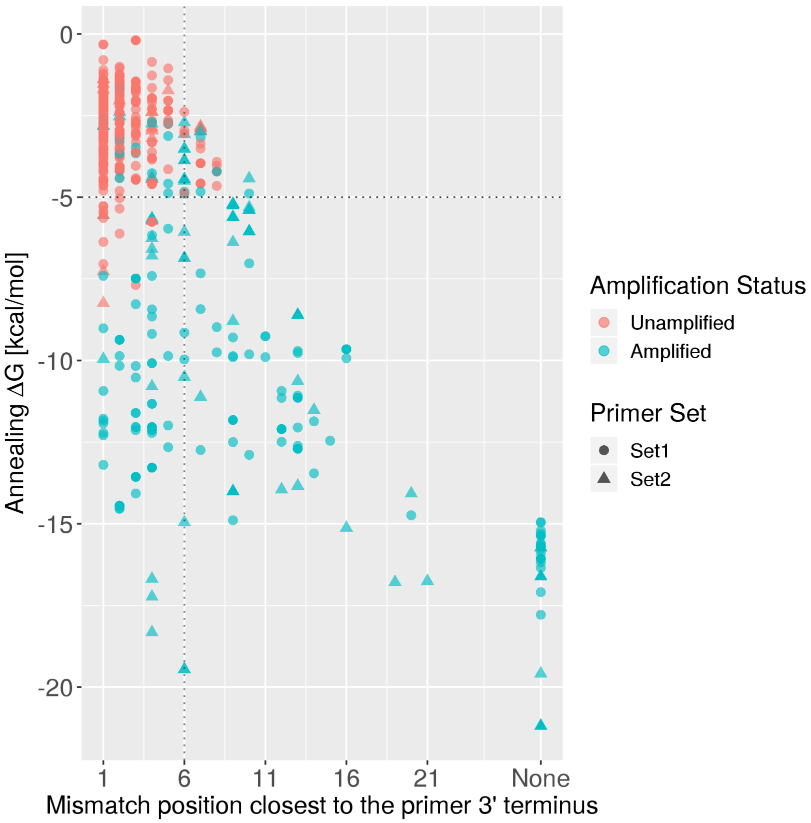


Figure 7.3: Impact of 3′ mismatches and free energy of annealing on amplification status. Every point represents a primer-template pair. Pairs that are labeled as *Amplified* are shown in blue, while those that are labeled as *Unamplified* are shown in red. Observations from *Set1* are indicated by circles and those from *Set2* by triangles. The vertical dashed line indicates the end of the 3′ hexamer, while the horizontal dashed line indicates a free energy of -5 kcal/mol.

Table 7.3 shows the distribution of the physicochemical properties of the aligned PTPs in the data set. The primers from *Set2* and *Set1* are characterized by contrasting rates of amplification. While 165 of 188 PTPs (87.8%) in *Set2* were labeled as *Amplified*, only 217 of 720 (30.1%) observations in *Set1* had a positive amplification status. Accordingly, observations from *Set1* exhibited a greater number of mismatches and higher free energies. The aligned PTPs from *Set1* had an average of 2.3 mismatches in the 3′ hexamer, while the primers from *Set2* had an average of 0.5 mismatches in this region. Moreover, while samples from *Set2* had a ΔG inter-quartile range (IQR) of [-8.6 kcal/mol, -5.2 kcal/mol], the primers from *Set1* set were associated with a higher range of [-4.9 kcal/mol, -2.0 kcal/mol].

A pronounced relationship between the number of mismatches

Property	*Set1*	*Set2*
N	720	188
ΔG [kcal/mol]	[-4.9, -2.0]	[-8.6, -5.2]
i_X	[2,6]	[0, 1]
X_N	[1, 3]	[0, 1]
$\lvert GC \rvert$	[1,2]	[1,1]
ΔG_f [kcal/mol]	[-1.53, -0.24]	[-1.24, -0.76]
ΔG_{sd} [kcal/mol]	[-2.1, -0.7]	[-1.2, -0.8]
$\lvert\{y_i \vert y_i = \mathit{Amplified}\}\rvert$	217 of 720 (30.1%)	165 of 188 (87.8%)
$\sum_{i=1}^{N} z_{i,1}$	271	25
$\sum_{i=1}^{N} z_{i,2}$	226	4
$\sum_{i=1}^{N} z_{i,3}$	272	31
$\sum_{i=1}^{N} z_{i,4}$	246	11
$\sum_{i=1}^{N} z_{i,5}$	308	12
$\sum_{i=1}^{N} z_{i,6}$	308	12

Table 7.3: Primer set properties. Values shown in brackets indicate inter-quartile ranges of observed values. N indicates the number of PTPs from each primer set. $z_{i,j}$ indicates the value of z_j for the i-th observation.

Number of mismatches	ΔG [kcal/mol]	i_X	Amplification rate
0	[-16.616, -15.696]	[0, 0]	100%
1	[-14.353, -12.1]	[0, 3]	100%
2	[-12.0455, -9.656]	[0, 3]	100%
3	[-11.607, -7.9185]	[0, 4]	100%
4	[-10.796, -7.409]	[2, 6]	92.31%
5	[-7.047, -6.047]	[0, 3]	88.89%
6	[-8.603, -5.11325]	[0, 0]	83.33%
7	[-5.39, -4.212]	[0, 3]	67.19%
8	[-5.56075, -2.539]	[3, 6]	34.04%
9	[-3.5335, -2.1325]	[4, 6]	23.08%
10	[-4.09, -1.724]	[4, 6]	18.02%
11	[-3.74, -1.695]	[4, 6]	10.53%
12	[-2.624, -1.413]	[6, 6]	3.75%

Table 7.4: Impact of primer binding properties on the rate of amplification.

and the rate of amplification was identified (Table 7.4). In this IGHV data set, all primers binding with at most three mismatches successfully amplified their templates. Even primers binding with six mismatches successfully amplified their templates in 83.3% of cases.

Comparing amplified and unamplified PTPs (Figure 7.3) revealed that the ΔG IQR of unamplified observations was higher and more concentrated ([-2.17 kcal/mol, -1.69 kcal/mol]) than for amplified observations ([-12.70 kcal/mol, -5.21 kcal/mol]). Amplified samples generally exhibited fewer mismatches in the 3′ hexamer (X_N IQR of [0,1] vs [2,4]) and mismatches occurred further from the 3′ terminus (i_X IQR of [0,3] vs [5,6]) than for unamplified samples. Applying two-sided Wilcoxon rank-sum tests (Section 3.6.4) revealed that the distributions of ΔG (p-value 1.68e-107) and i_X (p-value 1.51e-91) were significantly different when comparing *Amplified* (N = 382) and *Unamplified* (N = 526) observations.

7.2.2 *Logistic Regression Models*

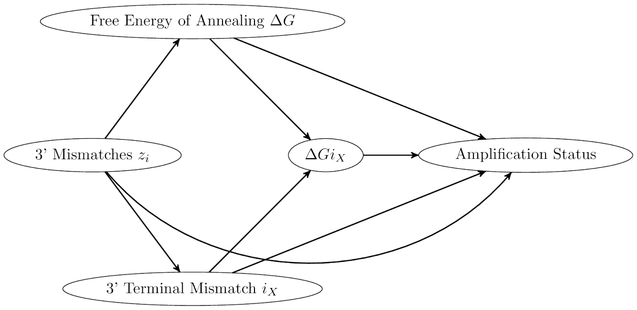

Free Energy of Annealing ΔG
3' Mismatches z_i
ΔGi_X
Amplification Status
3' Terminal Mismatch i_X

Figure 7.4: Relationship between variables in LR_2. Arrows indicate causal relationships.

Based on the properties of aligned PTPs in the IGHV data set, three logistic regression models for estimating the likelihood of template amplification were developed: LR_1, LR_2, and *TMM*. *TMM* was constructed by best subset selection on the features from LR_2. This procedure reduced the AIC of the initial model from 112.34 to 98.41 by eliminating all features except for the intercept, ΔG, and the interaction term ΔGi_X.

An overview of all generated models is shown in Table 7.5. The significance of features was evaluated based on an initial significance threshold of 0.05, which was adjusted to 0.05/9 = 0.0056 (LR_1), 0.05/11 = 0.0045 (LR_2), and 0.05/4 = 0.0125 (*TMM*) via Bonferroni correction (Section 3.6.5). According to LR_1, the features ΔG, z_3, z_4, and z_6 were significantly predictive of the amplification status. When correcting for the association between ΔG and i_X by including the ΔGi_X cross-term in LR_2 (Figure 7.4), only ΔG and ΔGi_X were found to be significantly predictive of the amplification status (Table 7.5).

7.2.3 *Comparison of Model and Classifier Performance*

The predictive performance of *TMM* was compared with the model *DE* from DECIPHER[22] and *FE*, a baseline model that relies only on ΔG. Quantitative model responses were compared with categorical amplification statuses from gel electrophoresis according to the area under the AUC (Section 3.3.3). *TMM* achieved the highest AUC (0.953) but was closely followed by *FE* (0.941) and *DE* (0.896). Testing the statistical significance of the AUCs from *TMM* and *DE* using the approach from DeLong et al. (1988) led to the rejection of the null hypothesis that there is no difference between the AUCs at a significance level of 0.05 (p = 0.0013).

[22] Wright et al. 2014

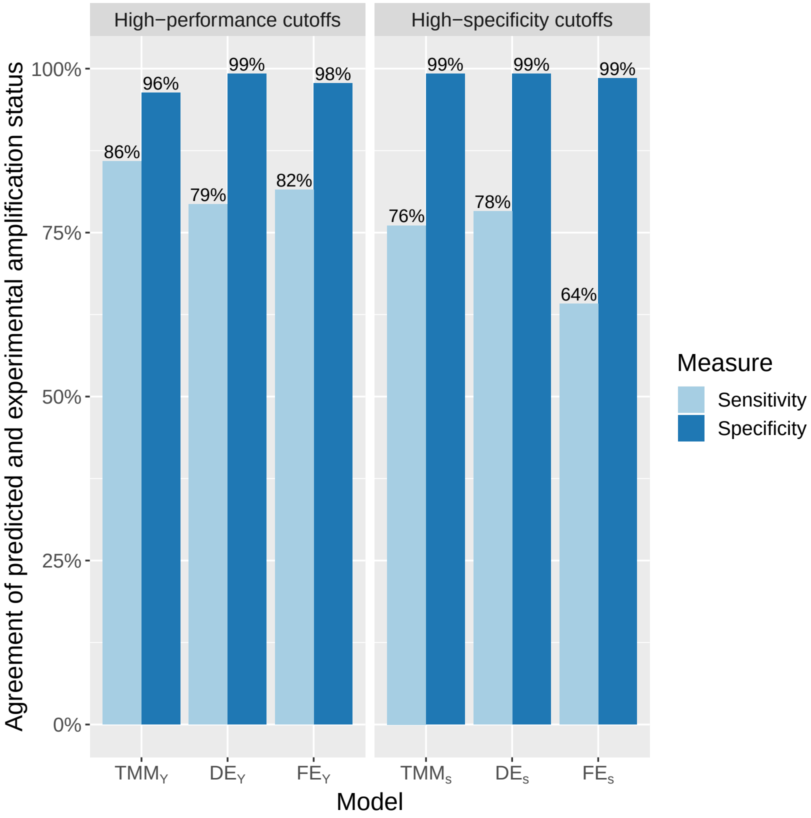


Figure 7.5: Performance of models identifying amplification events. *TMM* indicates our newly developed logistic regression model, *DE* refers to the approach from DECIPHER, and *FE* is solely based on the free energy of annealing. Models subscripted with s use cutoffs optimized for high specificity, while models subscripted with Y use cutoffs optimized for overall performance.

	LR_1		LR_2		TMM	
Feature	Estimate	p-value	Estimate	p-value	Estimate	p-value
Intercept	-2.86	1.56e-12 *	-5.76	6.16e-08 *	-3.99	<2e-16 *
z_1	-0.50	0.058	-0.187	0.4929	–	–
z_2	-0.00	0.977	-0.144	0.6164	–	–
z_3	-0.92	0.0005 *	-0.424	0.1359	–	–
z_4	-0.97	0.001 *	-0.46	0.1340	–	–
z_5	0.04	0.894	0.574	0.1085	–	–
z_6	-1.57	8.25e-08 *	-0.659	0.1069	–	–
X_N	*NA*	*NA*	*NA*	*NA*	–	–
ΔG	-0.83	$< 2e - 16$ *	-1.576	1.78e-11 *	-1.19	<2e-16 *
i_X	–	–	0.400	0.0829	–	–
$\Delta G i_X$	–	–	0.180	5.12e-05 *	0.11	<2e-16 *

Table 7.5: Comparison of the logistic regression models LR_1, LR_2, and *TMM*. *NA*s indicates features that could not be estimated due to singular matrices. Dashes indicate features that were not considered by a model. Asterisks indicate significant features.

Model	Outcome	High-specificity cutoff s	High-performance cutoff Y
TMM	$\Pr(y_i = \textit{Amplified})$	83.9%	46.1%
DE	Efficiency of PCR η	9.71e-05	1.88e-05
FE	ΔG [kcal/mol]	-6.05	-4.83

Table 7.6: Optimized cutoffs for the considered models for predicting PCR amplification. The column *Outcome* indicates the values on which cutoffs were applied.

To compare classifier performances, two types of cutoffs were determined for each model (Table 7.6). Figure 7.5 shows the predictive performance of high-performance and high-specificity classifiers. Among models using cutoffs optimized for overall performance, TMM_Y (balanced accuracy of 90.1%) slightly outperformed DE_Y (89%) and FE_Y (89.9%). Among high-specificity classifiers, TMM_s and DE_s outperformed FE_s with respect to sensitivity (77% and 78% vs 64%).

7.2.4 *Interpretation of TMM*

For interpreting *TMM*, a final model was trained on the full data set. Given ΔG and i_X, the model estimates $\hat{p} = \hat{\Pr}(y_i = \textit{Amplified})$ using the coefficients $\beta_0 = -3.99$, $\beta_1 = -1.19$, and $\beta_2 = 0.11$. The model can be formulated in the following way:

$$\begin{aligned}\ln \frac{\hat{p}}{1-\hat{p}} &= \beta_0 + \beta_1 \Delta G + \beta_2 \Delta G i_X \\ &= \beta_0 + (\beta_1 + \beta_2 i_X)\Delta G \\ &= -3.99 + (-1.19 + 0.11 i_X)\Delta G\end{aligned}$$

Since the intercept of the model is $\beta_0 = -3.99$, the odds of template amplification are low if the other terms in the model are negligible (i.e. for $\Delta G \to 0$ and $i_X \to 0$). The second term, $(-1.19 + 0.11 i_X)\Delta G$, is dominated by the free energy of annealing. For typical (negative) values of ΔG, the odds of amplification increase with decreasing ΔG because $-1.19 + 0.11 i_X$ is always negative since $0 \leq i_X \leq 6$. However, if 3′ terminal mismatches are present (i.e. $i_X \neq 0$), the extent to which the odds of amplification increase is attenuated.

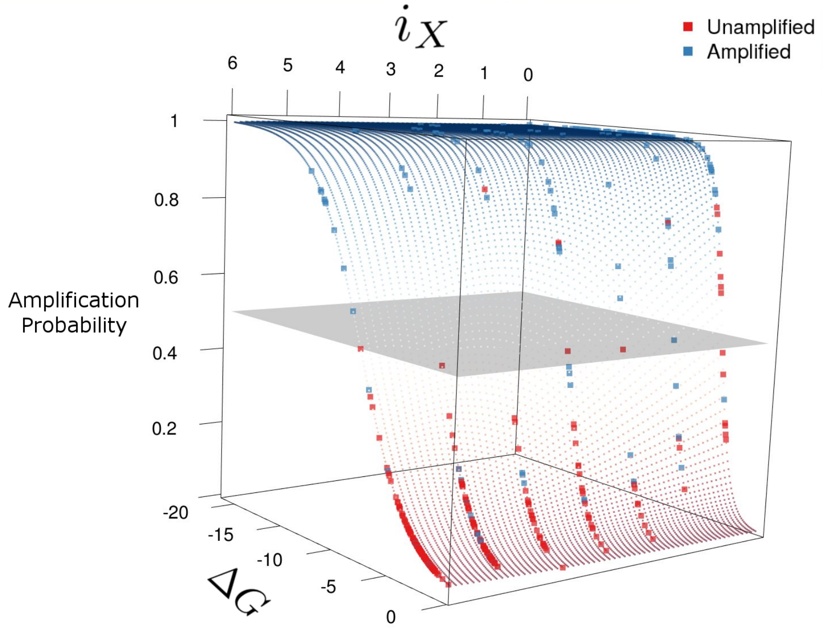


Figure 7.6: Visualization of *TMM*. Small dots indicate samples from the prediction function of the model. Red dots indicate low probabilities of amplification while blue dots indicate high probabilities. Large squares show the model estimates for the observations contained in the IGHV data set. Here, red squares indicate primer-template pairs that are labeled as *Unamplified*, while blue squares indicate observations labeled as *Amplified*.

Figure 7.6 shows the decision surface of *TMM*. The illustration demonstrates that for high and low values of ΔG (e.g. at -20 and -3 kcal/mol), $\hat{p}$ is hardly influenced by i_X. At intermediate values of ΔG (e.g. between -5 and -10 kcal/mol), however, high values of i_X considerably reduce $\hat{p}$.

7.3 Discussion

In this chapter, I analyzed the molecular conditions under which primers allow for the amplification of templates using a novel data set indicating the amplification status for all combinations of 47 immunoglobulin sequences and 20 primers. The following findings were made. First, statistical analyses revealed that the amplification status of primer-template pairs is governed by the interplay of the free energy of annealing and the presence of 3′ terminal mismatches. Second, *TMM*, a new logistic regression model for estimating the likelihood of amplification was developed. The predictive performance of *TMM* was found to be favorable when compared with the thermodynamic model *DE* from DECIPHER[23] as well as *FE*, a simple approach based only on the free energy of annealing. In the following paragraphs, these findings are discussed in more detail.

With regard to the factors influencing the amplification of primer-template pairs, the analysis of the IGHV data at hand largely confirmed the established factors. The data revealed that primer-

[23] Wright et al. 2014

template pairs whose amplification could not be detected via gel electrophoresis exhibit high free energies of annealing, an increase in the number of mismatches within the 3′ hexamer, and a tendency for displaying mismatches close to the 3′ terminus. Logistic regression modeling, however, revealed that terminal mismatches by themselves are not significantly predictive of the amplification status but only when considered in concert with the free energy of annealing. Thus, it is possible that previously reported effect sizes relating to 3′ terminal mismatches may not be solely attributable to the inhibition of polymerase attachment but also to the inhibition of primer annealing. This suggests that future studies investigating the impact of 3′ mismatches on polymerase elongation could benefit from considering how 3′ mismatches influence the free energy of annealing.

Several approaches for predicting the likelihood of PCR amplification achieved comparably high predictive performances on the IGHV data set. Still, *TMM* achieved the largest AUC and its AUC was found to be significantly different from the AUC of *DE*. Although the predictive performance of FE_Y was surprisingly high, the considerably lower performance of FE_s indicates that the free energy of annealing by itself lacks robustness. In contrast to *DE*, which estimates the efficiency of polymerase elongation according to the impact of position- and base-specific effects in the 3′ region, *TMM* considers only the position of 3′ mismatches. The following three observations could explain why the consideration of base-specific effects by *DE* did not provide an advantage over *TMM*. First, none of the primers contained in the IGHV data set displayed terminal nucleotides other than G or C (Table 3). Second, since base-specific differences in amplification efficiencies were reported only for RT-PCR[24], these difference may simply not be observable in the present data set due to the high sensitivity of gel electrophoresis. Third, terminal mismatches seem to have the greatest influence on the amplification status indicated by gel electrophoresis at low to intermediate free energies of annealing ($\Delta G \in [-10, -5\}$). Since these values of ΔG are associated with multiple mismatches, the computational placement of primers relative to templates is subject to a certain level of uncertainty. Incorrectly aligned primer-template pairs may exhibit 3′ mismatches that are not actually observed. In this case, it would be challenging to find an association between specific types of 3′ mismatches and the amplification status.

The newly developed *TMM* model for predicting amplification events has several advantages over *DE*. First, since the model is based only on ΔG and i_X, it is easily interpretable and it is unlikely that the model suffers from overfitting. Second, the model computes an intuitive measure, the probability of amplification. Third, the model

[24] Wright et al. 2014

performs at least as well as *DE* on the IGHV data set. A limitation of *TMM* is that it does not consider as many predictors as a previous model[25], which is not publicly available. Since *TMM* was developed using primers exhibiting favorable properties such as the absence of self-dimers and the presence of a GC clamp (Table 7.3), it is likely to overestimate the amplification status for primers exhibiting less favorable properties or when templates exhibit secondary structures[26]. Therefore, the model should not be applied on primers that exhibit unfavorable physicochemical properties that may preclude their amplification.

Finally, I would like to discuss the choice of models for multiplex primer design. In multiplex primer design, false positive predictions should be avoided at all costs as they may preclude the amplification of templates that are not redundantly covered. False negative predictions, on the other hand, are much more tolerable because they merely result in larger primer sets in which primers exhibit redundant coverage. Therefore, amplification events should be detected with high specificities. The present data suggest that even simple approaches can be used for this purpose. For example, a specificity of 100% can be obtained by designing primers with at most 3 mismatches (Table 7.4). Models such as TMM_s and DE_s have the advantage that they provide higher sensitivities at similar specificities. Therefore, these models could assist in designing primer sets for template libraries for which no set of appropriate size has been found yet[27].

To conclude, this chapter demonstrated that the interplay of the free energy of annealing and the presence of 3′ terminal mismatches are the main factors influencing the success of PCR amplification. Based on this insight, *TMM*, a simple logistic regression model for predicting amplification events was developed. On the present data, *TMM* performs at least as well as other models for predicting PCR amplification and is freely available through openPrimeR, which can be found at openprimer.mpi-inf.mpg.de.

[25] Yuryev et al. 2002

[26] Lvovsky et al. 1998

[27] Gardner et al. 2009

Part III:

Synthesis

This part connects the contributions that were presented in Part II. Chapter 8 summarizes and critically discusses the developed methods. Chapter 9 reflects on the use of intelligent systems in medicine and comments on the fight against viral infections.

8
Conclusions

Everything should be made as simple as possible, but no simpler.

Albert Einstein

In this dissertation, I developed computational methods with the aim of improving treatment and prevention of viral infections. In the following paragraphs, I draw conclusions for the individual scientific contributions that were presented in Chapters 4, 5, 6, and 7. Thereafter, I reflect on the impact and implications of my contributions.

Chapter 4 introduced the geno2pheno[ngs-freq] web service, which allows for the detection of drug resistance from NGS samples of HIV-1 or HCV. In this way, the tool supports the selection of antiviral treatment regiments. The main contributions of geno2pheno[ngs-freq] are threefold. First, the tool allows for the detection of drug resistance in minority populations, which is not possible with approaches that accept only Sanger sequences as an input. Second, in contrast to other web services, the approach decouples NGS data processing and drug resistance interpretation through the use of frequency files. Third, geno2pheno[ngs-freq] eases the interpretation of drug resistance through novel visualizations. geno2pheno[ngs-freq] could be expanded to support additional viral regions (e.g. HIV-1 integrase) or other viral species. Currently, drug resistance predictions for minor viral populations should be considered with particular care due to biological and technical reasons. In the future, novel prediction models based on NGS data could improve the interpretation of drug resistance in the following way. Multi-task learning, with which multiple drugs can be considered simultaneously, could improve the performance of predictive models based on clinical data. Cutoff-based consensus generation could be eliminated via encoding amino acids by their frequencies (e.g. as provided by frequency files) or through multi-instance learning, in which sets of individual NGS reads are considered. Finally, quasispecies reconstruction may allow for the consideration of resistance mutations that occur on the same strand while taking viral replicative competence into account.

In Chapter 5, I investigated HIV-2 coreceptor usage and trained an

SVM model that has advantages over the only available rules-based approach for HIV-2 coreceptor identification[1]. The model was integrated into geno2pheno[coreceptor-hiv2], the first web server for the identification of HIV-2 coreceptor usage. The service can support the management of HIV-2 infection by assisting clinicians considering the prescription of coreceptor antagonists and enables the realization of large-scale epidemiological studies on HIV-2 coreceptor usage. The work confirmed that the V3 is the major determinant of HIV-2 coreceptor usage, for which novel molecular markers were revealed. Moreover, geno2pheno[coreceptor-hiv2] improves upon the interpretability of previous work through a new visualization for coreceptor predictions based on V3 amino-acid sequences. A limitation of the approach is that it considers only amino acids in the V3 and not those in other regions of the viral surface glycoprotein. The analyzed data suggest that it would be important to develop a standardized assay for the phenotypic determination of HIV-2 coreceptor usage, to shed more light on HIV-2 intra-host evolution, and to investigate the V1 and V2 regions in more detail.

Chapter 6 introduced openPrimeR, a reliable, open-source software that provides a novel approach for designing, evaluating, and comparing primer sets for mPCR. The application of openPrimeR on human immunoglobulin sequences led to novel primer sets for IGHV, IGKV, and IGLV. The analysis of the IGHV primers revealed that primers, which bind to the early leader region, allow for the enhanced amplification of highly mutated antibody sequences. Most importantly, it could be shown that the newly designed primer sets enable the isolation of bNAbs against HIV-1 that are especially heavily mutated. In the future, samples from elite neutralizers should be re-analyzed using the newly designed primer sets in order to investigate whether novel bNAbs can be discovered. openPrimeR could be further enhanced by providing an efficient, stand-alone library for multiplex primer design (e.g. in C++) and improving the algorithmics of the primer design procedures.

Chapter 7 investigated the features that are predictive of successful PCR amplification events and introduced *TMM*, a new model for the prediction of amplification events, which performs similarly to the thermodynamic model from DECIPHER[2]. The free energy of annealing and the simultaneous presence of 3′ mismatches were identified as the key drivers of amplification success. Since the free energy of annealing seems to be the main factor controlling whether amplification takes place, simple strategies such as calling amplification events with at most three mismatches or with less than -10 kcal/mol can already reach sufficient specificities for primer design. More intricate methods such as *TMM*, however, allow for greater

[1] Visseaux et al. 2011

[2] Wright et al. 2014

sensitivities. *TMM* should be useful to all practitioners of PCR because the model enables improved primer design. Due to the high quality of the primers in the training data, *TMM* should not be applied to estimate the amplification of primers with less favorable physicochemical properties. The findings of this work are limited by the scarcity of PCR data. To obtain a more generalizable model for the prediction of PCR amplification events, it would be necessary to perform a dauntingly large number of PCR measurements because two criteria need to be considered simultaneously. First, the impact of all types of 3′ mismatches would need to be systematically analyzed across all possible values of the free energy of annealing. Second, to consider the physicochemical properties of primers, it would be necessary to experimentally measure the amplification of primers with varying physicochemical properties.

The work presented in this dissertation constitutes advances for treatment and prevention of viral infections. Only a few months after its inception, geno2pheno[ngs-freq] has already entered clinical routine and is in constant use by the virologists from the University of Cologne. Since the tool provides a more detailed account of drug resistance than previously used approaches such as geno2pheno[resistance] and geno2pheno[hcv], geno2pheno[ngs-freq] allows for the further personalization of treatments against HIV-1 and HCV infection. The unique feature of geno2pheno[coreceptor-hiv2] is that it represents the only web service for the identification of HIV-2 coreceptor usage. Thus, geno2pheno[coreceptor-hiv2] is the first place for clinicians to go to when they consider prescribing coreceptor antagonists to HIV-2-infected patients. The primer sets that were designed by openPrimeR enhance the isolation of bNAbs to HIV-1. The use of these primer sets may enable the discovery of novel antibodies that could be used for treating and preventing HIV-1 infection in the future.

To conclude, I would like to highlight the significance of geno2pheno[ngs-freq] and openPrimeR. The availability of antiviral drugs with increased potency has reduced the rate of treatment failure due to drug resistance. Nevertheless, drug resistance is a major concern of antiviral treatment because drug resistant strains are still frequently transmitted and resistant variants emerge especially commonly in countries of the global South, which lack well-developed clinical infrastructures and access to potent drugs. The existence of freely available resistance interpretation engines is particularly important for guiding the treatment of infected persons from these countries. It is important that available interpretation engines are not only maintained but also further developed. geno2pheno[ngs-freq] constitutes such an advancement. The treatment of HIV infection could undergo substantial changes in the coming years due to the implementation of antibody-based therapies. Antibodies are especially promising treatment options for patients that are infected with highly resistant viral strains, patients with organ damage, or patients that struggle with daily medication. Due to their long half life, antibodies are also promising agents for PrEP. The primer sets that were designed by openPrimeR could aid the discovery of antibodies that may, at one point, enter clinical routine.

Having recapitulated the significance of my scientific contributions, I would like to use the outlook (Chapter 9) to comment on the requirements for the further digitalization of medicine and to reflect on our progress towards the elimination of viral infectious disease.

9 Outlook

I will follow that system of regimen which, according to my ability and judgment, I consider for the benefit of my patients, and abstain from whatever is deleterious and mischievous.

The Hippocratic Oath, written 400 BCE

In this final chapter of the main matter, I would like to comment on two aspects that are relevant for my work on computational approaches for improving treatment and prevention of viral infections. The first aspect concerns how intelligent systems can support medical decision making and the roadblocks that may prevent their widespread use. The second aspect deals with the progress that has been made in the fight against viral infectious disease and the challenges that still need to be overcome.

On the Use of Intelligent Systems in the Biomedical Sector The ever-increasing amount of medical data is accompanied by rising technological requirements for their storage, processing, and interpretation. There is no sign that this trend is going to decline, which is why it is crucial that the necessary infrastructures for smart information management are established as soon as possible. Only then can data from several sources be effectively connected, investigated, and used for deploying machine learning models. Naturally, to obtain such models, we need machine learning experts. However, the success of these models does not only hinge on the technical expertise of the developers but also on their domain knowledge and understanding of the user requirements. On the one hand, domain knowledge is important in order to prevent fallacies during predictive modeling. For example, if a machine learning expert would create a model that predicts HCV drug resistance without accounting for distinct HCV geno- and subtypes, the model would perform poorly. On the other hand, a deep understanding of the user requirements is necessary to ensure that the resulting system will be accepted by the medical community. Clinicians in particular favor interpretability over predictive performance. To ensure that both conditions (i.e. domain knowledge and understanding of the user requirements) are met, it is necessary to include medical experts in the development of predictive engines, in order to allow for an exchange of information. Additionally, by including clinicians during the development process, they will have a better grasp of how predictive engines function, which should improve their acceptance in the community. For this, clinicians need to be open to learning new concepts. A successful example for a fruitful exchange between developers and users is geno2pheno[coreceptor][1]. Because clinicians were involved in the development process, it was possible to produce an interpretable tool that introduced the statistical concept of false positive rates into the virological community.

A roadblock that may prevent the establishment of intelligent systems in the biomedical field is constituted by prediction engines that supplant human decision making rather than supporting it. Since personal decision making informs individualism, a fundamental constituent of Western culture, it is unlikely that automated decision making will replace human decision making without a profound cultural shift. Maintaining the paradigm of expert decision making is also in the self-interest of physicians. If expert decision making is deemed to be the ne plus ultra by society, automated computational systems do not pose a threat to the medical profession and it is acceptable to disregard them. This could be the reason why treatment recommendation systems such as THEO[2] or the EuResist engine[3] have not been adopted by the medical community even though they have been shown to be more capable at selecting effective antiviral treatments than a selection of physicians[4]. It seems plausible that intelligent systems that do not compete with humans but rather facilitate human decision making, will find a greater level of acceptance. For example, genotypic resistance testing systems, which assist the drug selection procedure and increase the agency of physicians, are widely used.

Another factor is the complexity of the problem and that of the predictive model. When a statistical model solves a complex problem, users may be skeptical of the model, wondering whether all factors were correctly taken into account. For example, resistance interpretation systems deal with a manageable problem whose solution can be found based on the amino-acids of the viral sequence alone. Treatment optimization, on the other hand, is a more complex problem, as it should take characteristics relating to the patient (e.g. ethnicity or lifestyle)[5], treatment information (e.g. previous and current treatments or polypharmacy[6]), and data from clinical monitoring (e.g. VLs or CD4 counts)[7] into account. Even state-of-the-art treatment recommendation systems such as geno2pheno[integrated] do not take all of the aforementioned, potentially relevant features into account but are solely based on viral amino-acid sequences. It is natural that physicians are skeptical when an intelligent system outputs treatment

[1] Lengauer et al. 2007

[2] Altmann et al. 2007

[3] Zazzi et al. 2011

[4] Altmann et al. 2009; Zazzi et al. 2011

[5] Genetic and environmental factors can influence the way in which drugs are metabolized, potentially calling for adjusted drug dosing schemes.

[6] Polypharmacy describes the simultaneous administration of multiple drugs. Since the concurrent use of multiple medications may lead to unwanted interactions, the goal of polypharmacy is to ensure that drug concentrations remain in their therapeutic range rather than falling below (ineffective treatment) or exceeding it (adverse events).

[7] Treatments could be simplified or intensified based on clinical measurements.

recommendations based on a subset of the relevant predictors that are available to physicians.

However, such skepticism should not be a general skepticism but skepticism that is based on the properties of the predictive engine. Let us consider the commercial treatment recommendation system Watson Oncology as an example. Watson's training data solely originates from Memorial Sloan Kettering Cancer Center in New York. Therefore, the system is not only biased towards the characteristics of American patients but also to the treatment strategies that are employed by the physicians from a single institution[8]. Accordingly, it could be shown that the system exhibits a low concordance with treatment decisions that are taken in hospitals located outside the USA[9]. As a consequence, it would be reasonable to avoid the use of Watson in countries other than the USA. However, whether Watson is used should be decided based on the facts (e.g. the properties of the training data) rather than general skepticism about intelligent systems.

In the following two paragraphs, I will sketch two scenarios for the future of clinical decision making, one where automated decision making permeates medicine and one where it is rejected. In the first scenario, clinicians have all necessary information at their fingertips due to advances in clinical information-technology infrastructures. However, rather than spending their valuable time pondering treatment decisions, they mostly use automated systems for this purpose. The time that is freed up in this way, is used for interacting with patients and dealing with challenging cases, which still require human intervention. Since open data is the norm, all clinical centers feed their anonymized data into a central, publicly available repository that is hosted by the World Health Organization. Prediction models for the most common diseases are provided without charge, in terms of cloud services hosted by the World Health Organization. Since these models have been developed with the utmost care in collaborations with clinicians, they are trustworthy, interpretable, and highly accurate as they incorporate all global health data. Commercial providers fill the gaps by providing models with additional features or models that target rare diseases.

In the second scenario, clinics still produce a large amount of data but sharing of data has come to a stop because clinics are concerned about lawsuits, which are becoming more and more common due to tight data protection laws. Since few data are publicly available, there exist only commercial prediction engines, which rely on privately generated data. Because every clinic only purchases the predictive system that performs best, companies are either pushed out of the business or are bought up, resulting in a monopoly on intelligent

[8] STAT 2018

[9] Lee et al. 2018

medical systems. Because these commercial systems are intransparent, outsiders cannot gauge whether they are biased or not. Moreover, the existence of a monopoly allows for arbitrarily high licensing fees for prediction systems such that only affluent individuals can afford to be treated at institutions that have access to the best-performing engines.

The future of clinical decision making probably lies somewhere between these two scenarios. To set the course towards the first scenario, the following steps should be undertaken. First, infrastructures have to be improved such that every clinic systematically collects and shares treatment-relevant data. This requires that funding for the digitalization of hospitals is increased. Second, clinicians need to become acquainted with modern, open approaches for dealing with data[10]. It should be the norm that data that have been generated using public funds, are made available to the public. Of course, to protect the privacy of the patient, these data should be stored in an anonymized fashion[11]. Third, when models are made publicly available, their properties should be clearly communicated. For example, stored models should provide summary statistics about the training data, performance on test sets, as well as meta information. Currently, there is already an active online exchange of data sets as well as (deep learning) models using online platforms such as Kaggle. However, statistical models and data for medical purposes are still not widely available. Fourth, clinicians and model developers need to learn from each other. In order to make intelligent systems more accessible to clinicians, clinicians should receive more training in statistics and computer science. On the other hand, model developers need to learn from clinicians in order to avoid fundamental modeling fallacies and to ensure that models are interpretable. Fifth, model developers should stringently validate models to ensure high generalizability. Finally, to make sure that patients are compliant, patient data need to be handled with care and the conditions under which data are stored need to be clearly communicated. Most importantly, patients should be informed about how their data could improve the treatment of future patients.

Recent advances in the fields of machine learning combined with the increasing digitalization have the potential to revolutionize biomedical research and medicine.[12] However, while everyone is talking about industry 4.0 (i.e. the 4th industrial revolution)[13], a state in which industry is more digitalized and automated, the equivalent term for medicine, medicine 4.0, is rarely used[14]. One potential reason why medical digitalization seems to progress more slowly is that there are strict laws specifying the requirements for medical devices. For example, decision support software qualifies

[10] Wilkinson et al. 2016

[11] Privacy-preserving machine learning methods (Mohassel and Zhang, 2017) could play an important role for training models that require data from several data sources.

[12] For example, Recursion Pharmaceuticals is currently working on an automated, high-throughput pipeline for drug discovery that combines cell imaging with machine learning (Marketwired, 2017).

[13] The 1st industrial revolution (1760–1820) marked the transition from manual production to production by machines. The 2nd industrial revolution (1870–1914) is characterized by rapid industrialization and the implementation of new technologies and manufacturing processes. The 3rd industrial revolution is the digital revolution that began in the 1980s and heralded the information age. In this ongoing period, personal computers and the internet began to be widely adopted.

[14] This may also be because the notion of *medical revolution* has never reached wide publicity, in contrast to the notion of *industrial revolution*.

as a medical device according to the guidelines of the European Commission[15]. Making the vision of medicine 4.0 a reality requires that we close the gap between biomedical researchers, physicians, and experts from quantitative disciplines such as bioinformatics, statistics, and computer science.

On The Fight Against Viral Infections The last years have brought tremendous changes to the treatment of HIV and HCV infection. Research into antiretrovirals has led to novel medications with greater potency at reduced side effects. The improved durability of antiretroviral regimens is showcased by US patients enrolled in the HOPS study. Within roughly ten years, the median duration of first-line treatments increased from one year (1996–1999 period) to four and a half years (2008–2011 period)[16]. The increased potency of antiretroviral agents has led to increased rates of viral suppression. Treatment data from eight American hospitals show that viral suppression (defined as less or equal to 400 HIV RNA copies per mL) increased from 32% in 1997 to 86% in 2015. Similarly, the introduction of new compounds has revolutionized the treatment of HCV. Recently introduced pangenotypic DAAs targeting HCV allow for SVR rates of up to 99% within twelve weeks of treatment[17]. Although drug resistance is an issue that needs to be considered when treating patients infected with either of the two viruses, these challenges can be overcome using a wide range of computational tools.

The advancement of antiviral treatment and the availability of tools for handling drug resistance suggest that the main challenges for containing and eradicating HIV and HCV lie elsewhere. Let us first deal with HIV. The WHO has postulated the 90-90-90 goals in 2017[18]. According to these goals, by 2020, 90% of HIV-infected persons should be diagnosed, 90% of diagnosed persons should be on treatment, and 90% of those on treatment should have suppressed viral loads. For well-developed countries such as Sweden, Germany, or France, it is easily possible to achieve these values. For example, Sweden was the first country to meet the 90-90-90 targets, currently achieving 90-95-92[19], while Germany is close to reaching the targets, with 85-84-93[20]. This is because the aforementioned countries have strong healthcare systems, well-developed clinical infrastructures, a sufficient number of well-trained medical personnel, and the necessary resources for accessing the most potent drugs. However, countries that are struggling economically (e.g. Russia or the majority of African as well as South American nations), will have a hard time reaching these targets. For example, Russia is at 94-23-81[21] and Nigeria is at 53-48-81[22,23]. With regard to the total number of HIV-infected persons in Russia and Nigeria, these numbers correspond to

[15] European Commission 2016

[16] Sheth et al. 2016

[17] Puoti et al. 2018

[18] WHO 2017

[19] Gisslén et al. 2017

[20] ECDC 2017

[21] Pokrovskaya et al. 2014

[22] PEPFAR 2017

[23] Data on the global progress towards the 90-90-90 goals can be visualized on the 90-90-90 Watch website.

suppression rates of merely 17.5% and 20.6%, respectively.

Evidently, the rate of suppression relative to the treated population is similar for all countries, with a rate of 92% in Sweden, 93% in Germany, and 81% in both Russia and Nigeria. These data suggest that, if treatment is performed, viral suppression can be reached for most patients, independent of the country where they are treated. Thus, it seems that the main difference between high-income and low-income countries is that high-income countries have much higher rates of diagnosis (e.g. 90% in Sweden vs 53% in Nigeria) and treatment (e.g. 95% in Sweden vs 23% in Russia).

For HCV, similar conclusions can be drawn. Since anti-HCV treatments are highly effective, HCV can typically be eliminated as long as the infection is diagnosed and appropriate drugs are available. The 2030 WHO action plan for HBV and HCV treatment[24] suggests that the low rate at which HCV is diagnosed and treated is the main challenge[25].

All in all, the data suggest that it is necessary to invest more resources into aid programs that are active in low-income countries in order to educate persons at risk of infection, support the development of clinical infrastructures, and provide medical training. Only by assisting the countries that are hotspots of HIV and HCV infection will it be possible to contain the global epidemic of these infectious diseases.

[24] WHO 2016

[25] The combined diagnosis and treatment rates of HBV and HCV are only at 5% and 1%, respectively, which is much lower than for HIV.

Part IV:

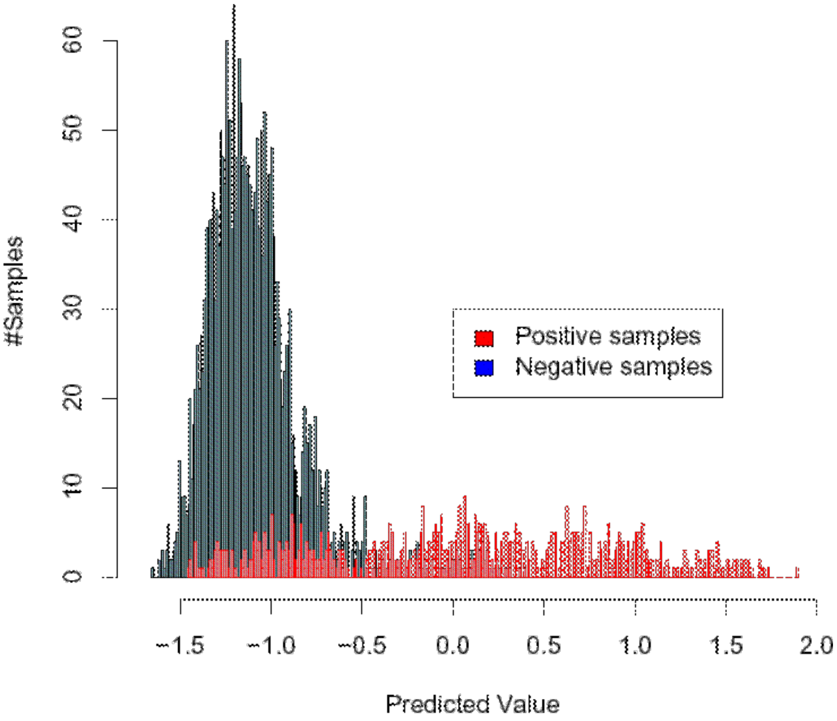
Appendix

This part comprises information that supports the contents of this dissertation. Chapter A provides supplemental information. Chapter B lists the scientific contributions that I made during my doctoral phase. Chapter C presents the results of evaluating this dissertation using plagiarism prevention software.

A
Supplementary Material

Table A.1: IUPAC nucleotide ambiguity codes.

IUPAC Code	Nucleotides
A	A
C	C
G	G
T	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	G or A or T or C



Retrieved from the help page of the geno2pheno[coreceptor] website on December 19th, 2018.

Figure A.1: Decision values from geno2pheno[coreceptor] for HIV-1 V3 sequences. The positive class corresponds to X4-capable variants, while the negative class corresponds to viruses using CCR5.

	IGHV		IGKV		IGLV	
Property	Setting	Limit	Setting	Limit	Setting	Limit
Binding region	[-60,-20]	NA	[-60,-20]	NA	[-60,-10]	NA
Primer length	[18,28]	NA	[18,28]	NA	[18,30]	NA
Max mismatches	1	NA	3	NA	3	NA
Max 3′ mismatches	0	NA	0	NA	0	NA
Prevent stop codons	✓	NA	✓	NA	✓	NA
Prevent substitutions	✗	NA	✗	NA	✗	NA
Free energy of annealing	✗	NA	✗	NA	✗	NA
Amplification efficiency	✗	NA	✗	NA	✗	NA
TMM model	✗	NA	✗	NA	✗	NA
Specificity	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]
|GC|	[1, 3]	[1, 3]	[1, 3]	[1, 3]	[1, 3]	[1, 3]
GC%	[0.4, 0.6]	[0.3, 0.7]	[0.4, 0.6]	[0.3, 0.7]	[0.4, 0.6]	[0.3, 0.7]
Runs	[0, 4]	[0, 6]	[0, 4]	[0, 6]	[0, 4]	[0, 6]
Repeats	[0, 4]	[0, 6]	[0, 4]	[0, 6]	[0, 4]	[0, 6]
ΔG_{sd} [kcal/mol]	[-5, ∞]	[-5, ∞]	[-5, ∞]	[-5, ∞]	[-5, ∞]	[-5, ∞]
T_m [°C]	[60, 75]	[57, 78]	[60, 75]	[57, 78]	[60, 75]	[57, 78]
ΔT_m [°C]	[0, 3]	[0, 3]	[0, 3]	[0, 3]	[0, 3]	[0, 3]
ΔG_{cd} [kcal/mol]	[-5, ∞]	[-5, ∞]	[-5, ∞]	[-5, ∞]	[-5, ∞]	[-5, ∞]

Table A.2: Primer design settings for IGH, IGK, and IGL.

Identifier	Reference	Year	Size
Brezinschek1995_1st	(Brezinschek et al., 1995)	1995	6 (6)
Brezinschek1995_2nd	(Brezinschek et al., 1995)	1995	6 (14)
Ippolito2012	(Ippolito et al., 2012)	2012	8 (23)
Marks1991	(Marks et al., 1991)	1991	6 (6)
Glas1999	(Glas et al., 1999)	1999	43 (43)
Persson1991	(Persson et al., 1991)	1991	4 (4)
Cardona1995	(Cardona et al., 1995)	1995	6 (6)
vanEs1991	(van Es et al., 1991)	1991	5 (5)
Rubinstein1998	(Rubinstein et al., 1998)	1998	7 (7)
vanEs1991	(van Es et al., 1991)	1991	5 (5)
Chong2002	(Chong et al., 2002)	2002	6 (6)
Weng1992	(Weng et al., 1992)	1992	1 (1)
Glamann1998	(Glamann et al., 1998)	1998	9 (9)
Manheimer1991	(Manheimer-Lory et al., 1991)	1991	4 (4)
Verhagen2000	(Verhagen et al., 2000)	2000	7 (7)
Szczepa2001	(Szczepański et al., 2001)	2001	6 (6)
vanDongen2003_A	(Van Dongen et al., 2003)	2003	6 (6)
vanDongen2003_B	(Van Dongen et al., 2003)	2003	7 (7)
vanDongen2003_C	(Van Dongen et al., 2003)	2003	7 (7)
vanDongen2003_single	(Van Dongen et al., 2003)	2003	1 (1)
Kueppers1993_1st	(Küppers et al., 1993)	1993	6 (10)
Lim2010	(Lim et al., 2010)	2010	8 (23)
Murugan2015_1st	(Murugan et al., 2015)	2015	9 (11)
Murugan2015_2nd	(Murugan et al., 2015)	2015	1 (2)
Sblattero1998	(Sblattero and Bradbury, 1998)	1998	7 (22)
Scheid2011_1st	(Scheid et al., 2011)	2011	21 (21)
Tiller2008_1st	(Tiller et al., 2008)	2008	4 (5)
Tiller2008_2nd	(Tiller et al., 2008)	2008	6 (6)
Wardemann2003_1st	(Wardemann et al., 2003)	2003	4 (5)
Wu2010	(Wu et al., 2010b)	2010	6 (6)

Table A.3: Overview of IGHV primer sets provided by openPrimeR. *Size* indicates the number of primers where the number of disambiguated primers is shown in brackets.

Identifier	Reference	Year	Size
Atkinson1996	(Atkinson et al., 1996)	1996	3 (3)
Beishuizen1997	(Beishuizen et al., 1997)	1997	3 (3)
Cardona1995	(Cardona et al., 1995)	1995	4 (4)
Chen1986	(Chen et al., 1986)	1986	1 (1)
Chen1987	(Chen et al., 1987)	1987	1 (1)
Cox1994	(Cox et al., 1994)	1994	6 (19)
Giachino1995	(Giachino, 1995)	1995	2 (2)
Glamann1998	(Glamann et al., 1998)	1998	5 (5)
Huber1993	(Huber et al., 1993)	1993	4 (4)
Ippolito2012	(Ippolito et al., 2012)	2014	4 (27)
Juul1997	(Juul et al., 1997)	1997	1 (1)
Kueppers1993_1st	(Küppers et al., 1993)	1993	6 (26)
Lim2010	(Lim et al., 2010)	2010	6 (30)
Manheimer1991	(Manheimer-Lory et al., 1991)	1991	4 (4)
Marks1991	(Marks et al., 1991)	1991	6 (6)
Murugan2015_1st	(Murugan et al., 2015)	2015	3 (6)
Murugan2015_2nd	(Murugan et al., 2015)	2015	1 (24)
Padyukov2001	(Padyukov et al., 2001)	2001	2 (2)
Persson1991	(Persson et al., 1991)	1991	2 (2)
Pongers1999	(Pongers-Willemse et al., 1999)	1999	8 (8)
Rubinstein1998	(Rubinstein et al., 1998)	1998	4 (4)
Sblattero1998	(Sblattero and Bradbury, 1998)	1998	4 (27)
Tiller2008	(Tiller et al., 2008)	2008	3 (6)
Timmers1993	(Timmers et al., 1993)	1993	5 (5)
vanBurg2001	(van der Burg et al., 2001)	2001	9 (9)
vanDongen2003	(Van Dongen et al., 2003)	2003	6 (6)
vanEs1991	(van Es et al., 1991)	1991	4 (4)
Wardemann2003_1st	(Wardemann et al., 2003)	2003	3 (6)

Table A.4: Overview of IGKV primer sets provided by openPrimeR. *Size* indicates the number of primers where the number of disambiguated primers is shown in brackets.

Identifier	Reference	Year	Size
Cardona1995	(Cardona et al., 1995)	1995	1 (1)
Farner1999_1st	(Farner et al., 1999)	1999	8 (27)
Farner1999_2nd	(Farner et al., 1999)	1999	9 (20)
Ippolito2014	(Ippolito et al., 2012)	2014	9 (49)
Lim2010	(Lim et al., 2010)	2010	10 (51)
Marks1991	(Marks et al., 1991)	1991	7 (7)
Moraes2003	(Junta and Passos, 2003)	2003	3 (3)
Murugan2015_1st	(Murugan et al., 2015)	2015	7 (11)
Murugan2015_2nd	(Murugan et al., 2015)	2015	2 (10)
Rubinstein1998	(Rubinstein et al., 1998)	1998	1 (1)
Sblattero1998	(Sblattero and Bradbury, 1998)	1998	9 (49)
Stiernholm1994	(Stiernholm et al., 1994)	1994	4 (4)
Tiller2008	(Tiller et al., 2008)	2008	7 (11)
vanBurg2001	(van der Burg et al., 2001)	2001	7 (7)
vanDongen2003	(Van Dongen et al., 2003)	2003	2 (2)
Wardemann2003	(Wardemann et al., 2003)	2003	7 (11)

Table A.5: Overview of IGLV primer sets provided by openPrimeR. *Size* indicates the number of primers where the number of disambiguated primers is shown in brackets.

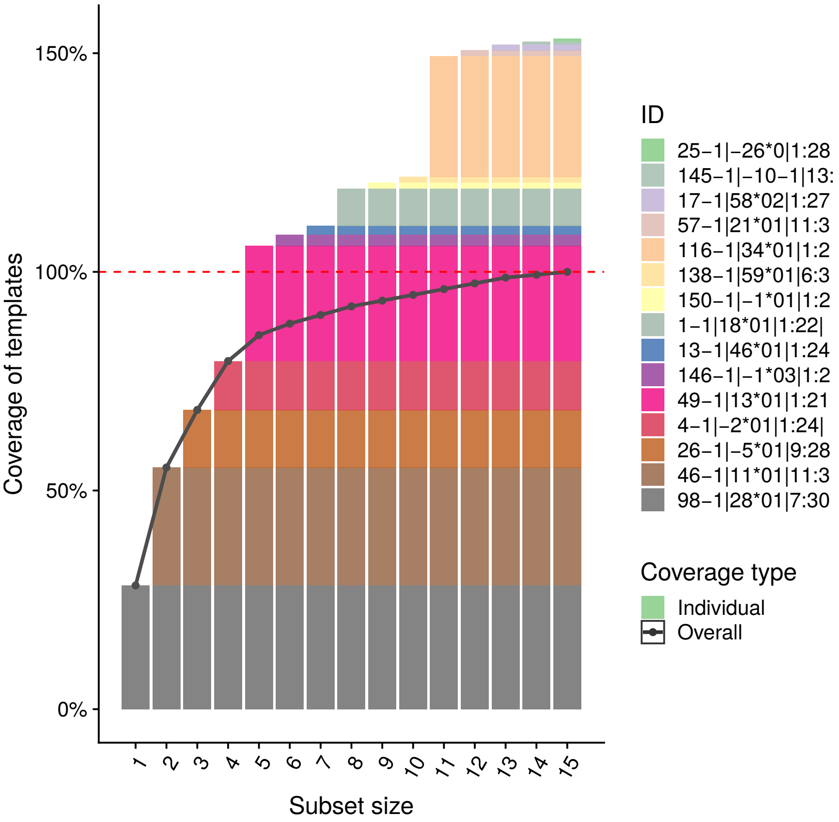


Figure A.2: Subset coverage of IGHV primers. Bars indicate the coverage of individual primers. The line indicates the overall coverage of optimal primer subsets for the indicated set sizes.

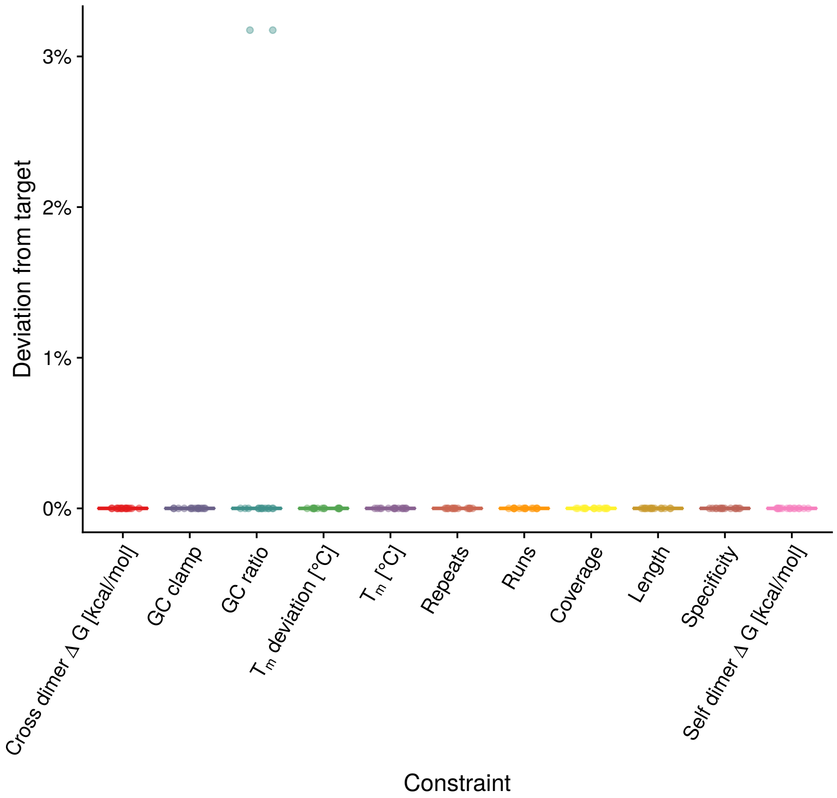


Figure A.3: Constraint deviation of IGHV primers. Each dot indicates the absolute deviation from the target constraints in percents.

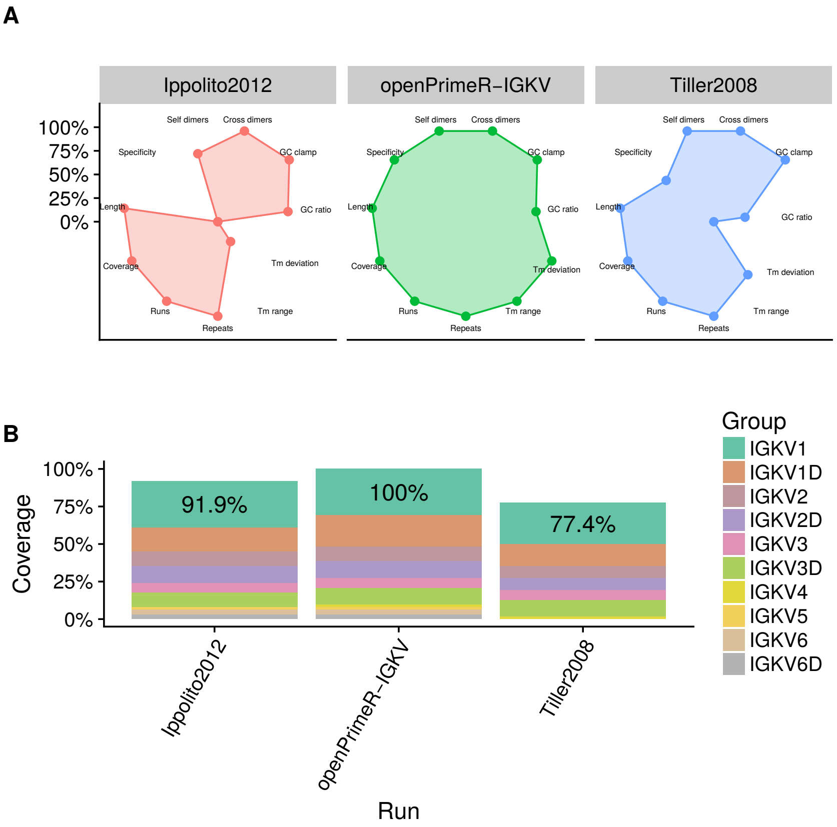


Figure A.4: IGKV primer set constraint fulfillment and coverage. (A) Rate of constraint fulfillment. (B) Percentage of covered templates.

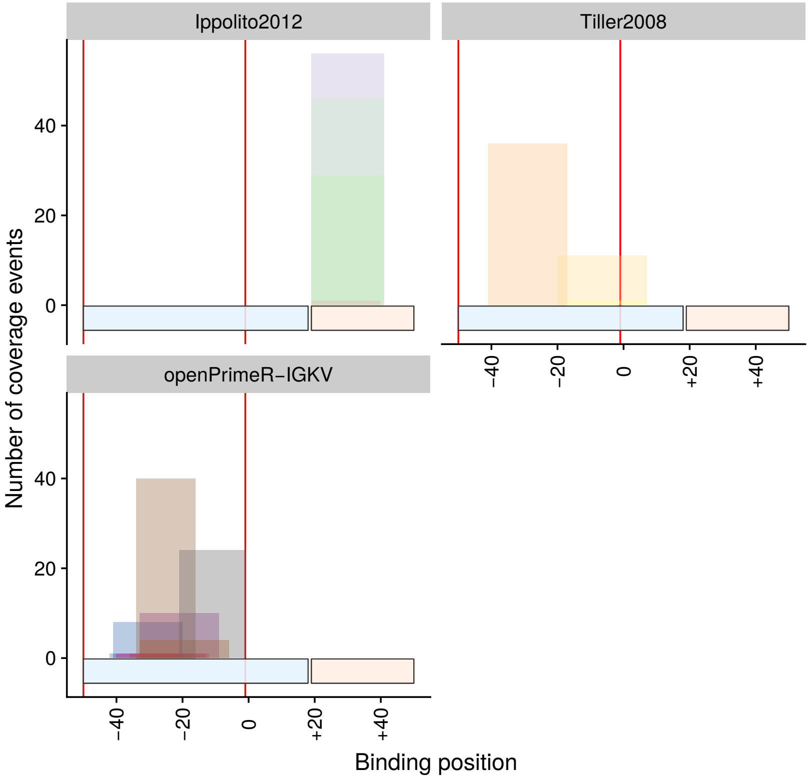


Figure A.5: IGKV primer binding regions. The leader region is indicated by the horizontal blue bar, while the horizontal red bar indicates the variable region. The region for which the new primers were designed is indicated by vertical red lines. The vertical bars indicate the number of coverage events for individual primers.

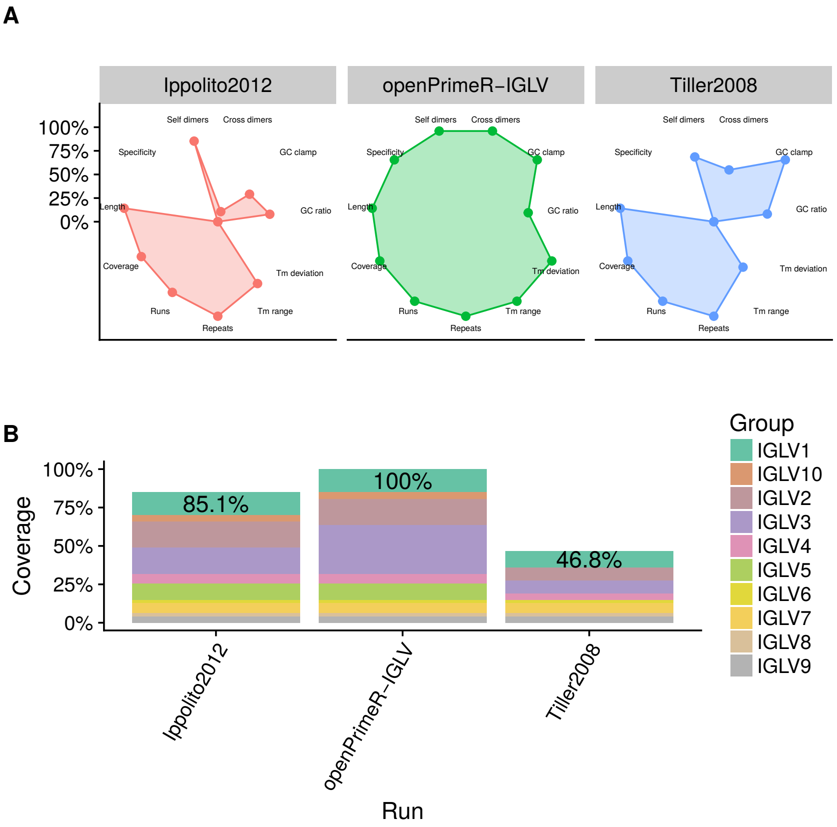


Figure A.6: IGLV primer set constraint fulfillment and coverage. (A) Rate of constraint fulfillment. (B) Percentage of covered templates.

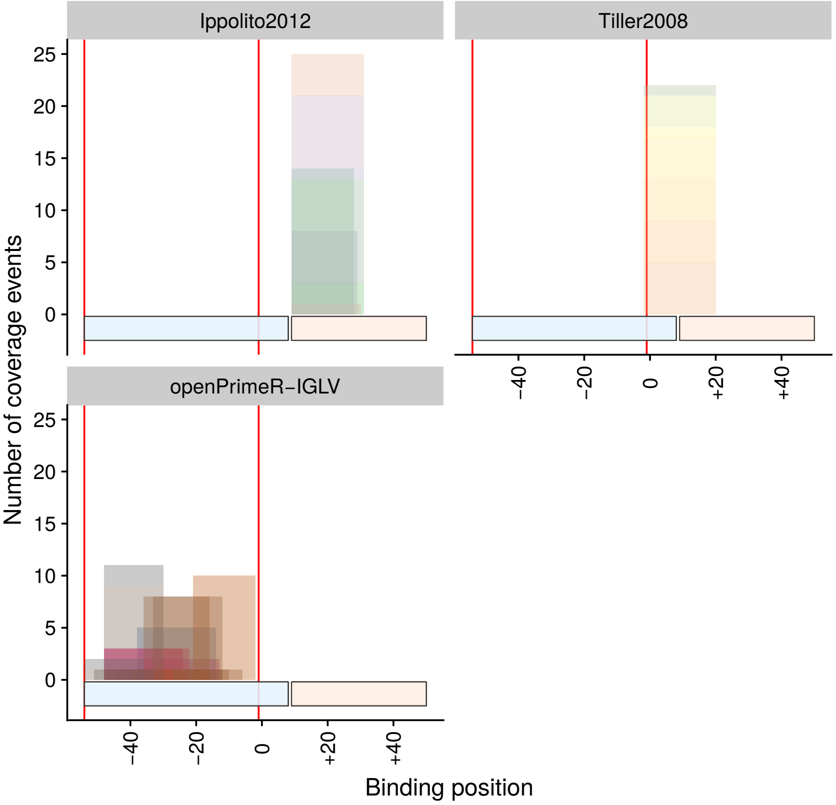


Figure A.7: IGLV primer binding regions. The leader region is indicated by the horizontal blue bar, while the horizontal red bar indicates the variable region. The region for which the new primers were designed is indicated by vertical red lines. The vertical bars indicate the number of coverage events for individual primers.

B Publications

Journal Articles

- Matthias Döring, Joachim Büch, Georg Friedrich, Alejandro Pironti, Prabhav Kalaghatgi, Elena Knops, Eva Heger, Martin Obermeier, Martin Däumer, Alexander Thielen, et al. geno2pheno [ngs-freq]: a genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. *Nucleic Acids Research*, 2018.

- Matthias Döring, Pedro Borrego, Joachim Büch, Andreia Martins, Georg Friedrich, Ricardo Jorge Camacho, Josef Eberle, Rolf Kaiser, Thomas Lengauer, Nuno Taveira, and Nico Pfeifer. A genotypic method for determining HIV-2 coreceptor usage enables epidemiological studies and clinical decision support. *Retrovirology*, 13(1):85, 2016.

- Vladimir Kartashev, Matthias Döring, Leonardo Nieto, Eleda Coletta, Rolf Kaiser, Saleta Sierra, and HCV EuResist Study group. New findings in HCV genotype distribution in selected West European, Russian and Israeli regions. *Journal of Clinical Virology*, 81: 82–89, 2016.

- Kai-Henrik Peiffer, Lisa Sommer, Simone Susser, Johannes Vermehren, Eva Herrmann, Matthias Döring, Julia Dietz, Dany Perner, Caterina Berkowski, Stefan Zeuzem, and Christoph Sarrazin. Interferon lambda 4 genotypes and resistance-associated variants in patients infected with hepatitis C virus genotypes 1 and 3. *Hepatology*, 63(1):63–73, 2016.

- Matthias Döring, Gasparoni Gilles, Jasmin Gries, Karl Nordström, Pavlo Lutsik, Jörn Walter, and Nico Pfeifer. Identification and analysis of methylation call differences between bisulfite microarray and bisulfite sequencing data with statistical learning techniques. *BMC Bioinformatics*, 16(Suppl 3):A7, 2015.

Award-Winning Posters

- Matthias Döring, Pedro Borrego, Joachim Buech, Andreia Martins, Georg Friedrich, Ricardo Jorge Camacho, Josef Eberle, Rolf Kaiser, Thomas Lengauer, Nuno Taveira, and Nico Pfeifer. geno2pheno[coreceptor-hiv2]: a computational tool for the prediction of HIV-2 coreceptor usage. In *Reviews in Antiviral Therapy & Infectious Diseases*, volume 4. 14th European Meeting on HIV & Hepatitis, 2016.

C
Plagiarism Prevention

The contents of this dissertation were evaluated using the plagiarism prevention software iThenticate on December 19th, 2018. The analysis was performed using two corpora. To determine overall similarity, I used the full text corpus of iThenticate. To determine the similarity to text that was not written by myself, I excluded the publications on which Chapters 4 and 5 are based[1], giving rise to the adjusted corpus. For evaluating the similarity index for the whole dissertation, the bibliography section was excluded from the analysis. The evaluations were performed using the default settings of iThenticate; the results of running iThenticate on individual chapters as well as the whole dissertation are shown in Table C.1.

For the unadjusted corpus, there are high similarity indices for Chapter 4 (24%), Chapter 5 (41%), and the discussion in Chapter 8 (18%) because these chapters are based on work that I have published previously. Note that the similarity indices for these chapters drop considerably when my own published work is excluded from the corpus: For the adjusted corpus, the similarity indices are merely 2%, 6%, and 3%, for Chapter 4, Chapter 5, and Chapter 8, respectively.

[1] Döring et al. 2018; Döring et al. 2016

Chapter	Similarity Index	Similarity Index (Adjusted Corpus)
Chapter 1	5%	5%
Chapter 2	8%	8%
Chapter 3	18%	17%
Chapter 4	24%	2%
Chapter 5	41%	6%
Chapter 6	4%	4%
Chapter 7	2%	2%
Chapter 8	16%	3%
Chapter 9	2%	2%
Overall	17%	9%

Table C.1: Results from running the plagiarism detection software iThenticate on this dissertation.

General Terms

gag denotes the HIV group-specific antigen, which encodes the structural proteins of the virus. 39, 40, 63

A refers to the purine nucleobase adenine. 67

C refers to the pyrimidine nucleobase cytosine. 67

env denotes the HIV envelope gene that encodes the transmembral gp41 protein and the gp120 protein, which are exposed on the cellular suface. 39, 40, 63

FASTA is the standard, text-based format that is used for storing the results from first-generation sequencing approaches (e.g. from Sanger sequencing). 158

FASTQ is the standard, text-based format that is used for storing the results from second-generation sequencing approaches. In contrast to FASTA, FASTQ stores not only sequences but also their corresponding nucleotide qualities in terms of Phred quality scores. 108, 118, 135

G refers to the purine nucleobase guanine. 67

IN denotes the HIV integrase, which allows for the integration of reverse-transcribed viral DNA into host DNA. 38, 40, 63

NS3 refers to non-structural protein 3, the HCV protease. 55, 58, 59, 116, 124

NS5A refers to non-structural protein 5 A, which is an important factor for HCV replication. 55, 58, 59, 116, 124, 132, 134

NS5B refers to non-structural protein 5 B, the HCV polymerase. 55, 58, 116, 124

pol is the HIV gene that encodes the structural proteins (PR, RT, and IN) of the virus. 39, 63

PR denotes the HIV protease, which is required to produce mature virions from an infected cell. 38, 40, 48, 63, 110, 112

RT denotes the HIV reverse transcriptase, which transcribes viral RNA to DNA. 38, 40, 48, 62, 63, 110, 112, 118, 125

SIR refers to a scheme of drug resistance classification that relies on three levels of drug resistance: susceptible, intermediate, and resistant. 65, 110, 111, 125, 127

T refers to the pyrimidine nucleobase thymine. 67

Drugs against HIV

3TC is the antiretroviral drug lamivudine. It belongs to the class of nucleoside-reverse transcriptase inhibitors. Since 3TC and FTC are associated with the same resistance mutations, these drugs are typically not differentiated when interpreting resistance. 112, 130

ABC is the antiretroviral drug abacavir. It belongs to the class of nucleoside-reverse transcriptase inhibitors. 112, 130, 136

APV is the antiretroviral drug amprenavir. It belongs to the class of protease inhibitors. 112, 134

ATV is the antiretroviral drug atazanavir. It belongs to the class of protease inhibitors. 112

d4T is the antiretroviral drug stavudine. It belongs to the class of nucleoside-reverse transcriptase inhibitors. Use of d4T can lead to thymidine-analog mutations. 112, 132

ddI is the antiretroviral drug didanosine. It belongs to the class of nucleoside-reverse transcriptase inhibitors. 112, 130

DRV is the antiretroviral drug darunavir. It belongs to the class of protease inhibitors. 50, 112, 114, 132

DTG refers to dolutegravir, an integrase strand-transfer inhibitor. 113

EFV is the antiretroviral drug efavirenz. It belongs to the class of non nucleoside-reverse transcriptase inhibitors. 112

ETR is the antiretroviral drug etravirine. It belongs to the class of non nucleoside-reverse transcriptase inhibitors. 112, 128

FTC is the antiretroviral drug emtricitabine. It belongs to the class of nucleoside-reverse transcriptase inhibitors. Since 3TC and FTC are associated with the same resistance mutations, these drugs are typically not differentiated when interpreting resistance. 110, 113, 114, 132

IDV is the antiretroviral drug indinavir. It belongs to the class of protease inhibitors. 110, 112

LPV is the antiretroviral drug lopinavir. It belongs to the class of protease inhibitors. 50, 112

NFV is the antiretroviral drug nelfinavir. It belongs to the class of protease inhibitors. 112

NVP is the antiretroviral drug nevirapine. It belongs to the class of non nucleoside-reverse transcriptase inhibitors. 112

RPV is the antiretroviral drug rilpivirine. It belongs to the class of non nucleoside-reverse transcriptase inhibitors. 112, 128

RTV refers to ritonavir, a booster for protease inhibitors. Protease inhibitors that are boosted through ritonavir are identified by appending */r*. For example, darunavir boosted with ritonavir is denoted by DRV/r. 48, 52

SQV is the antiretroviral drug saqinavir. It belongs to the class of protease inhibitors. 50, 112

T-20 is the antiretroviral drug enfuvirtide. It is an entry inhibitor that prevents HIV cell fusion. 53

TDF is the antiretroviral drug tenofovir. It belongs to the class of nucleoside-reverse transcriptase inhibitors. 52, 65, 112, 114, 132

TPV is the antiretroviral drug tipranavir. It belongs to the class of protease inhibitors. 112

ZDV is the antiretroviral drug zidovudine (3′-azidothymidine, AZT). It belongs to the class of nucleoside-reverse transcriptase inhibitors. Use of ZDV can lead to thymidine-analog mutations. 65, 112, 132

Drugs against HCV

ASV refers to asunaprevir, an NS3 inhibitor. 59

BOC refers to bocprevir, an NS3 inhibitor. 59

DCV refers to daclatasvir, an NS5A inhibitor. 59, 132

DSV refers to dasabuvir, an NS5B inhibitor. 59

EBR refers to elbasvir, an NS5A inhibitor. 59, 132

GLE refers to glecaprevir, an NS3 inhibitor. 59

GZR refers to grazoprevir, an NS3 inhibitor. 59

LDV refers to ledipasvir, an NS5A inhibitor. 59, 132

OBV refers to ombitasvir, an NS5A inhibitor. 59, 132

PIB refers to pibrentasvir, an NS5A inhibitor. 59

PTV refers to paritaprevir, an NS3 inhibitor. 59

SMV refers to simeprevir, an NS3 inhibitor. 59

SOF refers to sofosbovir, an NS5B inhibitor. 59

TVR refers to telaprevir, an NS3 inhibitor. 59

VEL refers to velpatasvir, an NS5A inhibitor. 59, 132

VOX refers to voxilaprevir, an NS3 inhibitor. 59

Acronyms

AAVF amino-acid variant format. 134

ADCC antibody-dependent cell-mediated cytotoxicity. 37

AIC Akaike information criterion. 78, 79, 82, 219, 223

AIDS acquired immunodeficiency syndrome. 21, 29, 37, 38, 46, 51

APC antigen-presenting cell. 33, 34

API application programming interface. 130, 135

APOBEC apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like. 47, 109, 118, 126, 127

ART antiretroviral therapy. 42, 47–49, 53, 114

ARV antiretroviral. 46, 52, 53, 63, 65, 136

AUC area under the receiver operating characteristic curve. 85, 86, 112, 116, 149, 152, 153, 223, 227

BAM binary alignment map. 133

bNAb broadly neutralizing antibody. 51–54, 140, 163, 164, 207, 209, 211, 232, 233

bp base pairs. 44, 68

CCR5 C-C chemokine receptor type 5. 42–44, 61, 139, 140, 145–147, 151, 159–162

CD cluster of differentiation. 34

CD4 cluster of differentiation antigen 4. 36, 40, 42, 43, 46, 50, 62, 236

CD4bs CD4 binding site. 51–53

CDC complement-dependent cytotoxicity. 37

cDNA complementary deoxyribonucleic acid. 40, 48, 163, 207–209

CDR complementarity-determining region. 34, 36

CL complete linkage. 96

CSV comma-separated values. 120, 158

CTL cytotoxic T lymphocyte. 34, 35

CV cross validation. 80, 81, 112, 149, 152, 156, 220

CXCR4 C-X-C chemokine receptor type 4. 42–44, 61, 139, 140, 145–147, 153, 159–162

DAA direct-acting antiviral. 55, 58, 59, 132, 137, 239

ddNTP didexobyribose nucleoside triphosphate. 67

DES drug-exposure score. 112, 114

DMEM Dulbecco's modified eagle medium. 144

DNA deoxyribonucleic acid. 30, 31, 40, 48, 67–69, 71, 109, 118, 119, 165, 171, 191

dNTP deoxyribose nucleoside triphosphate. 67, 69

DRM drug resistance mutation. 114

EACS European AIDS Clinical Society. 50

EAE expected amplification event. 171–174, 179, 190, 196, 200, 211

EI entry inhibitor. 48

Env envelope glycoprotein. 51, 140, 145

ER endoplasmatic reticulum. 40, 55, 57

ESTA enhanced-sensitivity Trofile assay. 61

Fab fragment antigen-binding. 36

Fc fragment crystallizable. 36, 37

FDA federal drug administration. 44, 53, 58

FDR false discovery rate. 85, 97, 100, 101, 154, 156

FN false negative. 83

FNR false negative rate. 84

FP false positive. 83, 150

FPR false positive rate. 84, 85, 142, 149, 150, 157–159

FWER familywise error rate. 100, 101

GA group average. 96

GALT gut-associated lymphoid tissue. 45

gp120 glycoprotein 120. 36, 40, 52

gp41 glycoprotein 41. 40, 48, 52

GSS genotypic susceptibility score. 116

GUI graphical user interface. 167, 168, 170, 209

HAART highly active antiretroviral therapy. 49

HBV hepatitis B virus. 135, 240

HCC hepatocellular carcinoma. 54, 58

HCV hepatitis C virus. 21–23, 29, 30, 53–55, 57–60, 107, 108, 116, 119, 121, 124–126, 128–130, 132–134, 231, 233, 235, 239, 240

HIV human immunodeficiency virus. 21–23, 29, 30, 36–40, 42, 44–48, 50–52, 54, 58, 60, 61, 63, 66, 109, 126, 135, 136, 139–141, 145, 147, 162, 233, 239, 240

HIV-1 human immunodeficiency virus type 1. 21–24, 37–40, 43, 44, 46, 50, 51, 53, 54, 57, 107, 108, 110–112, 115, 117–119, 121, 124–126, 128–130, 133–135, 140–143, 148, 149, 160–164, 199, 209, 211, 231–233

HIV-2 human immunodeficiency virus type 2. 22, 24, 37–40, 43, 44, 46, 50, 136, 140, 141, 143–145, 147, 148, 150, 152, 153, 157–162, 231–233

HLA human leukocyte antigen. 33, 136

HT high throughput. 168

IC50 half maximal inhibitory concentration. 51, 63–66, 109, 110

IgG immunoglobulin G. 164

IGH immunoglobulin heavy chain gene. 163, 200

IGHV immunoglobulin heavy chain variable region gene. 164, 169, 170, 202, 205, 207, 208, 214–216, 220, 223, 232

IGK immunoglobulin kappa gene. 163, 200

IGKV immunoglobulin kappa variable region gene. 164, 200, 208, 232

IGL immunoglobulin lambda gene. 163, 200

IGLV immunoglobulin lambda variable region gene. 164, 200, 208, 232

ILP integer linear program. 101, 102, 168, 170, 183, 185, 187, 190, 199, 200

IMGT international immunogenetics information system. 200

INSTI integrase strand-transfer inhibitor. 48, 49, 113, 135

IQR inter-quartile range. 221, 222

IRES internal ribosome entry site. 55

IRIS immune reconstitution inflammatory syndrome. 53

IU international units. 57

IUPAC International Union of Pure and Applied Chemistry. 120, 123, 167, 171, 180

IVDU intravenous drug use. 45, 57

JSON javasript object notation. 130

LANL Los Alamos National Laboratory. 126, 143, 145, 147

MAC membrane attack complex. 37

MHC major histocompatibility complex. 33, 34

MHC I major histocompatibility complex class I. 33–35

MHC II major histocompatibility complex class II. 33, 34

mPCR multiplex polymerase chain reaction. 23, 24, 60, 70, 163–165, 170, 179, 183, 208, 211, 232

MPER membrane proximal external region. 52

mRNA messenger RNA. 40, 164

MSE mean squared error. 76

MSM men who have sex with men. 45, 57

NAb neutralizing antibody. 51, 52

NCV nested cross validation. 81, 149, 153, 154

NGS next-generation sequencing. 22–24, 66, 68, 69, 108, 109, 117–119, 121, 133–136, 200, 231

NK natural killer. 32, 37

NNRTI non-nucleoside reverse transcriptase inhibitor. 48, 50, 112, 134

NPV negative predictive value. 85

NRTI nucleoside reverse transcriptase inhibitor. 48, 112, 114, 121, 130, 132

PBMC primary blood mononuclear cell. 43, 61, 62, 109, 160

PCR polymerase chain reaction. 30, 63, 68–71, 165, 166, 168, 171, 172, 178, 200, 205, 207, 213–215, 219, 228, 232, 233

PI protease inhibitor. 48, 50, 66, 112, 132

POS probability of susceptibility. 114, 115

PPV positive predictive value. 85

PrEP pre-exposure prophylaxis. 50, 54, 234

PTP primer-template pair. 216–223

qPCR quantitative polymerase chain reaction. 70, 71, 172, 213

RAS resistance-associated substitution. 59

RAV resistance-associated variant. 59

RBF radial basis function. 92, 148, 149, 153

REST representational state transfer. 130

RF resistance factor. 64–66, 110–115, 131

RNA ribonucleic acid. 30, 31, 38, 40, 47, 48, 55, 63, 109, 115

ROC receiver operating characteristic. 85

RT-PCR reverse transcription polymerase chain reaction. 63, 163, 205, 227

SCP set cover problem. 101–103, 167–170, 183, 185, 200, 210

SEVI semen-derived enhancer of viral infection. 45

SHM somatic hypermutation. 34

SL single linkage. 96

ss single-stranded. 38, 55

SVC support vector classification. 85, 112

SVM support vector machine. 86–89, 91, 92, 101, 112, 142, 143, 148–150, 152–160, 232

SVR sustained virologic response. 58, 59, 239

SVR support vector regression. 89–92, 112

TasP treatment as prevention. 46, 50

TCID50 half maximum tissue culture infectious dose. 144, 145

TDR transmitted drug resistance. 114

TE treatment episode. 115

Th T helper. 34, 35

TMM thermodynamic mismatch model. 219–221, 223, 225–228

TN true negative. 83, 150

TNR true negative rate. 84

TP true positive. 83

TPR true positive rate. 84, 157

V1 variable loop 1. 43, 145, 152, 160, 161, 232

V2 variable loop 2. 43, 140, 145, 152, 160, 161, 232

V3 variable loop 3. 43, 44, 51, 52, 140–153, 155–162, 232

V_H heavy chain variable. 164

VCF variant call format. 134

VL viral load. 45, 46, 49, 109, 115, 118, 130, 236

Bibliography

Carmen Aceijas, Gerry V Stimson, Matthew Hickman, Tim Rhodes, and United Nations Reference Group on HIV/AIDS Prevention and Care among IDU in Developing and Transitional Countries. Global overview of injecting drug use and HIV infection among injecting drug users. *AIDS*, 18(17):2295–303, November 2004.

Margaret E Ackerman and Galit Alter. Opportunities to exploit non-neutralizing HIV-specific antibody activity. *Current HIV Research*, 11 (5):365–77, July 2013.

Julius A Adebayo. *FairML: ToolBox for diagnosing bias in predictive modeling*. PhD thesis, Massachusetts Institute of Technology, 2016.

Vincent Agnello and Francesco G. De Rosa. Extrahepatic disease manifestations of HCV infection: some current issues. *Journal of Hepatology*, 40(2):341–352, February 2004.

Julie Yam Aguchi, Sushil G Devare, and Catherine A Brennan. Sequence Note: Identification of a New HIV-2 Subtype Based on Phylogenetic Analysis of Full-Length Genomic Sequence. *AIDS Research and Human Retroviruses*, 16(9):925–930, 2000.

Monica Airoldi, Mauro Zaccarelli, Luca Bisi, Teresa Bini, Andrea Antinori, Cristina Mussini, Francesca Bai, Giancarlo Orofino, Laura Sighinolfi, Andrea Gori, Fredy Suter, and Franco Maggiolo. One-pill once-a-day HAART: a simplification strategy that improves adherence and quality of life of HIV-infected subjects. *Patient Preference and Adherence*, 4:115–25, May 2010.

Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, and G. Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, New York, NY, New York, New York, USA, 1998.

André Altmann, Niko Beerenwinkel, Tobias Sing, Igor Savenkov, Martin Däumer, Rolf Kaiser, Soo-Yon Rhee, W Jeffrey Fessel, Robert W Shafer, and Thomas Lengauer. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy*, 12(2):169, 2007.

André Altmann, Martin Däumer, Niko Beerenwinkel, Yardena Peres, Eugen Schülter, Joachim Büch, Soo-Yon Rhee, Anders Sönnerborg, W Jeffrey Fessel, Robert W Shafer, et al. Predicting the response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database. *The Journal of Infectious Diseases*, 199(7):999–1006, 2009.

Matthew W. Anderson and Iris Schrijver. Next generation DNA sequencing and the future of genomic medicine. *Genes*, 1(1):38–69, 2010.

Adam Antiretroviral Therapy Cohort Collaboration, Margaret T May, Jorg-Janne Vehreschild, Niels Obel, M John Gill, Heidi M Crane, Christoph Boesecke, Sophie Patterson, Sophie Grabar, Charles Cazanave, Matthias Cavassini, Leah Shepherd, Antonella d'Arminio Monforte, Ard van Sighem, Mike Saag, Fiona Lampe, Vicky Hernando, Marta Montero, Robert Zangerle, Amy C Justice, Timothy Sterling, Suzanne M Ingle, and Jonathan A C Sterne. Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. *The Lancet HIV*, 4(8):e349–e356, aug 2017.

John Archer, Jan Weber, Kenneth Henry, Dane Winner, Richard Gibson, Lawrence Lee, Ellen Paxinos, Eric J. Arts, David L. Robertson, Larry Mimms, and Miguel E. Quiñones-Mateu. Use of Four Next-Generation Sequencing Platforms to Determine HIV-1 Coreceptor Tropism. *PLOS ONE*, 7(11):e49602, November 2012.

Andrew E. Armitage, Koen Deforche, Chih-hao Chang, Edmund Wee, Beatrice Kramer, John J. Welch, Jan Gerstoft, Lars Fugger, Andrew McMichael, Andrew Rambaut, and Astrid K. N. Iversen. APOBEC3G-Induced Hypermutation of Human Immunodeficiency Virus Type-1 Is Typically a Discrete "All or Nothing" Phenomenon. *PLOS Genetics*, 8(3):e1002550, March 2012.

Darius Armstrong-James, Justin Stebbing, Andrew Scourfield, Erasmus Smit, Bridget Ferns, Deenan Pillay, and Mark Nelson. Clinical outcome in resistant HIV-2 infection treated with raltegravir and maraviroc. *Antiviral Research*, 86(2):224–226, 2010.

Manit Arya, Iqbal S Shergill, Magali Williamson, Lyndon Gommersall, Neehar Arya, and Hitendra RH Patel. Basic principles of real-time quantitative PCR. *Expert Review of Molecular Diagnostics*, 5 (2):209–219, 2005.

Michael J Atkinson, Morton J Cowan, and Ann J Feeney. New alleles ofIGKV genes A2 and A18 suggest significant human IGKV locus polymorphism. *Immunogenetics*, 44(2):115–120, 1996.

AVERT. HIV and AIDS in East and Southern Africa regional overview | AVERT, 2016. URL `https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/overview`.

Srinivas Ayyadevara, John J Thaden, and Robert J Shmookler Reis. Discrimination of primer 3′-nucleotide mismatch by Taq DNA polymerase during polymerase chain reaction. *Analytical Biochemistry*, 284(1):11–18, 2000.

JM Azevedo-Pereira, Q Santos-Costa, K Mansinho, and J Moniz-Pereira. Identification and characterization of HIV-2 strains obtained from asymptomatic patients that do not use CCR5 or CXCR4 coreceptors. *Virology*, 313(1):136–146, 2003.

Jose M Azevedo-Pereira, Quirina Santos-Costa, and José Moniz-Pereira. HIV-2 infection and chemokine receptors usage-clues to reduced virulence of HIV-2. *Current HIV Research*, 3(1):3–16, 2005.

Masanori Baba, Osamu Nishimura, Naoyuki Kanzaki, Mika Okamoto, Hidekazu Sawada, Yuji Iizawa, Mitsuru Shiraishi, Yoshio Aramaki, Kenji Okonogi, Yasuaki Ogawa, et al. A small-molecule, nonpeptide CCR5 antagonist with highly potent and selective anti-HIV-1 activity. *Proceedings of the National Academy of Sciences*, 96(10): 5698–5703, 1999.

Suying Bao, Rui Jiang, WingKeung Kwan, BinBin Wang, Xu Ma, and You-Qiang Song. Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, 56(6):406–414, June 2011.

SW Barnett, M Quiroga, A Werner, D Dina, and JA Levy. Distinguishing features of an infectious molecular clone of the highly divergent and noncytopathic human immunodeficiency virus type 2 UC1 strain. *Journal of Virology*, 67(2):1006–1014, 1993.

SW Barnett, HS Legg, Y Sun, J Klinger, DJ Blackbourn, CP Locher, and JA Levy. Molecular cloning of the human immunodeficiency virus subtype 2 strain HIV-2 UC2. *Virology*, 222(1):257–261, 1996.

Ali Bashir, Yu Tsueng Liu, Benjamin J. Raphael, Dennis Carson, and Vineet Bafna. Optimization of primer design for the detection of variable genomic lesions in cancer. *Bioinformatics*, 23(21):2807–2815, November 2007.

Niko Beerenwinkel, Martin Däumer, Mark Oette, Klaus Korn, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, Joachim Selbig, and Hauke Walter. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Research*, 31(13):3850–5, 2003.

Bastian Beggel, Maria Neumann-Fraune, Matthias Döring, Glenn Lawyer, Rolf Kaiser, Jens Verheyen, and Thomas Lengauer. Genotyping hepatitis B virus dual infections using population-based sequence data. *Journal of General Virology*, 93(PART 9):1899–1907, 2012.

Georg Behrens, Pozniak Anton, Puoti Massimo, and Miro José M. EACS Guidelines Version 9.0. `http://www.eacsociety.org/guidelines/eacs-guidelines/eacs-guidelines.html`, 2017. Accessed: 2018-04-17.

A Beishuizen, MAC De Bruijn, MJ Pongers-Willemse, MA J Verhoeven, ER Van Wering, K Hählen, TM Breit, S de Bruin-Versteeg, H Hooijkaas, and JJM Van Dongen. Heterogeneity in junctional regions of immunoglobulin kappa deleting element rearrangements in B cell leukemias: a new molecular target for detection of minimal residual disease. *Leukemia*, 11(12):2200, 1997.

Beth P. Bell. Hepatitis A vaccine. *Seminars in Pediatric Infectious Diseases*, 13(3):165–173, July 2002.

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

Edward A Berger, RW Doms, E-M Fenyö, BTM Korber, DR Littman, JP Moore, QJ Sattentau, H Schuitemaker, J Sodroski, and RA Weiss. A new classification for HIV-1. *Nature*, 391(6664):240, 1998.

Michel Berkelaar, Kjell Eikland, Peter Notebaert, et al. lpsolve: Open source (mixed-integer) linear programming system. *Eindhoven U. of Technology*, 2004.

Neil Berry, Shabbar Jaffar, Maarten Schim van der Loeff, Koya Ariyoshi, Elizabeth Harding, Pa Tamba N'Gom, Francisco Dias, Andrew Wilkins, Dominic Ricard, Peter Aaby, Richard Tedder, and Hilton Whittle. Low Level Viremia and High CD4% Predict Normal Survival in a Cohort of HIV Type-2-Infected Villagers. *AIDS Research and Human Retroviruses*, 18(16):1167–1173, November 2002.

Chris Beyrer, Stefan D Baral, Frits van Griensven, Steven M Goodreau, Suwat Chariyalertsak, Andrea L Wirtz, and Ron Brookmeyer. Global epidemiology of HIV infection in men who have sex with men. *The Lancet*, 380(9839):367–377, July 2012.

Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for HIV Therapy screening. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 56–63, New York, New York, USA, 2008. ACM Press.

Miroslawa Bilska, Haili Tang, and David C Montefiori. Potential Risk of Replication-Competent Virus in HIV-1 Env-Pseudotyped Virus Preparations. *AIDS Research and Human Retroviruses*, 33(4):368–372, 2017.

A Björndal, Hongkui Deng, Marianne Jansson, Jose R Fiore, Claudia Colognesi, Anders Karlsson, Jan Albert, Gabriella Scarlatti, Dan R Littman, and Eva Maria Fenyö. Coreceptor usage of primary human immunodeficiency virus type 1 isolates varies according to biological phenotype. *Journal of Virology*, 71(10):7478–7487, 1997.

Hetty Blaak, PHM Boers, RA Gruters, H Schuitemaker, ME Van Der Ende, and ADME Osterhaus. CCR5, GPR15, and CXCR6 are major coreceptors of human immunodeficiency virus type 2 variants isolated from individuals with and without plasma viremia. *Journal of Virology*, 79(3):1686–1700, 2005.

Hetty Blaak, Marchina E van der Ende, Patrick HM Boers, Hanneke Schuitemaker, and Albert DME Osterhaus. In vitro replication capacity of HIV-2 variants from long-term aviremic individuals. *Virology*, 353(1):144–154, 2006.

Conrad C Bleul, Lijun Wu, James A Hoxie, Timothy A Springer, and Charles R Mackay. The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. *Proceedings of the National Academy of Sciences*, 94(5):1925–1930, 1997.

A Bobkov, M M Garaev, A Rzhaninova, P Kaleebu, R Pitman, J N Weber, and R Cheingsong-Popov. Molecular epidemiology of HIV-1 in the former Soviet Union: analysis of env V3 sequences and their correlation with epidemiologic data. *AIDS*, 8(5):619–24, May 1994a.

Aleksei Bobkov, Mansur M Garaev, Alla Rzhaninova, Pontiano Kaleebu, Richard Pitman, Jonathan N Weber, and Rachanee Cheingsong-Popov. Molecular epidemiology of HIV-1 in the former Soviet Union: analysis of env V3 sequences and their correlation with epidemiologic data. *AIDS*, 8(5):619–624, 1994b.

Diane L Bolton, Amarendra Pegu, Keyun Wang, Kathleen McGinnis, Martha Nason, Kathryn Foulds, Valerie Letukas, Stephen D Schmidt, Xuejun Chen, John Paul Todd, et al. Human immunodeficiency virus type 1 monoclonal antibodies suppress acute simian-human immunodeficiency virus viremia and limit seeding of cell-associated viral reservoirs. *Journal of Virology*, 90(3):1321–1332, 2016.

Mattia Bonsignori, David C Montefiori, Xueling Wu, Xi Chen, Kwan-Ki Hwang, Chun-Yen Tsao, Daniel M Kozink, Robert J Parks, Georgia D Tomaras, John A Crump, et al. Two distinct broadly neutralizing antibody specificities of different clonal lineages in a single HIV-1-infected donor: implications for vaccine design. *Journal of Virology*, pages JVI–07163, 2012.

Pedro Borrego, José Maria Marcelino, Cheila Rocha, Manuela Doroana, Francisco Antunes, Fernando Maltez, Perpétua Gomes, Carlos Novo, Helena Barroso, and Nuno Taveira. The role of the humoral immune response in the molecular evolution of the envelope C2, V3 and C3 regions in chronically HIV-2 infected patients. *Retrovirology*, 5(1):78, 2008.

Pedro Borrego, Rita Calado, José M Marcelino, Inês Bártolo, Cheila Rocha, Patrícia Cavaco-Silva, Manuela Doroana, Francisco Antunes, Fernando Maltez, Umbelina Caixas, et al. Baseline susceptibility of primary HIV-2 to entry inhibitors. *Antiviral Therapy*, 17:565–570, 2012.

Bluma G Brenner and Dimitrios Coutsinos. The K65R mutation in HIV-1 reverse transcriptase: genetic barriers, resistance profile and clinical implications. *HIV Therapy*, 3(6):583–594, November 2009.

Judith Breuer, Nigel W Douglas, Nick Goldman, and Rod S Daniels. Human immunodeficiency virus type 2 (HIV-2) env gene analysis: prediction of glycoprotein epitopes important for heterotypic neutralization and evidence for three genotype clusters within the HIV-2a subtype. *Journal of General Virology*, 76(2):333–345, 1995.

Hans P Brezinschek, Ruth I Brezinschek, and Peter E Lipsky. Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *The Journal of Immunology*, 155 (1):190–202, 1995.

Gary J Bridger, Renato T Skerlj, David Thornton, Sreenivasan Padmanabhan, Stephen A Martellucci, Geoffrey W Henson, Michael J Abrams, Naohiko Yamamoto, and Karen De Vreese. Synthesis and structure-activity relationships of phenylenebis (methylene)-linked bis-tetraazamacrocycles that inhibit HIV replication. Effects of macrocyclic ring size and substituents on the aromatic linker. *Journal of Medicinal Chemistry*, 38(2):366–378, 1995.

John A G Briggs, Thomas Wilk, Reinhold Welker, Hans-Georg Kräusslich, and Stephen D Fuller. Structural organization of authentic, mature HIV-1 virions and cores. *The EMBO Journal*, 22(7):1707–15, April 2003.

Romke Bron, PJ Klasse, David Wilkinson, Paul R Clapham, Annegret Pelchen-Matthews, Christine Power, TN Wells, Jin Kim, Stephen C Peiper, James A Hoxie, et al. Promiscuous use of CC and CXC chemokine receptors in cell-to-cell fusion mediated by a human immunodeficiency virus type 2 envelope protein. *Journal of Virology*, 71(11):8405–8415, 1997.

D Bru, F Martin-Laurent, and L Philippot. Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Applied and Environmental Microbiology*, 74(5):1660–1663, 2008.

Christopher J Bruno and Jeffrey M Jacobson. Ibalizumab: an anti-CD4 monoclonal antibody for the treatment of HIV-1 infection. *Journal of Antimicrobial Chemotherapy*, 65(9):1839–1841, 2010.

Fj Burpo. A critical review of PCR primer design algorithms and crosshybridization case study. *Biochemistry*, 86:1–12, 2001.

D. R. Burton, A. J. Hessell, B. F. Keele, P. J. Klasse, T. A. Ketas, B. Moldt, D. C. Dunlop, P. Poignard, L. A. Doyle, L. Cavacini, R. S. Veazey, and J. P. Moore. Limited or no protection by weakly or nonneutralizing antibodies against vaginal SHIV challenge of macaques compared with a strongly neutralizing antibody. *Proceedings of the National Academy of Sciences*, 108(27):11181–11186, July 2011.

Dennis R Burton, Ronald C Desrosiers, Robert W Doms, Wayne C Koff, Peter D Kwong, John P Moore, Gary J Nabel, Joseph Sodroski, Ian A Wilson, and Richard T Wyatt. HIV vaccine design and the neutralizing antibody problem. *Nature Immunology*, 5(3):233–236, March 2004.

Lei Cai, Wei Yuan, Zhou Zhang, Lin He, and Kuo-Chen Chou. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports*, 6(1):36540, December 2016.

Umbelina Caixas, Joana Ferreira, Aline T Marinho, Inês Faustino, Nádia M Grilo, Fátima Lampreia, Isabel Germano, Emília C Monteiro, and Sofia A Pereira. Long-term maraviroc use as salvage therapy in HIV-2 infection. *Journal of Antimicrobial Chemotherapy*, 67(10): 2538–2539, 2012.

Charles Calef, John Mokili, David H O'Connor, David I Watkins, and Bette Korber. Numbering positions in SIV relative to SIVMM239. *HIV Sequence Compendium*, 4:1, 2005.

Annapaola Callegaro, Elisa Di Filippo, Noemi Astuti, Paula Andrea Serna Ortega, Marco Rizzi, Claudio Farina, Daniela Valenti, and Franco Maggiolo. Early clinical response and presence of viral resistant minority variants: a proof of concept study. *Journal of the International AIDS Society*, 17(4 Suppl 3):19759, 2014.

Thomas B. Campbell, Nancy S. Shulman, Steven C. Johnson, Andrew R. Zolopa, Russell K. Young, Lane Bushman, Courtney V. Fletcher, E. Randall Lanier, Thomas C. Merigan, and Daniel R. Kuritzkes. Antiviral Activity of Lamivudine in Salvage Therapy for Multidrug-Resistant HIV-1 Infection. *Clinical Infectious Diseases*, 41 (2):236–242, July 2005.

Omobolaji T Campbell-Yesufu and Rajesh T Gandhi. Update on human immunodeficiency virus (HIV)-2 infection. *Clinical Infectious Diseases*, 52(6):780–787, 2011.

Alberto Caprara, Matteo Fischetti, and Paolo Toth. A Heuristic Method for the Set Covering Problem. *Operations Research*, 47(5):730–743, October 1999.

Ana Cardona, Otto Pritsch, Gérard Dumas, Jean-François Bach, and Guillaume Dighiero. Evidence for an antigen-driven selection process in human autoantibodies against acetylcholine receptor. *Molecular Immunology*, 32(16):1215–1223, 1995.

Timothy Cardozo, Tetsuya Kimura, Sean Philpott, Barbara Weiser, Harold Burger, and Susan Zolla-Pazner. Structural basis for coreceptor selectivity by the HIV type 1 V3 loop. *AIDS Research and Human Retroviruses*, 23(3):415–426, 2007.

A Carr and D A Cooper. Adverse effects of antiretroviral therapy. *Lancet (London, England)*, 356(9239):1423–30, October 2000a.

Andrew Carr and David A Cooper. Adverse effects of antiretroviral therapy. *The Lancet*, 356(9239):1423–1430, 2000b.

Séverine Carrère-Kremer, Claire Montpellier, Lazaro Lorenzo, Bénédicte Brulin, Laurence Cocquerel, Sandrine Belouzard, François

Penin, and Jean Dubuisson. Regulation of hepatitis C virus polyprotein processing by signal peptidase involves structural determinants at the p7 sequence junctions. *The Journal of Biological Chemistry*, 279(40):41384–92, October 2004.

Antonella Castagna, Anna Danise, Stefano Menzo, Laura Galli, Nicola Gianotti, Elisabetta Carini, Enzo Boeri, Andrea Galli, Massimo Cernuschi, Hamid Hasson, Massimo Clementi, and Adriano Lazzarin. Lamivudine monotherapy in HIV-1-infected patients harbouring a lamivudine-resistant virus: a randomized pilot study (E-184V study). *AIDS*, 20(6):795–803, April 2006.

Patricia Cavaco-Silva, Nuno C Taveira, Lino Rosado, Maria H Lourenço, José Moniz-Pereira, Nigel W Douglas, Rod S Daniels, and Maria O Santos-Ferreira. Virological and molecular demonstration of human immunodeficiency virus type 2 vertical transmission. *Journal of Virology*, 72(4):3418–3422, 1998.

CDC. HIV among Gay and Bisexual Men, 2017. URL https://www.cdc.gov/nchhstp/newsroom/docs/factsheets/cdc-msm-508.pdf.

Paula Cerejo, Quirina Santos-Costa, Marta Calado, Maria Espírito-Santo, Ricardo Parreira, and José Miguel Azevedo-Pereira. Characterization of Envelope Surface Glycoprotein from HIV-2 Primary Isolates with Different Coreceptor Usage Profile. *AIDS Research and Human Retroviruses*, 34(2):218–221, 2018.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Sui-Yuan Chang, Pi-Han Lin, Chien-Lin Cheng, Mao-Yuan Chen, Hsin-Yun Sun, Szu-Min Hsieh, Wang-Huei Sheng, Yi-Ching Su, Li-Hsin Su, Shu-Fang Chang, Wen-Chun Liu, Chien-Ching Hung, and Shan-Chwen Chang. Prevalence of Integrase Strand Transfer Inhibitors (INSTI) Resistance Mutations in Taiwan. *Scientific Reports*, 6(1):35779, December 2016.

Jim X Chen. The Evolution of Computing: AlphaGo. *Computing in Science & Engineering*, 18(4):4–7, 2016.

Pai-Hsuen Chen, Chih-Jen Lin, and Bernhard Schölkopf. A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry*, 21(2):111–136, 2005.

Pojen P Chen, Keith Albrandt, Norman K Orida, Victor Radoux, Emily Y Chen, Robert Schrantz, Fu-Tong Liu, and Dennis A Carson. Genetic basis for the cross-reactive idiotypes on the light chains of human IgM anti-IgG autoantibodies. *Proceedings of the National Academy of Sciences*, 83(21):8318–8322, 1986.

Pojen P Chen, Dick L Robbins, Frank R Jirik, Thomas J Kipps, and DA Carson. Isolation and characterization of a light chain variable region gene for human rheumatoid factors. *Journal of Experimental Medicine*, 166(6):1900–1905, 1987.

Stephen L Chen and Timothy R Morgan. The natural history of hepatitis C virus (HCV) infection. *International Journal of Medical Sciences*, 3(2):47–52, 2006.

Zhiwei Chen, Amara Luckay, Donald L Sodora, Paul Telfer, Patricia Reed, Agegnehu Gettie, James M Kanu, Ramses F Sadek, JoAnn Yee, David D Ho, et al. Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. *Journal of Virology*, 71(5):3953–3960, 1997.

Z G Chidgeavadze, R S Beabealashvilli, A M Atrazhev, M K Kukhanova, A V Azhayev, and A A Krayevsky. 2′,3′-Dideoxy-3′ aminonucleoside 5′-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Research*, 12 (3):1671–86, February 1984.

Yong Chong, Hideyuki Ikematsu, Masayuki Murata, Kouzaburo Yamaji, Shigeki Nabeshima, Seizaburo Kashiwagi, and Jun Hayashi. Two VH5 family genes expressed by human peripheral B cells display differential mutational frequencies in the VH region. *Molecular Immunology*, 39(1-2):31–38, 2002.

Qui-Lim Choo, George Kuo, Amy J Weiner, Lacy R Overby, Daniel W Bradley, and Michael Houghton. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*, 244(4902):359–362, 1989.

Vidita Choudhry, Mei-Yun Zhang, Ilia Harris, Igor A Sidorov, Bang Vu, Antony S Dimitrov, Timothy Fouts, and Dimiter S Dimitrov. Increased efficacy of HIV-1 neutralization by antibodies at low CCR5 surface concentration. *Biochemical and Biophysical Research Communications*, 348(3):1107–1115, 2006.

Li-Yeh Chuang, Yu-Huei Cheng, and Cheng-Hong Yang. URPD: a specific product primer design tool. *BMC Research Notes*, 5(1):306, 2012.

Tomas Cihlar and Marshall Fordyce. Current status and prospects of HIV treatment. *Current Opinion in Virology*, 18:50–56, June 2016.

Paul R Clapham and Áine McKnight. Cell surface receptors, virus entry and tropism of primate lentiviruses. *Journal of General Virology*, 83(8):1809–1829, 2002.

François Clavel, Mireille Guyader, Denise Guétard, Mireille Sallé, Luc Montagnier, and Marc Alizon. Molecular cloning and polymorphism of the human immune deficiency virus type 2. *Nature*, 324(6098): 691–695, 1986.

Myron Cohen, Ying Chen, Marybeth McCauley, Theresa Gamble, Mina Hosseinipour, Nagalingeshwaran Kumarasamy, James Hakim, Newton Kumwenda, Tania Brum, Beatriz Grinsztejn, et al. Final results of the HPTN 052 randomized controlled trial: antiretroviral therapy prevents HIV transmission. In *Journal of the International AIDS Society*, volume 18. IAS, 2015.

Myron S. Cohen, Cynthia L. Gay, Michael P. Busch, and Frederick M. Hecht. The Detection of Acute HIV Infection. *The Journal of Infectious Diseases*, 202(S2):S270–S277, October 2010.

Myron S. Cohen, Ying Q. Chen, Marybeth McCauley, Theresa Gamble, Mina C. Hosseinipour, Nagalingeswaran Kumarasamy, James G. Hakim, Johnstone Kumwenda, Beatriz Grinsztejn, Jose H.S. Pilotto, Sheela V. Godbole, Sanjay Mehendale, Suwat Chariyalertsak, Breno R. Santos, Kenneth H. Mayer, Irving F. Hoffman, Susan H. Eshleman, Estelle Piwowar-Manning, Lei Wang, Joseph Makhema, Lisa A. Mills, Guy de Bruyn, Ian Sanne, Joseph Eron, Joel Gallant, Diane Havlir, Susan Swindells, Heather Ribaudo, Vanessa Elharrar, David Burns, Taha E. Taha, Karin Nielsen-Saines, David Celentano, Max Essex, and Thomas R. Fleming. Prevention of HIV-1 Infection with Early Antiretroviral Therapy. *New England Journal of Medicine*, 365(6):493–505, aug 2011.

Myron S. Cohen, Ying Q. Chen, Marybeth McCauley, Theresa Gamble, Mina C. Hosseinipour, Nagalingeswaran Kumarasamy, James G. Hakim, Johnstone Kumwenda, Beatriz Grinsztejn, Jose H.S. Pilotto, et al. Antiretroviral Therapy for the Prevention of HIV-1 Transmission. *New England Journal of Medicine*, 375(9): 830–839, September 2016.

Ruth I Connor, Kristine E Sheridan, Daniel Ceradini, Sunny Choe, and Nathaniel R Landau. Change in coreceptor use correlates with disease progression in HIV-1–infected individuals. *Journal of Experimental Medicine*, 185(4):621–628, 1997.

A Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021–30, May 1985.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Jonathan PL Cox, Ian M Tomlinson, and Greg Winter. A directory of human germ-line Vχ segments reveals a strong bias in their usage. *European Journal of Immunology*, 24(4):827–836, 1994.

Alessandro Cozzi-Lepri, Marc Noguera-Julian, Francesca Di Giallonardo, Rob Schuurman, Martin Däumer, Sue Aitken, Francesca Ceccherini-Silberstein, Antonella D Arminio Monforte, Anna Maria Geretti, Clare L. Booth, Rolf Kaiser, Claudia Michalik, Klaus Jansen, Bernard Masquelier, Pantxika Bellecave, Roger D. Kouyos, Erika Castro, Hansjakob Furrer, Anna Schultze, Huldrych F. Günthard, Francoise Brun-Vezinet, Roger Paredes, Karin J. Metzner, and Norbert Brockmeyer. Low-frequency drug-resistant HIV-1 and risk of virological failure to first-line NNRTI-based ART: A multicohort European case-control study using centralized ultrasensitive 454 pyrosequencing. *Journal of Antimicrobial Chemotherapy*, 70(3):930–940, 2015.

Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

Yunji W Davenport, Anthony P West, and Pamela J Bjorkman. Structure of an HIV-2 gp120 in complex with CD4. *Journal of Virology*, 90 (4):2112–2118, 2016.

Katie L Davis, Frederic Bibollet-Ruche, Hui Li, Julie M Decker, Olaf Kutsch, Lynn Morris, Aidy Salomon, Abraham Pinter, James A Hoxie, Beatrice H Hahn, et al. HIV-2/HIV-1 envelope chimeras detect high titers of broadly reactive HIV-1 V3-specific antibodies in human plasma. *Journal of Virology*, 2008.

Joseph P Day, Francis Barany, Donald Bergstrom, and Robert P Hammer. Nucleotide analogs facilitate base conversion with 3′ mismatch primers. *Nucleic Acids Research*, 27(8):1810–1818, 1999.

Margaret O Dayhoff. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5:89–99, 1972.

Erik De Clercq, Naohiko Yamamoto, Rudi Pauwels, Jan Balzarini, Myriam Witvrouw, Karen De Vreese, Zeger Debyser, Brigitte Rosenwirth, Peter Peichl, and Roelf Datema. Highly potent and selective inhibition of human immunodeficiency virus by the bicyclam derivative JM3100. *Antimicrobial Agents and Chemotherapy*, 38(4):668–674, 1994.

Thushan I de Silva, Marlén Aasa-Chapman, Matthew Cotten, Stéphane Hué, James Robinson, Frederic Bibollet-Ruche, Ramu Sarge-Njie, Neil Berry, Assan Jaye, Peter Aaby, et al. Potent autologous and heterologous neutralizing antibody responses occur in HIV-2 infection across a broad range of infection outcomes. *Journal of Virology*, 86(2):930–946, 2012.

Michael Dean, Mary Carrington, Cheryl Winkler, Gavin A Huttley, Michael W Smith, Rando Allikmets, James J Goedert, Susan P Buchbinder, Eric Vittinghoff, Edward Gomperts, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science*, 273(5283):1856–1862, 1996.

Solène Debaisieux, Fabienne Rayne, Hocine Yezid, and Bruno Beaumelle. The Ins and Outs of HIV-1 Tat. *Traffic*, 13(3):355–363, March 2012.

Emma D Deeks. Cobicistat: a review of its use as a pharmacokinetic enhancer of atazanavir and darunavir in patients with HIV-1 infection. *Drugs*, 74(2):195–206, 2014.

Sylvie Delhalle, Jean-Claude Schmit, and Andy Chevigné. Phages and HIV-1: From display to interplay. *International Journal of Molecular Sciences*, 13(4):4727–4794, 2012.

Marie Laure Delignette-Muller and Christophe Dutang. fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, 64(4):1–34, mar 2015.

Pierre Delobel, Adrien Saliou, Florence Nicot, Martine Dubois, Stéphanie Trancart, Philippe Tangre, Jean-Pierre Aboulker, Anne-Marie Taburet, Jean-Michel Molina, Patrice Massip, Bruno Marchou, Jacques Izopet, and ANRS 106-Window Study Team. Minor HIV-1 Variants with the K103N Resistance Mutation during Intermittent Efavirenz-Containing Antiretroviral Therapy and Virological Failure. *PLOS ONE*, 6(6):e21655, June 2011.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

E. Delwart, E. Shpaer, J Louwagie, F. McCutchan, M Grez, H Rubsamen-Waigmann, J. Mullins, RM Hendry, N Dunlop, PL Nara, and al. Et. Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes. *Science*, 262(5137):1257–1261, November 1993.

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43 (5):491–498, May 2011.

Cynthia A Derdeyn, Julie M Decker, Jeffrey N Sfakianos, Xiaoyun Wu, William A O'Brien, Lee Ratner, John C Kappes, George M Shaw, and Eric Hunter. Sensitivity of human immunodeficiency virus type 1 to the fusion inhibitor T-20 is modulated by coreceptor specificity defined by the V3 loop of gp120. *Journal of Virology*, 74(18):8358–8367, 2000.

Karidia Diallo, Matthias Götte, and M A Wainberg. Molecular impact of the M184V mutation in human immunodeficiency virus type 1 reverse transcriptase. *Antimicrobial Agents and Chemotherapy*, 47(11): 3377–83, November 2003.

Adam S. Dingens, Hugh K. Haddox, Julie Overbaugh, and Jesse D. Bloom. Comprehensive Mapping of HIV-1 Escape from a Broadly Neutralizing Antibody. *Cell Host & Microbe*, 21(6):777–787.e4, June 2017.

Ron Diskin, Johannes F Scheid, Paola M Marcovecchio, Anthony P West, Florian Klein, Han Gao, Priyanthi NP Gnanapragasam, Alexander Abadir, Michael S Seaman, Michel C Nussenzweig, et al. Increasing the potency and breadth of an HIV antibody by using structure-based rational design. *Science*, page 1206727, 2011.

Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16): e105, September 2008.

Robert W Doms and Didier Trono. The plasma membrane as a combat zone in the HIV battlefield. *Genes & Development*, 14(21):2677–2688, 2000.

George A Donzella, Dominique Schols, Steven W Lin, José A Esté, Kirsten A Nagashima, Paul J Maddon, Graham P Allaway, Thomas P Sakmar, Geoffrey Henson, Erik DeClercq, et al. AMD3100, a small molecule inhibitor of HIV-1 entry via the CXCR4 co-receptor. *Nature Medicine*, 4(1):72, 1998.

Katie J. Doores. The HIV glycan shield as a target for broadly neutralizing antibodies. *FEBS Journal*, 282(24):4679–4691, December 2015.

Matthias Döring and Nico Pfeifer. openPrimeR, 2017. URL https://bioconductor.org/packages/release/bioc/html/openPrimeR.html.

Matthias Döring, Gasparoni Gilles, Jasmin Gries, Karl Nordström, Pavlo Lutsik, Jörn Walter, and Nico Pfeifer. Identification and analysis of methylation call differences between bisulfite microarray and bisulfite sequencing data with statistical learning techniques. *BMC Bioinformatics*, 16(Suppl 3):A7, 2015.

Matthias Döring, Pedro Borrego, Joachim Büch, Andreia Martins, Georg Friedrich, Ricardo Jorge Camacho, Josef Eberle, Rolf Kaiser, Thomas Lengauer, Nuno Taveira, and Nico Pfeifer. A genotypic method for determining HIV-2 coreceptor usage enables epidemiological studies and clinical decision support. *Retrovirology*, 13(1):85, 2016.

Matthias Döring, Pedro Borrego, Joachim Buech, Andreia Martins, Georg Friedrich, Ricardo Jorge Camacho, Josef Eberle, Rolf Kaiser, Thomas Lengauer, Nuno Taveira, and Nico Pfeifer. geno2pheno[coreceptor-hiv2]: a computational tool for the prediction of HIV-2 coreceptor usage. In *Reviews in Antiviral Therapy & Infectious Diseases*, volume 4. 14th European Meeting on HIV & Hepatitis, 2016.

Matthias Döring, Joachim Büch, Georg Friedrich, Alejandro Pironti, Prabhav Kalaghatgi, Elena Knops, Eva Heger, Martin Obermeier, Martin Däumer, Alexander Thielen, et al. geno2pheno [ngs-freq]: a genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. *Nucleic Acids Research*, 2018.

Patrick Dorr, Mike Westby, Susan Dobbs, Paul Griffin, Becky Irvine, Malcolm Macartney, Julie Mori, Graham Rickett, Caroline Smith-Burchnell, Carolyn Napier, Rob Webster, Duncan Armour, David Price, Blanda Stammen, Anthony Wood, and Manos Perros. 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. *Antimicrobial Agents and Chemotherapy*, 49(11):4721–32, November 2005.

Tatjana Dragic, Alexandra Trkola, Daniah AD Thompson, Emmanuel G Cormier, Francis A Kajumo, Elizabeth Maxwell, Steven W Lin, Weiwen Ying, Steven O Smith, Thomas P Sakmar, et al. A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembrane helices of CCR5. *Proceedings of the National Academy of Sciences*, 97(10):5639–5644, 2000.

Julia Drylewicz, Sophie Matheron, Estibaliz Lazaro, Florence Damond, Fabrice Bonnet, François Simon, François Dabis, Françoise Brun-Vezinet, Geneviève Chêne, and Rodolphe Thiébaut. Comparison of viro-immunological marker changes between HIV-1 and HIV-2-infected patients in France. *AIDS*, 22(4):457–68, February 2008.

Viktoriya Dubrovskaya, Javier Guenaga, Natalia de Val, Richard Wilson, Yu Feng, Arlette Movsesyan, Gunilla B. Karlsson Hedestam, Andrew B. Ward, and Richard T. Wyatt. Targeted N-glycan deletion at the receptor-binding site retains HIV Env NFL trimer integrity and accelerates the elicited antibody response. *PLOS Pathogens*, 13 (9):e1006614, September 2017.

Julie Dumonceaux, Sébastien Nisole, Chantal Chanel, Laurence Quivet, Ali Amara, Francoise Baleux, Pascale Briand, and Uriel Hazan. Spontaneous mutations in the env gene of the human immunodeficiency virus type 1 NDK isolate are associated with a CD4-independent entry phenotype. *Journal of Virology*, 72(1):512–519, 1998.

Marine Dumousseau, Nicolas Rodriguez, Nick Juty, and Nicolas Le Novère. MELTING, a flexible platform to predict the melting temperatures of nucleic acids. *BMC Bioinformatics*, 13:101, May 2012.

Tarn Duong. ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R. *Journal of Statistical Software*, 21 (7):1–16, October 2007.

J Eberle, C Noah, E Wolf, M Stürmer, P Braun, K Korn, M Däumer, A Thielen, T Berg, M Obermeier, et al. Empfehlungen zur Bestimmung des HIV-1-Korezeptor-Gebrauchs (DAIG Recommendations for the HIV-1 Tropism Testing), 2014.

ECDC. Thematic report: Continuum of HIV care. Technical report, European Centre for Disease Prevention and Control, 2017. URL `http://ecdc.europa.eu/sites/portal/files/documents/Continuum-of-HIV-care-2017.pdf`.

Natalia Echeverría, Gonzalo Moratorio, Juan Cristina, and Pilar Moreno. Hepatitis C virus genetic variability and evolution. *World Journal of Hepatology*, 7(6):831, 2015.

Manfred Eigen. Viral Quasispecies. *Scientific American*, 269(1):42–49, 1993.

Scott J Emrich, Mary Lowe, and Arthur L Delcher. PROBEmer: a web-based software tool for selecting optimal DNA oligos. *Nucleic Acids Research*, 31(13):3746–3750, 2003.

Alan Engelman and Peter Cherepanov. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nature Reviews Microbiology*, 10(4):279–290, April 2012.

Lucia Ercoli, Loredana Sarmati, Emanuele Nicastri, Giacomo Giannini, Clementina Galluzzo, Stefano Vella, and Massimo Andreoni. HIV phenotype switching during antiretroviral therapy: emergence of saquinavir-resistant strains with less cytopathogenicity. *AIDS*, 11 (10):1211–1217, 1997.

Zelda Euler and Hanneke Schuitemaker. Cross-reactive broadly neutralizing antibodies: timing is everything. *Frontiers in Immunology*, 3:215, 2012.

Zelda Euler, Marit J van Gils, Evelien M Bunnik, Pham Phung, Becky Schweighardt, Terri Wrin, and Hanneke Schuitemaker. Cross-reactive neutralizing humoral immunity does not protect from HIV type 1 disease progression. *The Journal of infectious diseases*, 201(7): 1045–1053, 2010.

Zelda Euler, Tom LGM van den Kerkhof, Marit J van Gils, Judith A Burger, Diana Edo-Matas, Pham Phung, Terri Wrin, and Hanneke Schuitemaker. Longitudinal analysis of early HIV-1-specific neutralizing activity in an elite neutralizer and in five patients who developed cross-reactive neutralizing activity. *Journal of Virology*, 86 (4):2045–2055, 2012.

European Association for the Study of the Liver. EASL Recommendations on Treatment of Hepatitis C 2018. *Journal of Hepatology*, 0(0), April 2018.

European Commission. MEDICAL DEVICES: Guidance document - Qualification and Classification of stand alone software. Technical report, European Commission, 2016. URL https://ec.europa.eu/docsroom/documents/17921/attachments/1/translations/en/renditions/native.

B Ewing, L Hillier, M C Wendl, and P Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8(3):175–85, March 1998.

Nancy L Farner, Thomas Dörner, and Peter E Lipsky. Molecular mechanisms and selection influence the generation of the human VλJλ repertoire. *The Journal of Immunology*, 162(4):2137–2145, 1999.

Beatriz Fernandez-Fernandez, Ana Montoya-Ferrer, Ana B Sanz, Maria D Sanchez-Niño, Maria C Izquierdo, Jonay Poveda, Valeria Sainz-Prestel, Natalia Ortiz-Martin, Alejandro Parra-Rodriguez, Rafael Selgas, Marta Ruiz-Ortega, Jesus Egido, and Alberto Ortiz. Tenofovir nephrotoxicity: 2011 update. *AIDS Research and Treatment*, 2011:354908, 2011.

S. Jane. Flint. *Principles of Virology: Molecular Biology, Pathogenesis, and Control of Animal Viruses*. ASM Press, 2004.

Donald N Forthal. Functions of Antibodies. *Microbiology Spectrum*, 2(4): 1–17, August 2014.

Solveig Fossum, Elliott Crooke, and Kirsten Skarstad. Organization of sister origins and replisomes during multifork DNA replication in Escherichia coli. *The EMBO journal*, 26(21):4514–4522, 2007.

Vincent Foulongne, Brigitte Montes, Marie-Noelle Didelot-Rousseau, and Michel Segondy. Comparison of the LCx human immunodeficiency virus (HIV) RNA quantitative, RealTime HIV, and COBAS AmpliPrep-COBAS TaqMan assays for quantitation of HIV type 1 RNA in plasma. *Journal of Clinical Microbiology*, 44(8):2963–6, August 2006.

Slim Fourati, Isabelle Malet, Sidonie Lambert, Cathia Soulie, Marc Wirden, Philippe Flandre, Djeneba B Fofana, Sophie Sayon, Anne Simon, Christine Katlama, et al. E138K and M184I mutations in HIV-1 reverse transcriptase coemerge as a result of APOBEC3 editing in the absence of drug exposure. *AIDS*, 26(13):1619–1624, 2012.

Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb. Accuracy of next generation sequencing platforms. *Next Generation, Sequencing & Applications*, 1, 2014.

Genoveffa Franchini, Kathleen A Fargnoli, Fabrizio Giombini, Linda Jagodzinski, Anita De Rossi, Marnix Bosch, Gunnel Biberfeld, Eva M Fenyo, Jan Albert, and Robert C Gallo. Molecular and biological characterization of a replication competent human immunodeficiency type 2 (HIV-2) proviral clone. *Proceedings of the National Academy of Sciences*, 86(7):2433–2437, 1989.

Christina Frank, Mostafa K Mohamed, G Thomas Strickland, Daniel Lavanchy, Ray R Arthur, Laurence S Magder, Taha El Khoby, Yehia Abdel-Wahab, Wagida Anwar, Ismail Sallam, et al. The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *The Lancet*, 355(9207):887–891, 2000.

Natalia T Freund, Haoqing Wang, Louise Scharf, Lilian Nogueira, Joshua A Horwitz, Yotam Bar-On, Jovana Golijanin, Stuart A Sievers, Devin Sok, Hui Cai, et al. Coexistence of potent HIV-1 broadly neutralizing antibodies and antibody-sensitive viruses in a viremic controller. *Science Translational Medicine*, 9(373), 2017.

George M Furnival and Robert W Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.

Joel E Gallant. The M184V mutation: what it does, how to prevent it, and what to do with it when it's there. *The AIDS Reader*, 16(10): 556–9, October 2006.

Joel E. Gallant, Ellen Koenig, Jaime Andrade-Villanueva, Ploenchan Chetchotisakd, Edwin DeJesus, Francisco Antunes, Keikawus Arastéh, Graeme Moyle, Giuliano Rizzardini, Jan Fehr, Yapei Liu, Lijie Zhong, Christian Callebaut, Javier Szwarcberg, Martin S. Rhee, and Andrew K. Cheng. Cobicistat Versus Ritonavir as a Pharmacoenhancer of Atazanavir Plus Emtricitabine/Tenofovir Disoproxil Fumarate in Treatment-Naive HIV Type 1-Infected Patients: Week 48 Results. *The Journal of Infectious Diseases*, 208(1): 32–39, July 2013.

Robert C Gallo, Syed Z Salahuddin, Mikulas Popovic, Gene M Shearer, Mark Kaplan, Barton F Haynes, Thomas J Palker, Robert Redfield, James Oleske, Bijan Safai, et al. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *science*, 224(4648):500–503, 1984.

Feng Gao, Ling Yue, David L Robertson, Sherri C Hill, Huxiong Hui, Robert J Biggar, Alfred E Neequaye, Thomas M Whelan, David D Ho, and George M Shaw. Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. *Journal of Virology*, 68(11): 7433–7447, 1994.

Ana Garcia-Diaz, Adele McCormick, Clare Booth, Dimitri Gonzalez, Chalom Sayada, Tanzina Haque, Margaret Johnson, and Daniel Webster. Analysis of transmitted HIV-1 drug resistance using 454 ultra-deep-sequencing and the DeepChek(®)-HIV system. *Journal of the International AIDS Society*, 17(4 Suppl 3):19752, 2014.

J. G. Garcia-Lerma, H. MacInnes, D. Bennett, H. Weinstock, and W. Heneine. Transmitted Human Immunodeficiency Virus Type 1 Carrying the D67N or K219Q/E Mutation Evolves Rapidly to

Zidovudine Resistance In Vitro and Shows a High Replicative Fitness in the Presence of Zidovudine. *Journal of Virology*, 78(14): 7545–7552, July 2004.

Javier Garcia-Perez, Isabelle Staropoli, Stéphane Azoulay, Jean-Thomas Heinrich, Almudena Cascajero, Philippe Colin, Hugues Lortat-Jacob, Fernando Arenzana-Seisdedos, Jose Alcami, Esther Kellenberger, et al. A single-residue change in the HIV-1 V3 loop associated with maraviroc resistance impairs CCR5 binding affinity while increasing replicative capacity. *Retrovirology*, 12(1):50, 2015.

S. N. Gardner, A. L. Hiddessen, P. L. Williams, C. Hara, M. C. Wagner, and B. W. Colston. Multiplex primer prediction software for divergent targets. *Nucleic Acids Research*, 37(19):6291–6304, October 2009.

Shea N. Gardner, Crystal J. Jaing, Maher M. Elsheikh, José Peña, David A. Hysom, and Monica K. Borucki. Multiplex Degenerate Primer Design for Targeted Whole Genome Amplification of Many Viral Genomes. *Advances in Bioinformatics*, 2014:101894, 2014.

R S Garfein, D Vlahov, N Galai, M C Doherty, and K E Nelson. Viral infections in short-term injection drug users: the prevalence of the hepatitis C, hepatitis B, human immunodeficiency, and human T-lymphotropic viruses. *American Journal of Public Health*, 86(5): 655–61, May 1996.

Lilit Garibyan and Nidhi Avashia. Polymerase chain reaction. *The Journal of Investigative Dermatology*, 133(3):1–4, March 2013.

Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv*, July 2012.

Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alex J. Smola. Multi-Instance Kernels. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann Publishers, 2002.

Rim Ghedira, Nina Papazova, Marnik Vuylsteke, Tom Ruttink, Isabel Taverniers, and Marc De Loose. Assessment of primer/template mismatch effects on real-time PCR amplification of target taxa for GMO quantification. *Journal of Agricultural and Food Chemistry*, 57 (20):9370–9377, 2009.

Claudia Giachino. Kappa+lambda+ Dual Receptor B Cells Are Present in the Human Peripheral Repertoire. *The Journal of Experimental Medicine*, 183(3):1245–1250, 1995.

Neil J. Gibson. The use of real-time PCR methods in DNA sequence variation analysis. *Clinica Chimica Acta*, 363(1-2):32–47, January 2006.

Robert Giegerich, Folker Meyer, and Chris Schleiermacher. GeneFisher-software support for the detection of postulated genes. In *ISMB*, pages 68–77, 1996.

Peter B Gilbert, Ian W McKeague, Geoffrey Eisen, Christopher Mullins, Aissatou Guéye-NDiaye, Souleymane Mboup, and Phyllis J Kanki. Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. *Statistics in Medicine*, 22(4): 573–593, 2003.

Christopher J Gill, John L Griffith, Denise Jacobson, Sarah Skinner, Sherwood L Gorbach, and Ira B Wilson. Relationship of HIV viral loads, CD4 counts, and HAART use to health-related quality of life. *Journal of Acquired Immune Deficiency Syndromes*, 30(5):485–92, August 2002.

M Gisslén, V Svedhem, L Lindborg, L Flamholc, H Norrgren, S Wendahl, M Axelsson, and A Sönnerborg. Sweden, the first country to achieve the Joint United Nations Programme on HIV/AIDS (UNAIDS)/World Health Organization (WHO) 90-90-90 continuum of HIV care targets. *HIV medicine*, 18(4):305–307, 2017.

V. Giudicelli, Denys Chaume, and Marie-Paule Lefranc. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Research*, 33(Database issue):D256–D261, December 2004.

Joakim Glamann, Dennis R Burton, Paul WHI Parren, Henrik J Ditzel, Karen A Kent, Caroline Arnold, David Montefiori, and Vanessa M Hirsch. Simian immunodeficiency virus (SIV) envelope-specific Fabs with high-level homologous neutralizing activity: recovery from a long-term-nonprogressor SIV-infected macaque. *Journal of Virology*, 72(1):585–592, 1998.

Annuska M Glas, Shu-Cai Huang, Erwin HN van Montfort, and Eric CB Milner. Motif-specific probes identify individual genes and detect somatic mutations. *Molecular Immunology*, 36(9):599–610, 1999.

Gaston H Gonnet, Mark A Cohen, and Steven A Benner. Exhaustive matching of the entire protein sequence database. *Science*, 256(5062): 1443–1445, 1992.

DD Goodman, NA Margot, DJ McColl, MD Miller, K Borroto-Esoda, and ES Svarovskaia. Pre-Existing Low-Levels of the K103N HIV-1 RT Mutation Above a Threshold is Associated with Virological

Failure in Treatment-Naïve Patients Undergoing EFV-Containing Antiretroviral Treatment. In *18th HIV Drug Resistance Workshop*, Fort Myers, Florida, 2009.

M K Gorny, J Y Xu, S Karwowska, A Buchbinder, and S Zolla-Pazner. Repertoire of neutralizing human monoclonal antibodies specific for the V3 domain of HIV-1 gp120. *Journal of Immunology*, 150(2): 635–43, January 1993.

Rainer Gosert, Denise Egger, Volker Lohmann, Ralf Bartenschlager, Hubert E Blum, Kurt Bienz, and Darius Moradpour. Identification of the hepatitis C virus RNA replication complex in Huh-7 cells harboring subgenomic replicons. *Journal of Virology*, 77(9):5487–92, May 2003.

Geoffrey S Gottlieb, Serge-Paul Eholié, John N Nkengasong, Sabelle Jallow, Sarah Rowland-Jones, Hilton C Whittle, and Papa Salif Sow. A call for randomized controlled trials of antiretroviral therapy for HIV-2 infection in West Africa. *AIDS*, 22(16):2069–72; discussion 2073–4, October 2008.

Elin S Gray, Maphuti C Madiga, Tandile Hermanus, Penny L Moore, Constantinos Kurt Wibmer, Nancy L Tumba, Lise Werner, Koleka Mlisana, Sengeziwe Sibeko, Carolyn Williamson, et al. The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4+ T cell decline and high viral load during acute infection. *Journal of Virology*, 85(10):4828–4840, 2011.

Moraima Guadalupe, Elizabeth Reay, Sumathi Sankaran, Thomas Prindiville, Jason Flamm, Andrew McNeil, and Satya Dandekar. Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy. *Journal of Virology*, 77(21):11708–17, November 2003.

Steven I Gubernick, Nuno Félix, Dolim Lee, Jing J Xu, and Bashar Hamad. The HIV Therapy market. *Nature Reviews Drug Discovery*, 15:451—452, June 2017.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316 (22):2402–2410, 2016.

Huldrych F. Günthard, Michael S. Saag, Constance A. Benson, Carlos del Rio, Joseph J. Eron, Joel E. Gallant, Jennifer F. Hoy, Michael J.

Mugavero, Paul E. Sax, Melanie A. Thompson, Rajesh T. Gandhi, Raphael J. Landovitz, Davey M. Smith, Donna M. Jacobsen, and Paul A. Volberding. Antiretroviral Drugs for Treatment and Prevention of HIV Infection in Adults. *JAMA*, 316(2):191, July 2016.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

Trevor T Hansel, Harald Kropshofer, Thomas Singer, Jane A Mitchell, and Andrew JT George. The safety and side effects of monoclonal antibodies. *Nature Reviews Drug discovery*, 9(4):325, 2010.

S. Hare, S. J. Smith, M. Metifiot, A. Jaxa-Chamiec, Y. Pommier, S. H. Hughes, and P. Cherepanov. Structural and Functional Analyses of the Second-Generation Integrase Strand Transfer Inhibitor Dolutegravir (S/GSK1349572). *Molecular Pharmacology*, 80(4):565–572, October 2011.

P R Harrigan, J S Montaner, S A Wegner, W Verbiest, V Miller, R Wood, and B A Larder. World-wide variation in HIV-1 phenotypic susceptibility in untreated individuals: biologically relevant values for resistance testing. *AIDS*, 15(13):1671–7, September 2001.

PR Harrigan, R McGovern, W Dong, A Thielen, M Jensen, T Mo, D Chapman, M Lewis, I James, and H Valdez. Screening for HIV tropism using population-based V3 genotypic analysis: a retrospective virological outcome analysis using stored plasma screening samples from MOTIVATE-1. In *Antiviral Therapy*, volume 14-4, pages A17–A17, 2009.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edition, 2009.

Barton F Haynes, Peter B Gilbert, M Juliana McElrath, Susan Zolla-Pazner, Georgia D Tomaras, S Munir Alam, David T Evans, David C Montefiori, Chitraporn Karnasuta, Ruengpueng Sutthent, et al. Immune-Correlates Analysis of an HIV-1 Vaccine Efficacy Trial. *New England Journal of Medicine*, 366(14):1275–1286, April 2012.

James M Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, January 2016.

Joris Hemelaar, Eleanor Gouws, Peter D Ghys, and Saladin Osmanov. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS*, 20(16):W13–W23, October 2006.

Craig W Hendrix, Charles Flexner, Ronald T MacFarland, Christen Giandomenico, Edward J Fuchs, Ella Redpath, Gary Bridger, and Geoffrey W Henson. Pharmacokinetics and safety of AMD-3100, a novel antagonist of the CXCR-4 chemokine receptor, in human volunteers. *Antimicrobial Agents and Chemotherapy*, 44(6):1667–1673, 2000.

K Hertogs, M P de Béthune, V Miller, T Ivens, P Schel, A Van Cauwenberge, C Van Den Eynde, V Van Gerwen, H Azijn, M Van Houtte, F Peeters, S Staszewski, M Conant, S Bloor, S Kemp, B Larder, and R Pauwels. A rapid method for simultaneous detection of phenotypic resistance to inhibitors of protease and reverse transcriptase in recombinant human immunodeficiency virus type 1 isolates from patients treated with antiretroviral drugs. *Antimicrobial Agents and Chemotherapy*, 42(2):269–76, February 1998.

Desmond G Higgins, Julie D Thompson, and Toby J Gibson. Using CLUSTAL for multiple sequence alignments. In *Methods in Enzymology*, volume 266, pages 383–402. Elsevier, 1996.

Hans H. Hirsch, Gilbert Kaufmann, Pedram Sendi, Manuel Battegay, and Manuel Battegay. Immune Reconstitution in HIV-Infected Patients. *Clinical Infectious Diseases*, 38(8):1159–1166, April 2004.

Martin S. Hirsch, Françoise Brun-Vézinet, Richard T. D'Aquila, Scott M. Hammer, Victoria A. Johnson, Daniel R. Kuritzkes, Clive Loveday, John W. Mellors, Bonaventura Clotet, Brian Conway, Lisa M. Demeter, Stefano Vella, Donna M. Jacobsen, and Douglas D. Richman. Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection. *JAMA*, 283(18):2417, May 2000.

Martin S. Hirsch, Huldrych F. Günthard, Jonathan M. Schapiro, Françoise Brun-Vézinet, Bonaventura Clotet, Scott M. Hammer, Victoria A. Johnson, Daniel R. Kuritzkes, John W. Mellors, Deenan Pillay, Patrick G. Yeni, Donna M. Jacobsen, and Douglas D. Richman. Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection: 2008 Recommendations of an International AIDS Society-USA Panel. *Clinical Infectious Diseases*, 47(2):266–285, July 2008.

Jacinta A Holmes and Alexander J Thompson. Interferon-free combination therapies for the treatment of hepatitis C: current insights. *Hepatic Medicine: Evidence and Research*, 7:51–70, 2015.

Thomas J. Hope. The Ins and Outs of HIV Rev. *Archives of Biochemistry and Biophysics*, 365(2):186–191, May 1999.

Tanya Hoskin. Parametric and nonparametric: Demystifying the terms. In *Mayo Clinic*, pages 1–5, 2012.

Michael Houghton and Sergio Abrignani. Prospects for a vaccine against the hepatitis C virus. *Nature*, 436(7053):961–966, August 2005.

Ming-Hua Hsieh, Wei-Che Hsu, Sung-Kay Chiu, and Chi-Meng Tzeng. An efficient algorithm for minimal primer set selection. *Bioinformatics*, 19(2):285–6, January 2003.

Jinghe Huang, Gilad Ofek, Leo Laub, Mark K. Louder, Nicole A. Doria-Rose, Nancy S. Longo, Hiromi Imamichi, Robert T. Bailer, Bimal Chakrabarti, Shailendra K. Sharma, S. Munir Alam, Tao Wang, Yongping Yang, Baoshan Zhang, Stephen A. Migueles, Richard Wyatt, Barton F. Haynes, Peter D. Kwong, John R. Mascola, and Mark Connors. Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature*, 491(7424):406–412, November 2012.

Mei-Mei Huang, Norman Arnheim, and Myron F Goodman. Extension of base mispairs by Taq DNA polymerase: implications for single nucleotide discrimination in PCR. *Nucleic Acids Research*, 20(17): 4567–4573, 1992.

Yaoxing Huang, William A Paxton, Steven M Wolinsky, Avidan U Neumann, Linqi Zhang, Tian He, Stanley Kang, Daniel Ceradini, Zhanqun Jin, Karina Yazdanbakhsh, et al. The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nature Medicine*, 2(11):1240–1243, 1996.

Yu-Cheng Huang, Chun-Fan Chang, Chen-hsiung Chan, Tze-Jung Yeh, Ya-Chun Chang, Chaur-Chin Chen, and Cheng-Yan Kao. Integrated minimum-set primers and unique probe design algorithms for differential detection on symptom-related pathogens. *Bioinformatics*, 21(24):4330–4337, 2005.

Yunda Huang, Lily Zhang, Julie Ledgerwood, Nicole Grunenberg, Robert Bailer, Abby Isaacs, Kelly Seaton, Kenneth H Mayer, Edmund Capparelli, Larry Corey, et al. Population pharmacokinetics analysis of VRC01, an HIV-1 broadly neutralizing monoclonal antibody, in healthy adults. *mAbs*, 9(5):792–800, 2017.

Christian Huber, Karlheinz F Schäble, Erwin Huber, Ralph Klein, Alfons Meindl, Rainer Thiebe, Rosemarie Lamm, and Hans G Zachau. The Vχ genes of the L regions and the repertoire of Vχ gene sequences in the human germ line. *European Journal of Immunology*, 23(11):2868–2875, 1993.

Michael Huber, Karin J Metzner, Fabienne D Geissberger, Cyril Shah, Christine Leemann, Thomas Klimkait, Jürg Böni, Alexandra Trkola, and Osvaldo Zagordi. MinVar: a rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. *Journal of Virological Methods*, 240:7–13, 2017.

Joel R. Huff. HIV protease: a novel chemotherapeutic target for AIDS. *Journal of Medicinal Chemistry*, 34(8):2305–2314, August 1991.

Stephen S Hwang, Terence J Boyle, H Kim Lyerly, and Bryan R Cullen. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science*, 253(5015):71–74, 1991.

David A. Hysom, Pejman Naraghi-Arani, Maher Elsheikh, A. Celena Carrillo, Peter L. Williams, and Shea N. Gardner. Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments. *PLOS ONE*, 7(4):e34560, April 2012.

Simona Alexandra Iacob and Diana Gabriela Iacob. Ibalizumab targeting CD4 receptors, an emerging molecule in HIV Therapy. *Frontiers in Microbiology*, 8:2323, 2017.

INSIGHT START Study Group. Initiation of antiretroviral therapy in early asymptomatic HIV infection. *New England Journal of Medicine*, 373(9):795–807, 2015.

Gregory C. Ippolito, Kam Hon Hoi, Sai T. Reddy, Sean M. Carroll, Xin Ge, Tobias Rogosch, Michael Zemlin, Leonard D. Shultz, Andrew D. Ellington, Carla L. VanDenBerg, and George Georgiou. Antibody Repertoires in Humanized NOD-scid-IL2Rγnull Mice and Human B Cells Reveals Human-Like Diversification and Tolerance Checkpoints in the Mouse. *PLOS ONE*, 7(4):e35497, April 2012.

Yoshitaka Isaka, Akihiko Sato, Shigeru Miki, Shinobu Kawauchi, Hitoshi Sakaida, Toshiyuki Hori, Takashi Uchiyama, Akio Adachi, Masanori Hayami, Tamio Fujiwara, et al. Small amino acid changes in the V3 loop of human immunodeficiency virus type 2 determines the coreceptor usage for CXCR4 and CCR5. *Virology*, 264(1): 237–243, 1999.

ISO. ISO 20776-1:2006(en), Clinical laboratory testing and in vitro diagnostic test systems - Susceptibility testing of infectious agents and evaluation of performance of antimicrobial susceptibility test devices - Part 1: Reference method for testing the in vitro activity of antimicrobial agents against rapidly growing aerobic bacteria involved in infectious diseases, 2006.

Omar J Jabado, Gustavo Palacios, Vishal Kapoor, Jeffrey Hui, Neil Renwick, Junhui Zhai, Thomas Briese, and W Ian Lipkin. Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Research*, 34(22):6605–6611, 2006.

Sushama Jadhav, Srikanth Tripathy, Smita Kulkarni, Kalpana Agnihotri, Arun Risbud, and Ramesh Paranjape. Molecular phylogenetics of nearly full-length HIV type 2 envelope gene sequences from West India. *AIDS Research and Human Retroviruses*, 25(1):115–121, 2009.

Elmar Jaeckel, Markus Cornberg, Heiner Wedemeyer, Teresa Santantonio, Julika Mayer, Myrga Zankel, Giuseppe Pastore, Manfred Dietrich, Christian Trautwein, and Michael P Manns. Treatment of acute hepatitis C with interferon alfa-2b. *New England Journal of Medicine*, 345(20):1452–1457, 2001.

A J Japour, D L Mayers, V A Johnson, D R Kuritzkes, L A Beckett, J M Arduino, J Lane, R J Black, P S Reichelderfer, and R T D'Aquila. Standardized peripheral blood mononuclear cell culture assay for determination of drug susceptibilities of clinical human immunodeficiency virus type 1 isolates. The RV-43 Study Group, the AIDS Clinical Trials Group Virology Committee Resistance Working Group. *Antimicrobial Agents and Chemotherapy*, 37(5):1095–101, May 1993.

Simon N Jarman. Amplicon: software for designing PCR primers on aligned DNA sequences. *Bioinformatics*, 20(10):1644–1645, 2004.

Mark A Jensen, Fu-Sheng Li, Angélique B van't Wout, David C Nickle, Daniel Shriner, Hong-Xia He, Sherry McLaughlin, Raj Shankarappa, Joseph B Margolick, and James I Mullins. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *Journal of Virology*, 77(24):13376–13388, 2003.

Hezhao Ji, Eric Enns, Chanson J Brumme, Neil Parkin, Mark Howison, Emma R Lee, Rupert Capina, Eric Marinier, Santiago Avila-Rios, Paul Sandstrom, et al. Bioinformatic data processing pipelines in support of next-generation sequencing-based HIV drug resistance testing: the Winnipeg Consensus. *Journal of the International AIDS Society*, 21(10):e25193, 2018.

J. A. Johnson and A. M. Geretti. Low-frequency HIV-1 drug resistance mutations can be clinically significant but must be interpreted with caution. *Journal of Antimicrobial Chemotherapy*, 65(7):1322–1326, July 2010.

Jeffrey A Johnson, Jin-Fen Li, Xierong Wei, Jonathan Lipscomb, David Irlbeck, Charles Craig, Amanda Smith, Diane E Bennett, Michael Monsour, Paul Sandstrom, E. Randall Lanier, and Walid Heneine. Minority HIV-1 Drug Resistance Mutations Are Present in Antiretroviral Treatment-Naive Populations and Associate with Reduced Treatment Efficacy. *PLOS Medicine*, 5(7):e158, July 2008.

Margaret I. Johnston and Anthony S. Fauci. An HIV Vaccine - Challenges and Prospects. *New England Journal of Medicine*, 359(9): 888–890, August 2008.

Daniel M Jones and John McLauchlan. Hepatitis C virus: assembly and release of virus particles. *The Journal of Biological Chemistry*, 285 (30):22733–9, July 2010.

Cristina Moraes Junta and Geraldo AS Passos. Genomic EcoRI polymorphism and cosmid sequencing reveal an insertion/deletion and a new IGLV5 allele in the human immunoglobulin lambda variable locus (22q11. 2/IGLV). *Immunogenetics*, 55(1):10–15, 2003.

L Juul, L Hougs, V Andersen, P Garred, L Ryder, A Svejgaard, B Høgh, L Lamm, B Graugaard, and T Barington. Population studies of the human Vk A18 gene polymorphism in Caucasians, blacks and Eskimos. *HLA*, 49(6):595–604, 1997.

Prabhav Kalaghatgi, Anna Maria Sikorski, Elena Knops, Daniel Rupp, Saleta Sierra, Eva Heger, Maria Neumann-Fraune, Bastian Beggel, Andreas Walker, Jörg Timm, Hauke Walter, Martin Obermeier, Rolf Kaiser, Ralf Bartenschlager, and Thomas Lengauer. Geno2pheno[HCV] - A Web-based Interpretation System to Support Hepatitis C Treatment Decisions in the Era of Direct-Acting Antiviral Agents. *PLOS ONE*, 11(5):e0155869, May 2016.

Ruslan Kalendar, David Lee, and Alan H Schulman. FastPCR Software for PCR Primer and Probe Design and Repeat Search. *Genes, Genomes and Genomics*, 3((Special Issue 1)):1–14, 2009.

Ruslan Kalendar, David Lee, and Alan H Schulman. FastPCR software for PCR, in silico PCR, and oligonucleotide assembly and analysis. In *DNA Cloning and Assembly Methods*, pages 271–302. Springer, 2014.

Olga V Kalinina, Nico Pfeifer, and Thomas Lengauer. Modelling binding between CCR5 and CXCR4 receptors and their ligands suggests the surface electrostatic potential of the co-receptor to be a key player in the HIV-1 tropism. *Retrovirology*, 10(1):130, 2013.

Thomas Kämpke, Markus Kieninger, and Michael Mecklenburg. Efficient primer design algorithms. *Bioinformatics*, 17(3):214–225, 2001.

Barry L Karger and András Guttman. DNA sequencing by CE. *Electrophoresis*, 30 Suppl 1(Suppl 1):S196–202, June 2009.

Gunilla B. Karlsson Hedestam, Ron A.M. Fouchier, Sanjay Phogat, Dennis R. Burton, Joseph Sodroski, and Richard T. Wyatt. The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nature Reviews Microbiology*, 6(2):143–155, February 2008.

Vladimir Kartashev, Matthias Döring, Leonardo Nieto, Eleda Coletta, Rolf Kaiser, Saleta Sierra, and HCV EuResist Study group. New findings in HCV genotype distribution in selected West European, Russian and Israeli regions. *Journal of Clinical Virology*, 81:82–89, 2016.

Nobuyuki Kato. Genome of Human Hepatitis C Virus (HCV): Gene Organization, Sequence Diversity, and Variation. *Microbial & Comparative Genomics*, 5(3):129–151, January 2000.

P Kellam and B A Larder. Recombinant virus assay: a rapid, phenotypic assay for assessment of drug susceptibility of human immunodeficiency virus type 1 isolates. *Antimicrobial Agents and Chemotherapy*, 38(1):23–30, January 1994.

Harry Kestler, Toshiaki Kodama, Douglas Ringler, Marta Marthas, Niels Pedersen, Andrew Lackner, Dean Regier, Prabhat Sehgal, Muthiah Daniel, Norval King, et al. Induction of AIDS in rhesus monkeys by molecularly cloned simian immunodeficiency virus. *Science*, 248(4959):1109–1112, 1990.

Eun-Young Kim, Ramon Lorenzo-Redondo, Susan J. Little, Yoon-Seok Chung, Prabhjeet K. Phalora, Irina Maljkovic Berry, John Archer, Sudhir Penugonda, Will Fischer, Douglas D. Richman, Tanmoy Bhattacharya, Michael H. Malim, and Steven M. Wolinsky. Human APOBEC3 Induced Mutation of Human Immunodeficiency Virus Type-1 Contributes to Adaptation and Evolution in Natural Infection. *PLOS Pathogens*, 10(7):e1004281, July 2014.

P. J. Klasse. Neutralization of Virus Infectivity by Antibodies: Old Problems in New Perspectives. *Advances in Biology*, 2014:1–24, September 2014.

Dieter Klein, Christian M Leutenegger, Claudia Bahula, Peter Gold, Regina Hofmann-Lehmann, Brian Salmons, Hans Lutz, and Walter H Gunzburg. Influence of preassay and sequence variations on viral load determination by a multiplex real-time reverse transcriptase-polymerase chain reaction for feline immunodeficiency virus. *Journal of Acquired Immune Deficiency Syndromes*, 26(1): 8–20, 2001.

F. Klein, H. Mouquet, P. Dosenovic, J. F. Scheid, L. Scharf, and M. C. Nussenzweig. Antibodies in HIV-1 Vaccine Development and Therapy. *Science*, 341(6151):1199–1204, September 2013.

A A Kolykhalov, E V Agapov, K J Blight, K Mihalik, S M Feinstone, and C M Rice. Transmission of hepatitis C by intrahepatic inoculation with transcribed RNA. *Science*, 277(5325):570–4, July 1997.

Bette Korber, Brian T Foley, C Kuiken, Satish K Pillai, Joseph G Sodroski, et al. Numbering positions in HIV relative to HXB2CG. *Human retroviruses and AIDS*, 3:102–111, 1998.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Anne Bruun Krøigård, Mads Thomassen, Anne-Vibeke Lænkholm, Torben A. Kruse, and Martin Jakob Larsen. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLOS ONE*, 11(3): e0151664, March 2016.

Smita Kulkarni, Srikanth Tripathy, Kalpana Agnihotri, Neena Jatkar, Sushama Jadhav, Winston Umakanth, Kavita Dhande, Prasad Tondare, Raman Gangakhedkar, and Ramesh Paranjape. Indian primary HIV-2 isolates and relationship between V3 genotype, biological phenotype and coreceptor usage. *Virology*, 337(1):68–75, 2005.

R Küppers, M Zhao, ML Hansmann, and K Rajewsky. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. *The EMBO Journal*, 12 (13):4955–4967, 1993.

Shigehiro Kuraku, Christian M. Zmasek, Osamu Nishimura, and Kazutaka Katoh. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. *Nucleic Acids Research*, 41(W1):W22–W28, July 2013.

S. Kwok, D. E. Kellogg, N. Mckinney, D. Spasic, L. Goda, C. Levenson, and J. J. Sninsky. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nucleic Acids Research*, 18(4):999–1005, February 1990.

Peter D Kwong, John R Mascola, and Gary J Nabel. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nature Reviews Immunology*, 13(9):693, 2013.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

B A Larder, S D Kemp, and P R Harrigan. Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy. *Science*, 269(5224):696–9, August 1995.

B A Larder, A Kohli, S Bloor, S D Kemp, P R Harrigan, R T Schooley, J M Lange, K N Pennington, and M H St Clair. Human immunodeficiency virus type 1 drug susceptibility during zidovudine (AZT) monotherapy compared with AZT plus 2′,3′-dideoxyinosine or AZT plus 2′,3′-dideoxycytidine combination therapy. The protocol 34,225-02 Collaborative Group. *Journal of Virology*, 70(9):5922–9, September 1996.

Max Lataillade, Jennifer Chiarella, Rong Yang, Steven Schnittman, Victoria Wirtz, Jonathan Uy, Daniel Seekins, Mark Krystal, Marco Mancini, Donnie McGrath, Birgitte Simen, Michael Egholm, and Michael Kozal. Prevalence and Clinical Significance of HIV Drug Resistance Mutations by Ultra-Deep Sequencing in Antiretroviral-Naïve Subjects in the CASTLE Study. *PLOS ONE*, 5(6):e10952, June 2010.

Anne G Laurent-Crawford, Bernard Krust, Sylviane Muller, Yves Rivière, Marie-Anne Rey-Cuillé, Jean-Marie Béchet, Luc Montagnier, and Ara G Hovanessian. The cytopathic effect of HIV is associated with apoptosis. *Virology*, 185(2):829–839, 1991.

J. E. Ledgerwood, E. E. Coates, G. Yamshchikov, J. G. Saunders, L. Holman, M. E. Enama, A. DeZure, R. M. Lynch, I. Gordon, S. Plummer, C. S. Hendel, A. Pegu, M. Conan-Cibotti, S. Sitar, R. T. Bailer, S. Narpala, A. McDermott, M. Louder, S. O'Dell, S. Mohan, J. P. Pandey, R. M. Schwartz, Z. Hu, R. A. Koup, E. Capparelli, J. R. Mascola, and B. S. Graham. Safety, pharmacokinetics and neutralization of the broadly neutralizing HIV-1 human monoclonal antibody

VRC01 in healthy adults. *Clinical & Experimental Immunology*, 182(3): 289–301, December 2015.

Chungnan Lee, Jain-Shing Wu, Yow-Ling Shiue, and Hong-Long Liang. MultiPrimer. *Applied Bioinformatics*, 5(2):99–109, 2006.

Wan-Ping Lee, Michael P. Stromberg, Alistair Ward, Chip Stewart, Erik P. Garrison, and Gabor T. Marth. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *PLOS ONE*, 9(3):e90581, March 2014.

Won-Suk Lee, Sung Min Ahn, Jun-Won Chung, Kyoung Oh Kim, Kwang An Kwon, Yoonjae Kim, Sunjin Sym, Dongbok Shin, Inkeun Park, Uhn Lee, and Jeong-Heum Baek. Assessing Concordance With Watson for Oncology, a Cognitive Computing Decision Support System for Colon Cancer Treatment in Korea. *JCO Clinical Cancer Informatics*, pages 1–8, mar 2018.

M.-P. Lefranc. IMGT, the international ImMunoGeneTics information system(R). *Nucleic Acids Research*, 33(Database issue):D593–D597, January 2004.

Marie Paule Lefranc, Véronique Giudicelli, Patrice Duroux, Joumana Jabado-Michaloud, Géraldine Folch, Safa Aouinti, Emilie Carillon, Hugo Duvergey, Amélie Houles, Typhaine Paysan-Lafosse, Saida Hadi-Saljoqi, Souphatta Sasorith, Gérard Lefranc, and Sofia Kossida. IMGT R, the international ImMunoGeneTics information system R 25 years on. *Nucleic Acids Research*, 43(D1):D413–D422, January 2015.

Thomas Lengauer and Tobias Sing. Bioinformatics-assisted anti-HIV Therapy. *Nature Reviews Microbiology*, 4(10):790–797, October 2006.

Thomas Lengauer, Oliver Sander, Saleta Sierra, Alexander Thielen, and Rolf Kaiser. Bioinformatics prediction of HIV coreceptor usage. *Nature Biotechnology*, 25(12):1407, 2007.

Marc C Levesque, M Anthony Moody, Kwan-Ki Hwang, Dawn J Marshall, John F Whitesides, Joshua D Amos, Thaddeus C Gurley, Sallie Allgood, Benjamin B Haynes, Nathan A Vandergrift, et al. Polyclonal B cell differentiation and loss of gastrointestinal tract germinal centers in the earliest stages of HIV-1 infection. *PLOS medicine*, 6(7):e1000107, 2009.

Baohui Li, Ibrahim Kadura, Dong-Jing Fu, and David E Watson. Genotyping with TaqMAMA. *Genomics*, 83(2):311–320, 2004.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, September 2010.

Haifeng Li and Tao Jiang. A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs. *Journal of Computational Biology*, 12(6):702–718, 2005.

Frédérick Libert, Pascale Cochaux, Gunhild Beckman, Michel Samson, Marina Aksenova, Antonio Cao, Andrew Czeizel, Mireille Claustres, Concepción De La Rúa, Maurizio Ferrari, et al. The Δ CCR5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. *Human Molecular Genetics*, 7(3):399–406, 1998.

Theam Soon Lim, Svetlana Mollova, Florian Rubelt, Volker Sievert, Stefan Dübel, Hans Lehrach, and Zoltan Konthur. V-gene amplification revisited–An optimised procedure for amplification of rearranged human antibody genes of different isotypes. *New Biotechnology*, 27(2):108–117, 2010.

Ming-Tseh Lin, Stacy L Mosier, Michele Thiess, Katie F Beierl, Marija Debeljak, Li-Hui Tseng, Guoli Chen, Srinivasan Yegnasubramanian, Hao Ho, Leslie Cope, et al. Clinical validation of KRAS, BRAF, and EGFR mutation detection using next-generation sequencing. *American Journal of Clinical Pathology*, 141(6):856–866, 2014.

Nina Lin, Oscar A Gonzalez, Ludy Registre, Carlos Becerril, Behzad Etemad, Hong Lu, Xueling Wu, Shahin Lockman, Myron Essex, Sikhulile Moyo, et al. Humoral immune pressure selects for HIV-1 CXC-chemokine receptor 4-using variants. *EBioMedicine*, 8:237–247, 2016.

Nina H Lin, Daniel M Negusse, Rameen Beroukhim, Francoise Giguel, Shahin Lockman, Myron Essex, and Daniel R Kuritzkes. The design and validation of a novel phenotypic assay to determine HIV-1 coreceptor usage of clinical isolates. *Journal of Virological Methods*, 169(1):39–46, 2010.

Chaim Linhart and Ron Shamir. The degenerate primer design problem. *Bioinformatics*, 18(suppl_1):S172–S181, 2002.

Nicholas J Loman, Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain, and Mark J Pallen.

Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30(5):434–439, May 2012.

Alan D Lopez, Colin D Mathers, Majid Ezzati, Dean T Jamison, and Christopher JL Murray. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, 367(9524):1747–1757, may 2006.

Todd C. Lorenz. Polymerase Chain Reaction: Basic Protocol Plus Troubleshooting and Optimization Strategies. *Journal of Visualized Experiments*, 63(e3998), May 2012.

Andrew J Low, Winnie Dong, Dennison Chan, Tobias Sing, Ronald Swanstrom, Mark Jensen, Satish Pillai, Benjamin Good, and P Richard Harrigan. Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS*, 21(14):F17–F24, 2007.

Andrew J Low, Rachel A McGovern, and P Richard Harrigan. Trofile HIV co-receptor usage assay. *Expert Opinion on Medical Diagnostics*, 3(2):181–191, 2009.

Lenette L. Lu, Todd J. Suscovich, Sarah M. Fortune, and Galit Alter. Beyond binding: antibody effector functions in infectious diseases. *Nature Reviews Immunology*, 18(1):46–61, October 2017.

Gregory M Lucas. Antiretroviral adherence, drug resistance, viral fitness and HIV disease progression: a tangled web is woven. *Journal of Antimicrobial Chemotherapy*, 55(4):413–416, 2005.

J. A. Luckey, H. Drossman, A. J. Kostichka, D. A. Mead, J. D'Cunha, T. B. Norris, and L. M. Smith. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research*, 18(15):4417–4421, August 1990.

L Lvovsky, I Ioshikhes, M C Raja, D Zevin-Sonkin, I a Sobolev, a Liberzon, J Shwartzburd, and L E Ulanovsky. Interdependence between DNA template secondary structure and priming efficiencies of short primers. *Nucleic Acids Research*, 26(23):5525–5532, December 1998.

Rebecca M Lynch, Eli Boritz, Emily E Coates, Adam DeZure, Patrick Madden, Pamela Costner, Mary E Enama, Sarah Plummer, Lasonji Holman, Cynthia S Hendel, et al. Virologic effects of broadly neutralizing antibody VRC01 administration during chronic HIV-1 infection. *Science Translational Medicine*, 7(319):319ra206–319ra206, December 2015.

Joseph D Ma, Kelly C Lee, and Grace M Kuo. HLA-B*5701 testing to predict abacavir hypersensitivity. *PLoS currents*, 2, 2010.

A. Macdonald and Mark Harris. Hepatitis C virus NS5A: tales of a promiscuous protein. *Journal of General Virology*, 85(9):2485–2502, September 2004.

Michael H Malim and Paul D Bieniasz. HIV Restriction Factors and Mechanisms of Evasion. *Cold Spring Harbor perspectives in medicine*, 2(5):a006940, May 2012.

Audrey Manheimer-Lory, Jessica B Katz, Michael Pillinger, Cybele Ghossein, Alan Smith, and Betty Diamond. Molecular characteristics of antibodies bearing an anti-DNA-associated idiotype. *Journal of Experimental Medicine*, 174(6):1639–1652, 1991.

Leigh J Manley, Duanduan Ma, and Stuart S Levine. Monitoring Error Rates In Illumina Sequencing. *Journal of biomolecular techniques : JBT*, 27(4):125–128, 2016.

M P Manns, H Wedemeyer, and M Cornberg. Treating viral hepatitis C: efficacy, side effects, and complications. *Gut*, 55(9):1350–9, September 2006.

Michael P Manns, John G McHutchison, Stuart C Gordon, Vinod K Rustgi, Mitchell Shiffman, Robert Reindollar, Zachary D Goodman, Kenneth Koury, Mei-Hsiu Ling, Janice K Albrecht, et al. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. *The Lancet*, 358(9286):958–965, 2001.

L M Mansky and H M Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology*, 69(8): 5087–94, August 1995.

José M Marcelino, Pedro Borrego, Charlotta Nilsson, Carlos Família, Helena Barroso, Fernando Maltez, Manuela Doroana, Francisco Antunes, Alexandre Quintas, and Nuno Taveira. Resistance to antibody neutralization in HIV-2 infection occurs in late stage disease and is associated with X4 tropism. *AIDS*, 26(18):2275–2284, 2012.

José Maria Marcelino, Pedro Borrego, Cheila Rocha, Helena Barroso, Alexandre Quintas, Carlos Novo, and Nuno Taveira. Potent and broadly reactive HIV-2 neutralizing antibodies elicited by a vaccinia virus vector prime-C2V3C3 polypeptide boost immunization strategy. *Journal of Virology*, 84(23):12429–36, December 2010.

David A Margolis, Juan Gonzalez-Garcia, Hans-Jürgen Stellbrink, Joseph J Eron, Yazdan Yazdanpanah, Daniel Podzamczer, Thomas Lutz, Jonathan B Angel, Gary J Richmond, Bonaventura Clotet, et al. Long-acting intramuscular cabotegravir and rilpivirine in adults with HIV-1 infection (LATTE-2): 96-week results of a randomised, open-label, phase 2b, non-inferiority trial. *The Lancet*, 390(10101):1499–1510, 2017.

B Marincovich, J Castilla, J Del Romero, S Garcia, V Hernando, M Raposo, and C Rodriguez. Absence of hepatitis C virus transmission in a prospective cohort of heterosexual serodiscordant couples. *Sexually Transmitted Infections*, 79(2):160–162, 2003.

Marketwired. Recursion Pharmaceuticals Raises 60 Million to Industrialize Drug Discovery Using Artificial Intelligence. http://www.marketwired.com/press-release/recursion-pharmaceuticals-raises-60-million-industrialize-drug-discovery-using-artificial-2235894.htm, 2017. [Online; accessed 14-Dec-2018].

Nicholas R. Markham and Michael Zuker. UNAFold. In *Bioinformatics*, pages 3–31. Humana Press, 2008.

Martin Markowitz, Bach-Yen Nguyen, Eduardo Gotuzzo, Fernando Mendo, Winai Ratanasuwan, Colin Kovacs, Guillermo Prada, Javier O Morales-Ramirez, Clyde S Crumpacker, Robin D Isaacs, Lucinda R Gilde, Hong Wan, Michael D Miller, Larissa A Wenning, and Hedy Teppler. Rapid and Durable Antiretroviral Effect of the HIV-1 Integrase Inhibitor Raltegravir as Part of Combination Therapy in Treatment-Naive Patients With HIV-1 Infection. *Journal of Acquired Immune Deficiency Syndromes*, 46(2):125–133, October 2007.

James D Marks, Hennie R Hoogenboom, Timothy P Bonnert, John McCafferty, Andrew D Griffiths, and Greg Winter. By-passing immunization: human antibodies from V-gene libraries displayed on phage. *Journal of Molecular Biology*, 222(3):581–597, 1991.

Richard Marlink. Lessons from the second AIDS virus, HIV-2. *AIDS*, 10(7):689–700, 1996.

M. Marmor, H. W. Sheppard, D. Donnell, S. Bozeman, C. Celum, S. Buchbinder, B. Koblin, and G. R. Seage. Homozygous and heterozygous CCR5-Delta32 genotypes are associated with resistance to HIV infection. *J. Acquir. Immune Defic. Syndr.*, 27(5):472–481, Aug 2001.

J. Martinez-Picado, L. Sutton, M.P. De Pasquale, A.V. Savara, and R.T. D'Aquila. Human immunodeficiency virus type 1 cloning vectors for antiretroviral resistance testing. *Journal of Clinical Microbiology*, 37(9):2943–51, September 1999.

A Martins, M Calado, P Borrego, J Marcelino, and J Azevedo-Pereira. Determinants of coreceptor use, tropism and susceptibility to antibody neutralization in the V3 region of HIV-2. In *Keystone Symp. Conf. X7 HIV Persistence Pathog. Erad*, 2016.

John R Mascola and Gary J Nabel. Vaccines for the prevention of HIV-1 disease. *Current Opinion in Immunology*, 13(4):489–494, August 2001.

E Mast, F Mahoney, M Kane, and H Margolis. Hepatitis B vaccine. *Vaccines*, 5:205–241, 2004.

Tom Matthews, Miklos Salgo, Michael Greenberg, Jain Chung, Ralph DeMasi, and Dani Bolognesi. Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. *Nature Reviews Drug Discovery*, 3(3):215–225, March 2004.

A M Maxam and W Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–4, February 1977.

K. H. Mayer, G. J. Hanna, and R. T. D'Aquila. Clinical Use of Genotypic and Phenotypic Drug Resistance Testing to Monitor Antiretroviral Chemotherapy. *Clinical Infectious Diseases*, 32(5):774–782, March 2001.

Sheena McCormack, David T Dunn, Monica Desai, David I Dolling, Mitzy Gafos, Richard Gilson, Ann K Sullivan, Amanda Clarke, Iain Reeves, Gabriel Schembri, et al. Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. *The Lancet*, 387(10013):53–60, January 2016.

John G McHutchison, Stuart C Gordon, Eugene R Schiff, Mitchell L Shiffman, William M Lee, Vinod K Rustgi, Zachary D Goodman, Mei-Hsiu Ling, Susannah Cort, and Janice K Albrecht. Interferon alfa-2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C. *New England Journal of Medicine*, 339(21): 1485–1492, 1998.

John G McHutchison, Michael Manns, Keyur Patel, Thierry Poynard, Karen L Lindsay, Christian Trepo, Jules Dienstag, William M Lee, Carmen Mak, Jean-Jacques Garaud, et al. Adherence to combination therapy enhances sustained response in genotype-1–infected patients with chronic hepatitis C. *Gastroenterology*, 123(4):1061–1069, 2002.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2): 153–157, 1947.

Luis Menéndez-Arias and József Tözsér. HIV-1 protease inhibitors: effects on HIV-2 replication and resistance. *Trends in Pharmacological Sciences*, 29(1):42–49, 2008.

Jane P. Messina, Isla Humphreys, Abraham Flaxman, Anthony Brown, Graham S. Cooke, Oliver G. Pybus, and Eleanor Barnes. Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology*, 61(1):77–87, January 2015.

Luis R Miranda, Matthias Götte, Fei Liang, and Daniel R Kuritzkes. The L74V mutation in human immunodeficiency virus type 1 reverse transcriptase counteracts enhanced excision of zidovudine monophosphate associated with thymidine analog resistance mutations. *Antimicrobial Agents and Chemotherapy*, 49(7):2648–56, July 2005.

Payman Mohassel and Yupeng Zhang. SecureML: A system for scalable privacy-preserving machine learning. In *2017 38th IEEE Symposium on Security and Privacy (SP)*, pages 19–38. IEEE, 2017.

Khayriyyah Mohd Hanafiah, Justina Groeger, Abraham D. Flaxman, and Steven T. Wiersma. Global epidemiology of hepatitis C virus infection: New estimates of age-specific antibody to HCV seroprevalence. *Hepatology*, 57(4):1333–1342, April 2013.

Brian Moldt, Eva G Rakasz, Niccole Schultz, Po-Ying Chan-Hui, Kristine Swiderek, Kimberly L Weisgrau, Shari M Piaskowski, Zachary Bergman, David I Watkins, Pascal Poignard, and Dennis R Burton. Highly potent HIV-specific antibody neutralization in vitro translates into effective protection against mucosal SHIV challenge in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 109(46):18921–5, November 2012.

Jean-Michel Molina, Catherine Capitant, Bruno Spire, Gilles Pialoux, Laurent Cotte, Isabelle Charreau, Cecile Tremblay, Jean-Marie Le Gall, Eric Cua, Armelle Pasquet, et al. On-Demand Preexposure Prophylaxis in Men at High Risk for HIV-1 Infection. *New England Journal of Medicine*, 373(23):2237–2246, December 2015.

JA Morgan and JF Tatar. Calculation of the residual sum of squares for all possible regressions. *Technometrics*, 14(2):317–325, 1972.

Andreas Mörner, Åsa Björndal, Jan Albert, Vineet N KewalRamani, Dan R Littman, Rie Inoue, Rigmor Thorstensson, Eva Maria Fenyö, and Ewa Björling. Primary human immunodeficiency virus type 2 (HIV-2) isolates, like HIV-1 isolates, frequently use CCR5 but show promiscuity in coreceptor usage. *Journal of Virology*, 73(3):2343–2349, 1999.

Andreas Mörner, Åsa Björndal, Ann-Charlotte Leandersson, Jan Albert, Ewa Björling, and Marianne Jansson. CCR5 or CXCR4 is required for efficient infection of peripheral blood mononuclear cells by promiscuous human immunodeficiency virus type 2 primary isolates. *AIDS Research and Human Retroviruses*, 18(3): 193–200, 2002.

L. Morris and T.A. Moody. Broadly Neutralizing Antibodies. *Human Vaccines*, pages 3–21, January 2017.

Roy Moscona, Daniela Ram, Marina Wax, Efrat Bucris, Itzchak Levy, Ella Mendelson, and Orna Mor. Comparison between next-generation and Sanger-based sequencing for the detection of transmitted drug-resistance mutations among recently infected HIV-1 patients in Israel, 2000-2014. *Journal of the International AIDS Society*, 20(1):21846, January 2017.

Hugo Mouquet, Florian Klein, Johannes F. Scheid, Malte Warncke, John Pietzsch, Thiago Y. K. Oliveira, Klara Velinzon, Michael S. Seaman, and Michel C. Nussenzweig. Memory B Cell Antibodies to HIV-1 gp140 Cloned from Individuals Infected with Clade A and B Viruses. *PLOS ONE*, 6(9):e24078, September 2011.

Hugo Mouquet, Louise Scharf, Zelda Euler, Yan Liu, Caroline Eden, Johannes F Scheid, Ariel Halper-Stromberg, Priyanthi NP Gnanapragasam, Daniel IR Spencer, Michael S Seaman, et al. Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proceedings of the National Academy of Sciences*, 109(47):E3268–E3277, 2012.

Jan Münch, Elke Rücker, Ludger Ständker, Knut Adermann, Christine Goffinet, Michael Schindler, Steffen Wildum, Raghavan Chinnadurai, Devi Rajan, Anke Specht, et al. Semen-derived amyloid fibrils drastically enhance HIV infection. *Cell*, 131(6):1059–1071, 2007.

John M Murray, Anthony D Kelleher, and David A Cooper. Timing of the components of the HIV life cycle in productively infected CD4+

T cells in populations of HIV infected individuals. *Journal of Virology*, pages JVI–05095, 2011.

Rajagopal Murugan, Katharina Imkeller, Christian E Busse, and Hedda Wardemann. Direct high-throughput amplification and sequencing of immunoglobulin genes from single human B cells. *European Journal of Immunology*, 45(9):2698–2700, 2015.

T Muster, F Steindl, M Purtscher, A Trkola, A Klima, G Himmler, F Rüker, and H Katinger. A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *Journal of Virology*, 67(11): 6642–7, November 1993.

Sandra Marcia Muxel, Sueli Donizete Borelli, Marla Karine Amarante, Julio Cesar Voltarelli, Mateus Nobrega Aoki, Carlos Eduardo Coral de Oliveira, and Maria Angelica Ehara Watanabe. Association study of CCR5 delta 32 polymorphism among the HLA-DRB1 Caucasian population in Northern Paraná, Brazil. *Journal of Clinical Laboratory Analysis*, 22(4):229–233, 2008.

Shinichi Nakagawa. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6): 1044–1045, 2004.

Hirotomo Nakata, Kenji Maeda, Toshikazu Miyakawa, Shiro Shibayama, Masayoshi Matsuo, Yoshikazu Takaoka, Mamoru Ito, Yoshio Koyanagi, and Hiroaki Mitsuya. Potent anti-R5 human immunodeficiency virus type 1 effects of a CCR5 antagonist, AK602/ONO4128/GW873140, in a novel human peripheral blood mononuclear cell nonobese diabetic-SCID, interleukin-2 receptor γ-chain-knocked-out AIDS mouse model. *Journal of Virology*, 79(4): 2087–2096, 2005.

Michael Niepmann. Hepatitis C Virus RNA Translation. In *Hepatitis C Virus: From Molecular Virology to Antiviral Therapy*, pages 143–166. Springer, Berlin, Heidelberg, 2013.

Galina N Nikolenko, Sarah Palmer, Frank Maldarelli, John W Mellors, John M Coffin, and Vinay K Pathak. Mechanism for nucleoside analog-mediated abrogation of HIV-1 replication: balance between RNase H activity and nucleotide excision. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):2093–8, February 2005.

Yoshiaki Nishimura, Rajeev Gautam, Tae-Wook Chun, Reza Sadjadpour, Kathryn E Foulds, Masashi Shingai, Florian Klein, Anna Gazumyan, Jovana Golijanin, Mitzi Donaldson, et al. Early antibody therapy can induce long-lasting immunity to SHIV. *Nature*, 543 (7646):559, 2017.

AMCTB No, Analytical Methods Committee, et al. PCR–the polymerase chain reaction. *Analytical Methods*, 6(2):333–336, 2014.

Ann Noë, Jean Plum, and Chris Verhofstede. The latent HIV-1 reservoir in patients undergoing HAART: an archive of pre-HAART drug resistance. *Journal of Antimicrobial Chemotherapy*, 55(4):410–412, April 2005.

Marc Noguera-Julian, Alessandro Cozzi-Lepri, Francesca Di Giallonardo, Rob Schuurman, Martin Däumer, Sue Aitken, Francesca Ceccherini-Silberstein, AD'Arminio Monforte, Anna Maria Geretti, Claire L Booth, et al. Contribution of APOBEC3G/F activity to the development of low-abundance drug-resistant human immunodeficiency virus type 1 variants. *Clinical Microbiology and Infection*, 22(2): 191–200, February 2016.

Marc Noguera-Julian, Dianna Edgil, P Richard Harrigan, Paul Sandstrom, Catherine Godfrey, and Roger Paredes. Next-generation human immunodeficiency virus sequencing for patient management and drug resistance surveillance. *The Journal of Infectious Diseases*, 216(suppl_9):S829–S833, 2017.

Martin Obermeier, Alejandro Pironti, Thomas Berg, Patrick Braun, Martin Däumer, Josef Eberle, Robert Ehret, Rolf Kaiser, Niels Kleinkauf, Klaus Korn, et al. HIV-GRADE: a publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge. *Intervirology*, 55(2):102–107, 2012.

Damien M O'Halloran. PrimerMapper: high throughput primer design and graphical assembly for PCR and SNP detection. *Scientific Reports*, 6:20631, 2016.

World Health Organization et al. Noncommunicable diseases now biggest killers. *Retrieved July 2018*, 3:2009, 2008.

Julie Overbaugh and Lynn Morris. The Antibody Response against HIV-1. *Perspectives in Medicine*, 2(1):a007039, January 2012.

Sherry M Owen, Dennis Ellenberger, Mark Rayfield, Stefan Wiktor, Philippe Michel, Michael H Grieco, Feng Gao, Beatrice H Hahn, and Renu B Lal. Genetically divergent strains of human immunodeficiency virus type 2 use multiple coreceptors for viral entry. *Journal of Virology*, 72(7):5425–5432, 1998.

Leonid Padyukov, Mirjana Hahn-Zoric, Sandra R Blomqvist, Marina Ulanova, Simon G Welch, Ann J Feeney, Yu Lung Lau, and

Lars Åke Hanson. Distribution of human kappa locus IGKV2-29 and IGKV2D-29 alleles in Swedish Caucasians and Hong Kong Chinese. *Immunogenetics*, 53(1):22–30, 2001.

Frank J. Palella, Rose K. Baker, Anne C. Moorman, Joan S. Chmiel, Kathleen C. Wood, John T. Brooks, and Scott D. Holmberg. Mortality in the Highly Active Antiretroviral Therapy Era. *Journal of Acquired Immune Deficiency Syndromes*, 43(1):27–34, September 2006.

Sarah Palmer, Nicolas Margot, Harold Gilbert, Nigel Shaw, Robert Buckheit, and Michael Miller. Tenofovir, Adefovir, and Zidovudine Susceptibilities of Primary Human Immunodeficiency Virus Type 1 Isolates with Non-B Subtypes or Nucleoside Resistance. *AIDS Research and Human Retroviruses*, 17(12):1167–1173, August 2001.

Wenjing Pan, Miranda Byrne-Steele, Chunlin Wang, Stanley Lu, Scott Clemmons, Robert J Zahorchak, and Jian Han. DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnology*, 14(1):10, 2014.

Saswati Panda and Jeak L Ding. Natural antibodies bridge innate and adaptive immunity. *Journal of Immunology*, 194(1):13–20, January 2015.

A. Panjkovich and F. Melo. Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics*, 21(6):711–722, March 2005.

Giuseppe Pantaleo, Cecilia Graziosi, James F. Demarest, Luca Butini, Maria Montroni, Cecil H. Fox, Jan M. Orenstein, Donald P. Kotler, and Anthony S. Fauci. HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease. *Nature*, 362(6418):355–358, March 1993.

Roger Paredes, Philip L Tzou, Gert van Zyl, Geoff Barrow, Ricardo Camacho, Sergio Carmona, Philip M Grant, Ravindra K Gupta, Raph L Hamers, P Richard Harrigan, et al. Collaborative update of a rule-based expert system for HIV-1 genotypic resistance test interpretation. *PLOS ONE*, 12(7):e0181357, July 2017.

N T Parkin, N S Hellmann, J M Whitcomb, L Kiss, C Chappey, and C J Petropoulos. Natural variation of drug susceptibility in wild-type human immunodeficiency virus type 1. *Antimicrobial Agents and Chemotherapy*, 48(2):437–43, February 2004.

Jean-Michel Pawlotsky. Treatment failure and resistance with direct-acting antiviral drugs against hepatitis C virus. *Hepatology*, 53(5): 1742–1751, May 2011.

Jean-Michel Pawlotsky. Hepatitis C Virus Resistance to Direct-Acting Antiviral Drugs in Interferon-Free Regimens. *Gastroenterology*, 151 (1):70–86, July 2016.

William R Pearson, Gabriel Robins, Dallas E Wrege, and Tongtong Zhang. On the primer selection problem in polymerase chain reaction experiments. *Discrete Applied Mathematics*, 71(1-3):231–246, 1996.

Kai-Henrik Peiffer, Lisa Sommer, Simone Susser, Johannes Vermehren, Eva Herrmann, Matthias Döring, Julia Dietz, Dany Perner, Caterina Berkowski, Stefan Zeuzem, and Christoph Sarrazin. Interferon lambda 4 genotypes and resistance-associated variants in patients infected with hepatitis C virus genotypes 1 and 3. *Hepatology*, 63(1):63–73, 2016.

Robert Pejchal, Katie J Doores, Laura M Walker, Reza Khayat, Po-Ssu Huang, Sheng-Kai Wang, Robyn L Stanfield, Jean-Philippe Julien, Alejandra Ramos, Max Crispin, et al. A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science*, 334(6059):1097–103, November 2011.

PEPFAR. Nigeria, Country Operational Plan (COP) 2017, Strategic Direction Summary. Technical report, PEPFAR, 2017. URL https://www.pepfar.gov/documents/organization/272254.pdf.

A S Perelson, A U Neumann, M Markowitz, J M Leonard, and D D Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–6, March 1996.

Carlo-Federico Perno and Ada Bertoli. *Clinical cut-offs in the interpretation of phenotypic resistance*. Mediscript, 2006.

MA Persson, Roger H Caothien, and Dennis R Burton. Generation of diverse high-affinity human monoclonal antibodies by repertoire cloning. *Proceedings of the National Academy of Sciences*, 88(6):2432–2436, 1991.

G Pesole, S Liuni, G Grillo, P Belichard, T Trenkle, J Welsh, and M McClelland. GeneUp: a program to select short PCR primer pairs that occur in multiple members of sequence lists. *BioTechniques*, 25 (1):112–123, 1998.

Kevin Peterson and Sarah Rowland-Jones. Novel agents for the treatment of HIV-2 infection. *Antivir Ther*, 17(3):435–8, 2012.

C J Petropoulos, N T Parkin, K L Limoli, Y S Lie, T Wrin, W Huang, H Tian, D Smith, G A Winslow, D J Capon, and J M Whitcomb. A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrobial Agents and Chemotherapy*, 44(4): 920–8, April 2000.

Arnolfo Petruzziello, Samantha Marigliano, Giovanna Loquercio, and Carmela Cacciapuoti. Hepatitis C virus (HCV) genotypes distribution: an epidemiological up-date in Europe. *Infectious Agents and Cancer*, 11:53, 2016.

Olivia Peuchant, Rodolphe Thiébaut, Sophie Capdepont, Valérie Lavignolle-Aurillac, Didier Neau, Philippe Morlat, François Dabis, Hervé Fleury, Bernard Masquelier, and ANRS CO3 Aquitaine Cohort. Transmission of HIV-1 minority-resistant variants and response to first-line antiretroviral therapy. *AIDS*, 22(12):1417–1423, July 2008.

Nico Pfeifer and Thomas Lengauer. Improving HIV coreceptor usage prediction in the clinic using hints from next-generation sequencing data. *Bioinformatics*, 28(18):i589–i595, 2012.

Nico Pfeifer, Hauke Walter, and Thomas Lengauer. Association between HIV-1 coreceptor usage and resistance to broadly neutralizing antibodies. *Journal of Acquired Immune Deficiency Syndromes*, 67 (2):107, 2014.

Sean Philpott, Barbara Weiser, Kathryn Anastos, Christina Michelle Ramirez Kitchen, Esther Robison, William A Meyer, Henry S Sacks, Usha Mathur-Wagh, Cheryl Brunner, and Harold Burger. Preferential suppression of CXCR4-specific strains of HIV-1 by antiviral therapy. *The Journal of Clinical Investigation*, 107(4): 431–438, 2001.

Anne Piantadosi, Dana Panteleeff, Catherine A Blish, Jared M Baeten, Walter Jaoko, R Scott McClelland, and Julie Overbaugh. Breadth of neutralizing antibody response to human immunodeficiency virus type 1 is affected by factors early in infection but does not influence disease progression. *Journal of Virology*, 83(19):10269–10274, 2009.

Satish Pillai, Benjamin Good, Douglas Richman, and Jacques Corbeil. A new perspective on V3 phenotype prediction. *AIDS Research and Human Retroviruses*, 19(2):145–149, 2003.

Steven D Pinkerton. How many sexually-acquired HIV infections in the USA are due to acute-phase HIV transmission? *AIDS*, 21(12): 1625–9, July 2007.

Alejandro Pironti, Nico Pfeifer, Hauke Walter, Björn-Erik O. Jensen, Maurizio Zazzi, Perpétua Gomes, Rolf Kaiser, and Thomas Lengauer. Using drug exposure for predicting drug resistance - A data-driven genotypic interpretation tool. *PLOS ONE*, 12(4): e0174992, April 2017a.

Alejandro Pironti, Hauke Walter, Nico Pfeifer, Elena Knops, Nadine Lübke, Joachim Büch, Simona Di Giambenedetto, Rolf Kaiser, and Thomas Lengauer. Determination of Phenotypic Resistance Cutoffs From Routine Clinical Data. *Journal of Acquired Immune Deficiency Syndromes*, 74(5):e129, 2017b.

Emily J Platt, Kathy Wehrly, Shawn E Kuhmann, Bruce Chesebro, and David Kabat. Effects of CCR5 and CD4 cell surface concentrations on infections by macrophagetropic isolates of human immunodeficiency virus type 1. *Journal of Virology*, 72(4):2855–2864, 1998.

Emily J Platt, Miroslawa Bilska, Susan L Kozak, David Kabat, and David C Montefiori. Evidence that ecotropic murine leukemia virus contamination in TZM-bl cells does not affect the outcome of neutralizing antibody assays with human immunodeficiency virus type 1. *Journal of Virology*, 83(16):8289–8292, 2009.

Emily J Platt, Michelle M Gomes, and David Kabat. Kinetic mechanism for HIV-1 neutralization by antibody 2G12 entails reversible glycan binding that slows cell entry. *Proceedings of the National Academy of Sciences of the United States of America*, 109(20):7829–34, May 2012.

Anastasia Pokrovskaya, Anna Popova, Natalia Ladnaya, and Oleg Yurin. The cascade of HIV care in Russia, 2011–2013. *Journal of the International AIDS Society*, 17:19506, 2014.

MJ Pongers-Willemse, T Seriu, F Stolz, E d'Aniello, P Gameiro, P Pisa, M Gonzalez, CR Bartram, ER Panzer-Grümayer, A Biondi, et al. Primers and protocols for standardized detection of minimal residual disease in acute lymphoblastic leukemia using immunoglobulin and T cell receptor gene rearrangements and TAL1 deletions as PCR targets Report of the BIOMED-1 CONCERTED ACTION: Investigation of minimal residual disease in acute leukemia. *Leukemia*, 13(1):110, 1999.

Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983, 2018.

Stephen J Popper, Abdoulaye Dieng Sarr, Karin U Travers, Aissatou Gueye-Ndiaye, Souleymane Mboup, Myron E Essex, and Phyllis J Kanki. Lower human immunodeficiency virus (HIV) type 2 viral load reflects the difference in pathogenicity of HIV-1 and HIV-2. *The Journal of Infectious Diseases*, 180(4):1116–1121, 1999.

Susana Posada-Cespedes, David Seifert, and Niko Beerenwinkel. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Research*, 239:17–32, July 2017.

Thierry Poynard, Patrick Marcellin, Samuel S Lee, Christian Niederau, Gerald S Minuk, Gaetano Ideo, Vincent Bain, Jenny Heathcote, Stefan Zeuzem, Christian Trepo, et al. Randomised trial of interferon α2b plus ribavirin for 48 weeks or for 24 weeks versus interferon α2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. *The Lancet*, 352(9138):1426–1432, 1998.

Katrien Princen, Sigrid Hatse, Kurt Vermeire, Stefano Aquaro, Erik De Clercq, Lars-Ole Gerlach, Mette Rosenkilde, Thue W Schwartz, Renato Skerlj, Gary Bridger, et al. Inhibition of human immunodeficiency virus replication by a dual CCR5/CXCR4 antagonist. *Journal of Virology*, 78(23):12996–13006, 2004.

Heino Prinz. Hill coefficients, dose-response curves and allosteric mechanisms. *Journal of Chemical Biology*, 3(1):37–44, March 2010.

Mattia C. F. Prosperi, Li Yin, David J. Nolan, Amanda D. Lowe, Maureen M. Goodenow, and Marco Salemi. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Scientific Reports*, 3(1):2837, December 2013.

Mattia CF Prosperi, Laura Bracciale, Massimiliano Fabbiani, Simona Di Giambenedetto, Francesca Razzolini, Genny Meini, Manuela Colafigli, Angela Marzocchetti, Roberto Cauda, Maurizio Zazzi, et al. Comparative determination of HIV-1 co-receptor tropism by Enhanced Sensitivity Trofile, gp120 V3-loop RNA and DNA genotyping. *Retrovirology*, 7(1):56, 2010.

Massimo Puoti, Graham R Foster, Stanley Wang, David Mutimer, Edward Gane, Christophe Moreno, Ting Tsung Chang, Samuel S Lee, Rui Marinho, Jean-Francois Dufour, et al. High SVR12 with 8-week and 12-week glecaprevir/pibrentasvir therapy: An integrated analysis of HCV genotype 1–6 patients without cirrhosis. *Journal of Hepatology*, 2018.

Shoukat H Qari, Richard Respess, Hillard Weinstock, Elise M Beltrami, Kurt Hertogs, Brendan A Larder, Christos J Petropoulos, Nicholas Hellmann, and Walid Heneine. Comparative analysis of two commercial phenotypic assays for drug susceptibility testing of human immunodeficiency virus type 1. *Journal of Clinical Microbiology*, 40(1):31–5, January 2002.

John Rachlin, Chunming Ding, Charles Cantor, and Simon Kasif. MuPlex: multi-objective multiplex PCR assay design. *Nucleic Acids Research*, 33(suppl_2):W544–W547, 2005.

Cristina Rada, Africa González-Fernández, John M. Jarvis, and César Milstein. The 5′ boundary of somatic hypermutation in a Vχ gene is in the leader intron. *European Journal of Immunology*, 24(6):1453–1457, June 1994.

Francois Raffi, Anita Rachlis, Hans-Jürgen Stellbrink, W David Hardy, Carlo Torti, Chloe Orkin, Mark Bloch, Daniel Podzamczer, Vadim Pokrovsky, Federico Pulido, Steve Almond, David Margolis, Clare Brennan, and Sherene Min. Once-daily dolutegravir versus raltegravir in antiretroviral-naive adults with HIV-1 infection: 48 week results from the randomised, double-blind, non-inferiority SPRING-2 study. *The Lancet*, 381(9868):735–743, March 2013.

Lee Ratner, William Haseltine, Roberto Patarca, Kenneth J. Livak, Bruno Starcich, Steven F. Josephs, Ellen R. Doran, J. Antoni Rafalski, Erik A. Whitehorn, Kirk Baumeister, Lucinda Ivanoff, Stephen R. Petteway, Mark L. Pearson, James A. Lautenberger, Takis S. Papas, John Ghrayeb, Nancy T. Chang, Robert C. Gallo, and Flossie Wong-Staal. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*, 313(6000):277–284, January 1985.

Dana N Raugi, Robert A Smith, Geoffrey S Gottlieb, and for the University of Washington-Dakar HIV-2 Study University of Washington-Dakar HIV-2 Study Group. Four Amino Acid Changes in HIV-2 Protease Confer Class-Wide Sensitivity to Protease Inhibitors. *Journal of Virology*, 90(2):1062–9, January 2016.

Stéphanie Raymond, Pierre Delobel, Maud Mavigner, Michelle Cazabat, Corinne Souyris, Karine Sandres-Saune, Lise Cuzin, Bruno Marchou, Patrice Massip, and Jacques Izopet. Correlation between genotypic predictions based on V3 sequences and phenotypic determination of HIV-1 tropism. *AIDS*, 22(14):F11–F16, 2008.

Lowell Jacob Reed and Hugo Muench. A simple method of estimating fifty per cent endpoints. *American Journal of Epidemiology*, 27(3): 493–497, 1938.

Jacqueline D Reeves, Sam Hibbitts, Graham Simmons, Áine McKnight, José M Azevedo-Pereira, José Moniz-Pereira, and Paul R

Clapham. Primary human immunodeficiency virus type 2 (HIV-2) isolates infect CD4-negative cells via CCR5 and CXCR4: comparison with HIV-1 and simian immunodeficiency virus and relevance to cell tropism in vivo. *Journal of Virology*, 73(9):7795–7804, 1999.

JD Reeves, E Coakley, CJ Petropoulos, and JM Whitcomb. An enhanced sensitivity Trofile HIV coreceptor tropism assay for selecting patients for therapy with entry inhibitors targeting CCR5: a review of analytical and clinical studies. *J Viral Entry*, 3(3):94–102, 2009.

Dean A Regier and Ronald C Desrosiers. The complete nucleotide sequence of a pathogenic molecular clone of simian immunodeficiency virus. *AIDS Research and Human Retroviruses*, 6(11):1221–1231, 1990.

Elizabeth C Reuman, Severine Margeridon-Thermet, Harrison B Caudill, Tommy Liu, Katyna Borroto-Esoda, Evguenia S Svarovskaia, Susan P Holmes, and Robert W Shafer. A classification model for G-to-A hypermutation in hepatitis B virus ultra-deep pyrosequencing reads. *Bioinformatics*, 26(23):2929–2932, 2010.

Soo-Yon Rhee, Kris Sankaran, Vici Varghese, Mark A Winters, Christopher B Hurt, Joseph J Eron, Neil Parkin, Susan P Holmes, Mark Holodniy, and Robert W Shafer. HIV-1 Protease, Reverse Transcriptase, and Integrase Variation. *Journal of Virology*, 90(13): 6058–6070, July 2016.

Ruy M Ribeiro, Mette D Hazenberg, Alan S Perelson, and Miles P Davenport. Naive and memory cell turnover as drivers of CCR5-to-CXCR4 tropism switch in human immunodeficiency virus type 1: implications for therapy. *Journal of Virology*, 80(2):802–809, 2006.

Mona Riemenschneider, Thomas Hummel, and Dominik Heider. SHIVA-a web application for drug resistance and tropism testing in HIV. *BMC Bioinformatics*, 17(1):314, 2016.

J D Roberts, K Bebenek, and T A Kunkel. The accuracy of reverse transcriptase from HIV-1. *Science*, 242(4882):1171–3, November 1988.

Nicola D. Roberts, R. Daniel Kortschak, Wendy T. Parker, Andreas W. Schreiber, Susan Branford, Hamish S. Scott, Garique Glonek, and David L. Adelson. A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*, 29(18):2223–2230, September 2013.

J Rockstroh, S Bhagani, R Bruno, et al. European AIDS Clinical Society (EACS) guidelines. Version 6.1. 2012, 2012.

Jürgen K Rockstroh. HCV in 2018: Success stories and remaining challenges? In *The International Liver Congress*, Paris, April 2018.

Alison J Rodger, Valentina Cambiano, Tina Bruun, Pietro Vernazza, Simon Collins, Jan Van Lunzen, Giulio Maria Corbelli, Vicente Estrada, Anna Maria Geretti, Apostolos Beloukas, et al. Sexual Activity Without Condoms and Risk of HIV Transmission in Serodifferent Couples When the HIV-Positive Partner Is Using Suppressive Antiretroviral Therapy. *JAMA*, 316(2):171, July 2016.

Arne Rodloff, Torsten Bauer, Santiago Ewig, Peter Kujath, and Eckhard Müller. Susceptible, intermediate, and resistant - the intensity of antibiotic action. *Deutsches Arzteblatt international*, 105(39):657–62, September 2008.

M Rodriguez, R J von Wedel, R S Garrett, P W Lampert, and M B Oldstone. Pituitary dwarfism in mice persistently infected with lymphocytic choriomeningitis virus. *Laboratory Investigation*, 49(1): 48–53, July 1983.

Carlos Romero. Extended lexicographic goal programming: a unifying approach. *Omega*, 29(1):63–71, February 2001.

T M Rose, E R Schultz, J G Henikoff, S Pietrokovski, C M McCallum, and S Henikoff. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic acids research*, 26(7):1628–35, apr 1998.

Timothy M Rose, Jorja G Henikoff, and Steven Henikoff. CODEHOP (COnsensus-DEgenerate hybrid oligonucleotide primer) PCR primer design. *Nucleic Acids Research*, 31(13):3763–3766, 2003.

Daniel B Rubinstein, Pierre Leblanc, Daniel G Wright, Thierry Guillaume, Alexei Strotchevoi, and Michael Boosalis. Anti-CD34+ fabs generated against hematopoietic stem cells in HIV-derived combinatorial immunoglobulin library suggest antigen-selected autoantibodiesfn2. *Molecular Immunology*, 35(14-15):955–964, 1998.

Manuel Ruiz, Véronique Giudicelli, Chantal Ginestoux, Peter Stoehr, James Robinson, Julia Bodmer, Steven G. E. Marsh, Ronald Bontrop, Marc Lemaitre, Gérard Lefranc, Denys Chaume, and Marie-Paule Lefranc. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research*, 28(1):219–221, January 2000.

Melanie A Ryan-Graham and Keith WC Peden. Both virus and host components are important for the manifestation of a Nef-phenotype in HIV-1 and HIV-2. *Virology*, 213(1):158–168, 1995.

Wojciech Rychlik. OLIGO 7 primer analysis software. In *PCR Primer Design*, pages 35–59. Springer, 2007.

Maame Efua S Sampah, Lin Shen, Benjamin L Jilek, and Robert F Siliciano. Dose-response curve slope is a missing dimension in the analysis of HIV-1 drug resistance. *Proceedings of the National Academy of Sciences*, page 201018360, 2011.

Michel Samson, Frédérick Libert, Benjamin J Doranz, Joseph Rucker, Corinne Liesnard, Claire-Michèle Farber, Sentob Saragosti, Claudine Lapouméroulie, Jacqueline Cognaux, Christine Forceille, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature*, 382 (6593):722, 1996.

Sarah Sandmann, Aniek O. de Graaf, Mohsen Karimi, Bert A. van der Reijden, Eva Hellström-Lindberg, Joop H. Jansen, and Martin Dugas. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports*, 7:43169, February 2017.

F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, December 1977.

Rafael Sanjuán and Pilar Domingo-Calap. Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, 73(23):4433–4448, 2016.

J SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 95(4): 1460–5, February 1998.

Teresa Santantonio, Massimo Fasano, Emanuele Sinisi, Angela Guastadisegni, Caterina Casalino, Michele Mazzola, Ruggiero Francavilla, and Giuseppe Pastore. Efficacy of a 24-week course of PEG-interferon α-2b monotherapy in patients with acute hepatitis C after failure of spontaneous clearance. *Journal of Hepatology*, 42(3): 329–333, 2005.

Quirina Santos-Costa, Maria Manuel Lopes, Marta Calado, and José Miguel Azevedo-Pereira. HIV-2 interaction with cell coreceptors: amino acids within the V1/V2 region of viral envelope are determinant for CCR8, CCR5 and CXCR4 usage. *Retrovirology*, 11(1): 99, 2014.

Loredana Sarmati, Emanuele Nicastri, Saverio G Parisi, Gabriella D'Ettorre, Giorgio Mancino, Pasquale Narciso, Vincenzo Vullo, and

Massimo Andreoni. Discordance between genotypic and phenotypic drug resistance profiles in human immunodeficiency virus type 1 strains isolated from peripheral blood mononuclear cells. *Journal of Clinical Microbiology*, 40(2):335–40, February 2002.

Paul E. Sax. Editorial Commentary: Can We Break the Habit of Routine CD4 Monitoring in HIV Care? *Clinical Infectious Diseases*, 56 (9):1344–1346, May 2013.

PE Sax. FDA approval: maraviroc. *AIDS Clinical Care*, 19(9):75–75, 2007.

Daniele Sblattero and Andrew Bradbury. A definitive set of oligonucleotide primers for amplifying human V regions. *Immunotechnology*, 3(4):271–278, 1998.

Timothy Schacker, Ann C Collier, James Hughes, Theresa Shea, and Lawrence Corey. Clinical and epidemiologic features of primary HIV infection. *Annals of Internal Medicine*, 125(4):257–264, 1996.

Grant D. Schauer, Kelly D. Huber, Sanford H. Leuba, and Nicolas Sluis-Cremer. Mechanism of allosteric inhibition of HIV-1 reverse transcriptase revealed by single-molecule and ensemble fluorescence. *Nucleic Acids Research*, 42(18):11687–11696, October 2014.

Sophie Schbath, Véronique Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, and Jean-François Gibrat. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*, 19(6):796–813, June 2012.

Johannes F. Scheid, Hugo Mouquet, Niklas Feldhahn, Michael S. Seaman, Klara Velinzon, John Pietzsch, Rene G. Ott, Robert M. Anthony, Henry Zebroski, Arlene Hurley, Adhuna Phogat, Bimal Chakrabarti, Yuxing Li, Mark Connors, Florencia Pereyra, Bruce D. Walker, Hedda Wardemann, David Ho, Richard T. Wyatt, John R. Mascola, Jeffrey V. Ravetch, and Michel C. Nussenzweig. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature*, 458(7238):636–640, April 2009.

Johannes F Scheid, Hugo Mouquet, Beatrix Ueberheide, Ron Diskin, Florian Klein, Thiago Y K Oliveira, John Pietzsch, David Fenyo, Alexander Abadir, Klara Velinzon, Arlene Hurley, Sunnie Myung, Farid Boulad, Pascal Poignard, Dennis R Burton, Florencia Pereyra, David D Ho, Bruce D Walker, Michael S Seaman, Pamela J Bjorkman, Brian T Chait, and Michel C Nussenzweig. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science*, 333(September):1633–1637, 2011.

R F Schinazi, R M Lloyd, M H Nguyen, D L Cannon, A McMillan, N Ilksoy, C K Chu, D C Liotta, H Z Bazmi, and J W Mellors. Characterization of human immunodeficiency viruses resistant to oxathiolane-cytosine nucleosides. *Antimicrobial Agents and Chemotherapy*, 37(4):875–81, April 1993.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as Treatments: Debiasing Learning and Evaluation. In *International Conference on Machine Learning*, pages 1670–1679, 2016.

Eileen Schneider, Suzanne Whitmore, M Kathleen Glynn, Kenneth Dominguez, Andrew Mitsch, and Matthew T McKenna. Revised surveillance case definitions for HIV infection among adults, adolescents, and children aged $<$ 18 months and for HIV infection and AIDS among children aged 18 months to $<$ 13 years-United States, 2008. *Morbidity and Mortality Weekly Report: Recommendations and Reports*, 57(10):1–12, 2008.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural Computation*, 12(5): 1207–1245, 2000.

Frederik Schulz, Natalya Yutin, Natalia N Ivanova, Davi R Ortega, Tae Kwon Lee, Julia Vierheilig, Holger Daims, Matthias Horn, Michael Wagner, Grant J Jensen, et al. Giant viruses with an expanded complement of translation system components. *Science*, 356(6333):82–85, 2017.

S A Schwartz and M P Nair. Current concepts in human immunodeficiency virus infection and AIDS. *Clinical and Diagnostic Laboratory Immunology*, 6(3):295–305, May 1999.

Eduardo Seclén, Carolina Garrido, María del Mar González, Juan González-Lahoz, Carmen de Mendoza, Vincent Soriano, and Eva Poveda. High sensitivity of specific genotypic tools for detection of X4 variants in antiretroviral-experienced patients suitable to be treated with CCR5 antagonists. *Journal of Antimicrobial Chemotherapy*, 65(7):1486–1492, 2010.

David Seifert, Francesca Di Giallonardo, Karin J Metzner, Huldrych F Günthard, and Niko Beerenwinkel. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. *Genetics*, 199(1):191–203, January 2015.

Robert W Shafer. Rationale and uses of a public HIV drug-resistance database. *The Journal of infectious diseases*, 194 Suppl 1(Suppl 1):S51–8, September 2006.

Robert W Shafer, Soo-Yon Rhee, Deenan Pillay, Veronica Miller, Paul Sandstrom, Jonathan M Schapiro, Daniel R Kuritzkes, and Diane Bennett. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS*, 21(2):215–23, January 2007.

Tim Shaw, Angeline Bartholomeusz, and Stephen Locarnini. HBV drug resistance: Mechanisms, detection and interpretation. *Journal of Hepatology*, 44(3):593–606, March 2006.

Zhiyong Shen, Wubin Qu, Wen Wang, Yiming Lu, Yonghong Wu, Zhifeng Li, Xingyi Hang, Xiaolei Wang, Dongsheng Zhao, and Chenggang Zhang. MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics*, 11(1):143, 2010.

Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, October 2008.

Anandi N Sheth, Ighovwerha Ofotokun, Kate Buchacz, Carl Armon, Joan S Chmiel, Rachel LD Hart, Rose Baker, John T Brooks, and Frank J Palella Jr. Antiretroviral regimen durability and success in treatment-naive and treatment-experienced patients by year of treatment initiation, United States, 1996–2011. *Journal of acquired immune deficiency syndromes (1999)*, 71(1):47, 2016.

C Shi and J W Mellors. A recombinant retroviral system for rapid in vivo analysis of human immunodeficiency virus type 1 susceptibility to reverse transcriptase inhibitors. *Antimicrobial Agents and Chemotherapy*, 41(12):2781–5, December 1997.

Yu Shi, Eleonor Brandin, Elzbieta Vincic, Marianne Jansson, Anders Blaxhult, Katarina Gyllensten, Lars Moberg, Christina Broström, Eva Maria Fenyö, and Jan Albert. Evolution of human immunodeficiency virus type 2 coreceptor usage, autologous neutralization, envelope sequence and glycosylation. *Journal of General Virology*, 86 (12):3385–3396, 2005.

Melissa D Simek, Wasima Rida, Frances H Priddy, Pham Pung, Emily Carrow, Dagna S Laufer, Jennifer K Lehrman, Mark Boaz, Tony Tarragona-Fiol, George Miiro, et al. Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. *Journal of Virology*, 83(14):7337–48, July 2009.

Birgitte B. Simen, Jan Fredrik Simons, Katherine Huppler Hullsiek, Richard M. Novak, Rodger D. MacArthur, John D. Baxter, Chunli Huang, Christine Lubeski, Gregory S. Turenchalk, Michael S. Braverman, Brian Desany, Jonathan M. Rothberg, Michael Egholm, and Michael J. Kozal. Low-Abundance Drug-Resistant Viral Variants in Chronically HIV-Infected, Antiretroviral Treatment-Naive Patients Significantly Impact Treatment Outcomes. *Journal of Infectious Diseases*, 199(5):693–701, March 2009.

P. Simmonds. Genetic diversity and evolution of hepatitis C virus - 15 years on. *Journal of General Virology*, 85(11):3173–3188, November 2004.

Tobias Sing, Andrew J Low, Niko Beerenwinkel, Oliver Sander, Peter K Cheung, Francisco S Domingues, J Buch, M Daumer, Rolf Kaiser, Thomas Lengauer, et al. Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antiviral Therapy*, 12(7): 1097, 2007.

Rita Sipos, Anna J Székely, Márton Palatinszky, Sára Révész, Károly Márialigeti, and Marcell Nikolausz. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiology Ecology*, 60(2):341–350, 2007.

Helena Skar, Pedro Borrego, Timothy C Wallstrom, Mattias Mild, José Maria Marcelino, Helena Barroso, Nuno Taveira, Thomas Leitner, and Jan Albert. HIV-2 genetic evolution in patients with advanced disease is faster than that in matched HIV-1 patients. *Journal of Virology*, 84(14):7412–7415, 2010.

Renato T Skerlj, Gary J Bridger, Al Kaller, Ernest J McEachern, Jason B Crawford, Yuanxi Zhou, Bem Atsma, Jonathon Langille, Susan Nan, Duane Veale, et al. Discovery of Novel Small Molecule Orally Bioavailable C-X-C Chemokine Receptor 4 Antagonists That Are Potent Inhibitors of T-Tropic (X4) HIV-1 Replication. *Journal of Medicinal Chemistry*, 53(8):3376–3388, 2010.

Katharina Skrabal, Virginie Trouplin, Béatrice Labrosse, Véronique Obry, Florence Damond, Allan J Hance, François Clavel, and Fabrizio Mammano. Impact of antiretroviral treatment on the tropism of HIV-1 plasma virus populations. *AIDS*, 17(6):809–814, 2003.

Katharina Skrabal, Andrew J Low, Winnie Dong, Tobias Sing, Peter K Cheung, Fabrizio Mammano, and P Richard Harrigan. Determining human immunodeficiency virus coreceptor use in a clinical setting: degree of correlation between two phenotypic assays and a bioinformatic model. *Journal of Clinical Microbiology*, 45(2):279–284, 2007.

Donald B Smith, Jens Bukh, Carla Kuiken, A Scott Muerhoff, Charles M Rice, Jack T Stapleton, and Peter Simmonds. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology*, 59(1):318–327, 2014.

Lloyd M. Smith, Jane Z. Sanders, Robert J. Kaiser, Peter Hughes, Chris Dodd, Charles R. Connell, Cheryl Heiner, Stephen B. H. Kent, and Leroy E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, June 1986.

Robert A Smith, Donovan J Anderson, Crystal L Pyrak, Bradley D Preston, and Geoffrey S Gottlieb. Antiretroviral drug resistance in HIV-2: three amino acid changes are sufficient for classwide nucleoside analogue resistance. *The Journal of Infectious Diseases*, 199 (9):1323–1326, 2009.

T Smith and M Waterman. Identification of Common Molecular Subsequences. *Molecular Biology*, 147:195–197, 1981.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

Vicente Soriano, Jose Vicente Fernandez-Montero, Laura Benitez-Gutierrez, Carmen de Mendoza, Ana Arias, Pablo Barreiro, José M. Peña, and Pablo Labarga. Dual antiretroviral therapy for HIV infection. *Expert Opinion on Drug Safety*, 16(8):923–932, August 2017.

Richard Souvenir, Jeremy Buhler, Gary Stormo, and Weixiong Zhang. Selecting degenerate multiplex PCR primers. In *International Workshop on Algorithms in Bioinformatics*, pages 512–526. Springer, 2003.

David H. Spencer, Manoj Tyagi, Francesco Vallania, Andrew J. Bredemeyer, John D. Pfeifer, Rob D. Mitra, and Eric J. Duncavage. Performance of Common Analysis Methods for Detecting Low-Frequency Single Nucleotide Variants in Targeted Next-Generation Sequence Data. *The Journal of Molecular Diagnostics*, 16(1):75–88, January 2014.

William R Spreen, David A Margolis, and John C Pottage Jr. Long-acting injectable antiretrovirals for HIV treatment and prevention. *Current Opinion in HIV and AIDS*, 8(6):565, 2013.

Gyan Prakash Srivastava and Dong Xu. Genome-scale probe and primer design with PRIMEGENS. In *PCR Primer Design*, pages 159–175. Springer, 2007.

Ralph Stadhouders, Suzan D Pas, Jeer Anber, Jolanda Voermans, Ted H M Mes, and Martin Schutten. The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5′ nuclease assay. *The Journal of Molecular Diagnostics*, 12 (1):109–17, 2010.

STAT. IBM pitched Watson as a revolution in cancer care. It's nowhere close, 2018. URL https://www.statnews.com/2017/09/05/watson-ibm-cancer/.

James J Steinhardt, Javier Guenaga, Hannah L Turner, Krisha McKee, Mark K Louder, Sijy O'Dell, Chi-I Chiang, Lin Lei, Andrey Galkin, Alexander K Andrianov, et al. Rational design of a trispecific antibody targeting the HIV-1 Env with elevated anti-viral activity. *Nature Communications*, 9(1):877, 2018.

Joanne D. Stekler, Giovanina M. Ellis, Jacquelyn Carlsson, Braiden Eilers, Sarah Holte, Janine Maenza, Claire E. Stevens, Ann C. Collier, and Lisa M. Frenkel. Prevalence and Impact of Minority Variant Drug Resistance Mutations in Primary HIV-1 Infection. *PLOS ONE*, 6(12):e28952, December 2011.

J Claiborne Stephens, David E Reich, David B Goldstein, Hyoung Doo Shin, Michael W Smith, Mary Carrington, Cheryl Winkler, Gavin A Huttley, Rando Allikmets, Lynn Schriml, et al. Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes. *The American Journal of Human Genetics*, 62(6): 1507–1515, 1998.

NB Stiernholm, Beata Kuzniar, and Neil L Berinstein. Identification of a new human V lambda gene family–V lambda X. *The Journal of Immunology*, 152(10):4969–4975, 1994.

Susan L. Stramer, Simone A. Glynn, Steven H. Kleinman, D. Michael Strong, Sally Caglioti, David J. Wright, Roger Y. Dodd, and Michael P. Busch. Detection of HIV-1 and HCV Infections among Antibody-Negative Blood Donors by Nucleic Acid-Amplification Testing. *New England Journal of Medicine*, 351(8):760–768, August 2004.

Julie M Strizki, Cecile Tremblay, Serena Xu, Lisa Wojcik, Nicole Wagner, Waldemar Gonsiorek, R William Hipkin, Chuan-Chu Chou, Catherine Pugliese-Sivo, Yushi Xiao, et al. Discovery and characterization of vicriviroc (SCH 417690), a CCR5 antagonist with potent activity against human immunodeficiency virus type 1. *Antimicrobial Agents and Chemotherapy*, 49(12):4911–4919, 2005.

Mark S. Sulkowski. Drug-Induced Liver Injury Associated with Antiretroviral Therapy that Includes HIV-1 Protease Inhibitors. *Clinical Infectious Diseases*, 38(Supplement_2):S90–S97, March 2004.

Ying Sun, Hong-Yan Liu, Ling Mu, and En-Jie Luo. Degenerate primer design to clone the human repertoire of immunoglobulin heavy chain variable regions. *World Journal of Microbiology and Biotechnology*, 28(1):381–386, January 2012.

Simone Susser, Johannes Vermehren, Nicole Forestier, Martin Walter Welker, Natalia Grigorian, Caterina Füller, Dany Perner, Stefan Zeuzem, and Christoph Sarrazin. Analysis of long-term persistence of resistance mutations within the hepatitis C virus NS3 protease after treatment with telaprevir or boceprevir. *Journal of Clinical Virology*, 52(4):321–7, December 2011.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement Learning: An Introduction*. MIT press, 1998.

T Szczepański, MJ Willemse, ER Van Wering, JF Van Weerden, WA Kamps, and JJM Van Dongen. Precursor-B-ALL with D H–J H gene rearrangements have an immature immunogenotype with a high frequency of oligoclonality and hyperdiploidy of chromosome 14. *Leukemia*, 15(9):1415, 2001.

Hakim Tafer, Christian Höner zu Siederdissen, Peter F Stadler, Stephan H Bernhart, Ivo L Hofacker, Ronny Lorenz, and Christoph Flamm. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6 (1):26, 2011.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

Yasuhiro Takeuchi, Myra O McClure, and Massimo Pizzato. Identification of gammaretroviruses constitutively released from cell lines used for human immunodeficiency virus research. *Journal of Virology*, 82(24):12585–12588, 2008.

Qiuxiang Tan, Ya Zhu, Jian Li, Zhuxi Chen, Gye Won Han, Irina Kufareva, Tingting Li, Limin Ma, Gustavo Fenalti, Jing Li, et al. Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science*, 341(6152):1387–1390, 2013.

Michele W. Tang and Robert W. Shafer. HIV-1 Antiretroviral Resistance. *Drugs*, 72(9):e1–e25, June 2012.

Michele W Tang, Tommy F Liu, and Robert W Shafer. The HIVdb system for HIV-1 genotypic resistance interpretation. *Intervirology*, 55(2):98–101, 2012.

Amalio Telenti, Levi C T Pierce, William H Biggs, Julia di Iulio, Emily H M Wong, Martin M Fabani, Ewen F Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A Perkins, Franz J Och, Yaron Turpaz, and J Craig Venter. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42):11901–11906, October 2016.

M Thali, J P Moore, C Furman, M Charles, D D Ho, J Robinson, and J Sodroski. Characterization of conserved human immunodeficiency virus type 1 gp120 neutralization epitopes exposed upon gp120-CD4 binding. *Journal of Virology*, 67(7):3978–88, July 1993.

Rodolphe Thiébaut, Sophie Matheron, Audrey Taieb, Francoise Brun-Vezinet, Geneviève Chêne, and Brigitte Autran. Long-term nonprogressors and elite controllers in the ANRS CO5 HIV-2 cohort. *AIDS*, 25(6):865–867, March 2011.

Alexander Thielen. One year of routine HIV-1 drug resistance testing by deep sequencing: insights from comparative Sanger sequencing. In *12th European HIV & Hepatitis Workshop*, Barcelona, Spain, 2014.

Alexander Thielen and Thomas Lengauer. Geno2pheno [454]: a Web server for the prediction of HIV-1 coreceptor usage from next-generation sequencing data. *Intervirology*, 55(2):113–117, 2012.

Julie D Thompson, Benjamin Linard, Odile Lecompte, and Olivier Poch. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PLOS ONE*, 6(3):e18093, March 2011.

Brenda Thornton and Chhandak Basu. Real-time PCR (qPCR) primer design using free online software. *Biochemistry and Molecular Biology Education*, 39(2):145–154, March 2011.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 411–423, 2001.

Caroline T Tiemessen, Neil Martinson, Mb Bch, Chris Hani, and Baragwanath Hospital. Elite controllers: understanding natural suppressive control of HIV-1 infection. *Continuing Medical Education*, 282(8), 2012.

Thomas Tiller, Makoto Tsuiji, Sergey Yurasov, Klara Velinzon, Michel C. Nussenzweig, and Hedda Wardemann. Autoreactivity in Human IgG+ Memory B Cells. *Immunity*, 26(2):205–213, February 2007.

Thomas Tiller, Eric Meffre, Sergey Yurasov, Makoto Tsuiji, Michel C. Nussenzweig, and Hedda Wardemann. Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *Journal of Immunological Methods*, 329(1-2):112–124, 2008.

Erik Timmers, Marcel M Hermans, Margriet EM Kraakman, Rudolf W Hendriks, and Ruud KB Schuurman. Diversity of immunoglobulin χ light chain gene rearrangements and evidence for somatic mutation in VχIV family gene segments in X-linked agammaglobulinemia. *European Journal of Immunology*, 23(3):619–624, 1993.

Rania A Tohme and Scott D Holmberg. Is sexual contact a major mode of hepatitis C virus transmission? *Hepatology*, 52(4):1497–1505, 2010.

Armin Töpfer, Tobias Marschall, Rowena A. Bull, Fabio Luciani, Alexander Schönhuth, and Niko Beerenwinkel. Viral Quasispecies Assembly via Maximal Clique Enumeration. *PLOS Computational Biology*, 10(3):e1003515, March 2014.

L Trinh, D Han, W Huang, T Wrin, J Larson, L Kiss, E Coakley, CJ Petropoulos, N Parkin, JM Whitcomb, and JD Reeves. Validation of an enhanced sensitivity Trofile HIV-1 co-receptor tropism assay for selecting patients for therapy with entry inhibitors targeting CCR5. *Journal of the International AIDS Society*, 11(Suppl 1):P197, November 2008.

A Trkola, M Purtscher, T Muster, C Ballaun, A Buchacher, N Sullivan, K Srinivasan, J Sodroski, J P Moore, and H Katinger. Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. *Journal of Virology*, 70(2):1100–8, February 1996.

Alexandra Trkola. Potency needs constancy. *Nature*, 514(7523):442–443, October 2014.

Athanasios C Tsiatis, Alexis Norris-Kirby, Roy G Rich, Michael J Hafez, Christopher D Gocke, James R Eshleman, and Kathleen M Murphy. Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. *The Journal of Molecular Diagnostics*, 12(4):425–432, 2010.

Edouard Tuaillon, Marie Gueudin, Véronique Lemée, Isabelle Gueit, Pierre Roques, Gary E Corrigan, Jean-Christophe Plantier, François Simon, and Joséphine Braun. Phenotypic susceptibility to non-nucleoside inhibitors of virion-associated reverse transcriptase from different HIV types and groups. *Journal of Acquired Immune Deficiency Syndromes*, 37(5):1543–1549, 2004.

Douglas H. Turner and David H. Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(suppl_1):D280–D282, January 2010.

Daniel L Tuttle, Jeffrey K Harrison, Cynthia Anders, John W Sleasman, and Maureen M Goodenow. Expression of CCR5 increases during monocyte differentiation and directly mediates macrophage susceptibility to infection by human immunodeficiency virus type 1. *Journal of Virology*, 72(6):4962–4969, 1998.

UNAIDS. UNAIDS reports a 52% reduction in new HIV infections among children and a combined 33% reduction among adults and children since 2001 | UNAIDS, 2013. URL http://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2013/september/20130923prunga.

UNAIDS. AIDS by the numbers - AIDS is not over, but it can be | UNAIDS. Technical report, UNAIDS, 2016. URL http://www.unaids.org/en/resources/documents/2016/AIDS-by-the-numbers.

UNAIDS. Fact sheet - Latest statistics on the status of the AIDS epidemic | UNAIDS, 2018. URL http://www.unaids.org/en/resources/fact-sheet.

Thijs JW van de Laar, Akke K van der Bij, Maria Prins, Sylvia M Bruisten, Kees Brinkman, Thomas A Ruys, Jan TM van der Meer, Henry JC de Vries, Jan-Willem Mulder, Michiel van Agtmael, et al. Increase in HCV incidence among men who have sex with men in Amsterdam most likely caused by sexual transmission. *The Journal of Infectious Diseases*, 196(2):230–238, 2007.

Mirjam van der Burg, Talip Tümkaya, Marjan Boerma, Sandra de Bruin-Versteeg, Anton W Langerak, and Jacques JM van Dongen. Ordered recombination of immunoglobulin light chain genes occurs at the IGK locus but seems less strict at theIGL locus. *Blood*, 97(4):1001–1008, 2001.

ME van Der Ende, Christophe Guillon, PHM Boers, RA Gruters, P Racz, K Tenner-Racz, ADME Osterhaus, and Martin Schutten. Broadening of coreceptor usage by human immunodeficiency virus type 2 does not correlate with increased pathogenicity in an in vivo model. *Journal of General Virology*, 81(2):507–513, 2000.

JJM Van Dongen, AW Langerak, Monika Brüggemann, PAS Evans, Michael Hummel, FL Lavender, E Delabesse, Frédéric Davi, E Schuuring, Ramon García-Sanz, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*, 17(12):2257, 2003.

Johan H van Es, FH Gmelig Meyling, WR van De Akker, Henk Aanstoot, RH Derksen, and Ton Logtenberg. Somatic mutations in the variable regions of a human IgG anti-double-stranded DNA autoantibody suggest a role for antigen in the induction of systemic lupus erythematosus. *Journal of Experimental Medicine*, 173 (2):461–470, 1991.

Kristel Van Laethem, Paul De Munter, Yoeri Schrooten, Rene Verbesselt, Marc Van Ranst, Eric Van Wijngaerden, and Anne-Mieke Vandamme. No response to first-line tenofovir+lamivudine+efavirenz despite optimization according to baseline resistance testing: Impact of resistant minority variants on efficacy of low genetic barrier drugs. *Journal of Clinical Virology*, 39(1): 43–47, May 2007.

Steven F L van Lelyveld, Annemarie M J Wensing, and Andy I M Hoepelman. The MOTIVATE trials: maraviroc therapy in antiretroviral treatment-experienced HIV-1-infected patients. *Expert Review of Anti-Infective Therapy*, 10(11):1241–7, 2012.

LPR Vandekerckhove, AMJ Wensing, Rolf Kaiser, Francoise Brun-Vezinet, Bonaventura Clotet, Andrea De Luca, Stephan Dressler, Federico Garcia, Anna Maria Geretti, Thomas Klimkait, et al. European guidelines on the clinical management of HIV-1 tropism testing. *The Lancet Infectious Diseases*, 11(5):394–407, 2011.

Carmen Vandelli, Francesco Renzo, Luisa Romanò, Sergio Tisminetzky, Marisa De Palma, Tommaso Stroffolini, Ezio Ventura, and

Alessandro Zanetti. Lack of evidence of sexual transmission of hepatitis C among monogamous couples: results of a 10-year prospective follow-up study. *The American Journal of Gastroenterology*, 99(5):855, 2004.

Bie MP Verbist, Kim Thys, Joke Reumers, Yves Wetzels, Koen Van der Borght, Willem Talloen, Jeroen Aerssens, Lieven Clement, and Olivier Thas. VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics*, 31(1):94–101, 2014.

Jurgen Vercauteren and Anne-Mieke Vandamme. Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral Research*, 71(2-3):335–342, 2006.

OJHM Verhagen, MJ Willemse, WB Breunis, AJM Wijkhuijs, DCH Jacobs, SA Joosten, ER Van Wering, JJM Van Dongen, and CE Van der Schoot. Application of germline IGH probes in real-time quantitative PCR for the detection of minimal residual disease in acute lymphoblastic leukemia. *Leukemia*, 14(8):1426, 2000.

E. Vicenzi and G. Poli. Novel factors interfering with human immunodeficiency virus-type 1 replication in vivo and in vitro. *Tissue Antigens*, 81(2):61–71, February 2013.

Gabriel D. Victora and Michel C. Nussenzweig. Germinal Centers. *Annual Review of Immunology*, 30(1):429–457, April 2012.

Benoit Visseaux, Margarita Hurtado-Nedelec, Charlotte Charpentier, Gilles Collin, Alexandre Storto, Sophie Matheron, Lucile Larrouy, Florence Damond, Françoise Brun-Vézinet, Diane Descamps, et al. Molecular determinants of HIV-2 R5-X4 tropism in the V3 loop: development of a new genotypic tool. *Journal of Infectious Diseases*, 205(1):111–120, 2011.

Benoit Visseaux, Charlotte Charpentier, Margarita Hurtado-Nedelec, Alexandre Storto, Romain Antoine, Gilles Peytavin, Florence Damond, Sophie Matheron, Françoise Brun-Vézinet, Diane Descamps, et al. In vitro phenotypic susceptibility of HIV-2 clinical isolates to CCR5 inhibitors. *Antimicrobial Agents and Chemotherapy*, 56(1): 137–139, 2012.

Dalma Vödrös, Charlotte Tscherning-Casper, Leonor Navea, Dominique Schols, Erik De Clercq, and Eva Maria Fenyö. Quantitative evaluation of HIV-1 coreceptor use in the GHOST (3) cell assay. *Virology*, 291(1):1–11, 2001.

Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, 55(4):641–58, April 2009.

Bram Vrancken, Nídia Sequeira Trovão, Guy Baele, Eric Van Wijngaerden, Anne-Mieke Vandamme, Kristel Van Laethem, and Philippe Lemey. Quantifying next generation sequencing sample pre-processing bias in HIV-1 complete genome sequencing. *Viruses*, 8(1):12, 2016.

Nienke Vrisekoop, Julia Drylewicz, Rogier Van Gent, Tendai Mugwagwa, Steven FL Van Lelyveld, Ellen Veel, Sigrid A Otto, Mariëtte T Ackermans, Joost N Vermeulen, Hidde H Huidekoper, et al. Quantification of naive and memory T-cell turnover during HIV-1 infection. *AIDS*, 29(16):2071–2080, 2015.

Laura M Walker, Sanjay K Phogat, Po-Ying Chan-Hui, Denise Wagner, Pham Phung, Julie L Goss, Terri Wrin, Melissa D Simek, Steven Fling, Jennifer L Mitcham, et al. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289, 2009.

Laura M Walker, Michael Huber, Katie J Doores, Emilia Falkowska, Robert Pejchal, Jean-Philippe Julien, Sheng-Kai Wang, Alejandra Ramos, Po-Ying Chan-Hui, Matthew Moyle, et al. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*, 477(7365):466, 2011.

R B Wallace, J Shaffer, R F Murphy, J Bonner, T Hirose, and K Itakura. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Research*, 6(11):3543–57, August 1979.

H Walter, B Schmidt, K Korn, A M Vandamme, T Harrer, and K Uberla. Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. *Journal of Clinical Virology*, 13(1-2):71–80, June 1999.

Gary P. Wang, Norah Terrault, Jacqueline D. Reeves, Lin Liu, Eric Li, Lisa Zhao, Joseph K. Lim, Giuseppe Morelli, Alexander Kuo, Josh Levitsky, Kenneth E. Sherman, Lynn M. Frazier, Ananthakrishnan Ramani, Joy Peter, Lucy Akuskevich, Michael W. Fried, and David R. Nelson. Prevalence and impact of baseline resistance-associated substitutions on the efficacy of ledipasvir/sofosbuvir or simeprevir/sofosbuvir against GT1 HCV infection. *Scientific Reports*, 8(1):3199, dec 2018.

Jiren Wang, Kuo-Bin Li, and Wing-Kin Sung. G-PRIMER: greedy algorithm for selecting minimal primer set. *Bioinformatics*, 20(15): 2473–2475, 2004a.

Kai Wang, Ram Samudrala, and John E Mittler. Antivirogram or phenosense: a comparison of their reproducibility and an analysis of their correlation. *Antiviral Therapy*, 9(5):703–12, October 2004b.

Qingguo Wang, Peilin Jia, Fei Li, Haiquan Chen, Hongbin Ji, Donald Hucks, Kimberly Dahlman, William Pao, and Zhongming Zhao. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Medicine*, 5(10):91, 2013.

Hedda Wardemann, Sergey Yurasov, Anne Schaefer, James W Young, Eric Meffre, and Michel C Nussenzweig. Predominant autoantibody production by early human B cell precursors. *Science*, 301(5638): 1374–1377, 2003.

Xiping Wei, Julie M Decker, Hongmei Liu, Zee Zhang, Ramin B Arani, J Michael Kilby, Michael S Saag, Xiaoyun Wu, George M Shaw, and John C Kappes. Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. *Antimicrobial Agents and Chemotherapy*, 46(6):1896–1905, 2002.

Milton C. Weinstein, Sue J. Goldie, Elena Losina, Calvin J. Cohen, John D. Baxter, Hong Zhang, April D. Kimmel, and Kenneth A. Freedberg. Use of Genotypic Resistance Testing To Guide HIV Therapy: Clinical Impact and Cost-Effectiveness. *Annals of Internal Medicine*, 134(6):440, March 2001.

Barbara Weiser, Sean Philpott, Thomas Klimkait, Harold Burger, Christina Kitchen, Philippe Bürgisser, Meri Gorgievski, Luc Perrin, Jean-Claude Piffaretti, Bruno Ledergerber, et al. HIV-1 coreceptor usage and CXCR4-specific viral load predict clinical disease progression during combination antiretroviral therapy. *AIDS*, 22(4): 469–479, 2008.

Nan-Ping Weng, James G Snyder, Li-Yuan Yu-Lee, and Donald M Marcus. Polymorphism of human immunoglobulin VH4 germ-line genes. *European Journal of Immunology*, 22(4):1075–1082, 1992.

Anthony P West, Ron Diskin, Michel C Nussenzweig, and Pamela J Bjorkman. Structural basis for germ-line gene usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120. *Proceedings of the National Academy of Sciences*, 109(30):E2083–E2090, 2012.

Anthony P. West, Louise Scharf, Johannes F. Scheid, Florian Klein, Pamela J. Bjorkman, and Michel C. Nussenzweig. Structural Insights on the Role of Antibodies in HIV-1 Vaccine and Therapy. *Cell*, 156 (4):633–648, February 2014.

David M Whiley and Theo P Sloots. Sequence variation in primer targets affects the accuracy of viral quantitative PCR. *Journal of Clinical Virology*, 34(2):104–107, 2005.

Jeannette M Whitcomb, Wei Huang, Signe Fransen, Kay Limoli, Jonathan Toma, Terri Wrin, Colombe Chappey, Linda DB Kiss, Ellen E Paxinos, and Christos J Petropoulos. Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism. *Antimicrobial Agents and Chemotherapy*, 51(2):566–575, 2007.

WHO. WHO Expands Recommendation on Oral Pre-Exposure Prophylaxis on HIV Infection (PrEP), 2015.

WHO. Combating hepatitis B and C to reach elimination by 2030: Advocacy Brief. Technical report, World Health Organization, 2016.

WHO. 90-90-90 - An ambitious treatment target to help end the AIDS epidemic. Technical report, WHO, 2017. URL `http://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf`.

William B Widhelm. Extensions of goal programming models. *Omega*, 9(2):212–214, January 1981.

Johannes Wiegand, Peter Buggisch, Wulf Boecher, Stefan Zeuzem, Cornelia M Gelbmann, Thomas Berg, Wolfgang Kauffmann, Birgit Kallinowski, Markus Cornberg, Elmar Jaeckel, et al. Early monotherapy with pegylated interferon alpha-2b for acute hepatitis C infection: The HEP-NET acute-HCV-II study. *Hepatology*, 43(2): 250–256, 2006.

Stefan Z Wiktor and Yvan JF Hutin. The global burden of viral hepatitis: better estimates to guide hepatitis elimination efforts. *The Lancet*, 388(10049):1030–1031, 2016.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.

Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan.

LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, 40(22):11189–11201, December 2012.

Bart Winters, Julio Montaner, P Richard Harrigan, Brian Gazzard, Anton Pozniak, Michael D Miller, Sean Emery, Frank van Leth, Patrick Robinson, John D Baxter, Marie Perez-Elias, Delivette Castor, Scott Hammer, Alex Rinehart, Hans Vermeiren, Elke Van Craenenbroeck, and Lee Bacheler. Determination of Clinically Relevant Cutoffs for HIV-1 Phenotypic Resistance Estimates Through a Combined Analysis of Clinical Trial and Cohort Data. *Journal of Acquired Immune Deficiency Syndromes*, 48(1):26–34, May 2008.

Bart Winters, Elke Van Craenenbroeck, Koen Van der Borght, Pierre Lecocq, Jorge Villacian, and Lee Bacheler. Clinical cut-offs for HIV-1 phenotypic resistance estimates: Update based on recent pivotal clinical trial data and a revised approach to viral mixtures. *Journal of Virological Methods*, 162(1-2):101–108, December 2009.

Marc Wirden, Cathia Soulie, Marc-Antoine Valantin, Slim Fourati, Anne Simon, Sidonie Lambert-Niclot, Manuela Bonmarchand, Cyril Clavel-Osorio, Anne-Genevieve Marcelin, Christine Katlama, et al. Historical HIV-RNA resistance test results are more informative than proviral DNA genotyping in cases of suppressed or residual viraemia. *Journal of Antimicrobial Chemotherapy*, 66(4): 709–712, 2011.

Myriam Witvrouw, Christophe Pannecouque, Kristel Van Laethem, Jan Desmyter, Erik De Clercq, and Anne-Mieke Vandamme. Activity of non-nucleoside reverse transcriptase inhibitors against HIV-2 and SIV. *AIDS*, 13(12):1477–1483, 1999.

Myriam Witvrouw, Christophe Pannecouque, William M Switzer, Thomas M Folks, Erik De Clercq, and Walid Heneine. Susceptibility of HIV-2, SIV and SHIV to various anti-HIV-1 compounds: implications for treatment and postexposure prophylaxis. *Antiviral Therapy*, 9(1):57–66, 2004.

Erik S. Wright, L. Safak Yilmaz, Sri Ram, Jeremy M. Gasser, Gregory W. Harrington, and Daniel R. Noguera. Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical DNA templates. *Environmental Microbiology*, 16(5):1354–1365, 2014.

Jer-Horng Wu, Pei-Ying Hong, and Wen-Tso Liu. Quantitative effects of position and type of single mismatch on single base primer extension. *Journal of Microbiological Methods*, 77(3):267–275, 2009.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005, 2004.

Xueling Wu, Zhi-Yong Yang, Yuxing Li, Carl-Magnus Hogerkorp, William R Schief, Michael S Seaman, Tongqing Zhou, Stephen D Schmidt, Lan Wu, Ling Xu, Nancy S Longo, Krisha McKee, Sijy O'Dell, Mark K Louder, Diane L Wycuff, Yu Feng, Martha Nason, Nicole Doria-Rose, Mark Connors, Peter D Kwong, Mario Roederer, Richard T Wyatt, Gary J Nabel, and John R Mascola. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science*, 329(5993):856–61, August 2010a.

Xueling Wu, Tongqing Zhou, Jiang Zhu, Baoshan Zhang, Ivelin Georgiev, Charlene Wang, Xuejun Chen, Nancy S Longo, Mark Louder, Krisha McKee, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science*, 333(6049):1593–602, September 2011.

Yu-Chang Wu, David Kipling, Hui Sun Leong, Victoria Martin, Alexander A Ademokun, and Deborah K Dunn-Walters. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood*, 116(7):1070–1078, 2010b.

Chang Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24, January 2018.

Huilei Xu, John DiCarlo, Ravi Satya, Quan Peng, and Yexun Wang. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*, 15(1):244, March 2014.

Gur Yaari, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N H Stern, Kevin C O'Connor, David A Hafler, Uri Laserson, Francois Vigneault, and Steven H Kleinstein. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in Immunology*, 4:358, 2013.

Tomoyuki Yamada, Haruhiko Soma, and Shinichi Morishita. PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome. *Nucleic Acids Research*, 34(suppl_2): W665–W669, 2006.

Xiao Yang, Patrick Charlebois, Alex Macalalad, Matthew R Henn, and Michael C Zody. V-Phaser 2: variant inference for viral populations. *BMC Genomics*, 14(1):674, October 2013.

Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40, July 2013a.

Jian Ye, Ning Ma, Thomas L Madden, and James M Ostell. Igblast: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40, 2013b.

Satoshi Yoshimi, Michio Imamura, Eisuke Murakami, Nobuhiko Hiraga, Masataka Tsuge, Yoshiiku Kawakami, Hiroshi Aikata, Hiromi Abe, C. Nelson Hayes, Tamito Sasaki, Hidenori Ochi, and Kazuaki Chayama. Long term persistence of NS5A inhibitor-resistant hepatitis C virus in patients who failed daclatasvir and asunaprevir therapy. *Journal of Medical Virology*, 87(11):1913–1920, November 2015.

Neal E Young. Greedy Set-Cover Algorithms. In *Encyclopedia of Algorithms*, pages 1–99. Springer, 2008.

A. Yuryev, JianPing Huang, Mark Pohl, Robert Patch, Felicia Watson, Peter Bell, Miriam Donaldson, Michael S. Phillips, and Michael T. Boyce-Jacino. Predicting the success of primer extension genotyping assays using statistical modeling. *Nucleic Acids Research*, 30(23): 131e–131, December 2002.

Osvaldo Zagordi, Arnab Bhattacharya, Nicholas Eriksson, and Niko Beerenwinkel. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, 12(1):119, April 2011.

Maurizio Zazzi, R Kaiser, A Sönnerborg, Daniel Struck, Andre Altmann, Mattia Prosperi, M Rosen-Zvi, A Petroczi, Y Peres, E Schülter, et al. Prediction of response to antiretroviral therapy by human experts and by the euresist data-driven expert system (the eve study). *HIV Medicine*, 12(4):211–218, 2011.

R. K. Zeldin and Richard A. Petruschke. Pharmacological and therapeutic properties of ritonavir-boosted protease inhibitor therapy in HIV-infected patients. *Journal of Antimicrobial Chemotherapy*, 53(1):4–9, December 2003.

Stefan Zeuzem, Masashi Mizokami, Stephen Pianko, Alessandra Mangia, Kwang-Hyub Han, Ross Martin, Evguenia Svarovskaia, Hadas Dvory-Sobol, Brian Doehle, Charlotte Hedskog, et al. NS5A resistance-associated substitutions in patients with genotype 1 hepatitis C virus: Prevalence and effect on treatment outcome. *Journal of Hepatology*, 66(5):910–918, May 2017.

Jie Zhang, Soo-Yon Rhee, Jonathan Taylor, and Robert W Shafer. Comparison of the precision and sensitivity of the Antivirogram and PhenoSense HIV drug susceptibility assays. *Journal of Acquired Immune Deficiency Syndromes*, 38(4):439–44, April 2005.

Jing Zhang, Tingjun Hou, Wei Wang, and Jun S Liu. Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 107(4):1321–6, January 2010.

Yi-jun Zhang, Bernard Lou, Renu B Lal, Agegnehu Gettie, Preston A Marx, and John P Moore. Use of inhibitors to evaluate coreceptor usage by simian and simian/human immunodeficiency viruses and human immunodeficiency virus type 2 in primary cells. *Journal of Virology*, 74(15):6893–6910, 2000.

Tongqing Zhou, Ivelin Georgiev, Xueling Wu, Zhi-Yong Yang, Kaifan Dai, Andrés Finzi, Young Do Kwon, Johannes F Scheid, Wei Shi, Ling Xu, Yongping Yang, Jiang Zhu, Michel C Nussenzweig, Joseph Sodroski, Lawrence Shapiro, Gary J Nabel, John R Mascola, and Peter D Kwong. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science*, 329(5993):811–7, August 2010.

Yong-Zhe Zhu, Xi-Jing Qian, Ping Zhao, and Zhong-Tian Qi. How hepatitis C virus invades hepatocytes: the mystery of viral entry. *World Journal of Gastroenterology*, 20(13):3457–67, April 2014.