Computational Methods for Personalized Cancer Therapy Based on Genomics Data

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät der
Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Dipl.-Inform. (Bioinformatik)

Magdalena Feldhahn

aus Filderstadt

Tübingen

2012

Tag der mündlichen Prüfung:	25.06.2013
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr.-Ing. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Jürgen Bauer

Abstract

Despite the considerable progress in understanding cancer biology and cancer development that has been made over the last decades, the treatment options for cancer are still insufficient. This can be attributed to the tremendous heterogeneity of cancers, with respect to appearance, clinical outcome, and underlying genetic alterations. In traditional concepts of drug design and drug administration, pathologically similar diseases are treated with the same drugs. These approaches are not adequate to face the complexity of cancer. Personalized or individualized approaches, targeting individual characteristics of tumors, are promising concepts to develop successful treatment options for cancer with little side-effects.

The human organism is equipped with a powerful system that is capable of targeting abnormal cells specifically and efficiently: the immune system. T cells can distinguish healthy cells from infected or aberrant cells by scanning peptides that are presented on the surface of other cells. Genetic alterations in cancer cells can lead to the presentation of cancer-specific peptides that drive a very specific immune reaction against the cancer cells. These peptides are called cancer-specific T-cell epitopes. Each patient's immune system is individual with respect to the peptides that can elicit an immune response. The design of tailor-made immunotherapies against individual tumors can thus be realized by using sets of patient- and tumor-specific T-cell epitopes in so-called epitope-based vaccines.

A first major challenge in the development of such individualized therapies lies in the analysis of genetic information of individual cancers, which is necessary to detect cancer-specific mutations. A second challenge is the correct identification and selection of T-cell epitopes resulting from these mutations. In this thesis, we present computational methods that address these challenges. Starting from next-generation sequencing data of cancer and normal tissue from individual patients, we identify those mutations that are uniquely present in the tumor. We integrate information from gene expression, biological pathways, and functional annotation of genes and proteins to select suitable mutations. These mutations form the basis for potential targets for individualized immunotherapies. We present prediction algorithms based on machine learning approaches that identify T-cell epitopes that are specific for a patient's tumor and immune system.

In order to bring the computational methods to clinical applications, results have to be obtained in a reliable, reproducible, and timely manner, and have to be made available to clinical researchers in an easy-to-use and intuitive way. An additional focus of this thesis is thus the development of pipelines, tools, and user-interfaces that facilitate a close integration between the computational analysis with the experimental application in a clinical setting.

We apply the presented methods to clinical data. The results show that a combination of high-throughput data, computational data analysis, and accurate prediction methods with clinical research can promote the development of new individualized treatment options for cancer.

Zusammenfassung

Die Fortschritte der letzten Jahrzehnte in der Krebsforschung haben zu einem deutlich verbesserten Verständnis der Ursachen und Entwicklung von Krebs geführt. Dieses Wissen konnte bisher allerdings nur in relativ geringem Maß in neue Therapieoptionen für Krebs umgesetzt werden. Eine Erklärung hierfür ist die große Heterogenität von Krebs in Bezug auf das Erscheinungsbild, den klinischen Verlauf und auf die dem Krebs zugrunde liegenden genetischen Veränderungen. Bei traditionellen Ansätzen der Wirkstoffentwicklung und der medikamentösen Therapie werden pathologisch ähnliche Krankheiten mit den gleichen Wirkstoffen behandelt. Diese Ansätze sind nicht ausreichend um der großen Komplexität von Krebs zu begegnen. Personalisierte oder individualisierte Ansätze, die gezielt individuelle Eigenschaften von Tumoren angreifen, sind dagegen ein vielversprechendes Konzept für die Entwicklung wirksamer und nebenwirkungsarmer Krebstherapien.

Der menschliche Organismus ist mit dem Immunsystem bereits mit einem System ausgestattet, das ist der Lage ist abnorme Zellen gezielt und effizient anzugreifen. Mit Hilfe von auf der Oberfläche von Körperzellen präsentierten Peptiden sind T-Zellen in der Lage, gesunde Zellen von entarteten zu unterscheiden. Welche Peptide dabei erkannt werden können unterscheidet sich von Patient zu Patient. Genetische Veränderungen in Krebszellen können zur Präsentation von krebsspezifischen Peptiden führen, die eine gezielte Immunantwort gegen die Krebszellen auslösen. Solche krebsspezifischen T-Zell-Epitope können in Form von epitopbasierten Impfstoffen zur Bekämpfung von Tumoren verwendet werden. Dieses Verfahren bietet einen Ansatzpunkt für die Entwicklung von maßgeschneiderten Immuntherapien.

Eine große Herausforderung bei der Entwicklung solcher individueller Therapieansätze ist die Auswertung genetischer Informationen von einzelnen Tumoren für die Detektion krebsspezifischer Mutationen. Eine weitere große Herausforderung ist die Identifikation von T-Zell-Epitopen, die durch diese Mutationen erzeugt werden. In dieser Arbeit stellen wir Algorithmen und Methoden zur Lösung dieser Herausforderungen vor. Ausgehend von Sequenzierungsdaten von Tumor- und Normalgewebe von einzelnen Patienten werden Mutationen identifiziert, die zwar im Tumor aber nicht im Normalgewebe vorkommen. Informationen über Genexpression, biologische Netzwerke und funktionelle Annotation von Genen und Proteinen werden in die Auswahl von Mutationen einbezogen, die als Angriffspunkt für eine Immuntherapie geeignet sind. Wir stellen Algorithmen zur Identifikation von T-Zell-Epitopen vor, die spezifisch für den Tumor und gleichzeitig auf Immunsystem des Patienten abgestimmt sind.

Damit computergestützte Methoden in der klinischen Forschung zum Einsatz kommen können müssen deren Ergebnisse zuverlässig und reproduzierbar sein und Kooperationspartnern in der Klinik zeitnah und verständlich zur Verfügung gestellt werden. Die Entwicklung von Analysepipelines und intuitiven Benutzeroberflächen, die eine enge Verknüpfung zwischen spezialisierten bioinformatischen Analysen und klinischer Forschung erleichtern, ist daher ein weiterer Schwerpunkt dieser Arbeit.

Die vorgestellten Methoden werden im zweiten Teil der Arbeit auf klinische Daten angewendet. Die Ergebnisse zeigen, dass die Kombination von Hochdurchsatzdaten, rechnergestützter Datenanalyse, zuverlässigen Vorhersagemethoden und klinischer Forschung einen wichtigen Beitrag zur Entwicklung individualisierter Krebstherapien leisten kann.

Acknowledgments

I want to thank my adviser Prof. Oliver Kohlbacher for guiding me through the exciting years of my PhD. Thank you for challenging me and encouraging me to follow my own path, but also leading me back to the right way when I seemed lost.

I thank Prof. Jürgen Bauer for reviewing this thesis and for the good collaboration.

Special thanks go to Pierre Dönnes, who waked my interest and enthusiasms for computational immunology many years ago and who guided my first steps in this area. I am glad that you joined the team again and grateful for all the fruitful discussions, your input, and for always having an open ear for me.

During my thesis, I had the chance to collaborate with many interesting and inspiring people who opened my eyes for the immunological, medical, or clinical aspects of our collaborative projects. Among those are Hans-Georg Rammensee, Stefan Stevanović, Jürgen Bauer, Moritz Menzel, Diana Meckbach, Karin Schilbach and many others. Thank you for the open-minded discussions and the good collaboration.

I am much obliged to Anne-Katrin Emde from the group of Prof. Knut Reinert at the FU Berlin who introduced me to and supported me in NGS data analysis.

Thanks go also to my current and former colleagues at the department of Applied Bioinformatics and the Center for Bioinformatics. I enjoyed working with you guys and am grateful for the friendly and pleasant atmosphere. The many chats, coffee breaks and discussions prevented me from loosing perspective numerous times. Nora, thank you for the good years together in the office. I also want to thank all students who I had the honor to accompany through their theses during the last years, especially Philipp Thiel, Benjamin Schubert, and Sebastian Boegel.

I would like to thank Kay Nieselt for her efforts in gender-related issues, her support and encouragement, many chats, and for being there whenever I needed a friendly advice. I also thank all *WSI-Forscherinnen* for the enjoyable and informative lunches.

I thank PhD comics, for giving me the feeling that I am not alone.

Special thanks go to my parents who always encouraged me and had no doubt that I will reach my goals. I cannot express how much I appreciate your empathy, your patience and your being there for me.

I thank all my friends for reminding me that there is a real world outside of science. Additional thanks go to Sascha, for his patient LaTeXsupport.

Most importantly, I want to thank Nico. Thank you for your support and care, for being with me, and for showing me what really matters in life.

To Andreas. To Jan.
Ihr fehlt.

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Chapter 1

Introduction

In 1971 the president of the United States Richard Nixon signed the National Cancer Act that was intended "to amend the Public Health Service Act so as to strengthen the National Cancer Institute in order to more effectively carry out the national effort against cancer" [1]. This law was enacted in a period of strong belief into technology: joint efforts of industrial and government-funded research had recently enabled men to land on the moon. Eradication of cancer as a major cause of death by joining all forces and promoting research seemed to be possible, and the National Cancer Act was often referred to as a declaration of war on cancer.

In the four decades that have passed since the National Cancer Act, immense progress has been made in terms of prevention and treatment, but most prominently in the understanding of cancer biology [2].

The first theories on carcinogenesis in the history of cancer research were viral infection, carcinogenous agents (tobacco smoke, radium) and inherited genetic alterations. It took a long time until these competing theories could be reconciled into one: cancer is a disease that is associated with variations in the genome. The genetic variations can be inherited, occur spontaneously, be induced by viral infections or by chemical agents which is in agreement with the three, previously competing, theories of cancer development.

Cancer is a clonally developing disease and cancer cells develop from normal cells by accumulating somatic mutations. Cancer cells are mutated versions of healthy self cells. They hijack pathways and mechanisms that are inherent to human life (cell growth, cell division, cell migration) and also abuse the perhaps most powerful mechanism of life: evolution. The evolution of cancer is even accelerated by the genomic instability of cancer cells. The clonal evolution of cancer cells enable the cancer to repeatedly evade recognition by the human immune system and chemotherapeutic anti-cancer treatment.

It has become apparent that cancer is not a single disease, but a collection of over 100 different and distinctive diseases with heterogeneous appearance, pathology and prognosis. The main common characteristics of cancers are uncontrolled cell growth, tissue invasion, and metastasis, the ability to spread to other tissues [3].

Cancer research was greatly promoted by the advent of new high-throughput technologies that allow the study of whole sets of genes, transcripts, proteins, or metabolites of cells. The progress in both, the availability of new high-throughput data and computational methods to analyze the data, have contributed to the major gain in the understanding of the mechanisms that underlie cancer development. Many different mutations in various genes have been identified in cancer. The genomic landscape of whole groups of cancer has been published [4]. Genes that are differentially expressed in cancer have been detected as well as whole deregulated cellular pathways.

It has also become apparent that not only cancer is a heterogeneous disease with respect to appearance, clinical outcome and the tissues it can affect, but also that clinically or pathologically similar cancers can be very diverse with respect to their genetic alterations. A histological, pathological, and clinical categorization of tumors has to be accompanied by a genetic profiling today. To make things even more complicated, single tumors can no longer be viewed as homogenous entities but, due to their clonal development, as a heterogeneous collections of cells that harbor different sets of genetic alterations. The picture we draw today is very complex, but still far from being complete.

Despite the huge progress made in understanding cancer biology, over 40 years after declaring the war on cancer, we are far from winning that war. Different forms of cancer are today the second most frequent cause of death overall. Cancer rates are rising, and cancer is predicted to become the number one killer worldwide in the near future. One out of three women and one out of two men will develop cancer in their lifetimes [5]. Cancer is also a huge economic burden. The global economic toll caused by cancer is at $ 895 billion annually [6], not accounting for the direct costs of cancer treatment. The goal of completely curing all cancers seems far out of reach today. Nevertheless, the incidence and economic burden of cancer underline the importance of continuing the fight against cancer. Perhaps not with the too ambitious goal of winning the war but with a new and more realistic goal of turning cancer into a controllable and chronic disease rather than a deadly one.

Due to their clonal development, cancers are able to develop resistance against chemotherapeutic treatments. To approach the goal of controlling cancer we therefore need to enlarge our arsenal to attack the cancer, i.e. to find new treatment options, and a way to select the right combination of treatments for each single tumor and patient. The strategy of treating each patient individually, based on its particular genetic dispositions, is termed personalized medicine. Individualized or personalized cancer treatment is a subset of personalized medicine and is a promising approach to develop new therapeutic options.

The classical treatment options for cancer are surgery, radiation and the administration of cytotoxic drugs, or a combination of those. It is not always possible to remove all tumor cells during surgery, remaining tumor cells can lead to a relapse. If the tumor has already spread into surrounding tissue, metastases can develop later on. Cytotoxic drugs target the mechanisms that are hijacked by cancer: cell growth and cell division. These mechanisms are highly active in cancer cells. But they are not specific for cancer cells, so normal cells are severely affected. The cytotoxic treatment thus has severe side effects and can be compared to weapons of mass destruction that cause immense collateral damage. The growing knowledge on cancer biology permits the identification of new targets for cancer treatment, preferably ones that are specifically present in the cancer but not in healthy cells. The detection of new cancer-specific targets increases the overall number of possible targets and thus treatment options while, at the same time, promises therapies with drastically reduced side effects.

One example for the successful development of targeted anti-cancer drugs is the kinase inhibitor Imatinib (Gleevec®, Novartis). Imatinib specifically inhibits a constitutively active tyrosine kinase, the product of a gene fusion (BCR-ABL), that occurs frequently in chronic myelogenous leukemia [7] and some other cancers. Trastuzumab (Herceptin®, Roche) is a second successful example for a targeted therapy. Trastuzumab is a monoclonal antibody that targets the HER2 receptor, a growth factor receptor that is found to be overexpressed in 30% of all breast cancers but also in other tumor types. Both are widely and successfully applied in the clinic today.

These two examples show that the identification of tumor-specific or tumor-associated targets can lead to targeted therapies. The advantage of these targeted therapies is that they are highly effective and have fewer side effects compared to classical cytotoxic treatments. Despite these success stories, the number of targeted therapies against cancer greatly lacks behind the theoretical knowledge on cancer biology. This can be attributed to two main factors. Targeted treatments are only effective if the target is present in a tumor and most approaches for targeted therapies rely on classical drug development.

Classical drug development is a lengthy and costly process and is therefore generally focused on targets that promise to be profitable, i.e. targets that are frequently found in patients. In order to increase the treatment options for cancer, however, we need to be able to also attack infrequent targets. In addition, not the primary tumors but metastatic disease is the major cause of death after cancer. Widening the focus of anti-cancer treatment from only targeting the growth of primary tumors to preventing metastasis would be of great benefit for a large number of patients [8].

Towards personalized immunotherapy

A promising strategy for both, targeting infrequent targets and prevention of metastasis is immunotherapy. The human immune system is specialized on specifically targeting abnormal cells. Indeed, the immune system constantly protects us against newly arising cancer cells, and a cancer can only develop if the immune system fails to detect and eliminate these aberrant cells. The power of the immune system in the control of tumors is demonstrated by the increased cancer incidence in immunocompromised and elderly people. Strengthening the immune system to attack cancer cells is a promising treatment option [9]. A well-established method to direct immune response towards specific targets is vaccination. Vaccines cannot only be used to teach the immune system to protect us from pathogenic infections, but also to train the immune system to recognize tumor cells. The main feature of the immune system is the discrimination between self and non-self. The first step towards an anti-cancer vaccine is thus the identification of properties that distinguish the tumor from the normal tissue of the patient. New concepts of vaccine design, namely epitope-based vaccines, make it possible to directly target well defined structures, the T-cell epitopes. A single point mutation in a coding region of a protein can lead to the presentation of a cancer-specific T-cell epitope on the tumor cells and thus enable the immune system to distinguish the tumor cells from healthy cells. Such tumor-specific T-cell epitopes can be applied in the form of epitope-based vaccines to train the immune system to react specifically to these epitopes and thus against the tumor cells that present these epitopes. A promising scenario of application for such vaccines is the prevention of a relapse or metastasis after surgical removal of a tumor.

The identification of tumor-specific T-cell epitopes involves two main steps. The first step is the identification of tumor-specific mutations. This can be accomplished by a genetic profiling of single tumors and by comparing these genetic profiles with healthy cells of the same patient. In the next step, those tumor-specific mutations have to be selected that are likely to elicit an immune response in the respective patient. Both steps are challenging problems in terms of the experiments that have to be performed, as well as the computational methods that are needed to analyze the experimental data.

Next-generation sequencing (NGS) is the method of choice for genetic profiling. It produces large amounts of sequence data from genomes, exomes, or transcriptomes. The data produced form NGS instruments are millions of short reads per sample that represent short stretches of DNA (or mRNA in case of transcriptome sequencing). In oder to gain information on the genetic profile of the sample the reads have to be mapped onto and compared to a reference genome. When investigating cancer genomes, the data has to be compared to the sequencing data from healthy cells of the same patient in order to distinguish somatic cancer mutations from normal genetic variation between individuals. Sophisticated algorithms and efficient software are indispensable to perform read mapping and variant detection. Analysis of NGS data has been a main field of research in computational biology during the last years, but there are still many open questions in the analysis of sequencing data from cancer samples.

The identification of suitable targets for cancer immunotherapy from tumor-specific mutations involves several tasks. Only mutations that lead to alterations in a protein can lead to tumor-specific epitopes. Whether a tumor-specific peptide can function as an epitope depends on the immune system of the respective patient. In order to go from a list of tumor-specific mutations to candidate T-cell epitopes two main steps have to accomplished. First, computational methods that assess the influence of genetic mutations on the corresponding proteins are needed. Second, we need computational prediction methods that predict T-cell epitopes from mutated proteins.

Genetic profiling and the identification of potential T-cell epitopes greatly depend on the availability of accurate and appropriate computational methods. The development of such computational methods and their application to clinical data offers completely new roads for the development of cancer immunotherapies. The computational analysis has to be closely integrated with clinical and biological research. Beside the development of new computational methods, promoting the collaboration between biological cancer research, computational biology, clinical research and the clinical application is therefore an important task in computational biology.

The computational methods presented in this thesis contribute to these main issues in the development of cancer immunotherapies in the following areas: 1) genetic profiling of cancers, 2) identification of targets for personalized immunotherapy, and 3) building a bridge between computational methods and clinical research.

Genetic profiling of cancers. We present approaches towards the genetic profiling of single tumors based on NGS data. The genetic profiling of cancers allows to gain insight to the genetic alterations that are responsible for the development of a tumor, but also to select the right treatment in the presence of molecular targets for existing therapies. We present methods to detect and thoroughly analyze and interpret genetic variations with a focus on viral integration, single nucleotide variants, and short insertions and deletions. A special focus is on the identification of tumor-specific alteration which can only be detected with respect to data from healthy tissue from the same patient.

Detection of targets for personalized immunotherapy. The tumor-specific genetic alterations obtained from genetic profiling, somatic mutations as well as viral sequences that are integrated in the genome of a cancer cell, are the basis for the identification of targets for epitope-based anti-tumor immunotherapies. We present accurate methods to predict two major steps that contribute to the generation of cytotoxic T-cell epitopes: HLA binding and T-cell reactivity. We integrate these prediction methods with genetic profiling. Together with immunotyping information of a patient, this allows for the identification of potential T-cell epitopes that are specific with respect to the tumor's somatic mutations and the patient's immune system. These epitopes can be administered in the form of epitope-based vaccines to induce specific anti-tumor immune reactions.

Building a bridge between computational methods and clinical research. The results of such computational methods need to be presented to biomedical researchers in a convenient and comprehensive manner. The impact of computational methods is limited if they are not used in biological research. A major focus of this thesis is thus a close and bidirectional collaboration between computational biology and biomedical and/or clinical research. We develop pipelines that integrate all steps, from NGS data analysis to an application of immunoinformatics prediction methods. These pipelines allow a reliable and reproducible processing of patient-related data in a timely manner. The pipelines are coupled to user interfaces that allow a direct and interactive presentation of the results to our partners in the clinic. We show how such pipelines can enable completely new treatment strategies. The pipelines are based on a very flexible workflow system to allow for reproducible results on the one hand, and, on the other hand, for the continuous improvement of the analysis pipelines when results from first clinical applications demand it.

Summing up, the methods and results presented in this thesis show how a combination of computational data analysis, accurate prediction methods, and clinical research can promote the development of new treatment options for cancer.

This thesis is structured as follows: In Part I, background information is presented that is needed to understand the methods and results presented later in this thesis. Chapter 2 focuses on the foundations of cancer biology, cancer immunology and cancer treatment. Chapter 3 introduces the high-throughput experimental methods that can be used to identify genetic variation in cancer, along with the state-of-the art analysis of the respective high-throughput data with a focus on next-generation sequencing data.

The second part presents the computational methods that were developed in this thesis. This part is structured according to the three main areas that are addressed in this thesis: In Chapter 4 we present methods for the integrated analysis of next-generation sequencing data with the aim of identifying and annotating genetic variations in single cancer genomes or transcriptomes. In Chapter 5 computational prediction methods for the identification of T-cell epitopes are presented, as well as a procedure to apply those to genetic variation obtained from genetic profiling. In Chapter 6, workflows are presented that allow the application of the methods presented above to clinical data in a timely manner. We also present a user interface that allows biomedical researchers to conveniently access and interpret the results of the computational analyses.

In Part III we demonstrate the application of the computational methods and workflows to clinical data. In Chapter 7 we analyze transcriptome sequencing data from 10 melanoma metastases for viral integration and short variations. The single nucleotide variants that we find in these melanomas are thoroughly analyzed and annotated. We identify tumor-specific mutations and analyze those for recurrence with respect to specific mutations, affected genes and pathways. In Chapter 8 we demonstrate how the large-scale application of computational prediction methods can be used to identify promising targets for graft-versus-leukemia reactions after stem cell transplantation. The use of donor-derived T-cells that target residual tumor cells in the patient is a promising new treatment modality for relapse in hematologic malignancies.

In the last part, Part IV, the methods and results presented throughout this thesis are discussed and the perspective of their application in clinical research is outlined.

Part I

Biological and Experimental Background

This part provides background information that is needed throughout this thesis. We first introduce the current state-of-the-research of cancer, cancer genomics and cancer immunology. It also presents classical approaches to cancer treatment and introduces the concept of targeted and personalized treatments. Chapter 3 describes high-throughput experimental methods that can be used to identify genetic variation in cancer, along with the state-of-the art of the analysis of the respective high-throughput data. The focus here lies in the analysis of next-generation sequencing data.

Chapter 2

Cancer

This chapter will give a short introduction to the current state-of-research of cancer, cancer genomics, and cancer immunology. Furthermore, the current status of cancer treatment, as well as innovative or personalized approaches to fight cancer and cancer immunotherapy are introduced. Cancer genomics and cancer biology are complex, vivid and evolving fields. A complete review of these fields goes beyond the scope of this thesis. We therefore focus on those parts that form the basis for understanding the computational approaches towards personalized cancer treatment that are presented in this thesis.

Cancer, medically also termed malignant neoplasm or malignancy, is not a single disease but a large group of (over 100) different and distinctive diseases. The appearance, pathology and prognosis of cancers is very heterogeneous. The main common characteristics of cancers are uncontrolled cell growth, tissue invasion, and metastasis, the ability to spread to other tissues [3]. These three characteristics distinguish cancer from benign neoplasms or benign tumors. Enormous effort was made during the last decades regarding the understanding and treatment of cancer, however, cancer is still the second most prominent cause of death in industrialized countries after cardiovascular diseases.

Cancer is caused by the accumulation of genomic alterations, and is therefore sometimes referred to as a genetic disease [10]. The genetic alterations can be triggered by environmental factors like exposure to tobacco smoke, sun light or viral infections. Alterations can be inherited but can also occur spontaneously. The genetic variations observed in cancer are diverse and range from mutations of single bases to translocations of whole chromosome arms [3]. The effects of these mutations on the biology and physiology of cancer cells are as diverse as the observed genetic alterations. Some alterations frequently occur in different types of cancers, others are strongly associated with special types of cancer (e.g. the *Philadelphia Chromosome* in chronic myelogenous leukemia [7] or the V600E mutation of BRAF in malignant melanoma [11]), other alterations are only found in single tumors or small subgroups of cancer patients. The current understanding is that cancer development is a multi-step process caused and accompanied by the accumulation of different genomic alterations.

In 2000, Hanahan and Weinberg proposed six hallmarks of cancer as biological capabilities that are acquired during the multi-step development of human tumors [12, 13]. The hallmark concept shifts the focus from the wide variety of single observed cancer mutations to the biological and physiological effects of these mutations. The causes of the hallmarks are still genetic, but the hallmarks regard cancer from a more phenomenological than mechanistic point of view. The proposed hallmarks are *sustaining proliferative signaling ,evading growth suppressors*, *activating invasion and metastasis*, *enabling replicative immortality* , *inducing angiogenesis*, and *resisting cell death*. In 2011, Hanahan and Weinberg added two emerging hallmarks, namely *avoiding immune destruction* and *deregulation cellular energetics* and two enabling characteristics, *tumor-promoting inflammation* and *genome instability*. These enabling characteristics underlie and foster the hallmarks of cancer. The concept of these hallmark capabilities implies that the acquisition of the hallmark properties is essential but specific genome changes are not. Or - as long as a specific hallmark property is maintained - it is irrelevant which gene is affected by mutations. Mutations in different genes can lead to the acquisition of the same hallmark capability. On the one hand, the hallmark concept implies that trivial analyses will fall short of understanding cancer, since it requires to look beyond single mutations or genes. Whole pathways, the interaction between pathways, and biological processes have to be investigated. On the other hand, the hallmark concept offers the chance to understand the basic mechanisms that underlie cancer development without getting lost in the sheer multitude and complexity of the observed mutations. Detecting genomic alterations in different cancers and combining the results for different patients and cancers is still the basic step of an in-depth analysis of the biological processes altered in cancers.

The classes of genetic alterations observed in cancer are described in more detail in the next section. Not all mutations observed in cancer are actually responsible for or involved in cancer development. To distinguish those mutations that drive or promote cancer, so-called driver mutations, from mutations that are not relevant for cancer development, so-called passenger mutations, is one of the key questions in cancer genomics and cancer biology.

2.1 Cancer genomics

The idea that the abnormal proliferation of cancer cells results from derangement in the genome was already proposed at the beginning of the 20^{th} century [14]. Decades of research and the advent of high-throughput technologies, such as microarrays and genome sequencing, led to the identification of many different genetic alterations in cancers. These alterations can be grouped into short genetic variations and structural variants. Cancer can also be caused by viral infection, which can lead to viral integration into the host genome. Cancer genomics is a very complex topic and has been reviewed in detail elsewhere [15, 16, 17]. The genetic alterations that are relevant for this thesis are described shortly in this section.

Short genetic variations. Short genetic variations comprise single nucleotide variants (SNVs) and small insertions and deletions (INDELs).

SNVs are point mutations, where one nucleotide is exchanged by another. Besides SNV, The term single nucleotide polymorphism (SNP) is widely but not consistently used in the scientific community. The most common definition is that the less frequent version of a polymorphism has to occur with a minimum frequency of 1 % of a given population [18]. For the investigation of single (cancer) genomes, a term is needed to describe point mutations without information about its frequency in a population. We will therefore use the term SNV to refer to a single nucleotide change, independent of its frequency in a population. INDELs are short insertions and deletions, where, compared to a reference sequence, new nucleotides are inserted or nucleotides are deleted from a genomic sequence.

The impact of a SNV or INDEL depends on its genomic context. SNVs in coding regions of genes have the potential to directly influence the corresponding protein sequence. For synonymous or silent SNVs the change in the nucleotide sequence does not result in a change in the protein sequence due to the degeneracy of the genetic code. Non-synonymous mutations, also called replacement mutations, lead to an alteration in the protein sequence. A non-synonymous substitution can either be missense, where one amino acid is exchanged by another, or nonsense, where a newly introduced stop codon leads to an earlier stop in the protein sequence. In nonsense mutations, the function of the respective protein is typically lost. The functional impact of missense mutations is more diverse: some missense mutations do not or barely influence the biological function of the protein, whereas other single amino acid substitutions can have a severe effect of regulatory pathways. The V600E mutation in BRAF for example mimics the phosphorylation of T599 and/or S602 in the activation segment and so BRAF stays constitutively active independent of its normal regulators [11, 19]. The resulting deregulation of the MAPK pathway effects, amongst others, cell division and differentiation. BRAF with V600E mutation is an important oncogene in malignant melanoma and some other cancers. INDELs in coding regions, unless the length of the INDEL is a multiple of three where the mutation leads to the insertion or deletion of some amino acids, lead to frameshift mutations. Frameshift mutations alter the whole protein sequence downstream of the mutation.

Mutations in non-coding regions do not have a direct influence on the protein sequence. They can be intronic (in non-coding regions of genes) or intergenic. Mutations in non-coding regions can still have an effect on gene splicing, transcription factor binding, messenger RNA degradation, or the sequence of non-coding RNAs. They can thereby have an effect on gene expression or splice variant expression.

Structural variants. In this thesis, structural variants (SVs) are defined as genomic rearrangements that affect 50+ bp of sequence. This group comprises larger deletions, duplications, novel insertions, inversions and translocations. SVs, particularly copy number variations (CNVs) and translocations, are associated with carcinogenesis [20, 15, 16, 17]. CNVs are imbalanced structural variants that lead to an altered number of copies of DNA segments (gains if the number of copies is increased, losses if the number is decreased). An altered copy number can result in the down- or upregulation of the corresponding genes, and thus in the deregulation of biological processes. Translocations can result in fusion genes, if the breakpoints lie in coding regions. The most prominent example of a gene fusion with carcinogenic effect is the *Philadelphia Chromosome* in chronic myelogenous leukemia [7], where the Abl1 gene on chromosome 9 is joined with a part of the BCR gene on chromosome 22. This translocation leads to a constitutively active BCR-Abl fusion transcript that disturbs cell cycle control and DNA repair [20].

Viral infection as cause for carcinogenesis. Infectious agents (viruses and bacteria) can presently be linked to about 20% of the global cancer incidence. One of the most prominent examples of a virus causing cancer in humans is the human papilloma virus (HPV) in cervical cancer [21]. There are many reasons that make it difficult to identify viral infections as causative factor in cancers. Amongst these is the fact that viral infection can contribute to carcinogenesis by indirect factors as virus-induced immunosuppression that activates other tumorviruses (HIV-1 and HIV-2), chronic inflammation (hepatitis virus B and C), prevention of apoptosis, and the induction of chromosomal instability and translocations. In this thesis we restrict the search for viral contribution to cancer development to two direct mechanisms of viral carcinogenesis: the introduction of viral oncogenes into host cells (as observed for high risk HPV) and the presence of modified viral oncogenes after integration into host cell DNA (as observed for Merkel cell polyomavirus [22]). In these two cases, viral sequences can be detected in the genome or transcriptome of the cancer.

2.2 Cancer immunology

2.2.1 The human immune system

The immune system is a highly complex network of tissues, cells, and organs that protects the body against pathogens and aberrant cells. The innate immune system as a first line of defense recognizes conserved and unspecific pathogenic structures like glycosylation patterns of bacterial cell walls. The adaptive immune system is responsible for the recognition of highly specific pathogenic structures and invokes specific and narrowly directed responses. The adaptive immune system is also able to establish an immunologic memory that leads to a life-long immunity against pathogens that were once encountered and successfully eliminated. In order to recognize and eradicate pathogens and abnormal cells the immune system must discriminate between self (normal and healthy structures of the body) and non-self (pathogens, but also virus infected or aberrant cells). B lymphocytes (B cells) and T lymphocytes (T cells) are the key players of the adaptive immune system. Specialized receptors on the surface of B and T cells achieve the self:non-self discrimination. Via a stochastic process, called somatic recombination, a large number of B and T cells is generated, each equipped with different highly specific receptors. Substances that are recognized by the adaptive immune system are called antigens. Lymphocyte receptors do not recognize whole antigens, but only small regions of the antigen, the so called epitopes.

The adaptive immune system can be divided into two main parts, the humoral and the cellular responses. The humoral immune response, mediated by B cells, results in the secretion of antibodies that neutralize and mark pathogens for destruction. The cellular immune response, mediated by T cells, is responsible for two main tasks: 1) the recognition and elimination of virus infected or cancer cells by cytotoxic T cells (CTLs) , and 2) the regulation of the adaptive immune response by various types of T helper cells.

Further details on the concepts presented in the following can be found in immunological textbooks (e.g. [23, 24]).

T-cell responses

T cells only bind parts of antigens (peptides) that are presented by so-called major histocompatibility complex (MHC) molecules on the surface of other cells. The cellular immune response is very complex and tightly regulated. There are three necessary conditions for a peptide to be a T cell epitope:

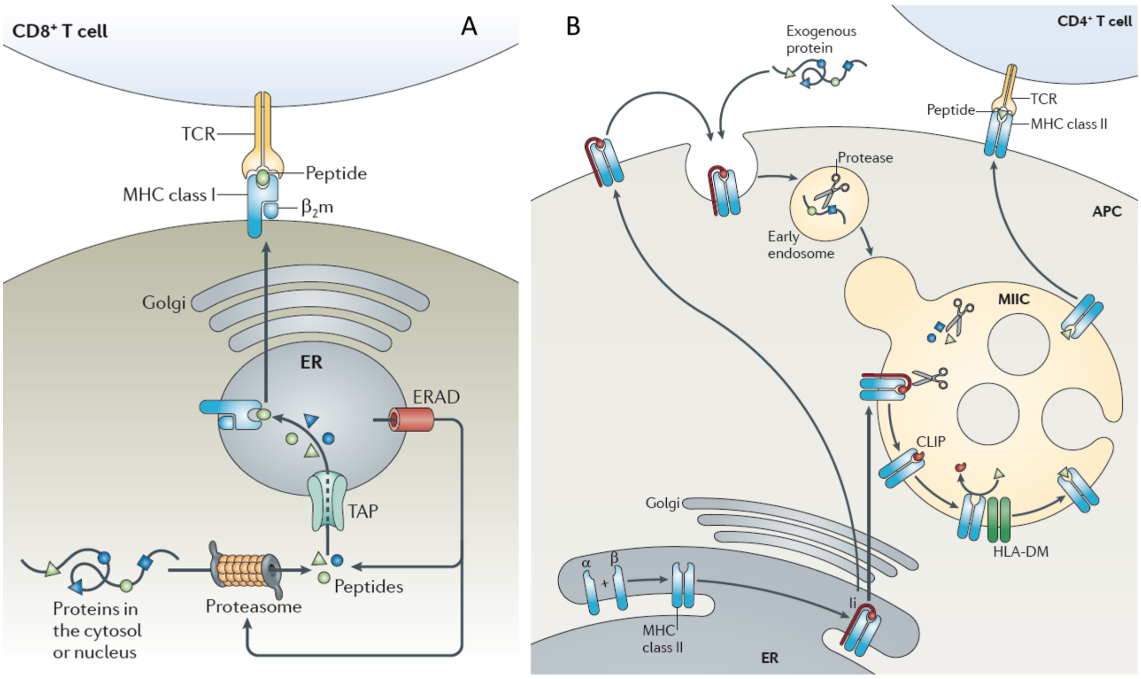


Fig. 2.1: Antigen processing. A: The endogenous pathway generates peptides from cytosolic proteins for the presentation by MHC class I. B: The exogenous pathway generates peptides from extracellular proteins for the presentation by MHC class II. Both figures are taken from [25], reprinted with permission.

1. The peptide has to be produced by the antigen processing machinery
2. The peptide has to bind to an MHC molecule and be presented on the cell surface
3. The T-cell repertoire of the individual has to contain a T cell with a T-cell receptor that matches the peptide:MHC complex.

Antigen processing. Two processes generate peptides for antigen presentation, the endogenous and the exogenous pathway. The antigen processing pathways are depicted in Fig. 2.1. In every cell intracellular proteins are degraded by the proteasome as a consequence of normal protein turnover. The length of the peptides produced by the proteasome varies between three and 30. Peptides of the appropriate length (around 9 to 15) can be transported into the endoplasmatic reticulum by a specialized transporter called TAP, where they can bind to MHC class I molecules. This process is called the endogenous antigen processing pathway. In the exogenous pathway extracellular proteins are ingested by special antigen presenting cells (mainly B cells, dendritic cells, and macrophages) and then degraded in endosomes. After the fusion with MHC class II containing vesicles, the peptides derived from extracellular proteins can be bound and presented by MHC class II.

MHC binding. MHC, the major histocompatibility complex, also called HLA for human leukocyte antigens in humans, is a genetic locus on chromosome 6 that encodes for key players of the immune system. Among those are the MHC class I and MHC class II molecules. MHC molecules are glycoproteins that present short antigenic peptides on the surface of cells to cytotoxic T lymphocytes. MHC class II is expressed on specialized immune cells, the antigen-presenting cells, and presents peptide fragments from the exogenous pathway to T helper cells. MHC class I is expressed in all nucleated cells and presents fragments of intracellular proteins to CTLs. The peptides bound to the MHC class I molecules on the cell surface represent a fingerprint of the proteins in the cell. Virus-infected and cancerous cells present non-self peptides on their surface and can thus be detected by CTLs. The MHC locus is polygenic and polymorphic, over 7.000 [26] different versions, or alleles, of MHC class I and class II molecules are known. The peptide binding repertoire depends on the amino acid sequence of the MHC molecule. Each version or allele of the MHC molecules binds a different set of peptides.

The peptide binding mode differs between MHC class I and MHC class II. MHC class I has a peptide binding groove that is closed at both ends. The length of the bound peptides is restricted to 8 to 12 amino acids, but the main part of the peptide bound to MHC class I is of length 9. The binding groove of MHC class II is open-ended and fits longer peptides of up to 20 amino acids.

T-cell receptor repertoire. A large variety of T-cell receptors is generated through a stochastic combinatorial step, the so called somatic recombination. To ensure a functional and non-autoreactive T-cell repertoire the T cells undergo a two-step selection process. Only T cells with receptors that can bind to MHC:peptide complexes get a survival signal (positive selection) and T cells that bind to self peptides presented by MHC are deleted from the repertoire (negative selection). The presentation of an antigenic peptide and the availability of a matching T-cell receptor are a prerequisite for a T-cell reaction. Other factors and mechanisms are involved in the decision if a T cell gets activated and elicits an immune response against the peptide. A T-cell reaction can be downregulated by regulatory T cells, or, in the absence of co-stimulatory factors like inflammation, T cells are directed towards anergy or deletion.

2.2.2 Cancer and the immune system

The immune system interacts closely with a tumor. The term immunoediting was introduced to describe the complex interplay of the immune system and a developing tumor[27, 10]. Immunoediting is viewed as a dynamic evolutionary process composed of three phases, elimination, equilibrium, and escape. Immunoediting is of dual nature since it comprises host-protecting as well as tumor-promoting effects. In the elimination phase, the immune system is able to identify and eliminate tumor cells specifically, based on their expression of tumor-specific antigens. In many cases, the immune reaction is able to fully destroy the cancerous cells. Cancerous cells undergo stochastic genetic alterations which, eventually, allow the cells to evade the immune response. If not all tumor cell variants are fully destroyed, they can enter an equilibrium phase where the immune system and the tumor cells are in a dynamic balance. An effective anti-tumor reaction of the immune system is still present and the net tumor growth is controlled. During the elimination and equilibrium phase the tumor is not clinically apparent. As a result of changes in the tumor cell population or in the host's immune system, some cancer cell clones can progress into the escape phase. These cell clones can then proliferate without being restricted by the immune system. The escape phase thus represents the failure of the immune system to control this specific tumor. Various complex mechanisms to evade immune surveillance of cancer cells are known, the most prominent ones are the down regulation of antigen processing and presentation (MHC), the secretion of factors that inhibit the immune response, or the recruitment of regulatory cells that downregulate an immune response. Re-enabling the immune system to fight a tumor in the escape phase is the major goal of active anti-cancer immunotherapies (see Section 2.3.1).

2.3 Cancer treatment

The standard treatment for cancers consists of surgery, radiation therapy, chemotherapy, or a combination of these. The aim of surgery is the complete removal of all tumor cells of a solid tumor. In theory, a complete removal of all tumor cells is a cure of the cancer. However, not all tumors can be removed without damaging important organs (e.g., tumors in the brain), and often some tumor cells remain in the body. Therefore surgery is often combined with radiation or chemotherapy to destroy remaining or metastatic cancer cells. Chemotherapy refers to the administration of cytotoxic drugs. These drugs kill or damage rapidly dividing cells and can thereby cause severe collateral damage in healthy tissues.

The targets of cytotoxic drugs are processes that are present in all rapidly dividing cells, not only in cancer cells. Targeting cancer-specific molecules or processes instead of general processes is one way to reduce the risk of severe side effects. The targets of choice for anti-cancer drugs are thus molecules or pathways that are specific to the cancer in the sense that they only occur in cancer cells. Since cancer cells evolve from normal cells, it is difficult to find such specific targets. More realistic are targets that are present in cancer cells but less abundant in normal cells, so-called cancer-associated targets. The rapidly growing knowledge on tumor biology and cancer genomics allows to mine for cancer-specific or cancer-associated targets and to develop so-called targeted therapies.

The application of targeted anti-cancer drugs is a very promising concept, however, the success of these new drugs lacks behind their theoretical potential. In many studies the new promising drugs only marginally improve the outcome of the treatment [28]. In order to be effective, the respective target has to be present in cancer cells. However, standard techniques for tumor classification like histology are not able to validate the presence of molecular targets like mutated kinases. In addition, the simple presence of a target does not necessarily imply efficacy of the drug. Tumor cells can acquire additional mutations that render the drug ineffective or even harmful. In addition, a tumor is not composed of homogenous cells but of heterogeneous cell populations [29]. Genetic profiling and subtyping of tumors is required, and the genetic profiles have to be correlated with drug response in order to distinguish groups of patients that benefit, not benefit or are potentially harmed by a targeted therapy. The immense advances in technologies to genetically profile tumors made in the last decade, together with advanced computational methods (see Chapter 3) will promote the development and hopefully the clinical success of targeted therapies in the near future.

2.3.1 Cancer immunotherapy

The idea of cancer immunotherapy is to manipulate the immune system of a cancer patient to enable it to eliminate already established cancers. Two ways of manipulation of the immune system are possible: passive and active immunization strategies.

Passive immunization

Passive immunization can be applied under the assumption that the patient's immune system is incapable of fighting the cancer by itself, even after the administration of immuno-stimulatory therapies. The patient is therefore supplied with immune products (e.g., antibodies) or immune effector cells that originate from a different organism. Therapeutic antibodies are one strategy in passive immunization and are widely used in the clinic today. A prominent example for therapeutic antibodies in cancer is Trastuzumab (Herceptin®, Roche), a monoclonal antibody against HER2, a cell surface receptor that is overexpressed in some cancers. Another passive immunization strategy is the transfer of a donor immune system. The graft-versus-tumor response was observed in bone marrow transplanted patients. The donor immunocytes are able to detect and destroy residual tumor cells that have survived radiation and chemotherapy. The graft-versus-tumor effect is mediated by minor histocompatibility antigens (miHAs). miHAs are peptides that are HLA-restricted peptides containing a SNP. If donor and recipient differ in a position they present different peptides. If the peptide presented in the recipient is not present in the donor, the T-cell repertoire of the donor can contain T cells that are reactive for that peptide. If these T cells are transferred to the donor they can attack cells that present the respective peptide.

Active immunization strategies

In contrast to passive immunization strategies, active strategies that aim at enabling or enhancing the endogenous anti-tumor immune response, rely on the conviction that the immune system is still capable of attacking and eliminating the tumor. The aim of these strategies is to elicit effective anti-tumor response by increasing the number and activity of cytotoxic immune cells. These immune-activating approaches are often termed anti-cancer “vaccines”, a term that can be misleading because vaccines are traditionally used to prevent and not to treat diseases. There is indeed an example of a classical vaccine that is applied to prevent cancer, the vaccine against the human papilloma virus. The human papilloma virus is strongly associated with the development of cervical cancer. The direct aim of the vaccine is however to prevent viral infection, and only by this indirect means preventing the development of cancer. In the following, the term anti-cancer vaccine will be used for approaches to treat an existing cancer with immunotherapy by directing an immune response against the cancer itself.

The general idea in the active immunization or vaccination approaches is to direct an immune response against tumor-specific antigens (TSA) or tumor-associated antigens (TAA). These are proteins that are affected by cancer-specific mutations or are overexpressed in a tumor, respectively.

Epitope-based approaches for cancer immunotherapy. Using only the smallest immunogenic regions of antigens, the epitopes, is a relatively new concept in vaccine design. One advantage of this concept is that the immune response can be specifically directed against highly immunogenic structures of antigens and that they can be tailored to the patient’s immune system. Epitope-based vaccines offer great opportunities for personalized immunotherapies against cancer. The selection of epitopes to be used should be based on the patient’s immune system and on the mutations specific for the cancer to treat. The concept of epitope-based vaccines was proven to be effective in phase I and phase II clinical trials [30, 31, 32, 33, 34]. There are, however, some problems with epitope-based vaccines. The first main issue is the administration of the vaccine. Different approaches for delivery of the vaccine exist (e.g., DNA vaccines or peptide-based vaccines [35]), but the capability of these formulations to elicit an immune response is limited since the epitopes are taken out of their normal context in the protein. A second issue, when applied as anti-tumor immunotherapy, is that immunologic tolerance has to be overcome before an effective anti-tumor reaction can be re-established. A problem with all personalized approaches is the approval of the potential therapy. Standard therapies have to be proven to be effective and safe in clinical trials with large patient cohorts. For therapies that are tailored to one patient's tumor and immune system new concepts for approval have to be established. If an epitope-based vaccine is designed for one patient based on the patient's tumor and MHC allele combination, there is no way to test the vaccine on different patients. If a vaccine is designed for a tumor type and larger groups of patients, the patients for clinical trials and for treatment have to be carefully selected (based on the presence of the mutated epitopes and on the related MHC alleles). Large efforts and advances have been made addressing these issues offering the perspective that this problem can be overcome in the future [36, 37].

Chapter 3

High-throughput experimental methods to detect genetic variation

The detection of genetic variation in cancer genomes is a key step in promoting the understanding of cancer. In the last decade experimental methods have been developed that allow high-throughput genetic profiling of cancers. Biological studies that are applied on a large or genome-wide scale are commonly characterized with the suffix *-omics*. Genomics is the study of genomes, transcriptomics refers to the investigation of whole transcriptomes. The most prominent technologies in genomics and transcriptomics are sequencing and array-based methods to detect point mutations, gene expression, or copy number variations. This chapter gives a short introduction to these technologies. The focus is not on the technical details, but on the type of data generated and the computational approaches and challenges associated with data analysis. The main part of this thesis deals with the analysis of NGS data, data from array technologies are only used as additional information. Array technologies are therefore only described briefly.

3.1 Next-generation sequencing

The history of DNA sequencing begins in 1977 when Sanger *et al.* introduced their dideoxy method [38]. Two big milestones in the sequencing history, the completion of cellular genomes [39, 40] in 1995 and of the human genome in 2001 [41, 42], were achieved with automated Sanger-based capillary sequencing [43]. Over 10 years and $ 3 billion were necessary to sequence the first human genome. With the advent of new sequencing technologies the sequencing costs began to drop drastically. The new massively parallel

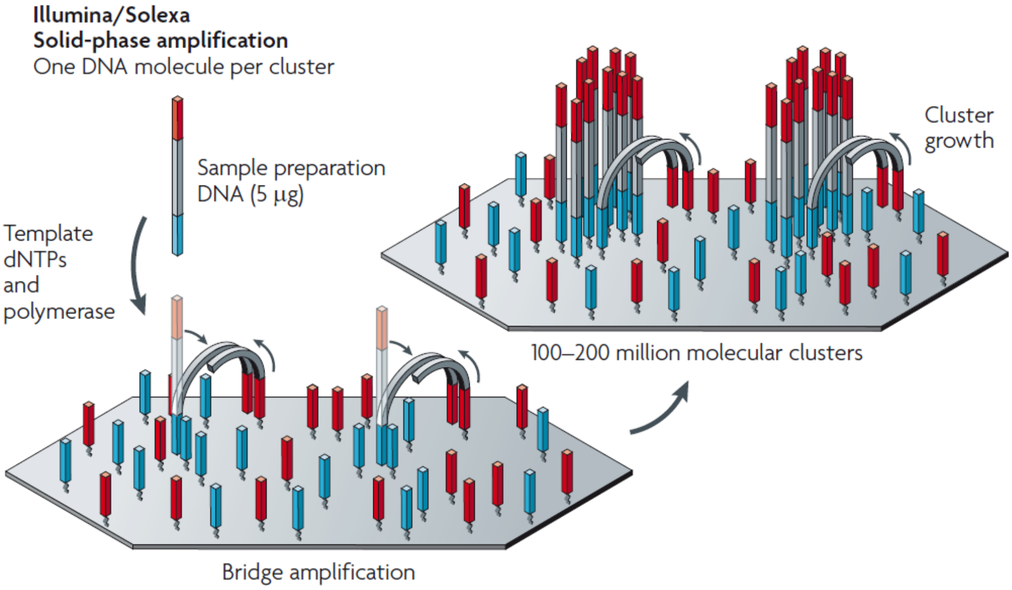


Fig. 3.1: Solid-phase template amplification for sequencing with the Illumina Genome Analyzer. The figure is based on [44]. Reprinted with permission.

sequencing technologies, also termed next-generation sequencing technologies (NGS), allow for cheap and quick production of enormous amounts of sequencing data. The sequencing of a personalized genome in 2011 takes about 10 days at the cost of $10,000 and the "$1,000 genome" is a realistic goal for the near future.

NGS technologies use various strategies for template generation, sequencing and imaging. The specific combination of protocols distinguishes the sequencing platforms and determines the type of data produced. The platform of choice for a study depends on the specific requirements regarding read length, sequencing depth, and error rate. The most widely used platforms are from Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences. A detailed technical description and comparison of the current NGS platforms is beyond the scope of this thesis and available elsewhere [44]. Sequencing with an NGS instrument is here described for the Illumina *Genome Analyzer IIx* (GA), since the data presented in this thesis was produced on this instrument.

The first step in the sequencing process is template preparation. Since the GA's imaging system is not designed to detect fluorescent events from single reactions, the templates are clonally amplified using solid-phase amplification. The amplification step produces 100-200 million spatially separated template clusters. The template amplification process is depicted in Fig. 3.1. The next step, sequencing and imaging, is based on *cyclic reversible termination* (CRT). CRT is a cyclic method comprised of four steps in each cycle, see Fig. 3.2 (a):

1. Fluorescently labeled nucleotides with reversible 3′-terminators are simultaneously added to the plate. Due to the 3′ terminators, only one nucleotide is incorporated, matching the template sequence.

2. Excess reagents are washed away.

3. The identity of the base that was incorporated in this cycle is detected using four color imaging (Fig. 3.2 (b)). The observed signal is a consensus of the nucleotides added to the identical templates in a given cycle.

4. The fluorescent dye and the terminators are cleaved and washed away and the process is repeated.

The slide is partitioned into eight channels allowing the simultaneous run of up to eight independent samples. The Illumina GA produces reads of length 36 to 100 bp. The most frequent error type of the GA are substitutions. The common error types have to be taken into account in data analysis.

Applications of NGS. NGS technologies produce large amounts of relatively low cost sequencing data and are therefore useful for many applications. NGS is broadly applied for resequencing projects, where a genome or parts of a genome are sequenced from an organism with known reference genome. The aim of resequencing projects is to identify variations between individuals of the same species, or, in the case of cancer genomics, differences between the genome of normal tissue and cancer cells. Resequencing projects can be applied to whole genomes or only to regions of interest, e.g. to all exons (the exome) of an individual. In order to identify variations between genomes of the same species, the sequenced reads are aligned and compared to a reference genome. For details on the algorithms and challenges for read mapping and variation detection see Section 3.2. Several large projects aim at sequencing many individuals in order to identify rare sequence variants in normal genomes (e.g., the 1,000 Genomes Project [45]) or genetic variations that are associated with major cancers (e.g., the Cancer Genome Atlas, `http://cancergenome.nih.gov`). If NGS technologies are used to sequence RNA (RNA-seq) an identification and quantitation of the transcripts of cells can by achieved. In metagenomics NGS is used to sequence an ensemble of different genomes from an environmental sample. In a later step one tries to assign the produced reads to different species to gain information of the composition of the sample. Other applications are *de novo* assemblies of smaller bacteria or lower eukaryotic organisms. Due to the short read length not all NGS technologies are suited for the assembly of larger genomes. The length of the reads produced by NGS platforms will increase in the foreseeable future, and thus open new opportunities for the application of NGS.

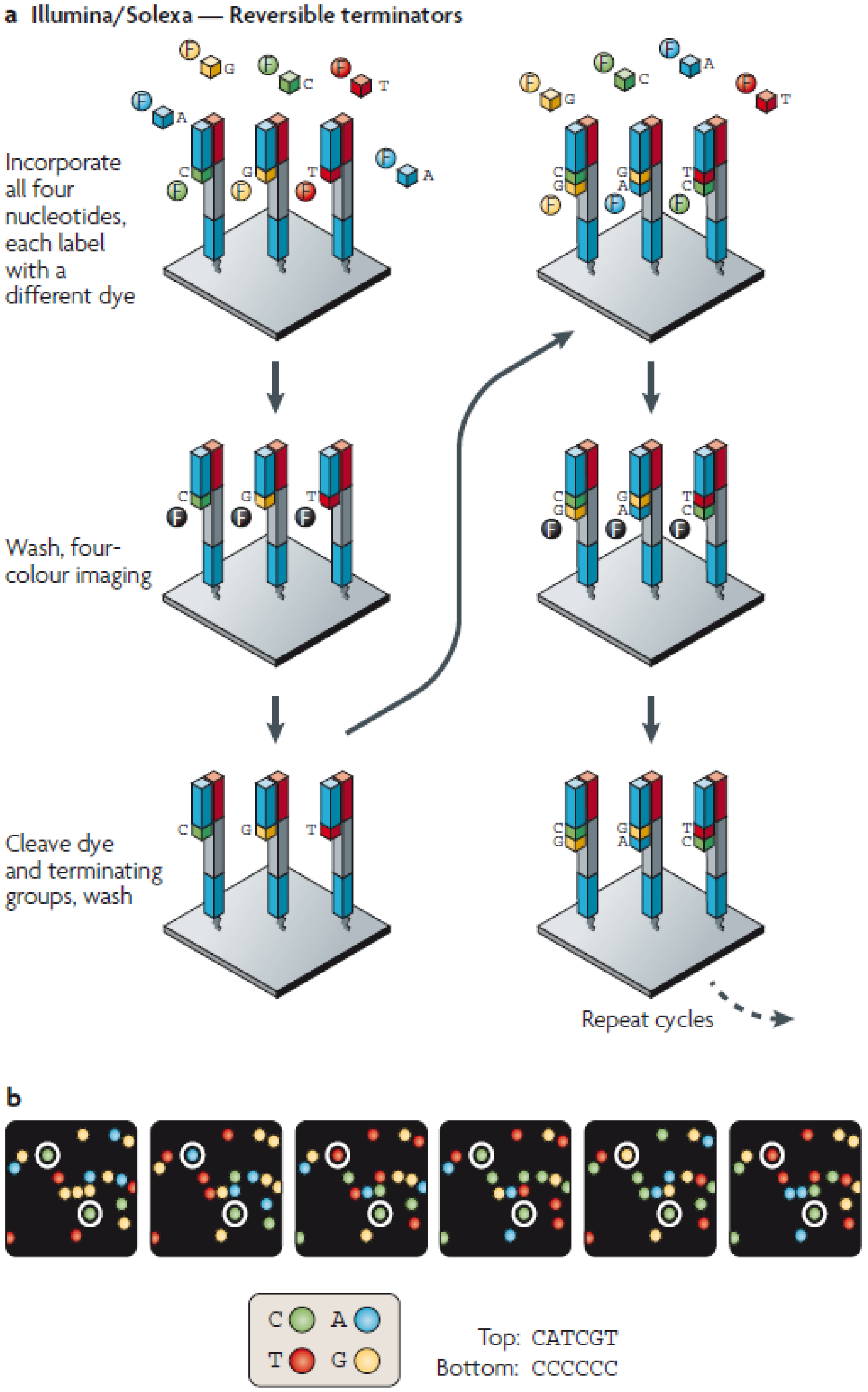

a Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
A
G
C
T
F
b
C
A
T
G
Top: CATCGT
Bottom: CCCCCC

Fig. 3.2: Sequencing (a) and imaging (b) process used by the Illumina Genome Analyzer. The figure is based on [44]. Reprinted with permission.

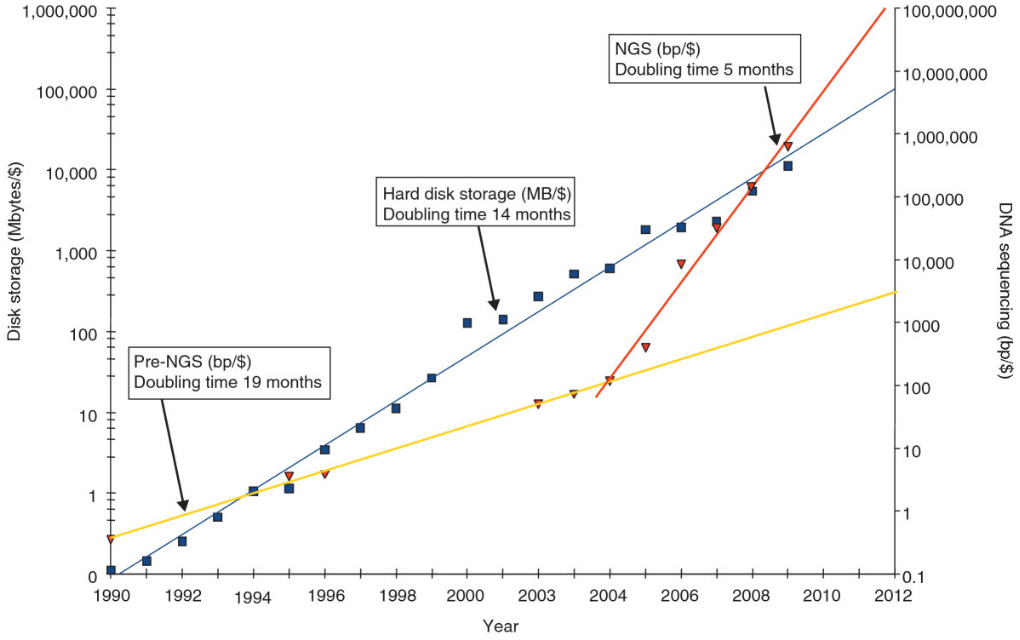


Fig. 3.3: Historical trends in storage versus DNA sequencing costs. With the advent of NGS technologies in the mid-2000s, the long term trend changed and sequencing data gets cheaper more quickly than disk storage. Taken from [49].

Emerging challenges. For many years data storage, data management, and compute power for analysis was not the major issue with DNA sequencing. The developments and improvements in computer technology still follow Moore's Law, a long-term trend first formulated in 1965 by Intel co-founder Gordon Moore [46]. Moore's Law states, informally speaking, that compute power doubles every 18 months. Similar laws exist for hard disk capacities [47] and the cost for sending data over optical networks [48]. For many years, the developments in sequencing technologies lagged behind or kept pace with the developments in the technical infrastructure. With the advent of NGS technologies, the long-term trend changed as depicted in Fig. 3.3.

The major challenges associated with sequencing projects are no longer time and money that have to be spent for data generation. Cheaper data is generated with higher throughput and challenges are shifting to data management and analysis. Massive amounts of data have to be archived, but also made available for analysis by collaborators. The increasing amounts of data, usually in the range of terabytes for a single study, can no longer be simply transferred over the internet. The requirements for the algorithms concerning memory and time efficiency and also for the compute servers used for computational analysis have drastically increased. Another new challenge is the need for new statistical and computational methods to make sense of the massive amounts of data collected in large sequencing projects.

3.2 Analysis of next-generation sequencing data

The details of the analysis of NGS data depends on the purpose of the respective project, as reviewed in [50]. There are however four steps that are performed in most NGS analysis pipelines, namely data preprocessing, read mapping, variation detection, and functional annotation of variants. Another application that is widely used in RNA-seq projects is the quantitation of transcripts. In this section the basic ideas and some of the available algorithms and tools for each of these steps are introduced.

Preprocessing

The quality and correctness of sequencing data is essential for the success of downstream analyses. Sequencing artifacts are quite common in NGS data. These sequencing artifacts include read errors (base calling errors or small insertions/deletions, depending on the platform), poor quality reads and contamination with primers and adapters. Reads containing parts of PCR primers or adapters from library preparation will not be mappable to a reference genome. Low quality reads and incorrectly called bases lead to problems during variant detection. Most commercial vendors of sequencing platforms provide a pipeline for quality control (QC) and filtering of the sequencing data. However, experience shows that a significant amount of sequencing artifacts still remains in the datasets. Several tools and software packages are available for QC of NGS data. The choice of an appropriate tool for quality control depends on the sequencing platform. One example for a QC tool that is suitable for data generated by Illumina platforms is FastQC [51]. FastQC provides several statistics relevant for quality assessment of NGS data like quality per base, mean sequence quality, or GC content. In addition overrepresented sequences are detected and compared to known adapter and primer sequences. Based on these statistics the user can decide which preprocessing steps like quality filtering, quality trimming, or clipping of adapter and primer sequences need to be applied.

Read mapping

Read mapping, the alignment of all reads to a reference genome, is fundamental to NGS data analysis. The read mapping problem can be formalized as follows: given a reference sequence G, a set of reads and a distance $k \in \mathbb{N}$, find all substrings g of G that are within distance k of a read. Not all reads are expected to match perfectly to the reference sequence. The divergence between the reference sequence and the reads can be attributed to natural divergence between the sequenced genome and the reference genome, but also to sequencing errors. The distance calculation used by the mapping algorithm should be able to account for the specific error distribution of the NGS platform in use. Efficient algorithms for read mapping are needed as NGS platforms can generate millions of reads per run. NGS reads are typically relatively short and have a specific error distribution. Many algorithms and tools exist for read mapping [52, 53, 54, 55, 56] using different search strategies and distance measures. It is not always possible to assign reads to exactly one position in the genome. Reads can originate from repetitive regions in the genome or from highly conserved substructures. For downstream analyses like variant detection it is important to resolve these ambiguities and/or restrict the analysis to uniquely mapped reads. Not all read mapping tools report if a read could be assigned to a unique position. Depending on the downstream analysis, if a tool provides information on unique mapping can be an important criterion for choosing a suitable mapping tool.

The mapping of RNA-seq data imposes additional problems. Reads that span exon boundaries will not map continuously to the reference genome. Mapping reads to the transcriptome is one way to solve this problem, however transcriptomes are incomplete even for well-studied species including human and mouse. When reads are mapped to the transcriptome it should be kept in mind that unique mapping has a different meaning than in mapping to a genome. A read can map ambiguously to several transcripts of the same gene but still have a unique assignment relative to the genome. An alternative to mapping to the transcriptome is to use a spliced mapping approach [57], some split mapping methods combine spliced mapping with novel transcript detection [58, 59, 60].

For mapping of both, DNA and RNA sequencing data, the choice of the right algorithm is a crucial and non-trivial step that depends on the goal of the project. In addition, the parameters must be fine-tuned to the respective projects, which is often difficult for the non-expert user.

Variation detection

A major goal of resequencing projects is the detection of variation between the sequenced genome and a reference genome. The major classes of variation are SNVs, short INDELs, copy number variations and larger genomic rearrangements, as depicted in Fig. 3.4. The strategies that can be applied to detect these classes of variations are described below.

SNV and INDEL calling. The input for the detection of SNVs and INDELs is a set of reads which are aligned to a reference genome. The main issue is to decide whether an observed difference between the sequenced reads and the reference sequence originates from sequencing errors or from real variation. A variety of tools exist to detect SNVs in sequencing data from single or multiple samples [54, 61, 62, 63, 64]. The majority of the SNV and INDEL detection tools relies on digital allelic counts to infer a consensus sequence and the allelic abundance in the sample. Early tools relied on simple cutoff rules for calling SNVs, more recent methods incorporate statistical models to infer the

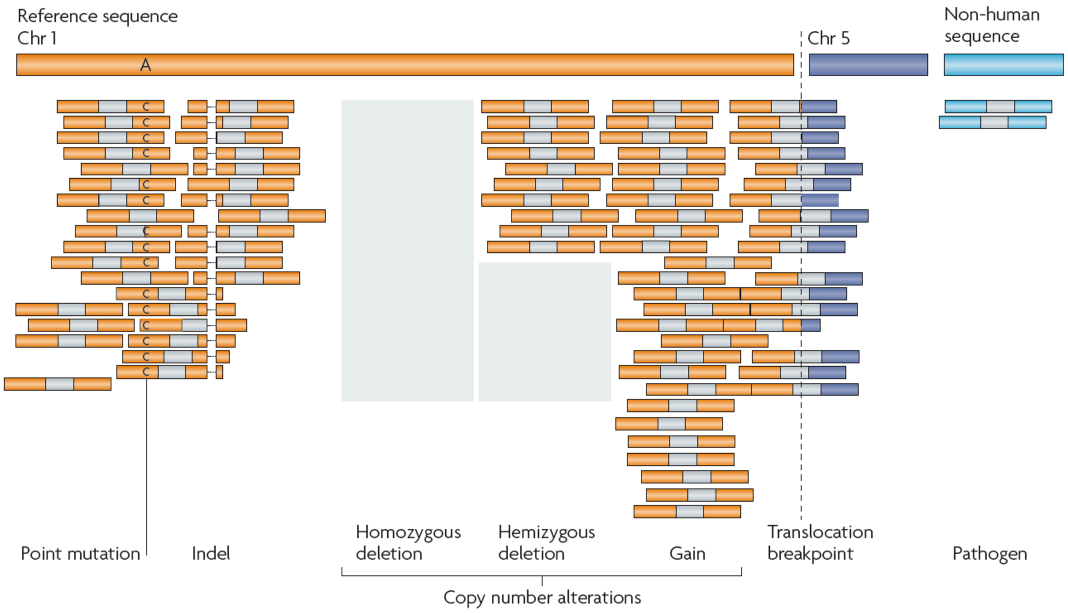


Fig. 3.4: Types of genetic variations in cancer that can be detected by NGS. Paired-end sequencing reads aligned to a reference sequence are depicted as bars. Taken from [50]. Reprinted with permission.

most likely genotype.

Of special interest in cancer genomics is the identification of somatic mutations, mutations that are present in the tumor but not in the germline of the patient. A common way to identify somatic mutation is to call variants from a sequenced tumor sample and from a sequenced sample of normal tissue. The mutations found in the normal tissue can be subtracted from the mutations found in the tumor sample to obtain the mutations that occur only in the tumor. Some approaches simultaneously process the tumor sample and the normal tissue sample to call somatic mutations [65, 66]. Detection of somatic mutation, however, requires a sufficient sequencing depth for tumor and normal tissue to be statistically reliable. This is a major issue in RNA-seq data, where the sequencing depth correlates with transcript abundance, which can differ significantly between samples. The analysis of somatic mutations must also consider the purity of the tumor sample. If a sample contains 50% DNA from tumors cells, and a mutation is present in one of the four copies of the chromosome, the observed frequency of that mutation in the sample will be 12.5%. In comparison, a heterozygous mutation in the germline is expected to be observed in 50% of the DNA reads that map to the respective position.

Detection of genomic rearrangements. Next-generation sequencing of whole genomes and transcriptomes has shown to be suitable for the systematic detection of rearrangements of cancer genomes [67, 68, 69, 70]. Reads that partly map to different locations or paired-end reads where the two ends map to different locations are evidence for rearrangements. Whole-genome sequencing is the most comprehensive approach for rearrangement detection but also the most expensive one. RNA-seq is more cost efficient, but restricted to rearrangement events in coding regions.

Copy number variations. NGS data can be used to detect copy number changes with high resolution and precise definition of the breakpoints [67, 71, 65, 72]. The detection of CNVs is based on changes in the coverage of the reference sequence. CNVs can be detected from whole-genome sequencing and from exome sequencing, but not from RNA-seq. In RNA-seq, the coverage of the reference sequence correlations with the abundance of the transcripts and therefore cannot be attributed to CNVs.

Analysis of unmapped reads. Reads that do not align to the human reference sequence can occur for different reasons. The first, most likely, scenario is that unmappable reads are caused by contamination of the sample or by errors during sample preparation and sequencing. These reads can be discarded as they do not carry additional information. Two other sources of unmappable reads are more interesting. These sequences can stem from parts of the human genome that are still missing in the reference sequence or they can come from pathogens. Pathogenic sequences can be identified by comparing the sequencing data to collections of known pathogenic sequences.

Quantitation of transcripts

RNA-seq can be used to quantify transcript expression [73, 74]. Expression levels are frequently estimated as RPKM, Reads Per Kilobase of exon model per Million mapped reads, as defined in [73]. The primary advantages of RNA-seq compared to array technologies (see Section 3.3) is the large dynamic range, the low background noise, the requirement of less sample RNA and the ability to detect novel transcripts, even in the absence of a sequenced genome. Major challenges of this technology are the handling of mapping uncertainty, non-uniform sequencing depth and potential new isoforms.

Functional analysis

The identification of genetic alterations or differentially expressed genes is the first step in cancer genomics. The next step is to make sense of the genetic variation. Therefore one needs to estimate the impact and effect of the observed variations. Not all mutations observed in a tumor are responsible or even related to the disease. One method to detect disease-related genetic alterations is to correlate the genetic profile with disease or disease outcome. A high correlation of a genetic variation with a disease (outcome) alone is no proof for a casual association. Additional evidence, e.g., from functional analyses, is needed to identify the disease-relevant mutations. Functional and biological annotation and analysis can help to separate driver mutations from passenger mutations and provide valuable insight. The functional and biological annotation of genetic variation can be divided into three areas: investigating the direct change in the genomic sequence, comparison with known variations, and evaluation of the biological context of the affected genes.

Investigation of the direct change in the genomic sequence. The fist thing to consider is the genomic context of the mutation. In intergenic regions a mutation can affect transcription factor binding sites or other regulatory regions. Intronic variation can for example affect splice sites. A comparison with prior knowledge on transcription factor binding sites or splice sites can provide valuable information. For mutations in coding regions the effect of the resulting transcript and protein sequence can be examined. Mutations in functional domains, highly conserved regions or around phosphorylation sites are more likely to affect protein function. Different tools are available that predict the functional impact of protein mutations [75, 76, 77].

Comparison with known variations. During the last decade numerous databases have been made publicly available that provide information about mutations and their disease association. The *dbSNP* [78] collects information about known SNPs in the human population. The *OMIM* database provides useful information about known inherited or Mendelian disorders [79]. In the context of cancer, databases like the *Sanger Institute Catalogue Of Somatic Mutations In Cancer (COSMIC)* [80, 81] or the *Roche Cancer Genome Database (RCGD)* [82] can give hints if mutations, or at least the mutated genes, have been previously associated with special types of cancer.

Biological context of the affected genes and proteins. The biological context of mutated genes and the corresponding proteins is of great value for assessing the functional impact of a mutation. The biological context is of special interest when comparing cancer genomes. The same biological pathway can be deregulated in two different tumors that do not share mutations. The function of one protein can be altered in the same way by different mutations. It can therefore be useful to shift the analysis from single mutations to mutated genes, especially in comparison studies. The interpretation can also be done on the pathway level to identify common pathways that are involved in cancer development. An annotated biological function of a mutated gene can also give useful hints whether a mutation could be associated with the disease of interest. Web resources on pathways [83, 84] and biological function annotation [85] are of great value in this context.

Functional investigation and annotation of mutations and mutated genes are essential tools when trying to make sense of observed mutations. This step is however not trivial and there are still many unresolved issues. Reliably predicting the functional impact of mutations is a very challenging task. Predicting a loss of function for a frameshift mutation in a protein is possible, but evaluating the impact of single amino acid substitutions on e.g., protein-protein interactions is still not feasible today. Data integration, data consistency, and the reliability of the data in public databases is a big issue when integrating prior knowledge on mutations. As in all prior steps of NGS data analysis, the functional annotation and especially its interpretation have to be closely adopted to the aim of the study and be performed with care. While there are some standard questions where standard analysis tools exist (e.g., gene set enrichment analysis for gene expression), new and innovative problems cannot be addressed with standard tools and thus introduce new challenges for computational biology.

3.3 Array technologies

The basic principle behind microarrays is hybridization between two complementary DNA single strands. Specific DNA sequence (probes) are attached to a solid surface. The surface is organized in spots, with each spot containing picomols of the same probe. Fluorescently labeled targets can bind to probes with a complementary sequence. The signal that is emitted from a spot depends on the amount of target sequences that bind to the probes in the spot. The identity of the spot is determined by its position on the chip. Microarrays do not allow for the direct quantitation but use relative quantitation in which the intensity of a feature is compared to the intensity of the same feature under a different condition. In two-channel microarrays, two samples, labeled with different colors, can be measured simultaneously and allow for the comparison of two conditions on the same chip. Microarrays allow for testing tens of thousands of genetic features in parallel. Different fields of applications exist for microarrays [86], the most widely used are gene expression profiling, SNP detection [87], the detection of alternative splicing of fusion genes, and comparative genomic hybridization (aCHG) to measure copy number variations [88].

As for NGS, data analysis, data management, and statistical data interpretation are challenging tasks. Statistical and computational methods are needed in order to make sense of the high-throughput raw data.

Microarrays are a very valuable approach for high-throughput, parallel genomic testing. The major drawback is that they can only capture features that are present on the chip, so one only gets what one is looking for. Novel mutations or transcripts cannot be detected with microarrays.

Part II

Integrated analysis of genomic variation and identification of targets for immunotherapy

In the first part of this thesis different hypotheses on tumorigenesis where presented, namely viral integration, somatic mutation, and deregulation. The experimental high-throughput methods that can produce large amounts of information about cancer genomes were introduced. The second part of this thesis addresses the question how high-throughput genomics data can be exploited to shed light onto the development of individual tumors. The combination of data from individual tumors, prior knowledge of biological mechanisms, and computational methods are first steps towards personalized cancer treatment and immunotherapies. This part describes the computational and theoretical methods of high-throughput analysis of genetic variation. All steps described here are applied to clinical data and the results are presented in Part III.

In Chapter 4 we present approaches for in-depth and individualized analysis of the genetic variation in cancer to assess the contribution of viral integration and somatic mutation on tumorigenesis. We show how mutations can be analyzed with respect to their functional impact and to their biological context.

In Chapter 5 we describe how information on tumor-specific mutations can be exploited to identify targets for personalized immunotherapy. We focus on the prediction of tumor- and patient-specific T-cell epitopes that can be applied as epitope-based vaccines. We present computational prediction methods for the identification of T-cell epitopes and show how these can be applied to genetic variations detected with the methods described in Chapter 4.

Chapter 6 addresses the major challenge of how we can make the different computational approaches available in an integrated, flexible and reproducible pipelining system. Large amounts of data need to be analyzed in genomic studies. Processing and storage of the data cannot be done on a single desktop computer. Compute cluster or grid infrastructures are needed in order to process the data in a timely manner. The application in a clinical setting requires high reliability and reproducibility of the data analysis pipelines. However, individualized approaches to cancer treatment are still in their infancy. Computational analysis systems therefore need to be flexible in order to adopt to new findings and new questions. To promote the transfer of new findings to the clinical application the computational results have to be made easily available to biomedical researchers and clinicians and to be presented in a comprehensive way. We present user-interfaces that address this critical issue.

Chapter 4

Integrated analysis of NGS data

We implement a workflow for an integrated analysis of NGS data. Our workflow follows the general analysis steps for NGS data described in the background section (Section 3.2). After quality control and, if necessary, adapter removal reads are mapped to a reference genome. In case of RNA-seq analysis unmapped reads are additionally mapped to a set of transcripts. Reads that can be mapped to the reference genome (and transcriptome) are used for variation detection. Reads that cannot be mapped to the reference genome or transcriptome are analyzed for potential viral sequences as described in more detail in Section 4.2. An overview of our NGS analysis workflow is given in Fig. 4.1. In the following sections we first describe how read mapping is performed. We then describe a method for the detection of viral sequences in the unmapped reads. Then we describe how we detect genetic variation from mapped reads and how to annotate and analyze genetic variation. Later in this part (Section 6.3) we also present VariationDB as a web-based tool for interactive analysis of genetic variations identified with this pipeline.

4.1 Read mapping

Prior to read mapping we first perform a quality control using FastQC [51]. If FastQC detects adapter sequences in the read data the adapter sequences are trimmed from the reads.

We first map all reads against the positive and the negative strand of a reference genome (*GRCh37/hg19* in most cases). For read mapping we use the tools RazerS [55] and SplazerS [57]. Both tools were developed based on SeqAn [89], an efficient, generic C++ library for sequence analysis. RazerS tags found mappings of reads with *unique*, *multi*,

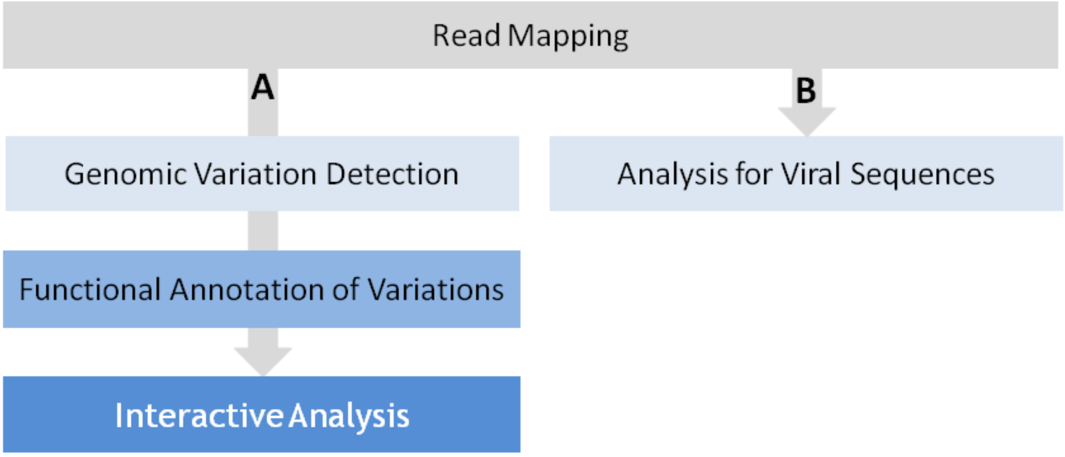

Read Mapping
A
B
Genomic Variation Detection
Analysis for Viral Sequences
Functional Annotation of Variations
Interactive Analysis

Fig. 4.1: General NGS analysis workflow. After quality control reads are mapped to a reference genome (and transcriptome). A: Mapped reads are submitted to variation detection. Found variations are annotated. We provide a tool, VariationDB (see Section 6.3) for the interactive analysis of the genetic variations. B: Unmapped reads are submitted to the pipeline for detection of viral sequences.

and *suboptimal*, indicating that the match is the unique best match found for that read, one of multiple equally good matches for that read, or that better hits were found for that read. We only use reads that are tagged as *unique* for variation detection. However, we include *multi* and *suboptimal* matches for the identification of unmapped reads. Only reads with no match below the detection threshold are considered *unmapped*.

If we apply out pipeline to RNA-seq data, we additionally map all reads that could not be mapped to the reference genome to all human transcript sequences that are available form RefSeq [90]. Since we want to combine the information from read mapping to the reference genome and to transcript sequences we map the transcript positions back to genomic positions. We therefore use transcript position annotation provided by the UCSC Genome Browser [91]. Mappings are split at exon boundaries, leading to partial mappings of reads. We then also split the read sequences and consider the partial reads as new reads. If several transcripts are known for one gene, reads can have multiple matches on transcripts but a unique position on the genome. We resolve matches on multiple transcripts with the following strategy:

1. If a read maps to several transcripts from more than one gene we reject the reads due to multiple matches.

2. If a read has several *multi* hits on different transcripts from the same gene we map the matches to genomic positions. When several (partial) matches cover the same genomic coordinates we only keep one match and consider that match unique.

This strategy allows us to adhere to the definitions for *unique* and *unmapped* reads that we introduced for mapping against the reference genome.

4.2 Viral integration

If a viral infection contributed to tumorigenesis via the introduction of viral oncogenes into host cells (as observed e.g. for high risk HPV) or the presence of modified viral oncogenes after integration into host cell DNA (as observed for Merkel cell polyomavirus [22]) traces of the viral infection can be detected in the human genome or transcriptome. The general idea is to analyze sequencing data from a human tumor for sequences with viral origin. The analysis pipeline is outlined in Fig. 4.2. The basic steps in the identification of sequences with potential viral origin are 1) removing sequences with human origin, 2) comparing the remaining sequences with known viral sequences and 3) analyzing the results of the comparison. This concept has been proven to be a useful strategy for identifying viral sequences [22, 92]. Depending on the scientific question, the identification of potential viral sequences can be performed on DNA or RNA sequencing data. Both strategies have advantages and drawbacks that are discussed at the end of this section.

4.2.1 Removing sequences with human origin: Digital Sequence Subtraction

The aim of this step is to obtain a set of sequences with non-human origin. This process was originally applied to transcriptome sequencing data and termed digital transcriptome subtraction (DTS)[93]. The basic principle is to subtract known sequences of the host (human) from the sequencing data to obtain a set of candidate viral sequences. This procedure can also be applied to genome and exome sequencing. We term this process digital sequence subtraction (DSS) instead of DTS to underline that is not restricted to transcriptome sequencing.

Mapping sequencing data to the reference genome is often the first step in an NGS analysis pipeline. When we search for traces of pathogenic genomes, only reads that cannot be mapped to the host genome are of interest. Many analysis steps downstream of read mapping rely on high-confidence mapping, so the criteria for read mapping are usually very stringent. In the context of variation detection, very reliable mapping results are indispensable. However, many sequences that cannot be mapped with stringent criteria are still of human origin. Reasons, that human reads cannot be mapped include sequencing errors, variable regions in the genome, an incomplete reference genome, or a too large divergence between the genome under investigation and the reference genome. For the purpose of identifying sequences that very likely do not have a human origin additionally applying less stringent criteria for the comparison to

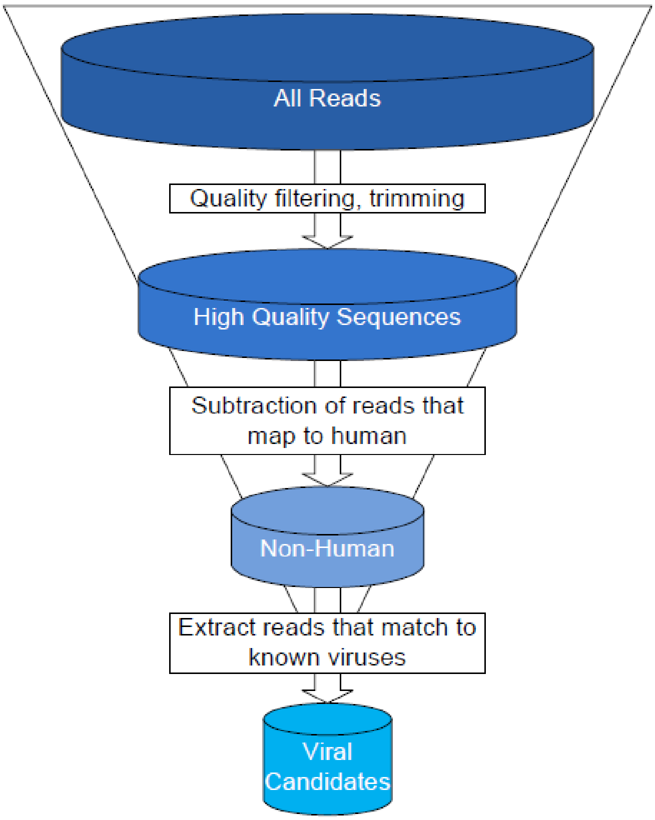

All Reads
Quality filtering, trimming
High Quality Sequences
Subtraction of reads that map to human
Non-Human
Extract reads that match to known viruses
Viral Candidates

Fig. 4.2: Detection of viral sequences in sequencing data. Low quality sequences and sequences that match to human are removed from the dataset. The remaining reads are compared to known viral sequences. Sequences that match to a known virus are candidates for viral sequences and are further analyzed.

known human sequences is necessary. We apply SeqClean [94], a tool that was originally designed to screen sequencing data for various contaminants, low-quality and low-complexity sequences. As contaminant sequences we use the human reference genome (assembly version *GRCh37/hg19*), a set of human repeats [95], a set of know human immunoglobulines, and the RefSeq collection of known human transcripts. Information on download locations and download dates for the databases can be found in Tab. 4.1. For a hit to a contaminant we require a minimal length of 30 bp and a minimal identity of 94%. If a read only partially hits on a contaminant, the matching part is removed but the non-matching part is kept. We thereby ensure that we do not discard reads completely that span the potential breakpoint where a virus was integrated into the human genome. SeqClean also removes low-complexity sequences and trims polyA/T stretches of the reads. SeqClean requires sequences in fasta format. If read data is given in fastq format, the sequences have to be converted to fasta.

4.2.2 Comparison to known viruses

The set of reads or partial reads remaining after DSS is compared to known viral sequences. The search strategy and databases to use depend on the type of viral sequences one aims to identify. One possible application is to search for a specific virus type, e.g., investigating sequence data from cervix carcinoma for the presence of human papilloma virus (HPV). When searching for a specific virus using all known sequences of that virus species and relatively stringent matching criteria are suitable. In contrast, if we try to identify new viruses with sequence similarity to any known virus using the complete set of known viral sequences and lower stringency are more suitable as used in [22]. When applied to transcriptome sequencing comparison against know viral protein sequences instead of known viral nucleotide sequences is also an option to identify possibly related proteins.

We use BLAST (blastx for comparison to protein sequences or blastn for comparison to nucleotide sequences) to compare the read data to known viral sequences. BLAST results are filtered for sequence identity. Reads with BLAST hits that pass the filtering step are considered as candidate viral sequences. As a first validation, the candidates are additionally blasted against the GenBank nonredundant collection (NR). Reads that have equal or better hits to non-viral species are discarded. The remaining reads are then further analyzed.

4.2.3 Analysis and interpretation of the results

To ensure that we do not miss potential viral sequences, the search criteria, especially when searching for viruses that are only similar to sequences represented in the viral

Tab. 4.1: Databases used for the identification of potential viral sequences.

Database Name	Download information	Download date	Used for
Human reference genome (GRCh37/hg19)	ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/H_sapiens/ Assemled chromosomes and mitochondrial chromosome in fasta format	2009-08-10	DSS
Human repeats	http://www.girinst.org/ RepBase version 14.9 in fasta format	2009-08-10	DSS
Human immunoglobulines	ftp://ftp.ncbi.nih.gov/blast/db/FASTA/ igSeqNt.gz	2009-08-13	DSS
Human RefSeq RNA	ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot, Release 36	2009-08-07	DSS
NR protein sequences	ftp://ftp.ncbi.nih.gov/refseq/release/complete/, Release 36	2009-08-07	blastx
Viral Refseq	ftp://ftp.ncbi.nih.gov/refseq/release/viral/, Release 36	2009-08-07	blastx

Tab. 4.2: RNA-seq samples from Arron *et al.* [97] that were analyzed for viral sequences. Sample ILS1933631 and HeLa are known to be infected with HPV.

Sample ID	**Number of total Reads**	**Number of reads with viral hits**	**Percentage of reads with viral hits**
ILS1933631 (Cervical SCC)	4.1 mio	627	0.02
HeLa (cell line)	1.9 mio	2,823	0.15
Arron STA01-106 (SCC)	3.6 mio	9	0.0002
Arron STA01-094 (normal Skin)	5.3 mio	44	0.0008

databases, are not stringent. This strategy however implies a high number of false positive hits. For large datasets with millions of reads, analyzing the remaining viral candidates manually is not feasible.

To facilitate the analysis of the candidate viral reads we used the metagenomics software MEGAN4 [96]. The original aim in metagenomics is to understand and analyze the composition of complex microbial consortia in environmental samples through sequencing and analysis of their DNA. Similar approaches can be performed to analyze metatranscriptomes. The problem we are facing when searching for viral sequences in human samples is somewhat similar to metagenomics or metatranscriptomics, except that we are trying to assign the reads to human or viruses instead of different bacteria. MEGAN4 takes a set of aligned reads and assigns them to species in a taxonomic tree.

To illustrate and validate our approach we applied it to a transcriptome sequencing dataset published by Arron *et al.* [97]. The aim of the original study was to investigate whether HPV plays a role in cutaneous squamous cell carcinoma, a special type of skin cancer. The datasets consist of RNA-seq data from samples of normal skin, different squamous cell carcinomas (SCC), and HeLa cells. Some of the samples are known to be HPV positive. We applied our strategy to four of the samples from the Arron dataset, two with known HPV transcription and two with no HPV transcription. The HPV positive controls are generated from HPV type 18-infected HeLa cells and from a HPV type 16-positive cervical carcinoma. The HPV negative sample are generated from one SCC and from normal skin. The samples are summarized in Tab. 4.2, along with the number of reads that have valid hits on viral sequences and the percentage of reads with valid hits on viruses.

The number of reads with hits on a viral sequence after blastx against known viral proteins is 2.823 for the HPV infected HeLa cell line and 627 for the HPV positive cervical SCC. This equates to 0.15% and 0.02% of the total number of reads, respectively. For the two negative datasets we found 9 and 44 reads with hits to a viral sequence, which amounts to 0.0002% and 0.0008%. All reads with a valid hit on a virus were additionally compared to the GenBank nonredundant protein collection GenBank (NR) using BLSTX.

The results are then analyzed for their taxonomic composition using MEGAN4 [96]. The min-support LCA parameter in MEGAN4 was changed to 1 order to make sure not to miss species with just one hit. The results are displayed in Figure 4.3.

The taxonomic analysis revealed that most of the reads with viral hits for the HPV positive samples are assigned to papillomaviridae. Interestingly, while nearly all viral reads in the cervical carcinoma sample directly matched to HPV type 16, the HeLa sample contained many reads that matched papillomaviridae or human papillomavirus but could not be assigned specifically to HPV type 18 (Figure 4.3). Most of these reads matched HPV E1 sequences. Due to the conserved nature of the E1 gene, these reads matched not only HPV18 but also other HPV subtypes and were thus placed lower in the taxonomic tree. The samples declared as virus-free by Arron *et al.* showed a small number of potential viral hits in our analysis. These hits do not produce a clear signal for a specific virus and further manual inspection revealed that these hits were artifacts.

These results show that our approach is able to identify actively transcribed viruses. Metagenomic analysis is a versatile tool for virus detection in sequencing data. The resulting graphical representations facilitate a direct visual inspection of sequencing results for many samples.

4.2.4 Discussion

Digital sequence subtraction of genomes or transcriptomes followed by a metagenomic analysis is a useful and valid approach for the investigation of tumor samples for potential viral sequences. A viral contribution can however only be detected in cases were viral sequences are actually present in the sample. Some viruses are known to contribute to tumor development with a "hit-and-run" mechanism, where a viral infection initiates tumor growth, but expression of viral transcripts is not needed for tumor maintenance [98]. In such cases it is even possible that the viral genome gets lost again, leaving no easily detectable evidence for a viral presence. If the analysis is done on the mRNA level, the virus has to be actively transcribed in order to be detectable. Performing the analysis on RNA-seq data is however a valid approach since most of the known carcinogenic viruses are still transcribed in tumor cells.

It has also to be kept in mind that we can only detect viruses that share sufficient sequence similarity with viruses that are represented by the search databases. Our reads are compared to a database containing all known viral sequences, a viral read that has no similarity with a known virus would not be detected. To assure that we do not miss longer viral sequences we propose to additionally assemble all non-human reads [99]. Longer contigs with good support could hint at the presence of an unknown virus.

The approach presented here relies on databases with known viruses. The quality of the results clearly depends on the quality and reliability of sequences in the databases.

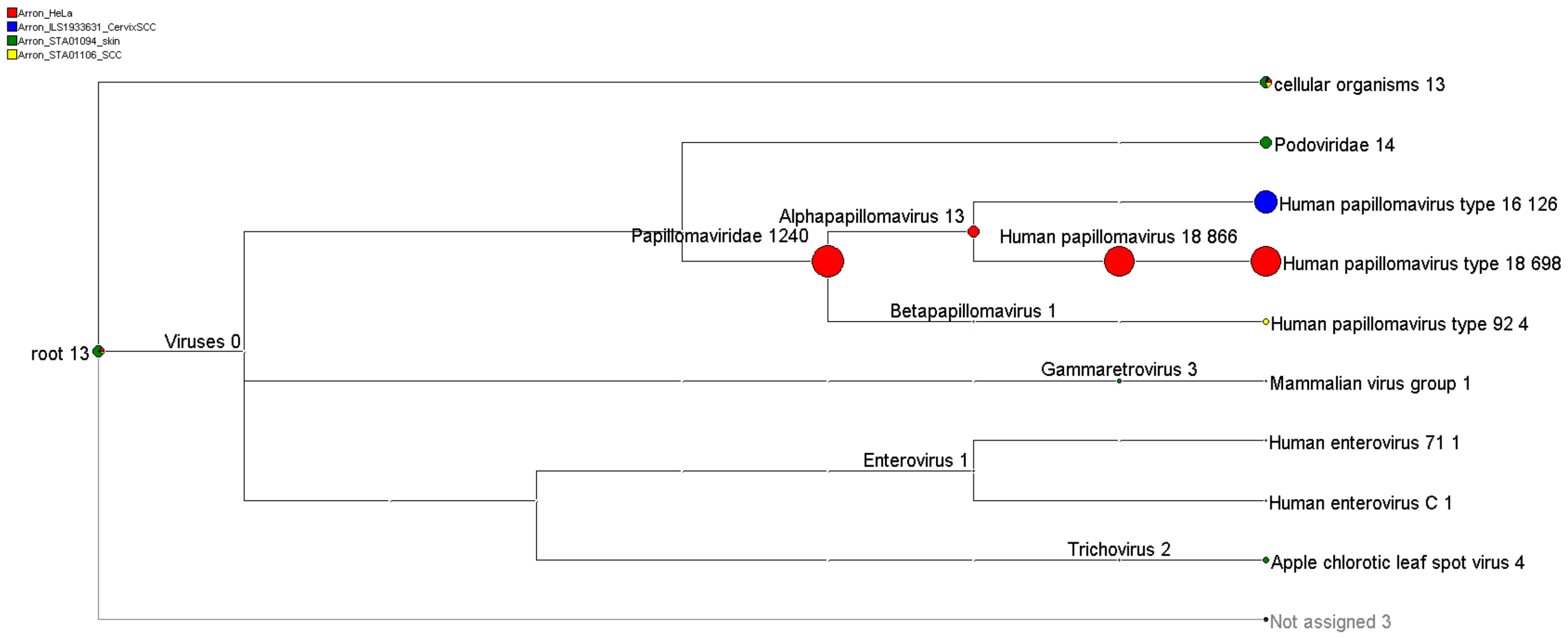


Fig. 4.3: Analysis of the viral content of four samples from the Arron dataset [97]. The taxonomic composition of all samples are assigned to a taxonomic tree. The number displayed after each taxon name represents the number of reads that were assigned to the taxon. The reads from each sample are represented by different colors. For the two HPV-positive samples we see a clear signal in the HPV subtree.

As shown in more detail in Chapter 7.3, possible “junk” sequences in the viral database can produce significant, but meaningless, hits on viral species. Potential hits on a viral sequence have to be treated with care and the possibility of invalid viral sequences in the reference databases has to be kept in mind.

Despite these limitations, combining digital sequence subtraction with metagenomics analysis tools is a useful and versatile approach for the identification of viral sequences in NGS data.

4.3 Detection of genetic variation

We use SnpStore for variation detection, a tool that was developed by Anne-Katrin Emde in the group of Prof. Knut Reinert, FU Berlin. SnpStore is based on SeqAN and detects SNVs and small INDELs in sequencing data. SnpStore takes mapped reads in GFF or SAM format as input.

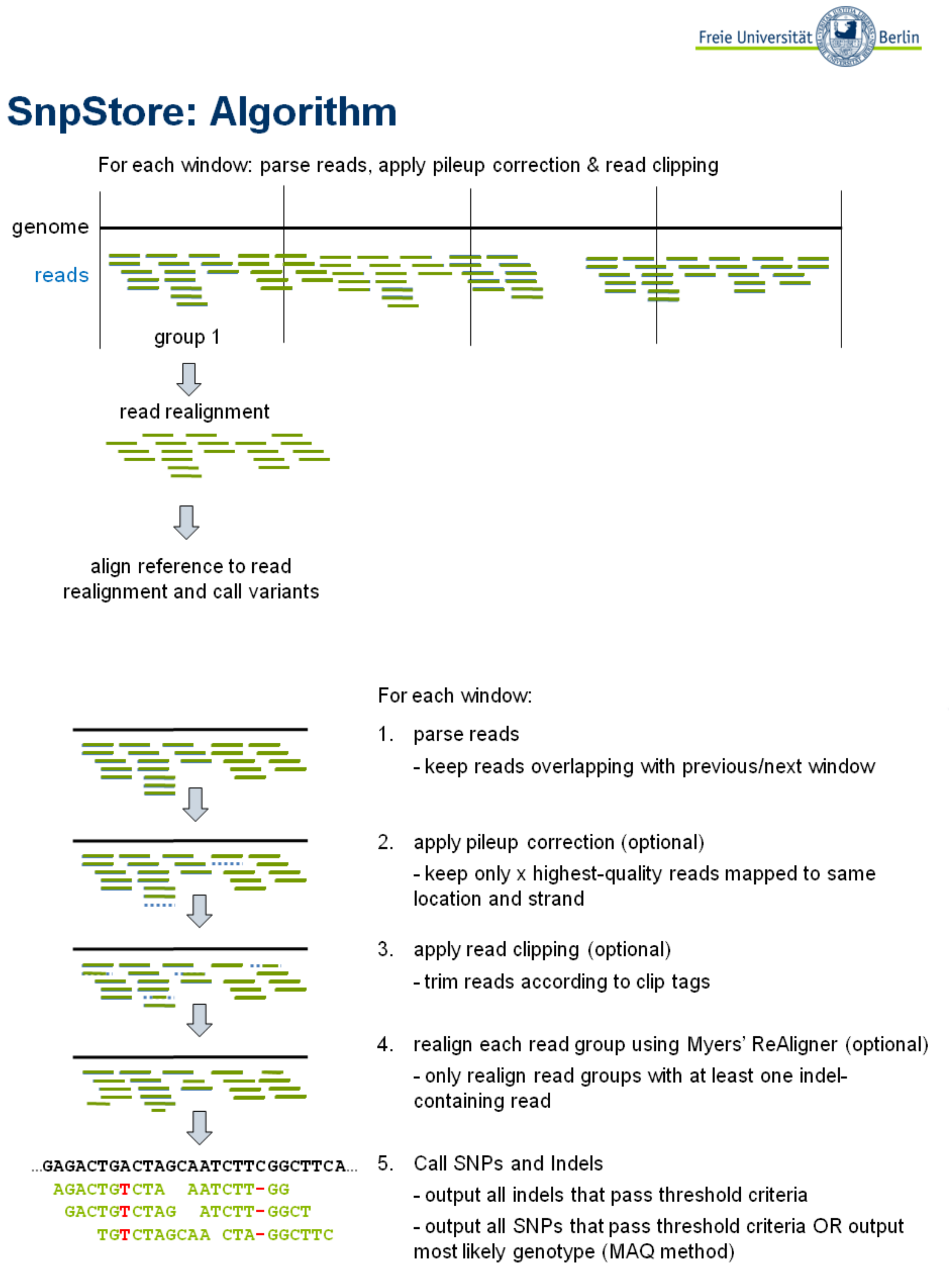
SNV detection can be performed using a Bayesian model based on the maq method [54] that outputs the most likely genotype for this positions. Alternatively SNVs can be called using a threshold model that is based on the minimum number of reads containing the non-reference base, the minimum fraction of reads containing the non-reference base and the minimum average quality of the non-reference base .

INDEL detection is performed using a threshold model using the minimum number of reads containing the non-reference base and the minimal fraction of reads containing the non-reference base. INDEL-contributing columns in the alignment are merged by taking the average with no phasing, i.e. only one INDEL is allowed per location. A realignment is performed around INDEL positions using the Anson-Myers ReAligner [100]. The SnpStore algorithm is outlined in Fig. 4.4.

4.4 Functional and differential analysis of genomic mutation

The detection of mutations in NGS data is a widely applied task in genomic studies. The post-processing and interpretation of the mutation data depends on the scientific question that underlies the study. In the context of personalized cancer genomics, the major questions are:

- Which of the observed mutations are somatic mutations of the tumor?
- What is the functional impact of the mutations?
- Are the mutation already known or known to be associated with cancer?
- Which pathways or biological processes are affected?



SnpStore: Algorithm

For each window: parse reads, apply pileup correction & read clipping

For each window:

1. parse reads
   - keep reads overlapping with previous/next window
2. apply pileup correction (optional)
   - keep only x highest-quality reads mapped to same location and strand
3. apply read clipping (optional)
   - trim reads according to clip tags
4. realign each read group using Myers' ReAligner (optional)
   - only realign read groups with at least one indel-containing read
5. Call SNPs and Indels
   - output all indels that pass threshold criteria
   - output all SNPs that pass threshold criteria OR output most likely genotype (MAQ method)

```
...GAGACTGACTAGCAATCTTCGGCTTCA...
   AGACTGTCTA   AATCTT-GG
     GACTGTCTAG   ATCTT-GGCT
        TGTCTAGCAA  CTA-GGCTTC
```

Fig. 4.4: Outline of the SnpStore algorithm used for SNV and INDEL detection. Figure provided by Anne-Katrin Emde, FU Berlin.

- Can we detect recurrent patters of single somatic mutations or affected genes, pathways, or biological pathways in different tumors?

In this section we propose approaches to give answers to these questions.

4.4.1 Somatic mutations

All healthy cells of an individual share basically the same DNA. During cancer development, cancer cells acquire and accumulate mutations. A cancer genome diverges significantly from the genome of the healthy cells. In order to identify somatic mutations, sequencing data from cancer cells and from control tissue (healthy tissue of the same individual) have to be compared. The identification of somatic mutations is not trivial and has to be adapted to the experimental setting. We want a method for the detection of somatic mutations that we can flexibly adapt to different experimental settings and that produces comprehensive results. We use an approach where we first detect mutations in the two samples separately and subtract the mutations observed in the healthy tissue from the mutations observed in the tumor. The remaining mutations are considered as candidate somatic mutations. There are several reasons why these candidates can be false positive somatic mutations. The mutation can also be present in the control tissue, but with a frequency that is just below the detection threshold. Another reason can be that the coverage for the same position in the two samples diverges. The latter is especially critical if RNA-seq data is used for the detection of mutations where coverage depends on transcript expression. To rule out at least some of the false positive somatic mutations that occur for these reasons we additionally analyze positions with candidate somatic mutations in the control tissue for coverage and observed bases. Mutations that are also observed in the control sample but with a frequency below the detection threshold are rejected as somatic mutations. Mutations in the tumor with low or no coverage in the control sample can be treated separately as low-confidence candidates for somatic mutations.

4.4.2 Functional annotation

Mutations can influence gene regulation, gene expression, protein expression, protein function, and protein-protein interaction. All these single levels are far from being fully described or understood. Prediction of the functional impact of mutation is therefore not yet feasible. Bringing mutation data together with known annotation of the genome can however give valuable hints to identify functionally interesting mutations (see Section 6.3).

Positional annotation of mutations

We use the software tool ANNOVAR [101] to generate a positional annotation of the mutations. The mutations are given relative to a reference genome. ANNOVAR reports the genomic context of a mutation, e.g., intergenic, intronic, or exonic. If a mutation falls within a gene, the gene name is reported. For mutations in coding regions the effect on the respective transcripts and protein sequences are also reported. For mutations that alter the sequence of a protein we additionally check whether the mutation lies within a specific domain or phosphorylation site. We use information on phosphorylation sites of proteins from the PhophoSite database [102] (`Phosphorylation_site_dataset.gz`, downloaded on February 11, 2011). Information on protein domains is based on PROSITE [103] and extracted from the *InterProXref* table provided by the UCSC genome Browser [104, 91]. The degree of evolutionary conservation of amino acids in protein sequences can give a hint for the severeness of a mutation. The conservation score for all proteins of interest is precomputed from multiple sequence alignments of homologous sequences [105].

Prior knowledge on mutations

dbSNP [78] is a database of genetic variations maintained by the National Center for Biotechnology Information (NCBI). dbSNP contains information on known mutations (SNVs and INDELs). Together with information on the mutation itself, dbSNP contains information on the frequency of the mutation in a population. If a mutation observed in a tumor is known to occur frequently in a population the chance of this mutation having a causal link to tumor development is rather slim. We associate the observed mutation or the genes that are affected with information form the *Sanger Institute Catalogue Of Somatic Mutations In Cancer database (COSMIC)* [80, 81] database and with information on inherited disorders taken from *OMIM* [79].

Affected biological pathways and processes

Once mutations are mapped to genes we can consider a broader context, namely the biological pathways or processes the respective genes are part of. The biological context of mutated genes can give insight into the functional impact on mutations. A mutation affecting a key player in a pathway that contributes to cell growth control for example is more likely to promote cancer development than a mutation in a gene with unknown function. We integrate KEGG [83, 84] as source for biological pathways and function annotation from Gene Ontology (GO) [85].

Recurrent mutations

The fact that a somatic mutation occurs in several different cancers can be interpreted as a hint that this mutation actually drives cancer development. Finding recurrent mutations is very interesting, but we propose to widen the comparative analysis to genes or even pathways and processes. Different mutations can severely affect the same pathways but the underlying mutations would still be missed if the comparison is restricted to specific mutations. Comparison on the gene, pathway, or process level can lead to the identification of recurrently affected genes, pathways, or processes. This approach is in agreement with the cancer hallmark theory proposes by Hanahan and Weinberg [12].

4.5 Discussion

The analysis of sequencing data can reveal valuable insights into the genetic alterations in cancer. A main interest in that field is the detection of somatic mutations that distinguish the tumor from the healthy tissue of the same individual. The detection of genetic variation relies on statistical models that try to identify the most likely genotype for a sequenced tissue. These models however rely on the assumption that the tissue is rather homogenous. Tissue samples obtained from biopsy or surgery are classified as tumor by the surgeon and pathologists. However, the percentage of real tumor cells in the sample is often unknown. The samples very likely contain surrounding non-tumor tissue, sometimes it is even unclear if the sample contains tumor cells at all. This is a general issue in the detection of somatic mutations, whether the variant detection is performed separately for the samples (as presented in our pipeline) or the variant detection is performed in parallel for tumor and control samples [106]. In addition, tumors themselves are heterogeneous collections of cells, which further complicates the picture. Single-cell sequencing [106] offers possibilities to address some of these issues, however, the experimental methods are still error-prone and expensive.

The annotation and interpretation of mutations relies on knowledge and annotation data, e.g., on gene annotations for genomes, biological pathways, mutation data or disease association. However, the knowledge in this field is still incomplete. Not all public resources that provide annotation information are curated, and integrating data from different resources is difficult. Statistical analyses are needed to separate important driver mutations from passenger mutations in cancer. These statistical methods still need to be developed and need to take into account more than just the mutation data. Integration of additional data like gene expression, copy number variations, or methylation will improve the detection of driver mutations in cancer, however the integration and statistical analysis of different types of omics data is still a challenging problem.

Our NGS analysis pipeline is based on a flexible workflow system (see Chapter 6.4 for details). If improved computational methods become available, we can integrate them into our pipelines.

Chapter 5

From mutation to targets for immunotherapy

Various approaches for tumor immunotherapy have been applied in the past [9]. We choose to focus on an approach that is individualized with respect to the tumor and to the patient's immune system. We identify cytotoxic T-cell epitopes that stem from somatic mutations of a cancer. The T-cell epitopes can be administered to the patient as an epitope-based vaccine.

Administering patient- and tumor-specific T-cell epitopes as vaccines is a fully individualized approach and a compelling idea. The development of new chemotherapeutic drugs is a lengthy and expensive procedure. A target has to be present in many different cancers in order to qualify as potential target for the development of a new drug. The individualized approach we present here does not require common targets. As long as a mutation is present in the tumor it can be used as target for a vaccine. The drawback of this approach is that the vaccine cannot be tested on large cohorts of patients, but has to be designed for each patient individually. Therefore each patient has to be genetically profiled, which is still a large experimental effort.

The first step in this approach is the identification of somatic mutations that lead to an alteration of a protein sequence as described in the last chapter. Also viral sequences that are integrated in the tumor genome and that are expressed are good targets for tumor-specific sequences. In this chapter we describe computational methods for the prediction of T-cell epitopes and how these can be applied to tumor-specific mutations.

In Section 2.2.1 we described the prerequisites for a peptide to be able to function as a T-cell epitope, i.e. to elicit an immune response: (1) the peptide has to be generated by the antigen processing pathway, (2) the peptide has to be presented by one of the patient's HLA molecules on the cell surface, and (3) the T-cell repertoire of the patient has to contain a T-cell receptor that is specific for the peptide:HLA complex. In this chapter we present the state-of-the-art methods that can be applied to predict these three steps and describe our own contribution to that field.

As a general concept, all presented methods use the concept of supervised learning. A prediction function is learned from examples with a known outcome and can then be applied to predict the outcome for new data points. Supervised learning can be applied to learn classification and regression models. Classification means separating data points into different classes, e.g. in the context of peptide-MHC binding prediction, the peptides are separated into binders and non-binders. For regression, the data points are associated with a value, e.g., the binding affinity of a peptide to an MHC molecule. The prediction model learns a function that predicts real value outcomes for new data points.

Different approaches are available for learning a prediction function, from simple linear models to advanced machine learning methods like artificial neural networks (ANNs) or Support Vector Machines (SVMs) [107].

The term "epitope prediction" is ambiguously used in the community. A stringent interpretation is that all of the before-mentioned requirements (antigen processing, MHC binding, availability of a suitable T-cell receptor, T-cell activation) have to be analyzed and predicted. In contrast to the antigen processing pathway and immunogenicity of peptide:HLA complexes, MHC binding is well defined and understood and reliable prediction methods are available. The term epitope prediction is sometimes used as a synonym for MHC binding prediction. For simplicity, in this thesis we use the term epitope prediction if we refer to the general attempt to predict whether a peptide has the potential to elicit an immune response. We are aware that not for all of the steps that contribute to immunogenicity prediction methods are available. However using the term epitope prediction allows us to describe the general concept. The advantage of this definition is that it is independent of the current status of prediction methods that are available for the different steps or that are actually included in the prediction process. As prediction methods improve or emerge they can be integrated without the need of adapting the general terms. If we refer to the details of the single sub-processes we use the term antigen processing prediction, MHC binding prediction and prediction of T-cell reactivity.

5.1 Antigen processing prediction

Cleavage by the proteasome and transport into the endoplasmatic reticulum of peptides by TAP are generally considered prerequisites for the generation of peptides that can bind to MHC class I. The goal of antigen processing prediction is to identify peptides that are available for MHC class I binding in the ER, meaning that they are produced by the proteasome and transported by TAP. The sequence specificity of proteasomal cleavage and TAP transport match the sequence specificity of many MHC alleles. Proteasomal cleavage and TAP transport are however less specific than MHC binding and are therefore harder to predict. Whereas MHC molecules are specialized to the presentation of a distinctive set of peptides to T-cell receptors and have a well defined binding specificity, the proteasome is involved in many other biological processes. Proteasomal cleavage shows some specificity for cleavage sites, but has rather broad specificity in general. The proteasome is responsible for the generation of the C-termini of peptides binding to MHC class I [108, 109]. Besides the relatively low specificity of proteasomal cleavage and TAP transport, the lack of appropriate experimental data is the main issue in developing reliable prediction methods. Without sufficient data, prediction methods tend to overfit the training data and therefore cannot be reliably applied to new datasets [110]. Different methods have been developed to predict proteasomal cleavage [111, 112, 113, 114] and TAP transport [115, 112, 113], the accuracy of these methods however leaves room for improvement. Nevertheless, integrating proteasomal cleavage and TAP transport prediction with MHC binding prediction has shown to be a promising approach to increase the specificity of the prediction of peptides that can be naturally presented by MHC class I [113, 114].

For the exogenous pathway that produces peptides for presentation by MHC class II little is known about specific sequence patters. A prediction of antigen processing in the context of MHC class II is therefore not yet possible.

5.2 MHC binding prediction

In this thesis we focus on the prediction of epitopes for cytotoxic T cells (CTLs). CTLs recognize peptides in complex with MHC class I. We therefore focus on the development of prediction methods for peptides that bind to MHC class I. The binding mode and interaction between MHC molecules and bound peptides is highly conserved. Amino acid side chains are involved in the interaction, the sequence of the peptide influences the binding. Most prediction methods available today for MHC binding prediction are sequence-based methods, which try to identify sequence patters correlated with binding affinity. The approaches for the identification of the sequence patters vary from simple motif-based methods [116, 117, 118, 119] to advanced machine learning methods [120, 121, 122, 123, 124, 125]. Some prediction methods that rely on structural information of the peptide:MHC complex have also been presented [126, 127, 128].

A major challenge in MHC binding prediction is MHC polymorphism. More than 7.000 different human MHC alleles are known today, and the different alleles display a wide spectrum of binding specificities. Classical approaches for MHC binding develop allele-specific methods and rely on the availability of a certain amount of allele-specific experimental binding data. Experimental binding data is only available for a small subset of all human MHC alleles, the number of alleles with sufficient training data to train robust prediction models is even smaller. Thus the number of "predictable alleles" lags behind the number of known alleles. However, for vaccine design and immunotherapy the binding specificity of all MHC alleles need to be known. One of the first approaches to tackle MHC polymorphism was the concept of MHC supertypes. Supertypes are groups of alleles that show a high overlap in their peptide binding repertoires [129, 119, 130, 131]. The concept of MHC supertypes is appealing because it greatly reduces the complexity of the MHC binding prediction problem. However, in-depth analyses have brought us to the conclusion that the overlap between peptide binding repertories of different MHC alleles is too slim to be of any value for the prediction of tumor-specific epitopes (data not shown).

Pan-specific approaches try to overcome the problem of the lack of allele-specific training data based on the idea that similarities in the binding groove can be exploited to leverage information across alleles. Among the first to address this problem for MHC class II were Sturniolo *et al.* [117]. They described the MHC binding groove to be composed of individual pockets. For each of the pockets a variety of compositions, the pocket variants, exists. These variants have been shown to be shared among different alleles. Using a modular matrix approach Sturniolo *et al.* determined the binding specificity of an MHC allele by combining the binding affinities of the individual pocket variants constituting the binding groove. DeLuca *et al.* [132] generalized this approach and applied it to MHC class I. Using the pocket definition of Chelvanayagam [133] in combination with a modular matrix approach they were able to increase the number of predictable MHC class I alleles by a factor of seven.

5.2.1 UniTope - Predict binding for all MHC class I alleles

We combine some of the ideas presented above into a more general method for the prediction of MHC class I binding peptides. Based on an analysis of crystal structures of MHC:peptide complexes, we determined which MHC residues contribute to a specific pocket. This information was used to determine pocket variants of all available MHC class I alleles allowing a modular representation. Since the majority of peptide:MHC structures available contains peptides of length nine, we only focus on nonameric peptides in this study. We incorporate physico-chemical properties to encode experimentally confirmed binding peptides taken from the Immune Epitope Database (IEDB) [134] and train an SVM model.

An overview of the UniTope approach is given in Fig. 5.1, details for the different steps are described in the following sections.

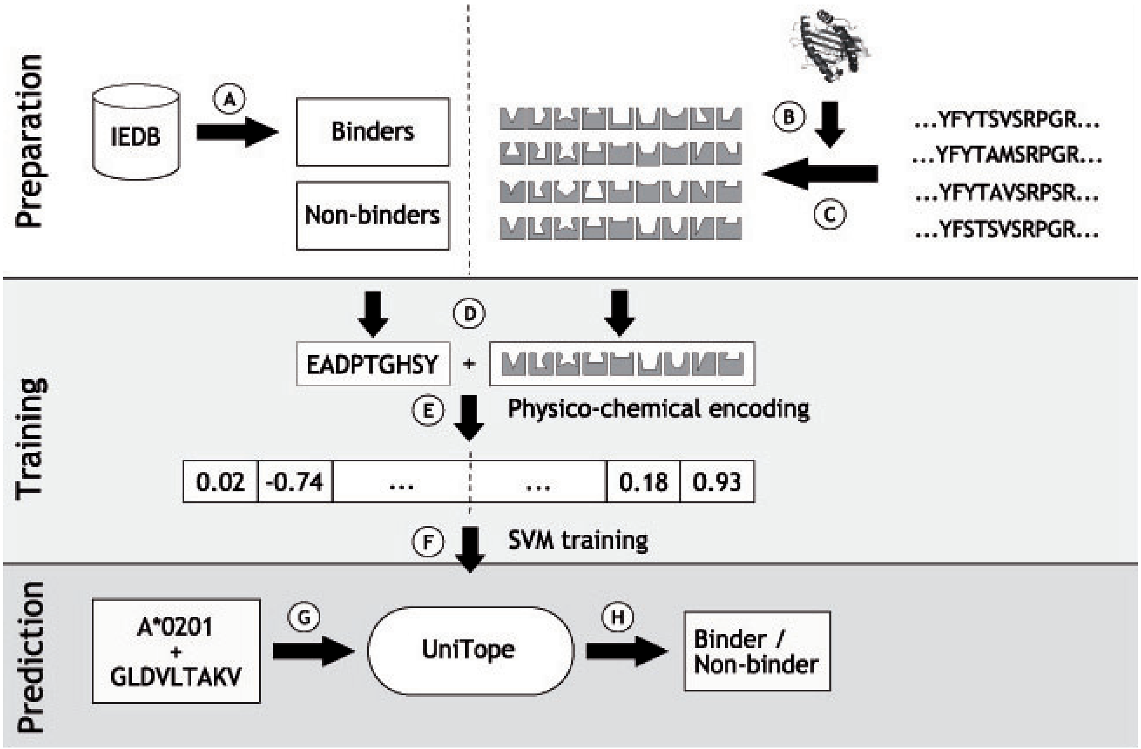


Fig. 5.1: An overview of the UniTope approach. **Preparation:** (A) Binding 9-mers are retrieved from the IEDB. Non-binders are generated randomly. (B) Pocket profiles are determined by analyzing structures of 9-mers bound to MHC. (C) The pocket profiles are mapped onto the MHC sequences to obtain the pocket variants. **Training**: (D) For each peptide-allele pair, the sequence of the peptide and the description of the corresponding MHC allele are merged. (E) A physico-chemical encoding is applied to the peptide-allele combinations to obtain the final input vectors. (F) The input vectors are used to train a single SVM model for all alleles. **Prediction:** (G+H) The SVM model (UniTope) classifies new peptide-allele pairs into one of the two classes "Binder" or "Non-binder".

Tab. 5.1: UniTope Pocket Profiles.

Pocket	MHC Residues													
Pocket 1	7	58	59	62	63	66	159	163	167	170	171			
Pocket 2	7	9	22	24	36	45	62	63	66	67	70	97	99	
Pocket 3	7	9	70	97	99	114	116	155	156	159				
Pocket 4	62	65	66	69	70	155	156	159	163					
Pocket 5	9	66	69	70	73	74	97	99	116	152	155	156	159	
Pocket 6	65	69	70	73	74	97	114	147	155	156				
Pocket 7	9	70	74	77	95	97	114	116	146	147	150	152	155	156
Pocket 8	72	73	76	77	80	146								
Pocket 9	74	77	80	81	95	97	116	123	124	143	147			

Retrieving peptide training data. Positive (binding peptides) as well as negative (non-binding peptides) examples are needed to train our model (see Fig. 5.1(A)). We retrieved all nonameric peptides binding to MHC class I from the IEDB [134]. We discard sequences that contain non-natural amino acids or that have ambiguous annotations within the IDEB (a peptide-allele combination is annotated as binder in one dataset and annotated as non-binder in another dataset). We then performed a homology reduction on the remaining set of binding peptides such that every two peptides binding the same allele differ in at least three positions. As negative examples we use random non-binders generated according to the amino acid distribution in SWISS-PROT [135], release 51.

Determination of pocket profiles and pocket variants. A pocket of the MHC class I binding groove is composed of all residues in contact with the corresponding amino acid of a bound ninemer. The MHC sequence indices of residues found to contribute to a specific pocket in any of the alleles are recorded in the pocket profile (Fig. 5.1 (B)). In order to determine the pocket profiles, 3D-structures of nonameric peptides bound to MHC class I molecules had to be analyzed. Seventy-five crystal structures of such MHC:peptide complexes were retrieved from the Protein Data Bank (PDB) [136]. These structures were analyzed using the BALL framework [137] and the "SS-SB" contact criteria from [127]. An MHC residue and a peptide residue are defined to be in contact when they are maximally 4 Å apart. Interactions with the MHC backbone as well as with the peptide backbone are omitted. The resulting pocket profiles are listed in Tab. 5.1.

To determine the pocket variants of the individual MHC alleles, all known amino acid sequences of human MHC alleles were retrieved from the IMGT/HLA database (release 2.16) [138]. Sequences of alleles which have been shown not to be expressed were discarded. Furthermore, all sequences with an incomplete binding groove were removed because they could not be used for the determination of pocket variants of the allele. A multiple alignment of the remaining sequences using ClustalW [139] showed a conserved sequence "SHSMRYF" at the beginning of the α-chain. To ensure consistent indexing all MHC sequences were truncated to begin with this conserved sequence. When the conserved sequence was incomplete, it was completed. After having adapted the pocket profiles to this indexing scheme, they were mapped onto the MHC sequences (Fig. 5.1 (C)). Thereby pocket variants could be determined for 1,348 alleles.

Building the prediction model. We use an SVM [107] for MHC class I binding prediction. For more information on SVMs and kernels see [140].

Our data are MHC allele-peptide pairs which are either labeled as binder (positive example) or non-binder (negative example). The generation of input vectors from this data is based on an amino acid encoding proposed by Venkatarajan *et al.* [141]. They derived five descriptors from a principal component analysis of over 200 physico-chemical properties of amino acids taken from the amino acid index database (AAindex) [142]. An amino acid can thus be described by a vector of length five. The composition of the input vectors for our SVM model is shown in Eq. 5.1. An input vector is composed of :

1. an encoding D of the amino acids AA_i of the peptide, and

2. an encoding $\overline{D}$ of the nine binding pockets P_i representing an allele.

The encoding D of an amino acid AA_i is composed of the five amino acid descriptors $d_k(AA_i)$. The encoding $\overline{D}$ of a pocket P_i is the mean over the 5-dimensional descriptors D of the contributing residues. This composition results in input vectors of length 90 (Fig. 5.1(D+E)).

$$\vec{x} = [D(AA_1), \ldots, D(AA_9), \overline{D}(P_1), \ldots, \overline{D}(P_9)] \tag{5.1}$$

with

$$\begin{aligned} D(AA_i) &= [d_1(AA_i), d_2(AA_i), d_3(AA_i), d_3(AA_i), d_5(AA_i)] \\ \overline{D}(P_i) &= \tfrac{1}{|P_i|} \textstyle\sum_{a \in P_i} D(a) \end{aligned}$$

Other encodings (binary sparse encoding [143], a simple physico-chemical encoding based on size, hydrophobicity, and charge) were also tested, but the encoding based on Venkatarajan's descriptors performed best.

Encoding both, the MHC allele and the peptide, enables us to train a single prediction model for all MHC alleles (Fig. 5.1(F)). An SVM with a Radial Basis Function (RBF) kernel was trained on a data set comprised of all peptides from the set of binders and the same number of random non-binders per allele. We used the LIBSVM library (Version 2.71)

[144] to implement the SVM. The parameters of the RBF kernel were optimized using the standard grid search method from LIBSVM.

Prediction performance. Our model is able to perform predictions for 'seen' and 'unseen' alleles (Fig. 5.1(G+H)). Seen alleles are those MHC alleles which were included in the training set, while unseen alleles are those not included. When measuring prediction performance these two cases have to be handled independently. We use the Matthew's Correlation Coefficient (MCC) as performance measure. The MCC is equivalent to the Pearson's correlation coefficient that can be applied to binary classifications. It measures the correlation between the actual class and the predicted class. For details on performance measures for binary classifiers see [145]. For all tests we require a minimum of 15 binders available for testing in order to obtain significant results.

To assess the performance on seen alleles we use standard performance measures and compare the performance to other prediction methods. We use a modification of a five-fold cross-validation (CV), where we account for allele-specific data. This allows for a separate CV result per allele. Sufficient binding data for allelewise cross-validation was available for 23 alleles. On these alleles UniTope yields an average MCC of 0.84 with a maximum of 0.95 (B*51:01) and a minimum of 0.58 (B*40:01). On all but one allele (95.7%) the MCC is greater than or equal to 0.73. On 17 alleles (73.9%) the MCC is to 0.8 or better.

Additional tests were run to compare our approach to SVMHC [124] and PeptideCheck [132]. We require a minimum of 20 peptides to train an SVM model [143]. All alleles that have at least 35 verified binders - 20 for training and 15 for testing - were included in this test. Allele-specific test sets were extracted from our data randomly. For alleles with more than 75 binders, one fifth of the allele-specific data was taken. For alleles with 35 to 75 known binders 15 peptides were removed for testing. The remaining data was used to train a model. This model was then used for comparison with other methods. UniTope was tested on all different test sets while the other methods were tested on all test sets concerning alleles for which prediction was offered. Because these methods score peptides in contrast to classifying them, a threshold separating binders from non-binders had to be chosen. For each method-allele combination we chose the threshold such that the MCC on the test set was maximized.

31 alleles met the requirement of having at least 35 known binders. For seven of these alleles an SVMHC model was available. PeptideCheck models were available for 25 of these test alleles. These two groups of alleles will be referred to as SVMHC alleles and PeptideCheck alleles, respectively. The results of the comparison of all three methods on SVMHC alleles are shown in Fig. 5.2. The average MCCs on the SVMHC alleles are 0.79 (UniTope), 0.74 (PeptideCheck) and 0.72 (SVMHC). On the PeptideCheck alleles UniTope achieves an average MCC of 0.82, and PeptideCheck an average MCC of 0.73. UniTope

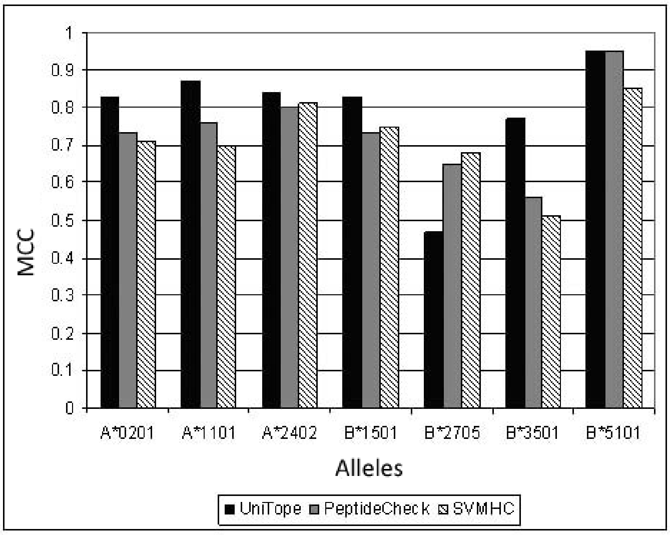


Fig. 5.2: Performance of UniTope, PeptideCheck and SVMHC on seen alleles. The average MCCs are 0.79, 0.74 and 0.72, respectively.

outperforms PeptideCheck in 19 of the 25 alleles.

To measure the prediction performance on unseen alleles blind prediction tests were performed. In these tests all binders and non-binders of a particular allele are removed from the training set. A model is trained on the remaining training set and tested on the extracted data. Every allele with at least 15 binders was included in this test.

Sufficient binding data for these tests was available for 34 alleles. On these alleles UniTope yields an average MCC of 0.80 with a maximum of 0.91 (B*54:01) and a minimum of 0.55 (B*27:05). On all but three alleles (91.2%) the MCC is greater than or equal to 0.71. On 20 alleles (58.8%) the MCC is greater than or equal to 0.8.

Further developments of UniTope. The original UniTope approach described above was further developed by N. C. Toussaint [146]. This new version uses a support vector regression (SVR) approach to predict binding affinities instead of performing a binary classification. In addition, the performance assessment and the comparison to other methods was improved. The regression version of UniTope outperforms existing allele-specific prediction methods on alleles that are included in the training data. For alleles not included in the training data UniTope performs on par with these methods, demonstrating the validity of the approach. However, netMHCpan [147], a very similar pan-specific approach performs better. For details on the implementation and evaluation and discussion for the new version of UniTope refer to [146].

Discussion. The polymorphic nature of the MHC I molecules together with the lack of sufficient experimental data for the majority of the known alleles is a key problem in

MHC I binding prediction. We have proposed UniTope, a pan-specific approach that is able to compensate for the lack of allele-specific data by sharing information across similar alleles. UniTope has proven to be a useful tool to increase the number of predictable alleles. However, in agreement with other pan-specific approaches [147], the additional analyses performed by N. C. Toussaint have shown that the availability of allele-specific training data also influences the performance of pan-specific approaches. In order to produce satisfying results for a specific allele with no binding data a sufficient amount of training data for related alleles has to be included in the training data. Pan-specific approaches are of great value for increasing the number of predictable alleles. For alleles that are not represented in the training data (by allele-specific data for that allele or from related alleles) the performance, however, leaves room for improvement.

5.3 Prediction of immunogenicity

As already introduced earlier, the prerequisites for a peptide to be a T-cell epitope are (1) the peptide has to be generated by the antigen processing pathway, (2) the peptide has to be presented by one of the patient's HLA molecules on the cell surface, (3) the peptide:HLA complex has to elicit an immune response. In this chapter we already presented computational approaches to predict antigen processing and MHC binding. While the accurate prediction of MHC binding for many MHC alleles is possible, the need for improvements in the prediction of actual T-cell reactivity is huge. The induction of an immune response requires the presence of a suitable T cell in the T-cell repertoire of the host. Negative selection is a key mechanism to ensure self-tolerance of the immune system. During negative selection the T-cell repertoire is shaped by the host proteome. The lack of appropriate data and the complex dependencies in self-tolerance render the prediction of T-cell reactivity a hard problem. Very few methods with limited performance are currently available. The more recent machine-learning based approaches [148] use different encodings of the peptide sequence but do not include additional information. The limited success in predicting T-cell reactivity can be attributed to three main factors: (1) The training data is not corrected for a bias towards MHC binding, it therefore cannot be excluded that the methods predict MHC binding rather than T-cell reactivity. (2) The relevance of the MHC context is disregarded. (3) The methods use only sequence information of the peptides and therefore make the implicit assumption that T-cell reactivity is solely dependent on the peptide sequence. The complex influence and dependencies of negative selection of the T-cell repertoire are neglected.

We present a prediction method for T-cell reactivity that addresses all three issues. We correct the bias towards MHC affinity in the training data, we consider the MHC restriction of T-cell reactivity by training an MHC allele-specific predictor and we incorporate a model of self-tolerance into our predictor. The focus in this thesis lies on the modeling of negative selection. For more details on the other aspects please refer to [146, 149].

5.3.1 Modeling self-tolerance

The distinction between self and non-self, also called self-tolerance, is a key concept of the immune system. Negative selection of the T-cell repertoire is one mechanism towards self-tolerance. Self-reactive clones, T cells that recognize self-peptides in complex with self MHC, are deleted from the T-cell repertoire. Peptides that are part of the proteome or very similar to self-peptides are therefore unlikely to elicit an immune response.

Representation of the self-proteome. Negative selection takes place in the thymus and the thymus proteome is a reasonable reference set for central tolerance. However, to also include peripheral tolerance in our model we need to take the complete host proteome into account. The complete human proteome was retrieved from the International Protein Index (IPI, version 3.47) [150]. The thymus proteome is derived from gene expression data for the thymus. 43% of all protein-encoding genes are represented by the probes of the employed microarrays [151, 152]. A consensus voting is used to resolve conflicting measurements and the covered proteins are divided into three groups: 45% are present in the thymus, 26% are marginally expressed, and 26% are absent from the thymus. We define two thymus proteomes, thymus-min, consisting of all proteins present in the thymus, and thymus-max, consisting of all proteins that are present or marginally expressed in the thymus. For details on the selection of the thymus proteomes and the majority-voting algorithm please refer to [146].

Only peptides that are presented by MHC are visible for T cells. To represent the proteome we therefore use a set of peptides that are predicted to bind to the MHC allele under consideration by netMHC [123]. Since the majority of all MHC-I-ligands are ninemers, we restrict all analyses to peptides of length nine.

Distance to Self. The distance of a peptide to a set of peptides is defined as the smallest pairwise distance of the peptide to one of the peptides in the set. In order to calculate this distance we need to define a measure for the distance between two peptide sequences. For computational reasons, we use a distance measure instead of a similarity measure. We propose a distance measure that is based on the BLOSUM50 [153] amino acid substitution matrix that originally represents similarities between amino acids. We convert the BLOSUM50 matrix into a 20x20 distance matrix where each element in the matrix represents the distance between two amino acids. The distance matrix is generated in four steps:

1. The original substitution matrix A is converted into a symmetric matrix $A^{'}$ by replacing the entries a_{ij} by $a^{'}_{ij} = \frac{a_{ij}+a_{ji}}{2}$
2. The matrix $A^{'}$ is shifted so that the resulting matrix $A^{''}$ contains no negative values: $a^{''}_{ij} = a^{'}_{ij} + |min(A^{'})|$.
3. The matrix is normalized by the maximum entry in $A^{''}$: $A^{'''} = \frac{a^{''}_{ij}}{max(A^{''})}$.
4. The final distance matrix M is obtained by subtracting the entries of the symmetric, non-negative and normalized matrix $A^{'''}$ from 1: $m_{ij} = 1 - a^{'''}_{ij}$

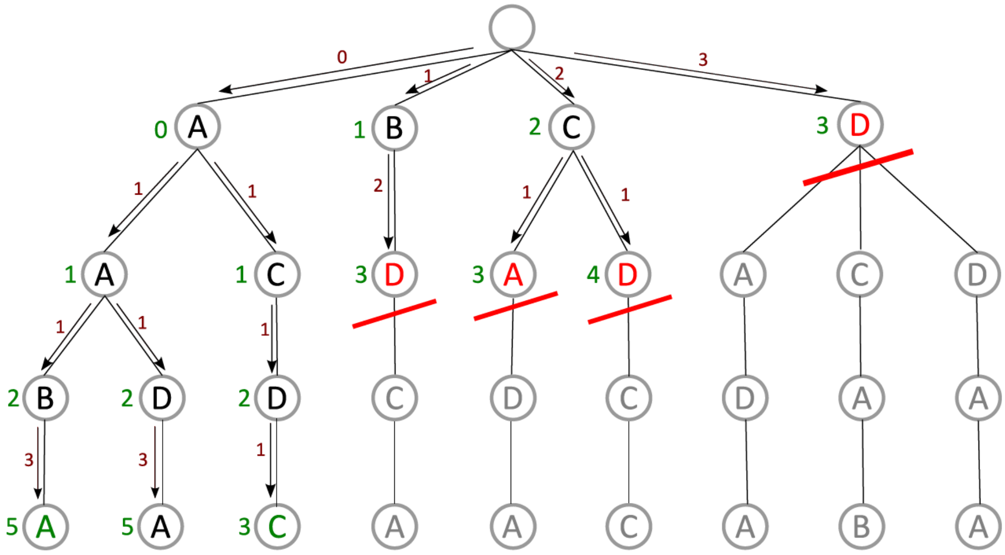
The distance between two peptides is computed as the sum of the matrix entries that correspond to the respective peptide sequences. This distance measure implies that all positions contribute individually to the overall distance. While the distance between two peptides can be calculated efficiently and in constant time, the search space gets very large when calculating of the closest distance between a peptide and a set of peptides that represent a whole proteome.

We represent all peptides of the proteome in a memory-efficient trie-based data structure. Each self-peptide is represented by a leaf in the trie. Peptides in a subtree that descend from one node have the same prefix. The depth of the trie corresponds to the peptide length and the path from the root to a leaf represents the peptide sequence. The distance between a peptide to all peptides in the trie are computed along the paths from the root to the leaves in the trie. The distance measure we use is additive so the distance can only increase in each step. This allows us to prune whole branches that only contain peptides that are too distant to be of interest. The search strategy is demonstrated in Fig. 5.3 for a reduced alphabet and words (peptides) of length four.

The tries are implemented in C++ using the boost libraries (`http://www.boost.org/`). This implementation allows us to quickly compute the distance between a peptide and a large set of peptides representing a proteome. The calculation of the distance of a target peptide to the IPI-based reference peptide set (861,352 HLA-B*35:01-binding ninemers) takes less than a second.

Not only the closest peptide but a set of close peptides contribute to negative selection. To account for this, the distance of a peptide to a set of peptides can alternatively be represented by the distances to the closest k peptides. Our implementation of the distance calculation in a trie can account for this approach by simultaneously keeping track of a set of the k closest peptides. Distance computation for a large set of peptides and for large reference peptide sets gets feasible using this implementation.

Feature encoding of self-tolerance. The *'distance to self'* of a peptide p^* is represented by the $k \gg 0$ closest peptides from the respective self proteome. We thereby can account



	A	B	C	D
A	0	1	2	3
B	1	0	1	2
C	2	1	0	1
D	3	2	1	0

Distance Matrix

Alphabet ($\simeq$ amino acids) used: A,B,C,D
Words ($\simeq$ peptides) represented by the trie:

AABA
AADA
ACDC
BDCA
CADA
CDCC
DADA
DCAB
DDAA

Word ($\simeq$ peptide) to compute distance for: **ABCD**
Minimal distance to a word in the trie: 3

Fig. 5.3: Calculation of the closest distance of a peptide to a set of peptides using a branch-and-bound strategy on a trie data structure. For simplicity, in this example we use a reduced alphabet containing the letters A,B,C,D instead of the 20 letters representing the amino acids. The distance is computed along the paths from the root of the tree to the leaves. We keep track of the smallest distance found so far. Since distance can only increase (or stay the same) when we proceed from one node to the next, we can stop at internal nodes that have equal or higher distance than the closest observed distance for a full peptide (indicated by red bars). The respective sub-branches can be pruned from the search tree (gray node labels), which reduces the search space.

for similarity distributions among the closest self-peptides. In concordance with the reported correlation between peptide-MHC affinity and peptide immunogenicity [154] we also take into account the binding affinity of the self-peptides to the MHC allele under consideration. For the target peptide p^* we first determine the distances of p^* to the k closest peptides $d(p_1), ..., d(p_k)$. We then determine the MHC binding affinities for p^* and the k peptides $b(p^*), b(p_1), ..., p(p_k)$ using netMHC [123]. The self-tolerance feature vector is given as

$$\phi(p^*) = [b(p^*), d(p_1), ..., d(p_k), b(p_1), ..., b(p_k)] \tag{5.2}$$

We choose k to be 100 and obtain a 201-dimensional feature vector.

5.3.2 Data

The selection of training data is a critical issue in the prediction of T-cell reactivity. We use a set of peptides derived from Epstein Barr Virus (EBV) antigens that are predicted to bind to HLA-B*35:01 using SYFPEITHI [116] or were selected by an expert. These peptides were tested for T-cell reactivity using ELISPOT assays. The resulting data set consists of 49 immunogenic peptides (positive examples) and 102 non-immunogenic peptides (negative examples). In oder to exclude the learning of MHC binding instead of T-cell reactivity we need to correct the dataset for MHC binding affinity. A subset of the dataset has to be chosen such that the MHC binding affinity distributions for the positive and the negative examples are very similar. For details on the selection process please refer to [149, 146]. We finally obtain a dataset that is unbiased with respect to MHC binding and consists of 49 immunogenic and 49 non-immunogenic peptides. The data set composition is depicted in Fig. 5.4.

5.3.3 Results

In order to assess the benefit of incorporating self-tolerance in the prediction of T-cell reactivity we first need a predictor that is based on the peptide sequence alone. We use support vector classification (SVC) with a Gaussian RBF kernel and a 20-dimensional amino acid encoding derived from the BLOSUM50 substitution matrix. We term this predictor blosum50 in the following. blosum50 achieves an average auROC [1] of 0.72. As a comparison, an approach based on POPI [148] (SVC with Gaussian RBF kernel,

[1] In order to measure the performance of a classification method on real data a threshold separating positive from negative examples has to be chosen. The performance depends on the choice of the threshold. The *Area under the Receiver Operating Curve (auROC)* is a threshold independent performance measure for classification. The ROC curve plots the sensitivity on the y-axis versus the 1-specificity on the x-axis as a function of the threshold. The auROC takes values from 0 to 1, a perfect classifier has an auROC of 1, a random classifier has an auROC of 0.5. All auROC values in this section are averaged over 100 runs of two-times nested five-fold cross validation.

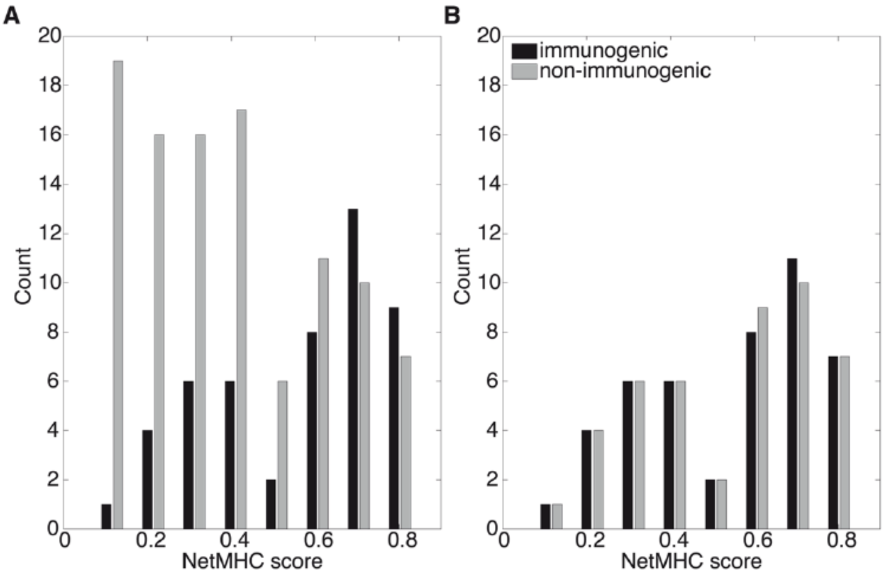


Fig. 5.4: Data set composition. netMHC score distribution for the immunogenic and non-immunogenic examples in (A) in the original dataset and (B) in the subset selected for training. Taken from [149].

physico-chemical encoding of the amino acid sequences) achieves an average auROC of 0.69.

To incorporate self-tolerance into the prediction, we add sequence-based and self-tolerance-based kernels. The self-tolerance-based kernel is a Gaussian RBF kernel on the 201-dimensional self-tolerance feature vectors as described above. The self-tolerance encodings are computed with respect to the IPI, the thymus-max and the thymus-min proteomes. The thymus-max-based self-tolerance leads to a considerable improvement (auROC=0.78), the IPI- and thymus-min-based self-tolerance models however did not improve the performance (auROCs of 0.70 and 0.71, respectively). Also in a purely self-tolerance-based prediction thymus-max outperforms the IPI and thymus-min predictors (auROCs of 0.58, 0.48 and 0.44, respectively).

All SVM computations were performed using the MATLAB interface for the freely available large-scale machine learning toolbox Shogun [155].

We compared the performance of the combined blosum50-&-thymus-max-based model to the purely sequence-based model blosum50, to POPI (a published, non-allele-specific method) and to a a HLA-B*35:01-specific reimplementation of POPI. The predictions for POPI are retrieved from the POPI-webserver (http://iclab.life.nctu.edu.tw/POPI). In contrast to our predictors, which assign the peptides to two classes (immunogenic and non-immunogenic), POPI assigns a peptide to one of four classes: non, little, moderate and high. In order to compare the POPI results to the results of our models, we consider all

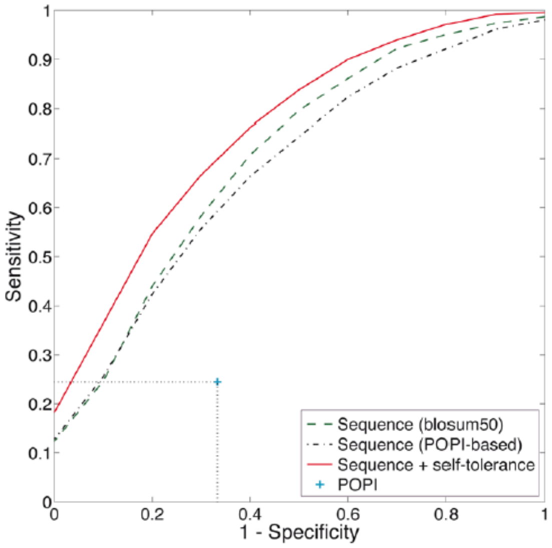


Fig. 5.5: Performance of T-cell reactivity predictors. The sequence-based predictors are the blosum50 (dashed line) and the HLA*B35:01-specific reimplementation of POPI (dash-dotted line). The sequence- and self-tolerance-based predictor uses the blosum50 and the thymus-max self-tolerance (solid line). The corresponding average performance of POPI is given (+). Taken from [149].

classes except none as immunogenic. We determine the mean sensitivity and specificity of all POPI-predictions for the peptides in our dataset. The prediction performances of all four models is shown in Fig. 5.5. It can be shown that our allele-specific predictors clearly outperform the non-allele-specific predictor POPI. The incorporation of self-tolerance further improves the prediction performance.

5.3.4 Discussion

The prediction of T-cell epitopes is of major interest for the design of epitope-based vaccines. While the prediction of MHC binding affinity as a prerequisite for T-cell epitopes performs well at least on MHC alleles with sufficient training data the correct prediction of T-cell reactivity to a given peptide is far from being a solved problem. We propose a prediction model for T-cell reactivity that takes into account MHC-restriction of T-cell epitopes and a model of self-tolerance. The incorporation of self-tolerance into an allele-specific model can considerably improve prediction performance.

Our model of self-tolerance is based on the similarity to a set of self-peptides derived from three different reference proteomes: thymus-min, thymus-max and IPI. The thymus proteomes model central tolerance and the IPI models peripheral tolerance. Modeling peripheral tolerance by the IPI proteome does not lead to an improvement of the prediction performance over the purely sequence based predictor. Modeling central tolerance by the thymus-max however improves the prediction performance. Our model of self-tolerance, that a T-cell repertoire does not contain self-reactive T cells, therefore seems appropriate to model central but not peripheral tolerance. Peripheral tolerance is more complex than central tolerance. Different mechanisms can contribute to peripheral tolerance [156]. In the periphery, clonal anergy of self-reactive T cells can be induced in the absence of costimulatory signals (e.g., B7) or inflammation. Regulatory T cells can specifically suppress T-cell reactions to antigens. These complex mechanisms of peripheral tolerance do not (solely) depend on the previous presentation of self-antigens and thus cannot be captured by our distance-to-self-based model of self-tolerance. They are, however, not yet well enough understood to be modeled explicitly.

5.4 Epitope prediction of mutated proteins

As described in the previous section computational methods are available that try to predict the potential of peptides to function as T-cell epitope. In order to identify patient- and tumor-specific potential epitopes we need the individualized tumor protein sequences and the patients HLA typing. The general procedure is depicted in Fig. 5.6.

The individualized proteins are produced as follows:

1. A list of somatic mutations of the tumor and a list mutations that also occur in the normal tissue, termed wildtype or germline mutations, of the patients are produced as described earlier (Section 4.4). These lists are filtered for mutations in coding regions that lead to an alteration of the corresponding protein sequence. The idea of tumor-specific epitopes implies that the tumor has a new epitope that is not present in normal tissue. For somatic mutations we therefore require that at least one new variant is present in the tumor compared to the normal tissue. Homozygous-to-heterozygous mutations (homozygous in the normal tissue, heterozygous in the tumor) are included, whereas loss-of heterozygosity mutations are discarded.

2. For each gene that is affected by a somatic mutation we obtain all corresponding RefSeq [90] transcripts. We then extract the protein sequences for the transcripts.

3. To account for the normal divergence between individuals we first apply all wildtype SNVs of the patient. For homozygous SNVs, we simply apply the amino acid substitution. For heterozygous SNVs, we apply all possible combinations. Each heterozygous wildtype SNV doubles the number of possible sequences. Only combining SNVs that actually occur on the same allele would be more correct. However, it is not always possible to define the correct haplotypes, especially if the distance between the SNVs is larger than the read length. In order not to miss possible tumor-specific peptides we decided to generate all possible combinations

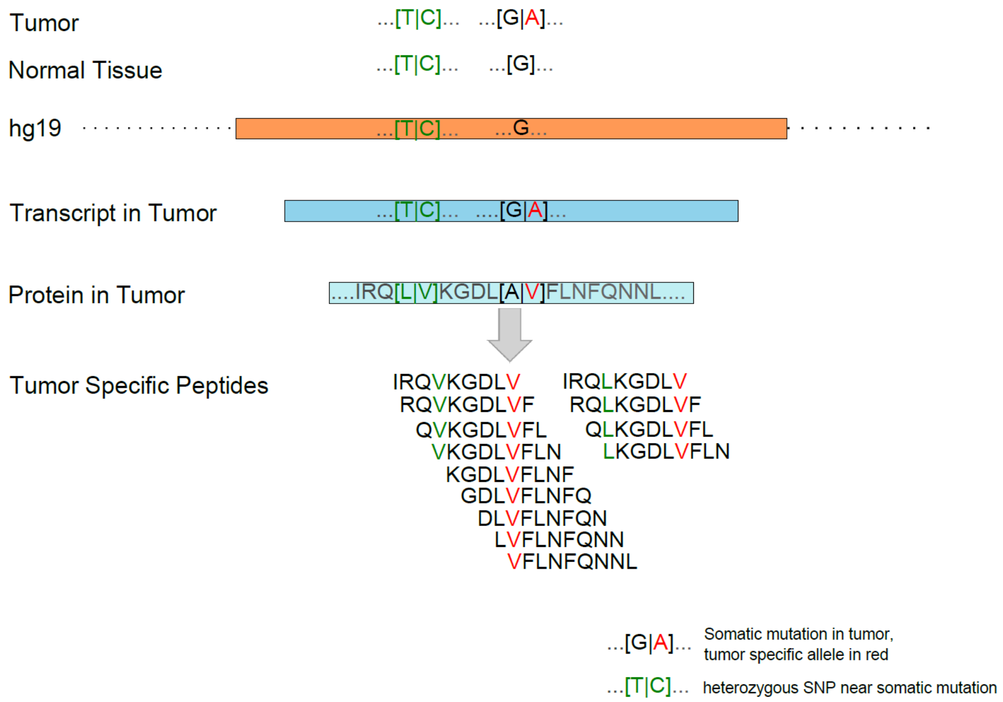


Fig. 5.6: Generation of mutated protein sequences. In the context of finding cancer-specific T-cell epitopes cancer-specific protein sequences have to be generated. We distinguish between somatic mutations and mutations that also occur in the normal tissue (for simplicity termed germline or wildtype mutations). All 9mer peptides that contain a somatic mutation are generated and submitted to epitope prediction. If a wildtype mutation occurs in proximity to a somatic mutation, this mutation is also included in the generated peptides. For heterozygous mutations all possible mutations are generated.

with the consequence of possibly generating some peptides that are not actually present in the tumor.

4. We submit the individualized sequences to an epitope prediction pipeline that accounts for the somatic mutations (see Section 6.1) and for the HLA types of the patient. We consider all peptides that contain a somatic mutation. For comparison we additionally perform the predictions for the corresponding peptide without the somatic mutation disregarding if the somatic mutation is heterozygous or homozygous.

5.4.1 Example: Prediction of tumor-specific T-cell epitopes

We integrated the generation of mutated peptides with our pipelines for epitope prediction (see Section 6.1). The results of the application of these pipelines to clinical data can be found in Section 7.4.3.

To demonstrate the prediction of potential epitopes on tumor-specific proteins, let us assume that we observe a *G* to *A* mutation at position *chr16:715,990*. This mutation is not observed in the normal tissue and is thus considered to be tumor-specific. We also observe a nearby mutation at position *chr16:716,002*, which was found in the tumor and the normal tissue and is thus considered to be a wildtype SNP. A comparison with genome annotations reveals that both mutations affect exon 36 of the coding region of gene *WDR90*. The effect of the tumor-specific SNV is a *G4475A* mutation in the respective transcript (RefSeq accession NM_145294) and a *R1492H* mutation in the respective protein. The wildtype mutation affects the same transcript and protein sequence. The mutations are *A4489T* and *Q1496L* in the transcript and protein, respectively.

The heterozygous wildtype SNP leads to two wildtype versions of the protein, one with a *Q* and one with a *L* at position 1,496. Both protein versions can also contain the tumor-specific mutation. This leads to four different versions of the resulting protein. Direct inspection of the reads mapped to the respective region could answer the question if really all different versions occur, or if the mutations occur in 'cis-' or 'trans-'conformation only. Without this information at hand, we consider all possible combinations. The four protein sequences that result from the two mutations and the resulting peptides are displayed in Fig. 5.7.

All peptides of length nine around the tumor-specific mutation are generated. This results in nine pairs (wildtype-peptide and tumor-specific-peptide) for each of the two wildtype sequences. We obtain 18 pairs of peptides (36 peptides in total, where 18 peptides are tumor-specific). Four of the peptide pairs (eight peptides) are duplicates that results from reading frames where the corresponding amino acid sequences are identical (the tumor-specific mutation is contained but not the wildtype SNP). We thus obtain 14

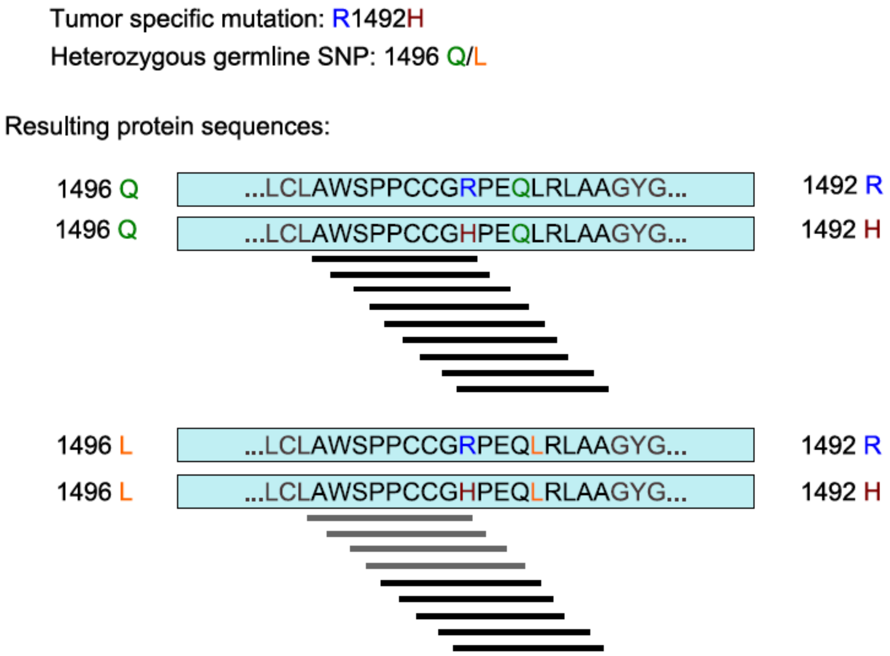

Tumor specific mutation: R1492H
Heterozygous germline SNP: 1496 Q/L
Resulting protein sequences:
1496 Q
...LCLAWSPPCCGRPEQLRLAAGYG...
1492 R
1496 Q
...LCLAWSPPCCGHPEQLRLAAGYG...
1492 H
1496 L
...LCLAWSPPCCGRPEQLRLAAGYG...
1492 R
1496 L
...LCLAWSPPCCGHPEQLRLAAGYG...
1492 H

Fig. 5.7: Generation of tumor-specific peptides from mutation data. This example includes one tumor-specific mutation and one heterozygous wildtype SNP. These mutations lead to four different protein sequences. In a reading frame of nine amino acids, all peptides that contain the tumor-specific mutation are generated. If the reading frame does not contain the wildtype SNP, duplicate peptides are generated (gray lines).

Tumor-specific peptide	Wildtype SNP	HLA binding
AWSPPCCG**H**	-	-
WSPPCCG**H**P	-	-
SPPCCG**H**PE	-	-
PPCCG**H**PEQ	-	-
PCCG**H**PEQQ	1496 Q	-
CCG**H**PEQQR	1496 Q	-
CG**H**PEQQRL	1496 Q	-
G**H**PEQQRLA	1496 Q	-
HPEQQRLAA	1496 Q	B*07:02 (10 nM)
PCCG**H**PEQL	1496 L	-
CCG**H**PEQLR	1496 L	-
CG**H**PEQLRL	1496 L	-
G**H**PEQLRLA	1496 L	-
HPEQLRLAA	1496 L	B*07:02 (21 nM)

Tab. 5.2: Tumor-specific peptides and epitope prediction results. All tumor-specific peptides that are generated around the *G* to *A* mutation at position *chr16:715,990* are listed, along with the version of the *A/T* wildtype SNP at position *chr16:716,002*. The tumor-specific peptide position is given in bold. If the tumor-specific peptide is predicted to bind to one of the patient's HLA alleles, the HLA allele and the predicted binding affinity (in nM) is listed.

tumor-specific peptides with the corresponding 14 wild-type peptides.

These peptides are submitted to HLA-binding prediction using netMHC [123]. Of the six HLA-types of the patient, only three are represented by a respective model in netMHC: A*03:01, B*07:02, and B*15:01. Two of the tumor-specific peptides are predicted to bind to one of the three HLA alleles. The tumor-specific peptides and the prediction epitope prediction results are summarized in Tab. 5.2

The peptides containing the tumor-specific amino acid at position one are predicted to bind to HLA-B*07:02. The wildtype SNP influences the predicted binding affinity slightly, but both peptides are predicted to be strong binders ($< 50\ nM$). The wild-type versions of both peptides (RPEQQRLAA and RPEQLRLAA) are also predicted to bind to HLA-B*07:02.

5.5 Discussion

In this chapter we presented computational methods for the prediction of antigen processing, binding to MHC class I and T-cell reactivity of peptides. All these steps contribute to the prediction of cytotoxic T-cell epitopes. The current prediction methods are not able to reliably predict the potential of a peptide to elicit an immune response. The main obstacles in the development of better prediction methods is the lack of experimental data needed to train prediction models and the insufficient understanding of the complex mechanisms that contribute to immunogenicity. Nevertheless, the prediction methods available today are valuable tools in the identification of promising candidates for T-cell epitopes. In combination with computational methods for the detection of genetic variation in cancer as presented in Chapter 4, these methods can be used to identify cancer-specific peptides that are potential T-cell epitopes with respect to the patients immune system, in particular to the patient's HLA types.

Good prediction methods are available for MHC binding, however other steps need further improvements. The model of self-tolerance presented in this thesis is based on a reference proteome. In a fully individualized approach, the model should be based on the patient's genome or, more precisely, on the patient's protein sequences. The information necessary for this kind of analysis is available from sequencing data of healthy tissue of the patients.

By including self-tolerance into a model to predict T-cell reactivity we have shown that including knowledge on the underlying mechanisms of immunogenicity can improve the prediction performance. Immunology is an evolving field of research and large efforts are undertaken to investigate immunogenicity and immune regulation. As the understanding of mechanisms underlying immunogenicity improves, new and improved prediction methods will be developed.

Epitope-selection strategies have been proposed for protective vaccines that are designed to cover natural variation of a pathogen and the HLA alleles present in large group of patients or even populations [157, 158]. Adapting these strategies to criteria suitable for personalized approaches is technically possible, however, the criteria have to be defined on the biological and medical level first. Including epitopes for helper T cells could be a useful approach to increase efficacy of the epitope-based vaccines.

Overall, prediction methods for immunoinformatics are valuable tools for the detection of cancer-specific epitopes, however their performance is limited by the current insight into the underlying immunological processes and by the availability of experimental data.

Chapter 6

Making immunoinformatics available for biomedical research

In the last chapters we described computational methods that can be used to detect, annotate, and interpret cancer mutations, as well as methods to identify targets for personalized immunotherapy. In order to promote cancer research, the existence of such methods alone does not suffice. A major challenge is finding a way to apply the computational methods, from basic data processing to the prediction of personalized T-cell epitopes, on a large scale and in a reliable and reproducible way. The amount of data to be processed introduce technical challenge concerning data transfer, data precessing, and data storage. The data analysis pipelines have to be reliable, reproducible, and able to perform the analyses in a timely manner. A second major challenge is that the results of the computational analysis have to be presented in a comprehensive and intuitive way to biomedical researchers. Results from different single computational analysis tools have to be combined and integrated.

In this chapter approaches to automate, facilitate, and integrate methods for epitope prediction (Section 6.1) and a pipeline for the high-throughput prediction of miHAs (Section 6.2) are introduced. We present VariationDB, a web-based tool for a comprehensive and integrated presentation of genetic variations in cancer (Section 6.3). In addition, we show how pipelining and workflow systems can be used to integrate NGS data analysis, epitope prediction and convenient presentation of the results (Section 6.4).

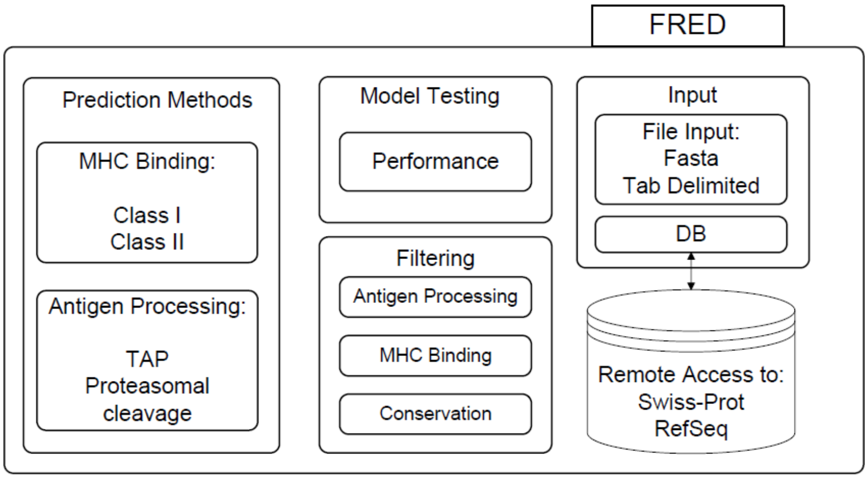


Fig. 6.1: FRED is organized in four main parts: sequence handling, application of prediction methods, filtering of the results, and model testing. Figure taken from [159].

6.1 Epitope prediction

6.1.1 FRED - a Framework for Epitope Detection

Many computational prediction methods for antigen processing, MHC binding, and T-cell reactivity are available today (see Section 5). Some of the methods are freely available over webservers or as stand-alone programs. Integrating and combining the prediction methods into larger pipelines is of great importance in immunological research. Web-based interfaces can easily and conveniently be applied in a small scale. On a large scale (e.g. in the prediction of MCH binding for a whole human proteome as we performed it for our model of self-tolerance in Section 5.3) is not feasible over web interfaces. The comparison and combination of prediction methods is challenging, since the different methods have different input and output formats and interfaces. To facilitate the access to and integration of computational prediction methods related to T-cell epitope prediction we developed the software framework FRED [159]. FRED offers a uniform interface for a variety of prediction methods and allows for an easy and quick implementation of tailor-made prediction pipelines.

Design and implementation of FRED

FRED is a software framework that provides methods for sequence input, sequence processing, filtering, comparison, and display of the prediction results. The framework is implemented in Python (www.python.org, release 2.6). The general organization of FRED is depicted in Fig. 6.1.

Tab. 6.1: Prediction methods currently integrated in FRED. * Installation of external software is required. Due to licensing issues, we could not include the standalone versions of these methods in the publicly available FRED package.

MHC Binding:		**Proteasomal Cleavage:**	
SYFPEITHI	[116]	PCM method from WAPP	[113]
SVMHC	[124]		
BIMAS	[160]	**TAP Transport:**	
NetMHCpan *	[147]	Additive Matrix Method	[161]
NetMHC *	[123]		
Hammer	[117]		
NetMHCIIpan *	[162]		

Sequence handling and sequence input are decoupled from the prediction methods. Access to the prediction methods is handled internally and the user is presented with a single, consistent interface that allows simultaneous use of prediction methods. Functionality for benchmarking and comparing prediction methods are also provided. FRED can handle polymorphic sequences to assess the influence of mutations on potential T-cell epitopes.

FRED integrates prediction methods for antigen processing, MHC binding and T-cell reactivity and is easily extensible to access external prediction methods. A list of prediction methods that are included in the first release of FRED is given in Tab. 6.1.

A key issue when comparing prediction methods for MHC binding is that the different methods use different scoring schemes and thresholds. We propose an approach to make thresholds comparable across different methods. For all prediction methods included in the original version of FRED we determined thresholds based on background score distributions that were computed on a large set of peptides from natural proteins. Based on these background distributions a threshold can be selected that gives percentage to which the background peptide set is predicted to bind to the respective HLA allele.

Application of FRED

Prediction pipelines can easily be implemented with short python scripts using FRED. We demonstrate the application of FRED for a vaccine design scenario based on a publication by Toussaint *et al.* [157]. The aim is to select a set of conserved peptides as candidates for an epitope-based vaccine against the hepatitis C virus. As input we use a set of sequences of the hepatitis C virus core protein that originate from four different subtypes. All peptides that occur in at least 90% of the input sequences are considered as candidate peptides. MHC binding predictions are performed for 29 HLA alleles using the BIMAS method [160]. This task can be implemented in a very short script based on FRED (see Listing 6.1).

Listing 6.1: Example python script that uses FRED to predict conserved epitopes in a set of viral sequences.

```
models = [(’MHC_I_BIMAS’,’A_0201_9’),(’MHC_I_BIMAS’,’A_1101_9’),...
prot_set= Fred.Fred.ProteinSet()
prot_set.readFasta("hcv-core-1a1b2a3a.fasta")
pep_set = protset.getPeptides(9, 0.9, 0)[0]
pep_set.predict(models)
cands = Fred.Fred.FindCandidates_HalfmaxThresholds(pep_set,models,0.9,1)
cands.writeTabDelimitedFile(’results_hcv_core.txt’)
```

6.1.2 EpiToolKit

While FRED is a useful tool for immunoinformaticians that want to implement large-scale predictions, the application of a python framework by biomedical researchers is not realistic. As a complementary tool we therefore implemented the intuitive web service EpiToolKit [163] (www.epitoolkit.org). The web service is based on FRED and provides similar functionality but is usable over the internet without the need for any software installation or programming. The user interface was developed in close cooperation with biomedical researchers in order to improve its usability.

EpiToolKit is a web server that offers simultaneous access to different state-of-the-art prediction methods. The general prediction workflow implemented in EpiToolKit is depicted in Fig. 6.2.

EpiToolKit offers two prediction pipelines, one for the prediction on normal protein sequences and one for predictions on polymorphic proteins (SNEPv2). The two pipelines have separate sequence input and result pages. Input sequences, together with known polymorphisms can directly be retrieved from public sequence databases (SWISS-PROT [135] and RefSeq [90]). Sequences and polymorphisms can be reviewed and selected after retrieval. Allele selection is conveniently realized in a tree structure. The tree of available HLA alleles is built for the prediction methods selected by the user. EpiToolKit offers different methods for filtering of the results and threshold generation. Prediction methods, alleles, and filtering thresholds can be adapted in the advanced options.

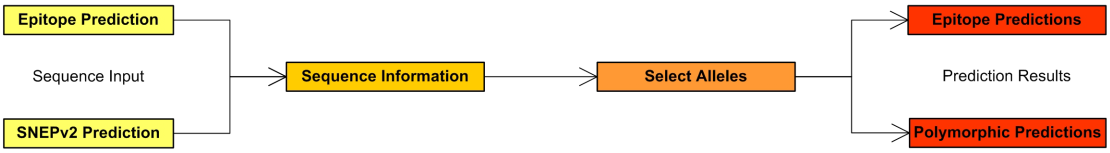
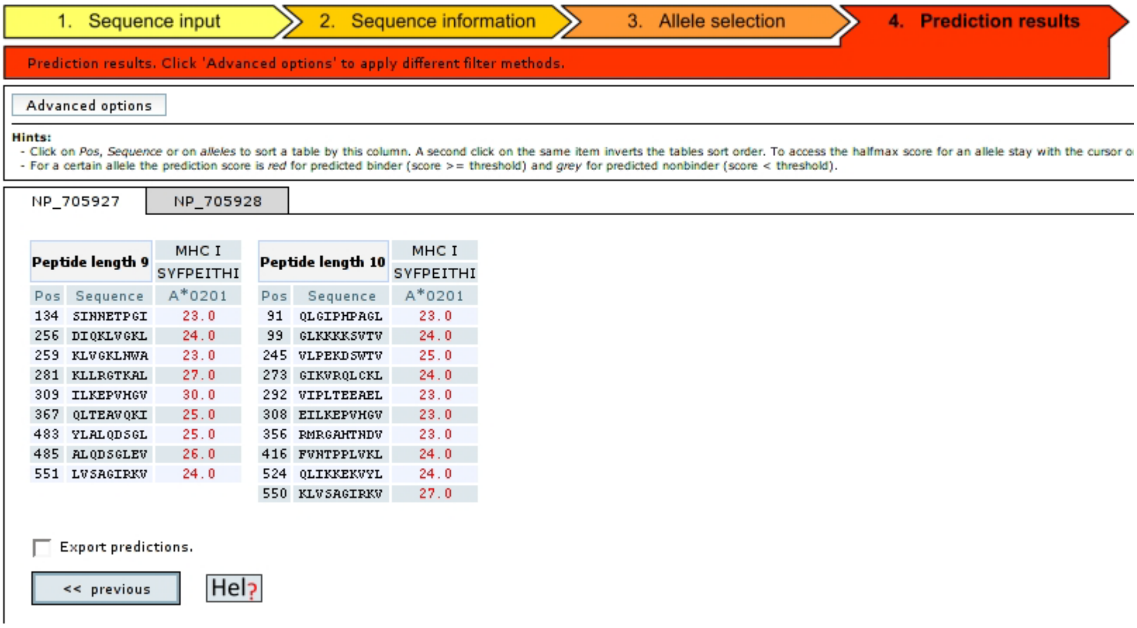


Fig. 6.2: EpiToolKit general prediction workflow. EpiToolKit offers two prediction pipelines, one for the prediction on normal protein sequences and one for predictions on polymorphic proteins. The two pipelines have separate sequence input and result pages.



Prediction results. Click 'Advanced options' to apply different filter methods.

Hints:
- Click on *Pos*, *Sequence* or on *alleles* to sort a table by this column. A second click on the same item inverts the tables sort order. To access the halfmax score for an allele stay with the cursor o
- For a certain allele the prediction score is *red* for predicted binder (score >= threshold) and *grey* for predicted nonbinder (score < threshold).

Peptide length 9		MHC I
		SYFPEITHI
Pos	Sequence	A*0201
134	SINNETPGI	23.0
256	DIQKLVGKL	24.0
259	KLVGKLNWA	23.0
281	KLLRGTKAL	27.0
309	ILKEPVHGV	30.0
367	QLTEAVQKI	25.0
483	YLALQDSGL	25.0
485	ALQDSGLEV	26.0
551	LVSAGIRKV	24.0

Peptide length 10		MHC I
		SYFPEITHI
Pos	Sequence	A*0201
91	QLGIPHPAGL	23.0
99	GLKKKKSVTV	24.0
245	VLPEKDSWTV	25.0
273	GIKVRQLCKL	24.0
292	VIPLTEEAEL	23.0
308	EILKEPVHGV	23.0
356	RMRGAHTNDV	23.0
416	FVNTPPLVKL	24.0
524	QLIKKEKVYL	24.0
550	KLVSAGIRKV	27.0

□ Export predictions.

Fig. 6.3: EpiToolKit result page for polymorphic predictions.

For predictions on polymorphic sequences the results can be displayed such that only peptides that are affected by the polymorphism are shown, as depicted in Fig. 6.3. Alternatively, the full predictions for the whole protein sequence disregarding the polymorphisms in the sequence can be displayed.

For large prediction jobs the user can request an email notification of job completion. Prediction results can be accessed for a week after starting the job. Results can be exported in CSV or Microsoft Excel format.

6.2 Identification of novel minor histocompatibility antigens

Minor histocompatibility antigens (miHAs) arise from differences in HLA-presented peptides between the donor and patient of organ transplants, often caused by SNPs affecting the amino acid sequence or altering the expression of proteins [164]. Allogenic hematopoietic stem cell transplantation (alloHCT) is a well established therapy for certain hematologic malignancies. After alloHCT residual T cells of the stem cell graft can lead to detrimental graft-versus-host-disease (GvHD) but also to beneficial graft-versus-tumor (GvT) or graft-versus-leukemia (GvL) effects. While alloHCT was initially used to support hematopoiesis after chemotherapy, GvL or GvT effects are now a main goal of alloHCT [22]. GvHD however is a severe side effect that needs to be avoided. The risk of GvHD depends for example on the tissue distribution of the antigen. miHAs presented exclusively on malignant cells are more likely to give precise GvL or GvT effects, whereas miHA expressed in many different tissues may lead to GvHD [165].

Several strategies have been applied in recent years for applying miHAs in order to elicit GvL or GvT effects [166], e.g., transfer of miHA-specific T cells and miHA peptide, protein, mRNA or DNA vaccination. One of the most feasible and efficient current approaches is vaccination with defined miHA peptides and longer peptides pulsed onto DC after alloHCT treatment [22].

A major challenge in this field of immunotherapy is the relatively low number of characterized miHAs in relation to the number of HLA alleles available and matching in donor-patient pairs. There is a substantial need for fast and accurate identification of novel miHAs to enable immunotherapy for a large number of patients. Computational methods from immunoinformatics can help to overcome the major bottlenecks in this process: access to existing SNP information, analysis of next-generation sequencing data to obtain polymorphisms, and determination of potentially presented HLA-binding peptides acting as miHAs even for infrequent HLA alleles.

In close collaboration with clinical experts we developed a computational pipeline to facilitate the identification of novel miHAs. The pipeline addresses several of the current problems in the identification of miHAs that are likely to be effective in GvL and GvT treatment. The pipeline is well-suited for both large-scale screening for novel miHAs based on existing SNP data, as well as for personalized settings where genomic differences between the donor and the patient are known. The screening can be used to design a customized genotyping assay that allows for a quick and cheap identification of miHAs that are relevant in a patient-donor pair.

Prediction of miHAs

We define a peptide to be a candidate miHA if 1) the peptide is affected by a SNP and 2) if the peptide is predicted to bind to at least one of the HLA alleles under consideration.

We first have a closer look at the first condition. Strictly speaking, a miHA can only be defined in a specific patient-donor pair. By definition, a miHA in the context of alloHCT is a peptide that is presented by HLA in the patient, but not in the donor. In the context of graft rejection, a miHA is a peptide that is presented in the graft but not in the patient. Since we assume patient and donor to be HLA-matched, the difference in presentation has to be caused by a genetic variation (usually a SNP) between patient and donor. This variation needs to change the peptide sequence (i.e., it has to be a non-synonymous SNP) and leads to a presented peptide not previously seen by donor T cells.

In Tab. 6.2 we show the influence of genotype combinations of donor and patient on the miHA relevance of SNPs.

There are two main settings for the computational identification of miHAs. The first one is personalized therapy, where information on genetic differences between a donor

Tab. 6.2: For SNP rs142901306 we show the combination of donor and patient genotypes decides if a SNP is relevant for miHA analysis for a specific donor-patient pair. For each genotype the corresponding amino acids are listed in brackets. In alloHTC settings a SNP is relevant if the patient expresses one peptide version that is not present in the donor.

Donor \ Patient	G/A (E/K)	G (E)	A (K)
G/A (E/K)	No	No	No
G (E)	**Yes**	No	**Yes**
A (K)	**Yes**	**Yes**	No

and a patient is at hand. The second setting is a large-scale screening for potential miHAs for a larger cohort of patients that can be turned into a rapid diagnostic test without the need for individualized genome sequencing. In the latter we computationally identify all potential miHAs for suitable set of genes, for example genes that are specifically expressed in hematopoietic cells. This strategy can be used to design diagnostic SNP-genotyping tests for the quick identification of relevant miHAs for specific donor-patient pairs. The screening identifies a feasible number of SNPs for genotyping. The results of the genotyping for a donor-patient pair can then be analyzed in the individualized setting. Such combined approaches can increase the chance of finding miHAs also for patients with infrequent HLA alleles while minimizing experimental effort to the most promising candidates.

We first describe our pipeline for the computational detection of miHAs and then show in an application case how the pipeline can be applied to clinical settings.

The second part of the definition of candidate miHAs is that the peptides are predicted to bind to at least one of the HLA alleles under consideration. In an individualized setting the HLA types of donor and patient are known. The relevant HLA alleles are those that match between donor and patient. In a screening setting the HLA alleles can be chosen to cover a patient cohort or a whole population.

6.2.1 Pipeline for the *in silico* prediction of miHAs

The workflow for the prediction of candidate miHAs is divided into three main steps, as illustrated in Fig. 6.4: (1) Retrieving polymorphisms of interest, (2) generating the peptides affected by the polymorphisms (3) and predicting HLA binding peptides for a given set of HLA alleles. A more detailed description of the three steps, along with a discussion of the different options and issues associated with each step, are given in the following.

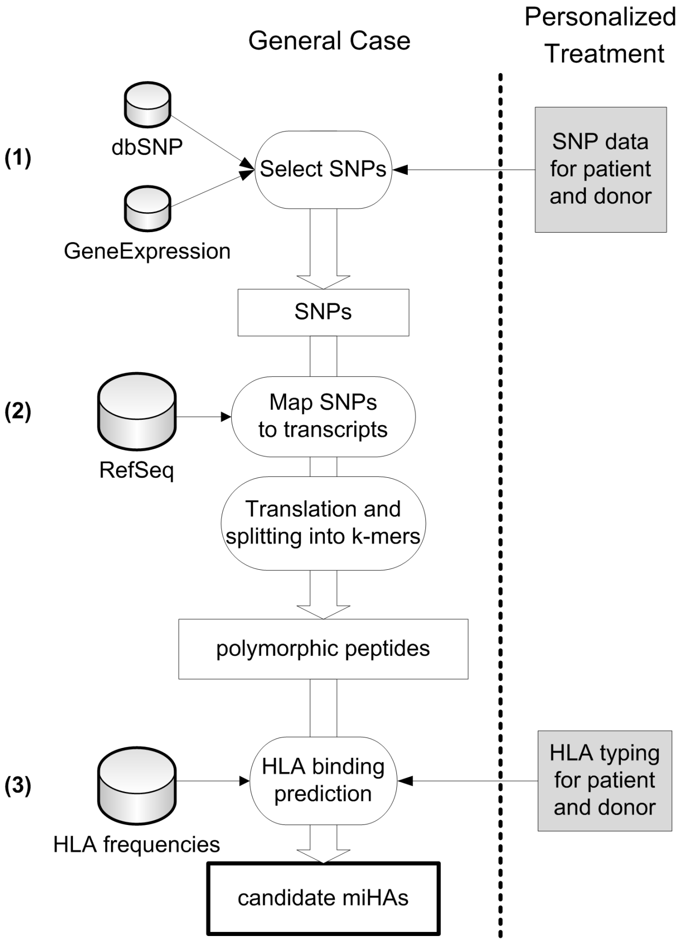


Fig. 6.4: Immunoinformatics pipeline for the identification of potential miHAs. (1) Obtain a set of polymorphisms of interest. In the general setting, SNPs can be retrieved from databases (e.g., dbSNP). In a personalized setting, a list of SNPs with genotyping information for patient-donor pairs is required. In the latter case only SNPs with disparity between patient and donor are considered. Gene expression data can be included to select tissue-specific SNPs. (2) The polymorphisms selected are mapped to the respective transcripts. All peptides containing the polymorphic position are generated from the transcript sequences. (3) Prediction of HLA binding for a set of HLA alleles (based on frequency information in the general case or on patient/donor HLA types) is performed on the polymorphic peptides.

Step 1: Obtain polymorphisms of interest. As input for the miHA analysis pipeline we need a set of polymorphisms that are of interest for the respective clinical setting. As outlined in the beginning of this section the selection of polymorphisms depends on the availability of genotyping information of donor-patient pairs.

In large-scale screenings where no genotyping information is available we use polymorphism data from dbSNP [78]. We start with a set of genes, e.g. genes that are specifically expressed in one tissue. For this set of genes we retrieve all known SNPs from dbSNP, along with additional information on allele frequencies and functional class of the SNP (for example whether the SNP is synonymous or non-synonymous). The SNPs can be filtered based on allele frequencies and functional class.

In individualized settings genotyping information for a donor-patient pair is available before the prediction process. We then restrict the analysis to those SNPs that differ between donor and patient as shown in Tab. 6.2. Excluding SNPs without disparity from the analysis can reduce experimental effort and increase the success rate of the *in silico* detection of miHAs [167].

Step 2: Generating polymorphic peptides. In the second step of the pipeline we generate all polymorphic peptides that result from the polymorphisms selected in the first step. For the genes affected by the polymorphisms we retrieve all transcript sequences from RefSeq [90]. Different transcript or protein isoforms can lead to different peptides. We therefore take all transcripts that are available for each gene into account. We then apply the polymorphisms *in silico* to the transcripts. We also take all possible combinations of polymorphisms into account if more than one polymorphism was chosen for the same transcript. From the polymorphic transcript sequences we generate all peptides that are affected by the polymorphic position. Per default we only translate in the normal reading frame. Alternative reading frame translation and the inclusion of synonymous SNPs is also possible. Duplicate peptides generated from different transcripts are removed before submitting the peptides to the prediction step. The length of the produced peptides depends on the methods used for HLA-binding prediction. Since the majority of all HLA binding peptides are ninemers the default peptide length is nine. For personalized studies we directly account for the patient-donor genotypes.

Step 3: Predicting candidate miHAs. In the last step of the pipeline we apply HLA binding-prediction to the set of peptides that was selected in the previous steps. By default we apply the MHC binding prediction method netMHCpan [147]. We choose netMHCpan since it provides predictions for a wide range of HLA alleles. Our pipeline is based on FRED, a framework for epitope prediction [159], and is therefore very flexible with respect to the prediction method to use.

In individualized studies, HLA typing of the patient and the donor is available. In such cases all HLA alleles shared by donor and patient are relevant. In a more general setting HLA alleles can be selected according to HLA allele frequencies in a population based on, e.g., dbMHC [168]). Our pipeline supports the use of compute clusters to perform large-scale screening for large numbers of SNPs and HLA alleles in a timely manner.

Binding to HLA is only a prerequisite for a peptide to function as a T-cell epitope. Additional factors are antigen processing and the availability of a T-cell receptor that matches the peptide:HLA complex. While good prediction methods for HLA binding are available, the prediction of antigen processing and T-cell reactivity leave room for improvement. We therefore restrict the epitope prediction to the prediction of HLA binding in the default settings. Incorporation of antigen processing and T-cell reactivity is technically possible if appropriate prediction methods are available.

6.2.2 Prioritizing miHAs

In personalized settings the number of disparate SNPs between patient and donor limits the number of predicted miHAs. A prioritization of miHAs in these cases is probably not necessary. In settings where SNPs for a gene list are extracted from a database the number of potential miHAs can exceed the number of miHAs that can be tested experimentally with acceptable effort. The number of SNPs known per gene drastically increases with each release of the dbSNP. In order to reduce experimental effort the miHAs need to be prioritized. We present criteria for filtering and ranking of miHAs. As for all other parameters and settings, SNP prioritization needs to be adapted to the clinical application.

miHAs can be prioritized on the SNP level. SNPs retrieved from dbSNP are annotated with additional information, e.g., the allele frequencies of the observed alleles. If the observed allele frequencies for a SNP are balanced, the chance of observing a disparity for this SNP in two individuals is increased. Another possible criterion to rank SNPs is the number of potential miHAs that are produced by a SNP. SNPs leading to many peptides that are presented by different HLA alleles are more likely to be relevant in a larger number of patient-donor pairs.

Prioritization can also be applied on the peptide level. Peptides can be ranked by the predicted binding affinity to the respective HLA allele. Promiscuous peptides that are predicted to bind to more than one of the patients alleles have a higher miHA potential than peptides that bind to just one allele. Including antigen processing prediction or T-cell reactivity prediction is also possible. However the usefulness of this strategy strongly depends on the availability of appropriate prediction methods.

Another criterion that could be included is the distance to self of the peptide. If a potential miHA is very close to other peptides in the self proteome, T cells recognizing

Tab. 6.3: Number of predicted miHAs for dnSNPv132 and dbSNPv135 for the 79 genes proposed in [167]. Genes and SNPs are counted if they lead to at least one peptide that is predicted to bind to at least one of the HLA alleles.

	dbSNPv132	**dbSNPv135**
SNPs found	654	2,036
9mers produced	14,562	62,655
SNPs with predicted miHAs (10 alleles)	493	1,598
SNPs with predicted miHAs (A*02:01)	201	651
Genes with predicted miHAs (10 alleles)	75	77
Genes with predicted miHAs (A*02:01)	61	73

this peptide:HLA complex will most probably be deleted from the T-cell repertoire. Approaches for assessing the distance to self for a peptide exist have been described in Section 5.3.1. However, computation of distance to self currently relies on a reference proteome and ignores natural variation between individuals.

In order to illustrate the need for miHA prioritization and the effects of the presented filters we apply our pipeline to a set of hematopoiesis-restricted genes proposed by Hombrink *et al.* [167]. For these genes we retrieve all non-synonymous SNPs from the dbSNP. We use and compare two releases of dbSNP, dbSNP132 and dbSNP135. Peptides of length nine are generated in normal reading frame translation. We select the five most frequent HLA-A and HLA-B alleles based on www.allelefrequencies.net and dbMHC [168]): HLA-A*02:01, HLA-A*01:01, HLA-A*03:01, HLA-A*24:02, HLA-A*11:01, HLA-B*07:02, HLA-B*08:01, HLA-B*44:02, HLA-B*35:01, HLA-B*15:01. We apply netMHCpan with a binding threshold of IC50 $\leq$ 500 nM for HLA binding predictions. The results of the prediction are summarized in Tab. 6.3.

These results show that the information on known SNPs provided by public databases like dbSNP is currently increasing. For the 79 genes in our analysis 654 non-synonymous SNPs were found in dbSNPv132, in dbSNPv135 we found 2,036 SNPs. If we consider the ten alleles listed above, we predict miHAs for almost all genes in the list. With an increasing number of SNPs that are included in the analysis we also obtain a larger number of predicted miHAs. The number of predicted miHAs is already too large for experimental testing and the number of known SNPs will keep on increasing in the future. SNPs and miHAs have to be filtered or prioritized to restrict experimental testing to the most promising candidates.

The presented filtering criteria can be adapted to different applications, e.g. the number of SNPs or miHAs that can be tested. In the following we demonstrate the effects of some of the filtering criteria. All numbers presented below refer to the prediction for the ten alleles and for SNPs retrieved from dbSNPv135.

SNP allele frequencies. A prerequisite for a miHA being present in a donor-patient pair is a disparity in a SNP between the two individuals. The chance of observing a disparity is higher if the allele frequencies of a SNP are balanced. Most SNPs in dbSNP are annotated with allele frequencies. SNPs can be filtered or prioritized based on the minor allele frequency. By requiring a minor allele frequency of at least 30% we can reduce the number of SNPs to consider from 1,598 to 55.

Number of predicted miHAs per SNP. The number of different peptides that are affected by a SNP and are predicted to bind to at least one of the HLA alleles under consideration varies between 1 and 40. A single SNP with two alleles can lead to at most 18 different peptides of length nine. (For each allele of the SNP there are nine peptides of length nine that can contain the SNP position.) A larger number of peptides per SNP is observed if two SNPs lie within a reading frame of nine amino acids. We then consider all possible peptides that arise from the combination of the two SNPs. By requiring a SNP to produce at least five predicted miHAs we can reduce the number of SNPs from 1,598 to 423.

Number of HLA alleles covered by a SNP. We define a HLA allele to be covered by a SNP if at least one of the peptides derived from the SNP is predicted to bind to this HLA allele. We count the number of different HLA alleles that are covered by a SNP following this definition. We find SNPs that cover nine out of the ten alleles, other SNPs only cover one allele. For some SNPs the number of covered alleles exceeds the number of predicted miHAs. This indicates that we observe promiscuous miHAs that are predicted to be relevant for more than just one HLA allele. HLA allele frequencies can be taken into account to select those SNPs that lead to peptides that are predicted to bind to the most prevalent HLA alleles.

Validation

To show the validity of our pipeline in a large-scale screening setting we compare the results to those presented in Hombrink *et al.* [167]. We only consider peptides of length nine that are translated in the normal reading frame and predictions for HLA-A*02:01 for comparison. We can reproduce 160 of the 177 (90.4%) predicted miHAs (Tab. S3 in [167]). Hombrink *et al.* also analyzed the predicted miHAs with respect to genotyping information for selected donor-patient pairs (Tab. 1 in [167]). We again exclude peptides that come from alternative reading frame translation or are not of length nine. We can reproduce four out of the remaining five peptides. The one peptide we miss is due to discrepancies between the used prediction methods. These results show the validity of our pipeline but also the influence of the used parameters and prediction methods.

To validate our pipeline in an individual setting we apply our miHA identification pipeline on data provided by van Bergen *et al.* [169]. For two patient-donor pairs we are given a list of SNP disparities and the HLA typing of the patients. We use the SNPs that were shown to be associated with a miHA (Tab. 3 in [169]). Polymorphic peptides of length nine are generated around the SNPs taking into account the genotypes of the respective patient and donor. Binding affinity predictions are performed using netMHCpan. We compare our predictions to the results reported by van Bergen *et al.* The results of the comparison are summarized in Tab. 6.4.

Van Bergen *et al.* found seven miHAs for patient H and three miHAs for patient Z. We excluded one of the miHAs for patient H (clone H2) from the analysis since van Bergen *et al.* could not find a peptide derived from the respective SNP that is predicted to bind of one of the patients HLA alleles. (The peptide reported in the column "Predicted HLA binding peptides" in Tab. 3 in [169] for this clone is actually not predicted to bind and was manually selected for testing due to anchor position that match a binding motif for HLA-A*02:01. The peptide could however not be confirmed as miHA. Personal communication with C. van Bergen). A second miHA for patient H (clone H10) was excluded since it is derived from an alternative reading frame translation and cannot be detected with our settings. The remaining five miHAs for patient H were correctly identified by our pipeline.
We could directly reproduce one of the miHAs for patient Z (Z1). We could not reproduce the peptides of length nine for Z2 and Z3. The nonameric peptide for Z2 is not predicted as a binder. These peptides therefore cannot be detected by our pipeline. We could, however, identify a predicted binder of length nine for the same allele that is contained in the peptides of length 10 and 11 reported by von Bergen. For miHA Z3 we did not find a binding ninemer. The peptides of length 10 and 11 reported by van Bergen however are predicted to bind to HLA-B*07:02 by netMHC and netMHCpan.

These results show that we are able to reproduce all miHAs identified by van Bergen *et al.* that match our criteria. The number of disparate SNPs given by van Bergen *et al.* is rather small. The general availability of information on genetic variation will increase in the future. With advances in sequencing technologies sequencing of complete exomes or even genomes for donor-patient pairs becomes a reachable goal. The broad application of the individualized setting described above will therefore gain importance.

In Chapter 8 we demonstrate the application of our pipeline in a clinical setting. We first perform a screening step to identify and select a feasible number of SNPs for genotyping. The results of the genotyping for a donor-patient pair can then be analyzed in the individualized setting. This approach can increase the chance of finding miHAs also for patients with infrequent HLA alleles while minimizing experimental effort to the

Tab. 6.4: Identification of the miHAs identified by van Bergen. The first six columns list the information taken from Table 3 in van Bergen *et al.* The last column lists the peptides identified as potential miHA by our pipeline. We can reproduce all miHAs that match our requirements.

Clone type	**SNP**	**Polymorphism**	**HLA**	**Predicted HLA binding peptides***	**miHA predicted by our pipeline**
H1	rs12828016	Ile/Met	A*02:01	TLSPEIITV	TLSPEIITV
H2	rs10004	Ser/Leu	A*02:01	SLAVAQDLT **	-
H3	rs2298668	Asp/Glu	A*02:01	FMWDVAEDL	FMWDVAEDL
H8	rs26653	Arg/Pro	B*07:02	HPRQEQIAL	HPRQEQIAL
H9	rs11548193	Arg/Thr	B*07:02	QPRRALLFV GVSQPRRAL	QPRRALLFV GVSQPRRAL
H10	rs4703	Arg/Pro	B*07:02	LPRACWREA ***	-
H11	rs2986014	Phe/Leu	B*07:02	GPDSSKTFL	GPDSSKTFL
Z1	rs4968104	Val/Glu	B*07:02	FPALRFVEV	FPALRFVEV
Z2	rs4740	Ile/Val	B*07:02	RPRARYYIQ **	RARYYIQVA
Z3	rs2076109	Lys/Glu	B*07:02	KPQYHAEMC **	-

*Peptides listed in the column "Predicted HLA binding peptides" from Table 3 in [169]. We only included peptides of length 9.
**Peptide not predicted to bind my netMHC, although they are listed in the predicted binder column by van Bergen.
*** Peptide generated from alternative reading frame.

most promising candidates.

The application of our pipeline is not restricted to stem cell transplantation. A similar approach could also be used to assess the risk of graft rejection based on miHAs that are presented on the transplanted tissue. Some modification in the settings of the pipeline are necessary for assessing the risk of, e.g., the rejection of a liver transplant in HLA-matched donors and patients. We now need to take into account all genes that are expressed in the liver, not only liver specific genes. The difference between these two sets is, that liver-specific genes are solely expressed in the liver, whereas the second set contains all genes that are expressed in the liver but can also be expressed elsewhere. In case of GvL or GvHD after stem cell transplantation the T cells of the donor react against tissue of the patient due to miHAs that are present in the patient but not in the donor. In graft rejection, the immune system of the patient reacts against miHAs that are presented on the transplanted organ but not in the patient. We simply need to invert the direction of the miHA definition: all SNPs where the donor has a variant that is not present in the patient are relevant. Today, this potential application is not relevant in the clinic. If genome sequencing becomes a standard step in clinical treatment, the data for such an analysis would be at hand, though.

6.3 VariationDB

The aim of VariationDB is to provide an easy-to-use and intuitive user-interface for the presentation and analysis of mutation data for cancer genomes. The intended users of VariationDB are biomedical researchers and clinicians. The challenging and computationally intensive data analysis normally cannot be performed by biomedical researchers directly. Sufficient technical equipment with respect to storage and computation resources and experience with the computational tools is needed in order to produce useful results. For the analysis of genomes of many patients the analysis has to be automatized to ensure reproducibility. However, even if such technical resources and analysis pipelines are at hand, the raw results produced by the pipelines - usually just lists of mutations in some file format that is easily readable for machines but not for humans - are not of great use for biomedical researchers. The large amounts of data produced by the computational analyses cannot be inspected without appropriate tools. What is needed to close the gap between raw results and the users is a platform to present the results conveniently, ideally providing additional annotation data at the same time.

To cover that need we developed VariationDB in which we present the results of our integrated mutation analysis of cancer genomes to the clinicians who work with the patients. Together with the information on observed mutations we provide annotation that is needed to interpret the mutations (see Section 4.4). Mutated protein sequences can

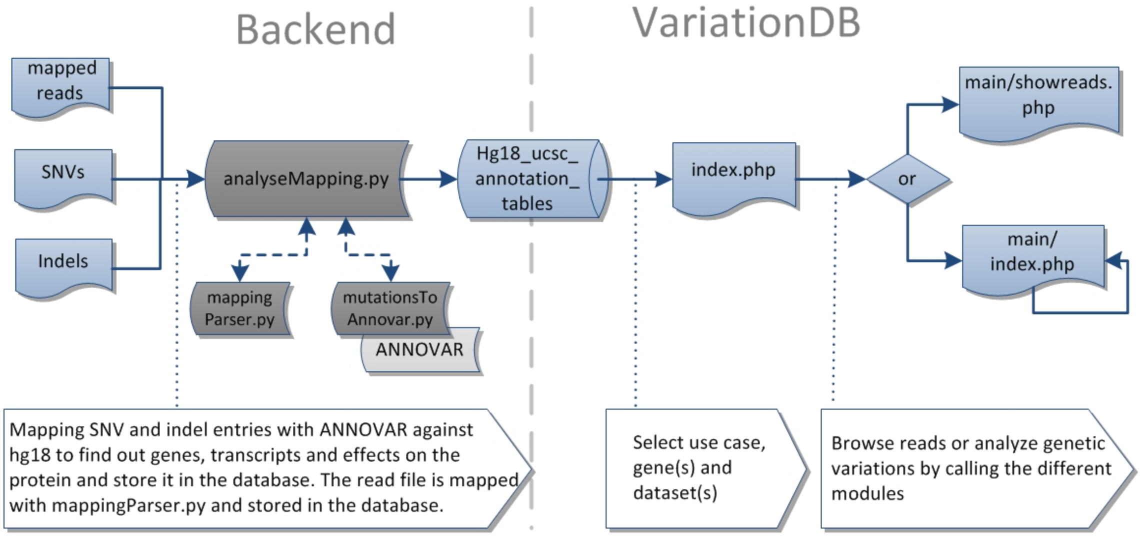

Backend
VariationDB
mapped reads
SNVs
Indels
analyseMapping.py
mapping Parser.py
mutationsTo Annovar.py
ANNOVAR
Hg18_ucsc_ annotation_ tables
index.php
or
main/showreads. php
main/ index.php
Mapping SNV and indel entries with ANNOVAR against hg18 to find out genes, transcripts and effects on the protein and store it in the database. The read file is mapped with mappingParser.py and stored in the database.
Select use case, gene(s) and dataset(s)
Browse reads or analyze genetic variations by calling the different modules

Fig. 6.5: The general workflow of the VariationDB project. Figure taken from [105].

directly be submitted to epitope prediction. The user interface was developed in close collaboration with colleagues in biomedical research to ensure that we meet the needs of the users.

6.3.1 Design and implementation

The technical constraints for the implementation of VariationDB were that the computational analysis and the presentation of the results to clinicians are locally separated. Computation is performed on compute clusters in the computer science department. The results, however, should be directly accessible to the users in the clinics without the need for massive data transfer or installation and maintenance effort. Another constraint was that multiple users should be able to work with the same data. To match these requirements we implemented VariationDB as a web-based tool. All the user needs is a web browser and an internet connection. All computation is performed on the server side. VariationDB is divided in two self-contained parts: a backend that is responsible for data processing and data storage, and the web application that is visible to the user. The backend can be integrated with the analysis pipelines for variation detection. The general workflow for VariationDB is depicted in Fig. 6.5.

VariationDB - Backend

The input for the VariationDB pipeline are lists of mutations that are generated from the analysis of sequencing data. These files are produced by our NGS data processing pipeline (see Section 6.4.2). The backend takes care of annotating the mutation data and integrating

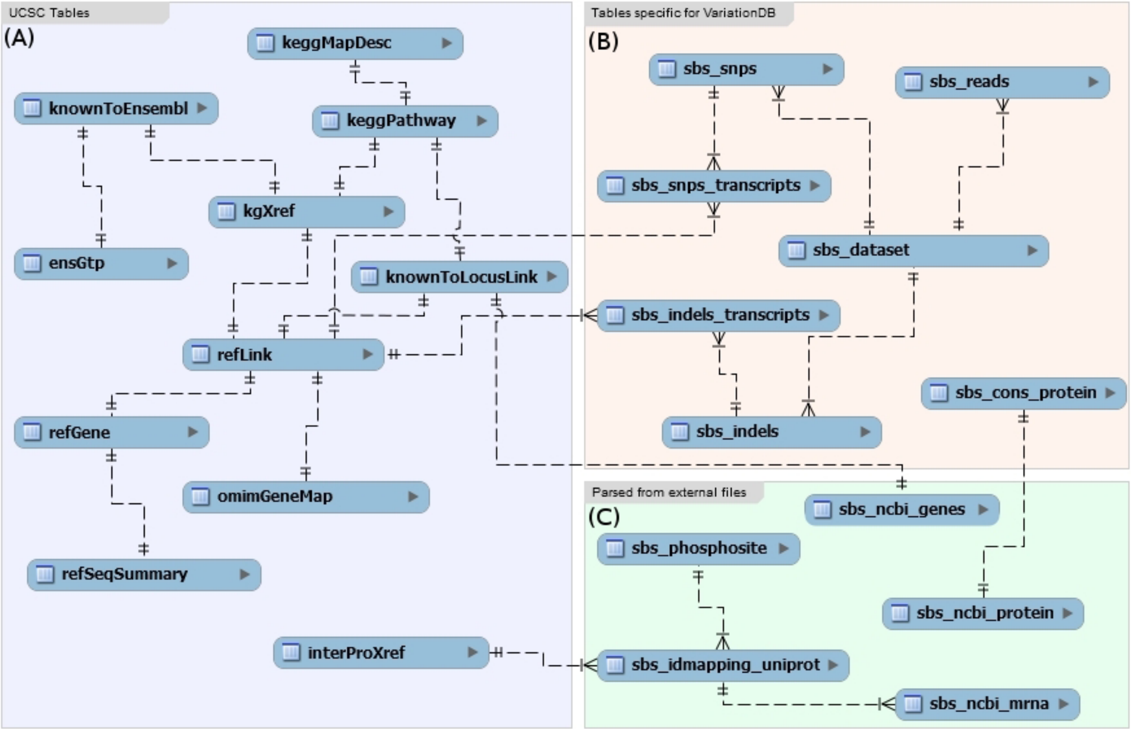


Fig. 6.6: Reduced schema for the central VariationDB database. The database consists of three logical units. (A) Pre-build annotation tables downloaded from the UCSC Genome Browser. (B) Tables that store the data set (mutations data, read data). (C) Sequence annotation tables that were parsed from different sources. Figure taken from [105].

it with prior knowledge on genes, proteins and pathways. The backend is implemented in Python and uses Biopython. It contains calls to external programs: BLAST [170] and T-coffee [171] are used for the computation of protein conservation scores. ANNOVAR [101] is used to assess the effect of genomic variations on the corresponding transcripts and proteins.

The mutation data along with annotation data is stored in a central database. The web frontend can access the data in the database. In the first version of VariationDB all annotation and mutation data was given relative to the reference genome hg18. VariationDB and all annotation data was later updated to *hg19*.

The central database. All data – patient related mutation data and annotation data to genes, transcripts and proteins – is stored in a central MySQL database. The database is structured in three logical units. A reduced schema of the central database is depicted in Fig. 6.6.

The first unit (A) contains pre-built annotation tables downloaded from the UCSC Genome Browser [91]. These tables contain information on genes and cross links to gene annotation from different sources (Ensemble [172], Omim [79], KEGG [84], UniProt [173] and RefSeq [90]).

Unit (C) contains sequence and annotation tables that were parsed from flat file downloads from external databases. These tables comprise information on phosphorylation sites [102], UniProt ID mapping, and RefSeq gene, protein and transcript annotation.

Unit (B) contains tables that were specifically designed for VariationDB. These tables store the datasets, the mutation data for the datasets and, in some cases, the original read data. Genetic variations are represented hierarchically. Each variation is associated with a dataset. Datasets are associated with users to ensure that users have only access to their own datasets. The tables *sbs_snps* and *sbs_indels* contain the genes affected by a point mutation or an INDEL. Each gene can be linked to different transcripts. The effects of each mutation on the respective sequences of the transcripts and proteins are stored in the tables *sbs_indel_transcripts* and *sbs_snps_transcripts*. The table *sbs_cons_protein* contains precomputed information about protein conservation. Conservation scores are computed for each protein using the following steps: A remote blastn search is performed to identify the 25 most similar protein sequences according to the blastn E-value. On these sequences, a multiple sequence alignment is performed using T-coffee. The information content of each column in the alignment is computed using Biopython. The information content can be interpreted as a degree of evolutionary conservation of the positions in the protein. For each protein the sequences used to compute conservation scores and the sequence alignment are stored and can be accessed via the frontend.

Information from the Gene Ontology project [85] is directly accessed over the MySQL interface. Information on naturally presented MHC ligands are taken from an offline version of the SYFPEITHI database [116].

VariationDB - Frontend

The frontend of VariationDB is a dynamic web application implemented in PHP, JavaScript and AJAX. It is designed according to a three-tier architecture. The three tiers (presentation tier, application tier, data tier) are independent from each other and communication is performed top down. The user interacts with the presentation tier, the application tier handles and processes the user requests and contains the business logic. The data tier stores information that is independent of the business logic, in our case the data tier is the central database.

VariationDB is structured in modules. A logical interpretation of a module is that it presents a view on entities. Technically a module can be interpreted as a set of scripts and programs in the application tier that are responsible for receiving and processing the user requests and returning the results in a specific way for each entity. The four entities implemented in VariationDB are: general information on a gene, information on SNVs, information on INDELs and reads that are mapped to genes. The modular design allows for new entities to be easily integrated into VariationDB. Compatibility of VariationDB was tested for Mozilla Firefox (version 3.6), Microsoft internet Explorer (version 8) and Google Chrome (version 9).

Use cases implemented in VariationDB. The use cases implemented in the frontend of VariationDB are strongly orientated towards important biomedical questions in cancer research. The use cases as well as the graphical display of the results are closely coordinated with our cooperation partners in clinical research.

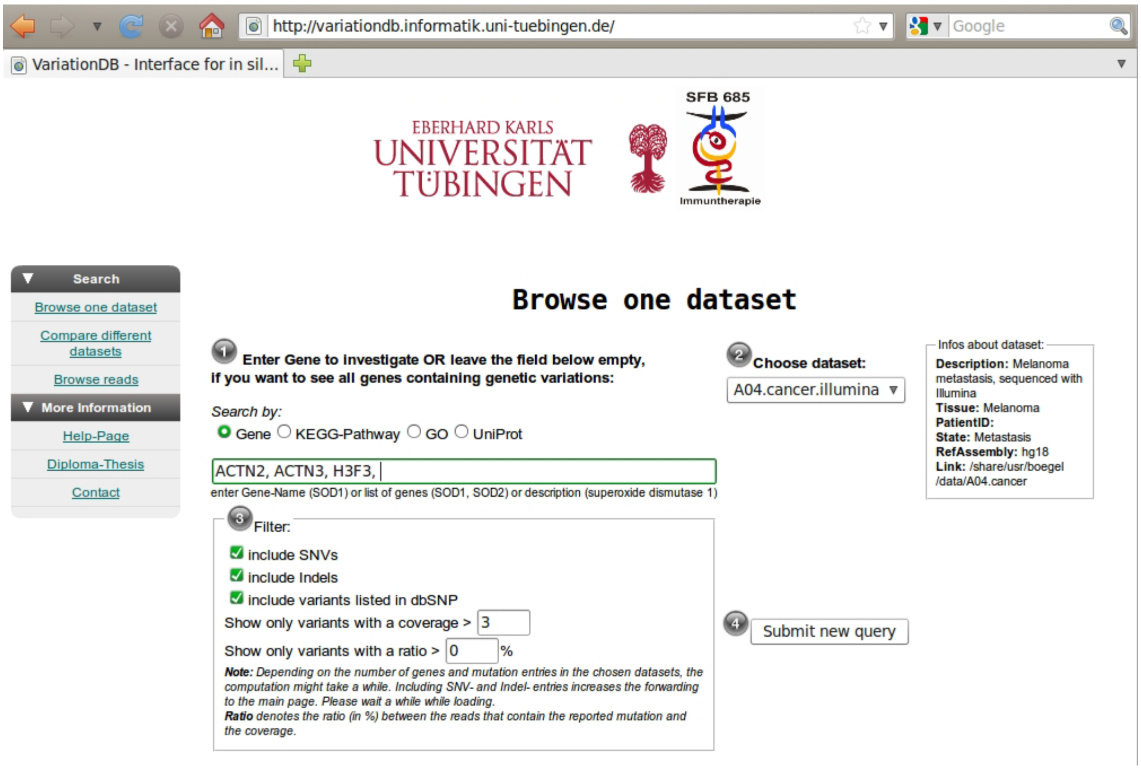
1. **In-depth analysis of a single dataset.** The user is interested in viewing mutations in a single dataset. The mutations can be accessed via search fields (gene names, pathways, GO terms). The user is presented with information on the selected genes and mutations can be visually inspected with respect to their impact on the transcript and protein sequences. The mutated protein sequences can be directly submitted to epitope prediction.

2. **Comparing datasets or groups of datasets.** The user is interested in comparing two datasets (e.g. normal tissue and tumor for one patient) or two groups of datasets (e.g. two different subgroups of tumor samples). The available datasets can be grouped by the user. Mutations can again be accessed via searches. The information presented to the user is similar to the first use case, however a direct comparison of the presence of single mutation in the different datasets is provided.

3. **Browse reads.** The user can also browse the reads that are mapped to a particular gene. This use case is only available for the first datasets produced with 454-sequencing. For the dataset generated with the Illumina Genome Analyzer the amount of read data available is too large to be conveniently be stored and accessed.

VariationDB user interface. Since data of genetic variation in cancer patients is very sensible, the access to VariationDB is restricted. Users have only access to certain datasets. After login, the user gets presented the ***landing page*** as shown in Fig. 6.7.

With the menu on the left the user can choose between the use cases described above. After selection of the use case, the first step is to select genes via a Google-like search with auto-completion. Possible search terms are gene names, KEGG pathways, GO annotation or UniProt IDs. In the next step, the user has to choose the dataset(s). In the "Browse one dataset" mode, the user can select one of the available datasets. Information on the dataset (patient ID, tissue type,...) is displayed on the right. In the "Compare different datasets"-mode the user can select datasets for two groups using two multiselect boxes. In the next step the user can define filters to be applied before display of the results.



Browse one dataset

1 Enter Gene to investigate OR leave the field below empty, if you want to see all genes containing genetic variations:

Search by:
● Gene ○ KEGG-Pathway ○ GO ○ UniProt

ACTN2, ACTN3, H3F3,

enter Gene-Name (SOD1) or list of genes (SOD1, SOD2) or description (superoxide dismutase 1)

2 Choose dataset: A04.cancer.illumina

Infos about dataset:
Description: Melanoma metastasis, sequenced with Illumina
Tissue: Melanoma
PatientID:
State: Metastasis
RefAssembly: hg18
Link: /share/usr/boegel/data/A04.cancer

3 Filter:
☑ include SNVs
☑ include Indels
☑ include variants listed in dbSNP
Show only variants with a coverage > 3
Show only variants with a ratio > 0 %

Note: *Depending on the number of genes and mutation entries in the chosen datasets, the computation might take a while. Including SNV- and Indel- entries increases the forwarding to the main page. Please wait a while while loading.*
Ratio *denotes the ratio (in %) between the reads that contain the reported mutation and the coverage.*

4

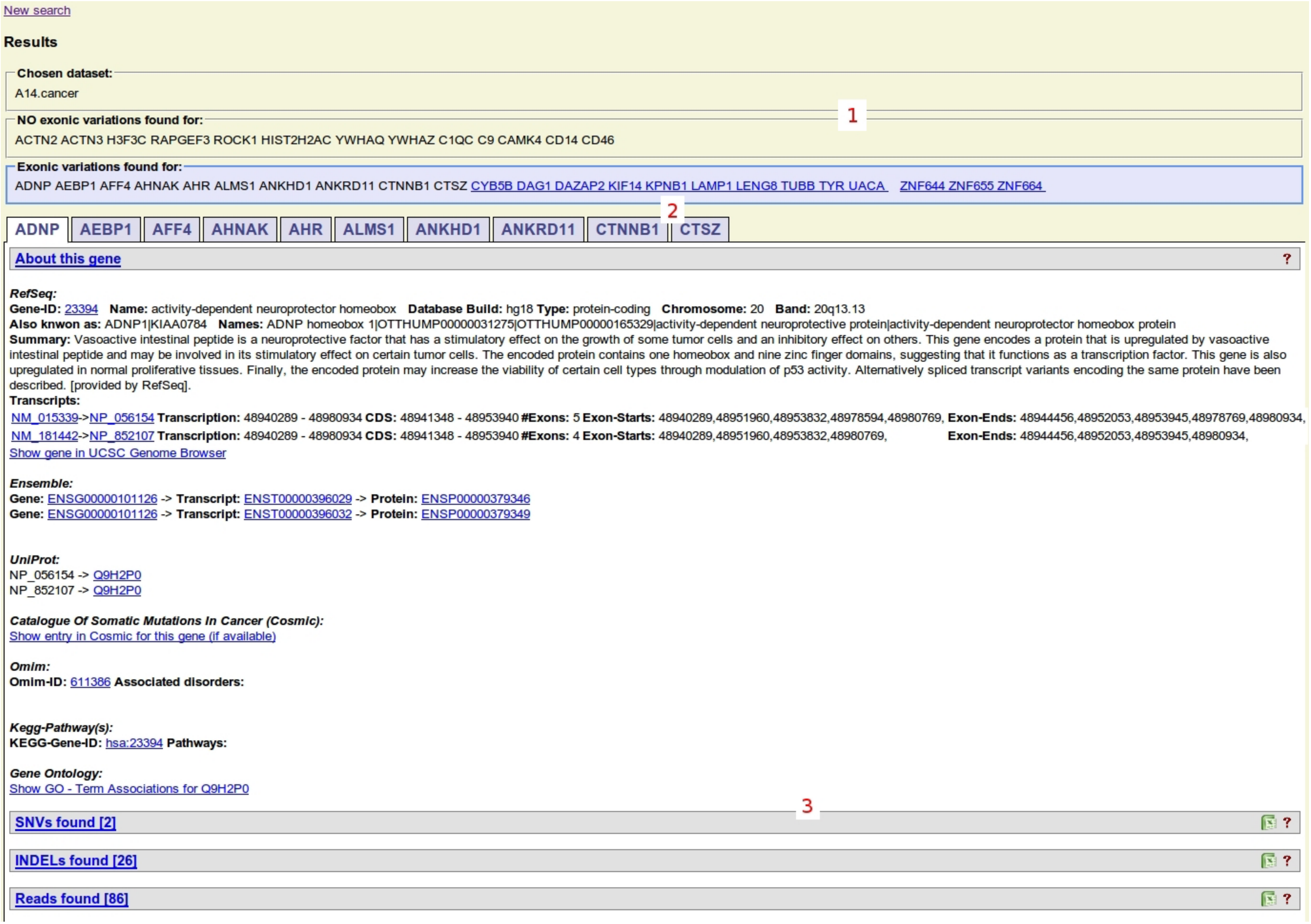
Fig. 6.7: Landing page of the VariationDB user interface. The menu on the left allows for the selection of the use case. The user can then select genes, datasets and filtering options.

Filtering options are for example the type of variation (SNV or INDELs) or coverage of the variant positions.

After selecting the use case, genes, datasets, and filter criteria the user is transferred to the result page. The general structure of the result page is displayed in Fig. 6.8. The results page is structured in three parts: a feedback section (1), a navigation section (2) and a the content section (3). The feedback section lists the selected dataset(s) as well as the search and filter criteria. The genes are divided into two groups, genes with variations and genes where no variation was found in the selected dataset(s). If more than 10 genes with variants are found, the genes are grouped. These groups then become the main navigation. VariationDB is gene-centered so the information is displayed per gene. Each gene is displayed in a single tab. Each tab contains the four modules that correspond to the technical modules of the backend: 1) The module "About the gene" displays general information about the gene: the header from the RefSeq gene entry, available transcript and protein sequences from different databases (RefSeq, Ensemble, UniProt) as well as information on pathways, GO annotations and disease association of the gene. 2) The module "SNVs found" handles the display and analysis of point mutations. 3) The module "INDELs found" handles the display of INDELs found. 4) The module "Reads found" is responsible for the display of all reads from the dataset(s) that were mapped to the gene (not available for all datasets).

Depending on the chosen use case (single dataset or comparative analysis) the display of the variants differs. All variants are listed, along with additional information on the genome position, matching dbSNP entries, codon change, position of the variation in the protein, effect of the variation and coverage of the variant position.

In an easy-to-use and convenient workflow the user can investigate the influence of selected mutations on the protein sequence. The protein sequence is displayed and annotated with conservation scores, known MHC ligands, domains, and phosphorylation sites. The selected mutations are also displayed along the protein sequence. The protein sequence together with the mutations can directly be submitted to an epitope prediction step. The implementation of the prediction is based on FRED (see Section 6.1.1), the allele selection and display of the results is similar to the approaches used in EpiToolKit (see Section 6.1.2). The integration of epitope prediction allows for the identification of tumor-specific epitopes. How VariationDB can be used to analyze mutation data from cancer patients is shown in detail in Section 7.4.3.



New search

Results

Chosen dataset:
A14.cancer

NO exonic variations found for:
ACTN2 ACTN3 H3F3C RAPGEF3 ROCK1 HIST2H2AC YWHAQ YWHAZ C1QC C9 CAMK4 CD14 CD46

Exonic variations found for:
ADNP AEBP1 AFF4 AHNAK AHR ALMS1 ANKHD1 ANKRD11 CTNNB1 CTSZ CYB5B DAG1 DAZAP2 KIF14 KPNB1 LAMP1 LENG8 TUBB TYR UACA ZNF644 ZNF655 ZNF664

About this gene ?

RefSeq:
Gene-ID: 23394 **Name:** activity-dependent neuroprotector homeobox **Database Build:** hg18 **Type:** protein-coding **Chromosome:** 20 **Band:** 20q13.13
Also knwon as: ADNP1|KIAA0784 **Names:** ADNP homeobox 1|OTTHUMP00000031275|OTTHUMP00000165329|activity-dependent neuroprotective protein|activity-dependent neuroprotector homeobox protein
Summary: Vasoactive intestinal peptide is a neuroprotective factor that has a stimulatory effect on the growth of some tumor cells and an inhibitory effect on others. This gene encodes a protein that is upregulated by vasoactive intestinal peptide and may be involved in its stimulatory effect on certain tumor cells. The encoded protein contains one homeobox and nine zinc finger domains, suggesting that it functions as a transcription factor. This gene is also upregulated in normal proliferative tissues. Finally, the encoded protein may increase the viability of certain cell types through modulation of p53 activity. Alternatively spliced transcript variants encoding the same protein have been described. [provided by RefSeq].
Transcripts:
NM_015339->NP_056154 **Transcription:** 48940289 - 48980934 **CDS:** 48941348 - 48953940 **#Exons:** 5 **Exon-Starts:** 48940289,48951960,48953832,48978594,48980769, **Exon-Ends:** 48944456,48952053,48953945,48978769,48980934,
NM_181442->NP_852107 **Transcription:** 48940289 - 48980934 **CDS:** 48941348 - 48953940 **#Exons:** 4 **Exon-Starts:** 48940289,48951960,48953832,48980769, **Exon-Ends:** 48944456,48952053,48953945,48980934,
Show gene in UCSC Genome Browser

Ensemble:
Gene: ENSG00000101126 -> **Transcript:** ENST00000396029 -> **Protein:** ENSP00000379346
Gene: ENSG00000101126 -> **Transcript:** ENST00000396032 -> **Protein:** ENSP00000379349

UniProt:
NP_056154 -> Q9H2P0
NP_852107 -> Q9H2P0

Catalogue Of Somatic Mutations In Cancer (Cosmic):
Show entry in Cosmic for this gene (if available)

Omim:
Omim-ID: 611386 **Associated disorders:**

Kegg-Pathway(s):
KEGG-Gene-ID: hsa:23394 **Pathways:**

Gene Ontology:
Show GO - Term Associations for Q9H2P0

SNVs found [2] ?

INDELs found [26] ?

Reads found [86] ?

Fig. 6.8: The result page of VariationDB is structured in a feedback section (1), a navigation section (2) and a content section (3). The content section contains the four modules "About the gene", "SNVs found", "INDELs found" and "Reads found".

6.4 Building immunoinformatics workflows

Immunoinformatics and its application in cancer research is a quickly evolving field. Different computational approaches from imunoinformatics have to be integrated with the analysis of high-throughput data for many patients. The challenges hereof are manifold. The computational analysis and especially the interpretation of personalized omics data leaves room for improvements. Several steps that contribute to the immunogenicity of peptides are not yet fully understood and new prediction methods need to be developed. The amount of data that needs to be processed, stored, and transferred introduces technical challenges. The application in clinical setting requires reliable and reproducible prediction pipelines. However, in order to establish new approaches for cancer treatment new methods and ideas have to be tested. The computational infrastructure needs to be flexible enough to evolve as new insights and methods emerge.

We therefore need a way to quickly build and test prototype workflows and apply them to large datasets. The workflow system also needs to be very flexible in order to adapt to new findings. The workflow system needs to support the easy and quick integration of various computational tools. Support for high-performance computation infrastructures is required since even prototype workflows need to be tested on large datasets. The workflow system should also allow for a close interplay with user interfaces. We chose Galaxy [174] as basis for developing workflows.

Galaxy is an open, web-based platform for data-intensive biomedical research. A free public Galaxy server is available at http://g2.bx.psu.edu/. It is also possible to run a local instance of Galaxy that can be extended and customized. The use of Galaxy is convenient since it comes with a large variety of software and tools for the analysis of sequencing data. External tools and software can quickly be integrated using tool wrappers defining the interface to the external tools. Galaxy is implemented in Python. A large community contributes to the further development of Galaxy as well as to the integration of new tools. Galaxy supports all standard file formats that are commonly used in the analysis of genomics data.

6.4.1 Architecture and general setup of the Galaxy server

The productive instance of our Galaxy server contains three webservers and one jobrunner, the load of the webservers is balanced over Apache. Jobs can be executed on the local machine or on a compute cluster over the Sun Grid Engine. The NFS file system of the department is embedded for data storage. The Galaxy server runs with a PostgreSQL database. Data backup is provided over the NFS backup system. For the Galaxy system and the databases full backup is performed monthly and incremental backup is performed on a daily basis.

To ease collaborative development our Galaxy system is under version control using mercurial (http://mercurial.selenic.com/). The integration of new tools and new pipelines can be implemented and tested on local instances. Our Galaxy server was set up and is maintained by Nico Weber at the group of Prof. Daniel Huson, University of Tübingen.

6.4.2 Workflows implemented in the Galaxy server

Integration of new tools in Galaxy. New tools can be integrated into a Galaxy server using tool definition files. Tool definition files are xml files that describe how to run the tool. Tool description files are used by the Galaxy server to generate the graphical user interface pages to run tools. The parameters and datasets, including data file formats, are handled within the tool description. Galaxy also provides functional tests for integrated tools. A tutorial how to integrate new tools into Galaxy can be found at http://wiki.g2.bx.psu.edu/Admin/Tools/AddToolTutorial.

We implemented several pipelines for the integrated analysis of NGS data into our Galaxy server. The three major pipelines along with the software and tools that had to be integrated first are presented in the following sections. The pipelines are orientated at the general NGS analysis pipeline presented in Fig. 4.1. While all steps of the analysis can also be combined into one single pipeline, we divided the analysis steps into modules corresponding to the steps in Fig. 4.1 for the sake of clarity. The transitions between the modules are explained. All figures of workflows are generated using the Galaxy workflow editor. For each tool, parameters can be set during workflow generation or - if stated in the workflow configuration - during run time.

Read mapping

The workflow we apply for read mapping of RNA-seq data is depicted in Fig. 6.9. Reads are first mapped against the human reference genome hg19. Unmapped reads are then mapped against human mRNA sequences. Matches to mRNA are converted to genomic positions. Reads that are not mapped to mRNA are used as input for the detection of viral sequences (Section 6.4.2). Matches to the genome and converted matches to mRNA are merged and sorted. The sorted gff-files are used as input for the detection of genetic variants (Section 6.4.2). The mapping results are converted to sam and bam format. bam format is a standard format that can be imported in external tools to visualize mapping results like the Integrative Genome Browser (http://www.broadinstitute.org/igv/).

For mapping of Exome-seq data the workflow is just composed of one read mapping step and the conversion to sam/bam (steps A, F, G).

An overview of the tools and software used in the workflow for mapping of RNA-

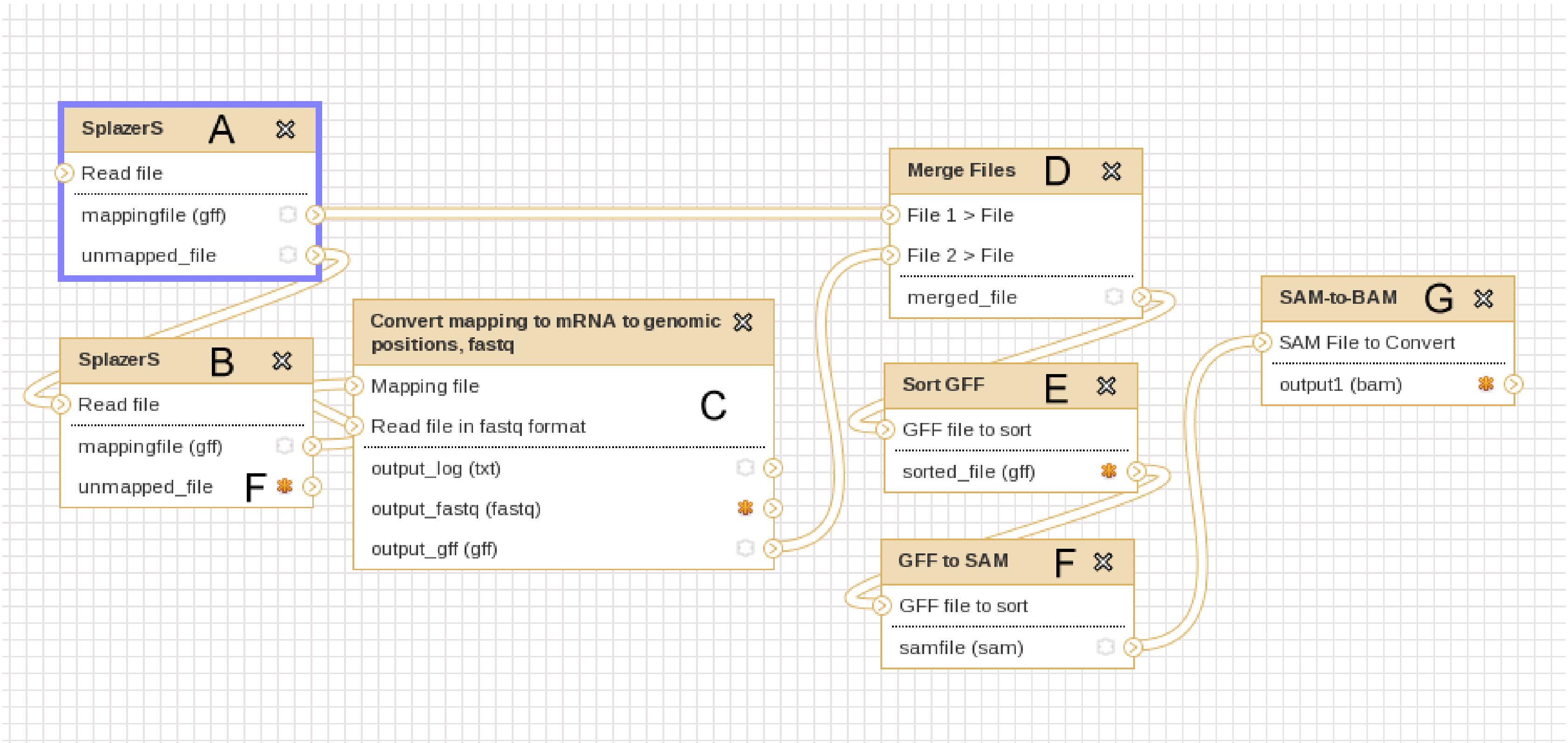


Fig. 6.9: Galaxy workflow for mapping of RNA-seq data. Reads are first mapped against hg19 using SplazerS [57], and the unmapped reads are mapped and against a collection of known human mRNA sequences (A+B). Mappings to mRNA are converted to genomic positions (C). Mappings to hg19 and converted mappings to mRNA are merged and sorted (D+E). Unmapped reads are used as input for the detection of viral sequences. Mapped reads are used as input for the detection of genetic variations. Mappings are converted to sam and bam formats (F+G). These file formats are convenient for graphical visualization of mapping results in external software.

seq data is given in the following. If it is not indicated otherwise, the tools had to be integrated into Galaxy. References are given for external tools.

- **SplazerS** [57] is used for read mapping.
- **Convert mapping to mRNA to genomic positions**: This tool is used to convert mappings to mRNA to genomic positions. It is implemented in Python. For a detailed description of the procedure see Section 4.1.
- **Merging** and **sorting** is performed using the UNIX command line tools *cat* and *sort*.
- **GFF-to-SAM**: Converts SplazerS output in gff format to sam format. (Provided by A.K. Emde, AG Prof. Knut Reinert, FU Berlin, based on SeqAN [89].)
- **SAM-to-BAM**: Based on SAMtools (http://samtools.sourceforge.net/, a wrapper for this tool is provided by Galaxy.

Detection of viral sequences in RNA-seq data

The method we use to detect viral sequences in NGS data is described in detail in Section 4.2. This workflow was also implemented in our Galaxy server as shown in Fig. 6.10. The first part of the workflow implements Digital Sequence Subtraction (see Section 4.2.1) to remove all remaining sequences of human origin and the first comparison to viral sequences (see Section 4.2.2). As an intermediate results we extract all sequences that have a valid hit on a known virus. In the second part the sequences with hits to a known virus are again compared to known viral sequences and to the GenBank nonredundant protein collection GenBank (NR). The blastxml files that are produced by blastx can directly be imported into MEGAN4 for metagenomics analyses. Based on the request by our collaborators from the clinic we also convert the results into two custom formats that are better suited for manual inspection than blastxml. We output a text file that lists all hits that pass the filtering criteria including the alignment information. Additionally we output the valid hits in a csv format that can be imported into MS Excel.

An overview over the tools and software used in the workflow to detect viral sequences is given in the following. If it is not indicated otherwise, the tools had to be integrated into Galaxy. References are given for external tools.

- **SeqClean** [94]. In our applications sequence data is given in fastq format. SeqClean works on sequences in fasta format, so sequences are converted to fasta (tool provided in Galaxy) and the clean sequences are converted back to fastq format (implemented in Python and Biopython).
- A wrapper for the **NCBI BLAST** tools is provided in Galaxy.

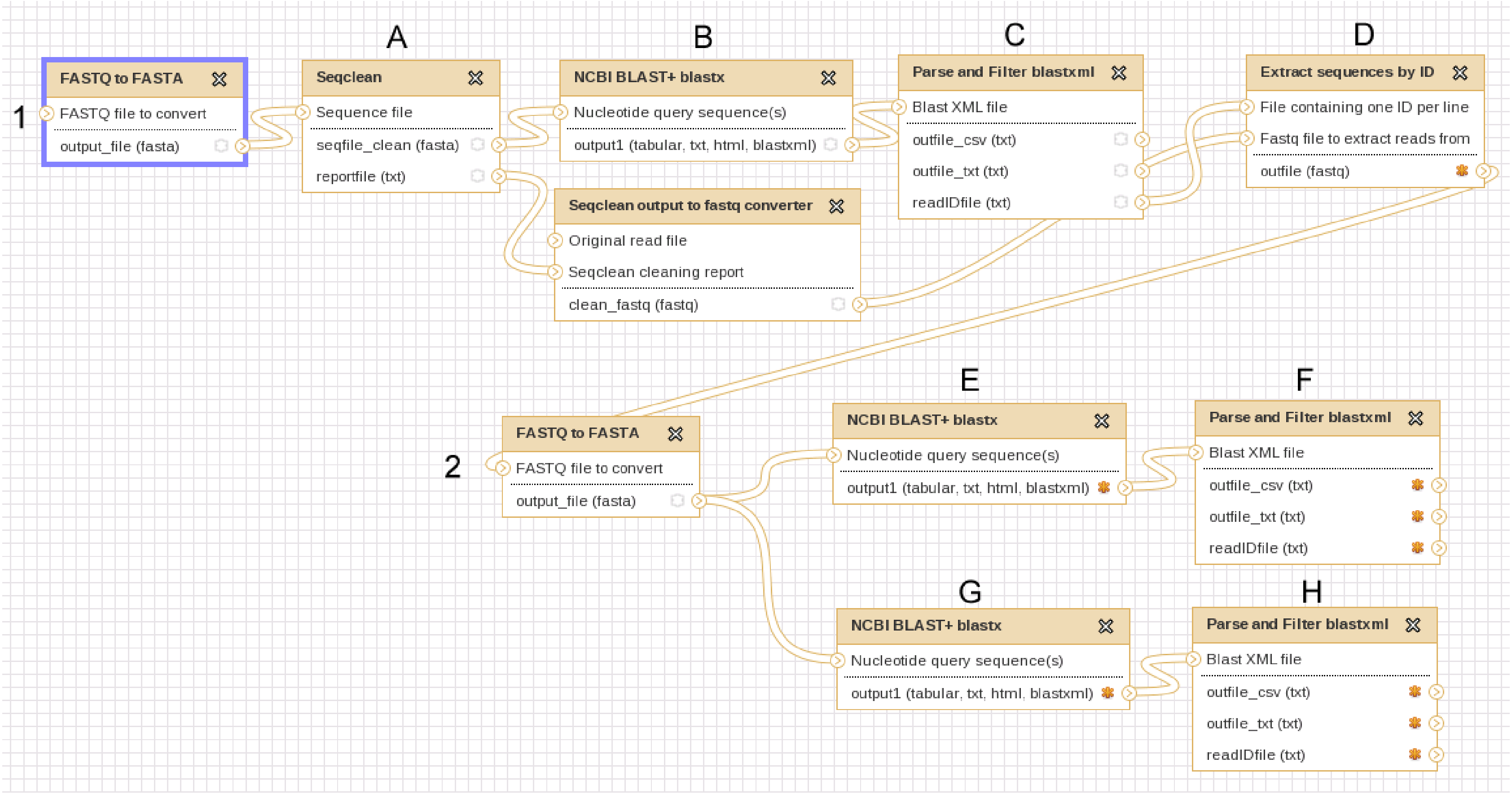


Fig. 6.10: Galaxy workflow for detection of viral sequences. 1: In the first part Digital Sequence Subtraction is performed (A). The clean sequences are then compared to known viral sequences using blastx (B). The blastx results are filtered for sequence identity, blastx E-value and length of the hit. All sequences that have a valid hit on a known virus are extracted. 2: In the second part the sequences with viral hits are compared to NR (E) to identify sequences that have additional or better hits on other species. The results of blastx against NR (E) and the viral database (G) can be imported into MEGAN4 for metagenomics analysis. The results are also converted into txt and csv format.

- **Extract sequences by ID**: Extracts sequences for a given list of fastq sequence IDs from a fastq file (implemented in Python and Biopython).

- **Filter and Parse blastxml**: Reads in blastxml files and filters the matches for minimal length of the alignment, bit score, E-value and sequence identity within the match (implemented in Python and Biopython).

Variation detection form sequencing data

In this workflow we integrate variant detection (see Section 4.3), annotation of the variants, the differential analysis of tumor and control tissue and the prediction of T-cell epitopes for the tumor-specific variants. The variants are also directly exported to VariationDB (see Section 6.3). The whole workflow is illustrated in Fig. 6.11.

Two samples (tumor and control tissue) are analyzed in parallel since we need information on both samples to perform a differential analysis. As input we use aligned reads from both samples (C-1, T-1) as generated in the read mapping workflow (SplazerS gff output). Read mapping nodes are included as input in this workflow to illustrate how the mapping files are further processed. For the control tissue (lower part of the workflow, steps C-2 to C-5) we perform variant detection using the tool SnpStore (C-2), and the found SNVs are annotated (see Section 4.4.2). Steps C-4 and C-5 are needed to export the variation data (SNVs and INDELs) into VariationDB.

The aligned reads from the tumor sample (T-1) are submitted to variation detection (T-2). The found SNVs and INDELs are exported to VariationDB (T-8, T-9). In order to perform a differential analysis for the two samples, i.e. to detect somatic SNVs, we include information on SNVs found in the control tissue (obtained from step C-2) as well as coverage information for the tumor SNV positions (T-3,T-4) in the annotation step (T-5). This information is then used to filter the tumor mutations. Possible filter criteria are: coverage of the SNV position, quality of the SNV call, non-synonymous SNVs, coverage for SNV position in control tissue and the existence of a tumor-specific variant. The filtered list of tumor variants is then submitted to an epitope prediction step (see Section 5.4). As additional input we need a file that contains the patient's HLA alleles. The output of the epitope prediction step consists of three files: One file contains a list of predictions for all peptides around the SNV position for all given HLA alleles. A second file that contains the annotated SNVs in a tabular format with additional columns with information of predicted epitopes. The third file contains the list of tumor-specific protein sequences. As described in Section 5.4, we can also include information on the patient's SNPs to generate patient and tumor-specific protein sequences. We can do so by using the output of C-3 as additional input for T-7 (not shown in Fig. 6.11).

An overview over the tools and software used in this workflow is given in the following.

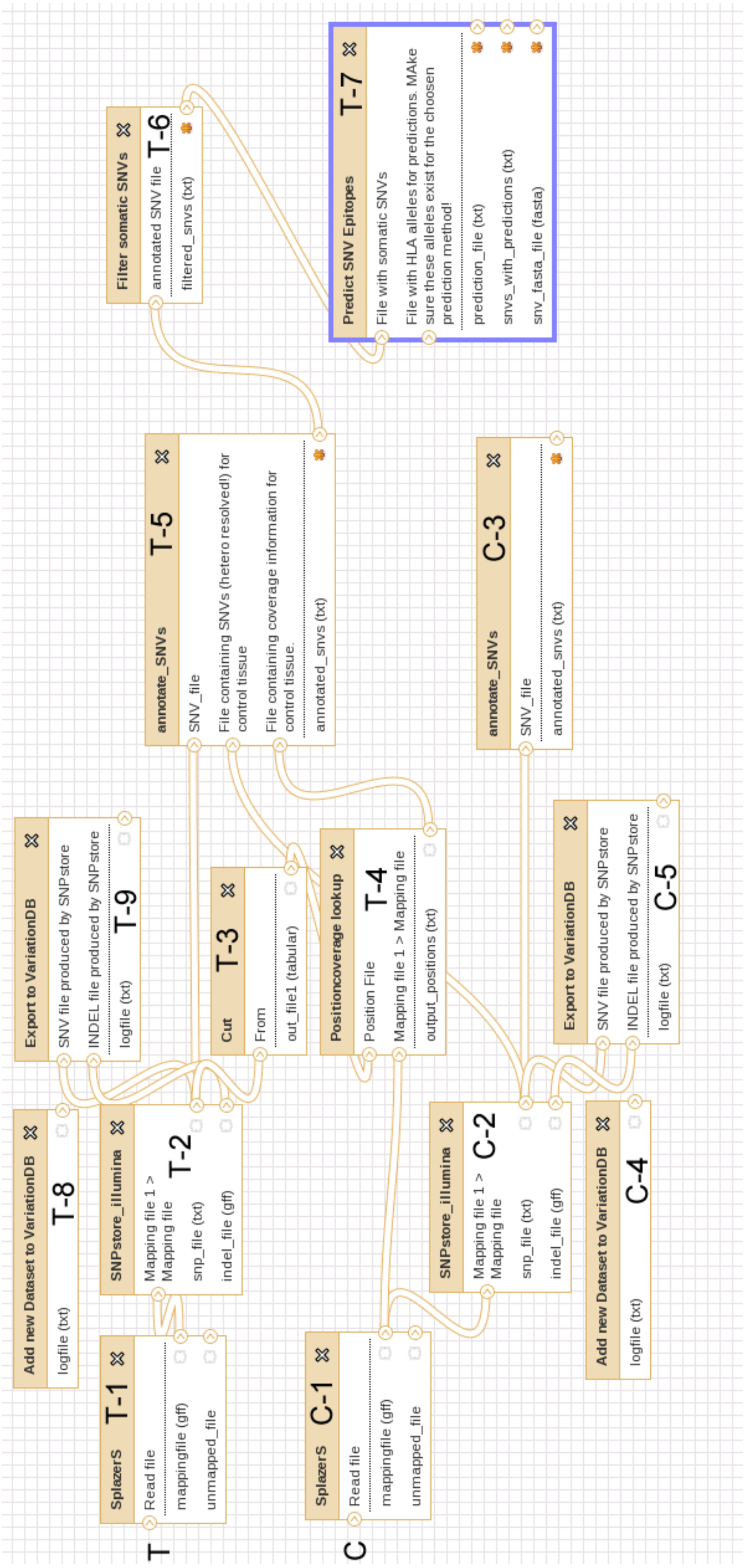


Fig. 6.11: Galaxy workflow for the detection and analysis of genetic variation.

If it is not indicated otherwise, the tools had to be integrated into *Galaxy*. References are given for external tools.

- **SnpStore**: Provided by A.K. Emde, AG Prof. Knut Reinert, FU Berlin, based on SeqAN [89]. See Section 4.3 for further details.

- **Position coverage lookup**: This tool is based on SnpStore and looks up the coverage in the control tissue for a list of tumor SNVs. It takes a list of genomic positions and a mapping file as input. For these positions the coverage on the mapping file is computed using the same parameters for realignment as in the variation detection mode of SnpStore. The tool was provided by A.K. Emde, AG Prof. Knut Reinert, FU Berlin. The list of tumor SNV positions is cut from the SnpStore output on the tumor sample using the *cut* tool provided in Galaxy.

- **annotate SNVs**: This tool is implemented in Python using Biopython. ANNOVAR [101] is called to annotate SNVs (see also Section 4.4.2). SNVs from the control tissue as well as coverage information in the control sample (see position coverage lookup) can be provided as optional input. The output of this tool is a column-based format that contains information on SNVs: position, coverage, base counts, functional annotation, influence on the protein sequence for SNVs in coding regions and information on control sample (optional). Filter criteria can be passed to this tool to only output SNVs that match certain criteria (e.g. minimum coverage).

- **Filter somatic SNVs**: Filters SNVs for coverage, quality, functional annotation, coverage in control and existence of a variant that is not present in the control sample (implemented in Python).

- **Predict SNV Epitopes**: Epitope prediction is performed as described in Section 5.4. This tool is implemented in Python using Biopython and is based on FRED. SNVs that are found in the control tissue (regarded as wildtype or germline SNVs of the patient) can be provided as additional input.

- **Add new Dataset to VariationDB**: A new dataset for the sample is added to VariationDB. Information on the patient ID, disease state and tissue (tumor/control) have to be provided. This tool directly interacts with the backend of VariationDB (see Section 6.3).

- **Export to VariationDB**: This tool directly interacts with the backend of VariationDB and imports the variants into VariationDB (see Section 6.3).

6.5 Discussion

In this section we presented two strategies to make MHC binding prediction applicable on a large scale. FRED is a software framework that allows immunoinformaticians to quickly implement larger epitope prediction pipelines with minimal coding effort. As a complementary tool we offer EpiToolKit, a webserver for epitope prediction that is based on FRED. Both systems permit the analysis of mutated sequences. Based on FRED we implemented a pipeline for the large-scale identification of miHA candidates. These systems help to bring prediction methods closer to a clinical application.

In addition, we presented VariationDB, a web-based tool for the integrated analysis of mutation data. VariationDB presents genetic variations in a comprehensive manner together with additional annotations for genes and proteins. It also allows the direct submission of mutated sequences to epitope prediction. Comparative analyses are possible as well as the analysis of single datasets. In the current implementation VariationDB, however, is a tool that facilitates manual inspection of variations found in NGS data. A useful extension of VariationDB would be to combine information on a tumor and a control tissue. A tumor and a control tissue can be compared using the "compare datasets"-mode of VariationDB, however, the direct search for somatic mutations is not possible. Providing more automated analyses and search criteria would also be useful, e.g., to only display somatic mutations that lead to a new predicted epitope in the tumor for a given set of HLA alleles. The integration of other omics data would also be desirable in oder to combine mutation data with data on gene expression, methylation or copy number variations. Due to its modular design the extension of VariationDB to account for these suggestions is quite simple.

We have implemented our workflows in Galaxy. Galaxy is only one of many possible ways to implement workflows, and portability between workflow systems (e.g. Galaxy and Grid infrastructures) would be a valuable feature. Workflows can then be tested and implemented locally and later be ported to Grid systems to be executed on a large scale on a grid. Efforts towards portability of workflows between workflow systems based on the Common Tool Description (CTD) used by OpenMS (www.openms.org) have already been made in our group for applications from structural bioinformatics and computational proteomics. The knowledge and experience gained here can be used to adapt the tools used in this thesis to be compliant with the CTD format. This would allow to port the workflows developed and tested on a local Galaxy server to a grid. The workflows are currently started manually for each dataset. Integration with data management systems like openBIS (http://www.cisd.ethz.ch/software/openBIS) to allow automated workflow execution are planned. Datasets can then also be associated with meta-information (on the patient, the disease state, results from treatments, ...). Such an automated system would facilitate the monitoring and assessment of our analyses, and, in the future, of clinical experiments based on the data analysis.

The workflows and methods presented here thus form a valuable basis for the large-scale application of integrated NGS data analysis together with methods from computational immunology and immunoinformatics in clinical settings. The close interaction between experimentalists, clinicians and computational analyses will promote insight to the complex mechanisms of immunogenicity and offer new possibilities for personalized immunotherapies based on T-cell epitopes.

Part III

Application to clinical data

In this part we demonstrate the application of the computational methods, tools, and workflows presented in in Part II to clinical data. In Chapter 7 we analyze RNA-seq data from ten melanoma patients. We first apply the pipeline for the detection of viral sequences and show that no evidence for the presence of a virus can be detected. We then apply the pipeline for the detection of SNVs to the tumor samples and the corresponding control samples. The mutations are analyzed for somatic mutations. The somatic mutations are analyzed with respect to the genes and pathways they occur in. We also search for recurrent mutations in all tumor samples. Somatic mutations are mapped to the respective transcript sequences translated into protein sequences. These protein sequences are analyzed for tumor-specific peptides that are predicted to bind to one of the patient's HLA alleles.

In Chapter 8 we show a clinical application of miHA identification pipeline. The aim of the study is to develop a new therapy that helps to prevent relapse of leukemia after allogenic stem cell transplantation. The approach comprises two main computational steps. In a first step we design a customized genotyping assay for hematopoiesis-restricted SNPs that are relevant as miHAs in a large group of patients. In a second step, the results of the genotyping assay allow for a quick and cheap detection of miHAs that are relevant in a patient-donor pair. We present the pipeline and describe and discuss where it has been adapted to the special clinical setting of allogenic hematopoietic stem cell transplantation.

Chapter 7

Genetic variation in melanoma metastases

7.1 Experimental data

Ten patients with metastatic melanoma were included in this study. For all ten patients RNA-seq was performed for samples derived from melanoma metastases. As control, RNA-seq was performed on peripheral blood mononuclear cells (PBMCs). HLA typing is available for all ten patients along with additional information. HLA typing is listed in Tab. 7.1, other information available on the patients and the tumors is summarized in Tab. 7.2.

RNA-seq data

RNA-seq data was generated at the Dept. of Medical Genetics at the University Hospital in Tübingen. Blood and tissue samples (PBMCs and melanoma metastases) were taken from the melanoma tissue bank, University of Tübingen. All patients donating to the tissue bank gave their written informed consent for the analysis of blood and tumor samples. The study was approved by the local ethics committee (identifier 609/2011BO2) and was carried out according to the Declaration of Helsinki.

Total RNAs were extracted, according to the manufacturer's instruction, from about 60 mg of tissues for each of the samples using the Qiagen RNAeasy Midi Kit (Qiagen). RNA yields were quantified by NanoDrop ND1000 (ThermoFisher Scientific, Waltham, MA) and the RNA quality was assessed by the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA). The RNA integrity number (RIN) of every RNA sample used for sequencing was greater than 7. cDNA libraries were constructed using mRNA-Seq Sample Prep Kit

Tab. 7.1: HLA typing for the ten melanoma patients. For HLA-A and HLA-B a four-digit resolution HLA typing was performed, HLA information on HLA-C is only available at a two-digit resolution.
(+) Alleles for which a SYFPEITHY prediction model is available.

Patient ID	HLA-A	HLA-B	HLA-C
Pat01	A*01:01 (+), A*68:01 (+)	B*08:01 (+), B*18:01 (+)	C*07, C*07
Pat02	A*03:01 (+), A*24:02 (+)	B*07:02 (+), B*35:01	C*04, C*07
Pat03	A*02:01 (+)	B*15:01 (+), B*35:03	C*04, c*04
Pat04	A*02:01 (+), A*25:01	B*44:02 (+), B*52:01	C*05, C*12
Pat05	A*01:01 (+), A*02:01 (+)	B*08:01 (+), B*15:01 (+)	C*03, C*07
Pat06	A*01:01 (+), A*26:01	B*08:01 (+), B*51:01 (+)	C*01, C*07
Pat07	A*03:01 (+), A*32:01	B*07:02 (+), B*15:01 (+)	C*03, C*07
Pat08	A*01:01 (+), A*32:01	B*15:01 (+), B*44:02 (+)	C*03, C*05
Pat09	A*01:01 (+), A*68:01 (+)	B*07:02 (+), B*08:01 (+)	C*03, C*07
Pat10	A*01:01 (+), A*02:01 (+)	B*08:01 (+), B*44:05	C*02, C*07

based on the Illumina Inc.'s guide. In brief, polyA-containing mRNA was purified using oligo-dT beads from 10 μg of total RNAs for each sample and fragmented using divalent cations at elevated temperature. The cleaved RNA fragments were reverse-transcribed into first strand cDNA using random primers (Invitrogen Inc.), followed by second-strand cDNA synthesis. After end-repair processing, a single 'A' base was added to cDNA fragments at the 3' end. cDNAs were then ligated to adapters, purified by 2% agarose gel, and enriched by PCR to create the final cDNA library. Finally, paired-end sequencing was performed on an Illumina Genome Analyzer IIx using the standard protocol. The cDNA library of each sample was loaded to a single lane of an Illumina flow cell. Sequenced reads were generated by base calling using the Illumina standard pipeline. Each sample produced on average 65 million of paired-end raw sequence reads with a length of at least 68 bp.

RNA-seq data was provided by Prof. Jürgen Bauer and Dr. Benjamin Weide from the Department of Dermatology, University Hospital Tübingen.

7.2 Analysis of RNA-seq data

We performed RNA-seq data analysis based on the methods and pipelines described in Chapter 4 and Section 6.4.2

We used *hg19* as reference genome (downloaded from ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/H_sapiens/Assembled_chromosomes/) and a set of know human transcripts from RefSeq (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot, download date: 2009-08-07). Read mapping against both strands of *hg19* was performed using SplazerS with a required minimal identify of a match of 94%. No split mapping is performed. The same parameters are used for mapping against the RefSeq mRNA collection.

Tab. 7.2: Patient data for ten patients with metastatic melanoma included in the study.

Patient ID	gender	Tumor origin	Age at surgery (year)	Last observation	Status last observation
Pat01	m	liver	55 (2007)	2009	dead (melanoma)
Pat02	m	skin	72 (2011)	2011	dead (melanoma)
Pat03	m	skin	63 (2010)	2011	dead (melanoma)
Pat04	f	skin	43 (2011)	2011	dead (melanoma)
Pat05	m	skin	61 (2010)	2011	alive
Pat06	m	skin	58 (2010)	2011	alive
Pat07	m	lymph node	36 (2008)	2009	dead (melanoma)
Pat08	m	skin	52 (2010)	2011	alive
Pat09	f	lymph node	45 (2007)	2011	dead (melanoma)
Pat10	f	skin	58 (2011)	2011	alive

A minimal identity of 94% and a minimal length of 10 for a split match are required during conversion of mapping to transcripts to genomic positions.

An overview over the raw datasets as well as the number of mapped and unmapped reads per dataset is given in Tab. 7.3.

7.3 Viral integration

We applied the pipeline for the detection of viral transcripts as described in Sections 4.2 and 6.4.2. After digital transcriptome subtraction, reads that match to known viral sequences were found in each sample. The number of reads with viral hits varies between 94 and 2,053 (0.0004-0.003%) for the tumor samples and between 32 and 1867 (0.0001-0.0044%) for the control samples (PBMC). The results are summarized in Tab. 7.4. We compared the results to datasets used in a study by Arron *at al.* [97] (see Section 4.2.3). We included two samples that are known to be infected by HPV as positive control (ILS1933631 and HeLa), and two samples that are virus negative (STA01-106 and STA01-094). For the virus-positive samples we observe 627 and 2823 reads with viral hits (0.0154% and 0.146 %), for the virus-negative samples we observe 9 and 44 reads with a hit on a viral sequence (0.0002% and 0.0008%). The percentage of reads with viral hits for our melanoma samples lies within the range of the virus-negative samples or at least by a factor of ten below the lower value of the virus-positive samples (0.0154%). The number of viral hits in the virus samples is not generally higher than the number of viral hits observed in the PBMC samples. These results suggest that no virus is actively expressed in our melanoma samples.

To assess whether the few observed hits on viral sequences are sequencing artifacts or hint at the presence of a known virus we compared the corresponding reads to the non-redundant protein sequence collection and subjected the results to a taxonomic analysis with MEGAN4 [96]. Most of the reads had better matches to bacteria or eukaryotes, leaving only few potential viral hits. The distribution of the reads on the taxonomic tree appeared random, suggesting that all of these hits are artifacts (Fig. 7.1). An exception was a cluster of hits on primates. These hits probably represent human reads of the patient samples with sequencing errors that prevented their recognition during DTS.

We then considered all reads as potential viral reads that are assigned to a virus by the metagenomic analysis on the full non-redundant protein sequence collection. These potential viral reads seemed to be randomly distributed with mostly singular hits on different viral species, including phages (Fig. 7.2).

Many of these matches were due to repetitive sequences found in some viral species and are not specific. Fig. 7.3 shows two viral species that are overrepresented in Pat02: *Glypta fumiferanae ichnovirus* with 529 reads assigned and *Taterapox virus* with 289 reads

Tab. 7.3: Information on raw data and statistics of mapping for the RNA-seq data of the ten melanoma patients. Paired-end sequencing was performed, the number of reads is the total number of reads not the number of paired reads. The read length varied between 68 and 100 bp. For some of the probes multiple sequencing runs were performed. Sequencing data from different runs for the same sample were pooled before the analysis.
*Unique mappings contain reads that map uniquely to the reference genome or uniquely mapped partial reads that result from mapping to mRNA. The number of unique mappings thus does not correspond to the number of original reads that could be mapped to a unique position. Only unique mappings were used for variation detection.
** Data from Pat02 was not analyzed for genetic variations other than viral integration.

Patient ID	Tissue	Reads total	Mapped reads	Unique mappings*	Reads unmapped	% unmapped reads
Pat01	Tumor	116.3 mio	99 mio	37 mio	17 mio	14,5%
	PBMC	43.1 mio	37 mio	34 mio	6 mio	14.0%
Pat02**	Tumor	180.0 mio	163 mio	163 mio	17 mio	9.2%
	PBMC	21.1 mio	18 mio	18 mio	2.5 mio	12.1%
Pat03	Tumor	85.2 mio	72 mio	68 mio	14 mio	16.0%
	PBMC	25.4 mio	23 mio	22 mio	3 mio	11.4%
Pat04	Tumor	27.0 mio	25 mio	23 mio	2 mio	7.5%
	PBMC	30.9 mio	30 mio	25 mio	1 mio	4.3%
Pat05	Tumor	20.0 mio	18 mio	17 mio	2 mio	8.7%
	PBMC	53.8 mio	48 mio	36 mio	6 mio	10.8%
Pat06	Tumor	13.9 mio	13 mio	12 mio	16 mio	7.3 %
	PBMC	40.8 mio	37 mio	35 mio	4 mio	9.7%
Pat07	Tumor	51.4 mio	45 mio	43 mio	7 mio	13.2%
	PBMC	41.2 mio	34 mio	33 mio	7 mio	17.0%
Pat08	Tumor	61.6 mio	58 mio	38 mio	3 mio	5.2%
	PBMC	28.7 mio	27 mio	16 mio	2 mio	7,11%
Pat09	Tumor	75.3 mio	67 mio	61 mio	8 mio	11.1%
	PBMC	12.4 mio	10 mio	10 mio	2 mio	15.0%
Pat10	Tumor	29.0 mio	26 mio	26 mio	3 mio	9.0%
	PBMC	31.0 mio	28 mio	28 mio	3 mio	9.6%

Tab. 7.4: Total number of reads per sample, along with the number of reads with a match on a virus and the percentage of reads with match on a virus. For comparison, read counts for samples from a study by Arron *at al.* [97] are also included (see also Section 4.2.3). Sample ILS1933631 and HeLa are known to be infected with HPV and serve as positive control. Arron STA01-106 and Arron STA01-094 are known to be negative for viral sequences and are used as a negative control.

Patient ID	Tissue	Reads total	Reads with viral hit	% reads with viral hits
Pat01	Tumor	116.3 mio	718	0.0006%
	PBMC	43.1 mio	1,471	0.0034%
Pat02	Tumor	180.0 mio	2,053	0.0011%
	PBMC	21.1 mio	328	0.0006%
Pat03	Tumor	85.2 mio	503	0.0006%
	PBMC	25.4 mio	165	0.0006%
Pat04	Tumor	27.0 mio	94	0.0003%
	PBMC	30.9 mio	32	0.0001%
Pat05	Tumor	20.0 mio	258	0.0013%
	PBMC	53.8 mio	1,745	0.0032%
Pat06	Tumor	13.9 mio	422	0.003 %
	PBMC	40.8 mio	232	0.0006%
Pat07	Tumor	51.4 mio	1,126	0.0022%
	PBMC	41.2 mio	1,867	0.0045%
Pat08	Tumor	61.6 mio	258	0.0004%
	PBMC	28.7 mio	98	0.0044%
Pat09	Tumor	75.3 mio	881	0.0012%
	PBMC	12.4 mio	456	0.0044%
Pat10	Tumor	29.0 mio	159	0.0005%
	PBMC	31.0 mio	168	0.0005%
ILS1933631	Cervical SCC	4.1 mio	627	0.0154%
HeLa	cell line	1.9 mio	2,823	0.146%
Arron STA01-106	SCC	3.6 mio	9	0.0002%
Arron STA01-094	Skin	5.3 mio	44	0.0008%

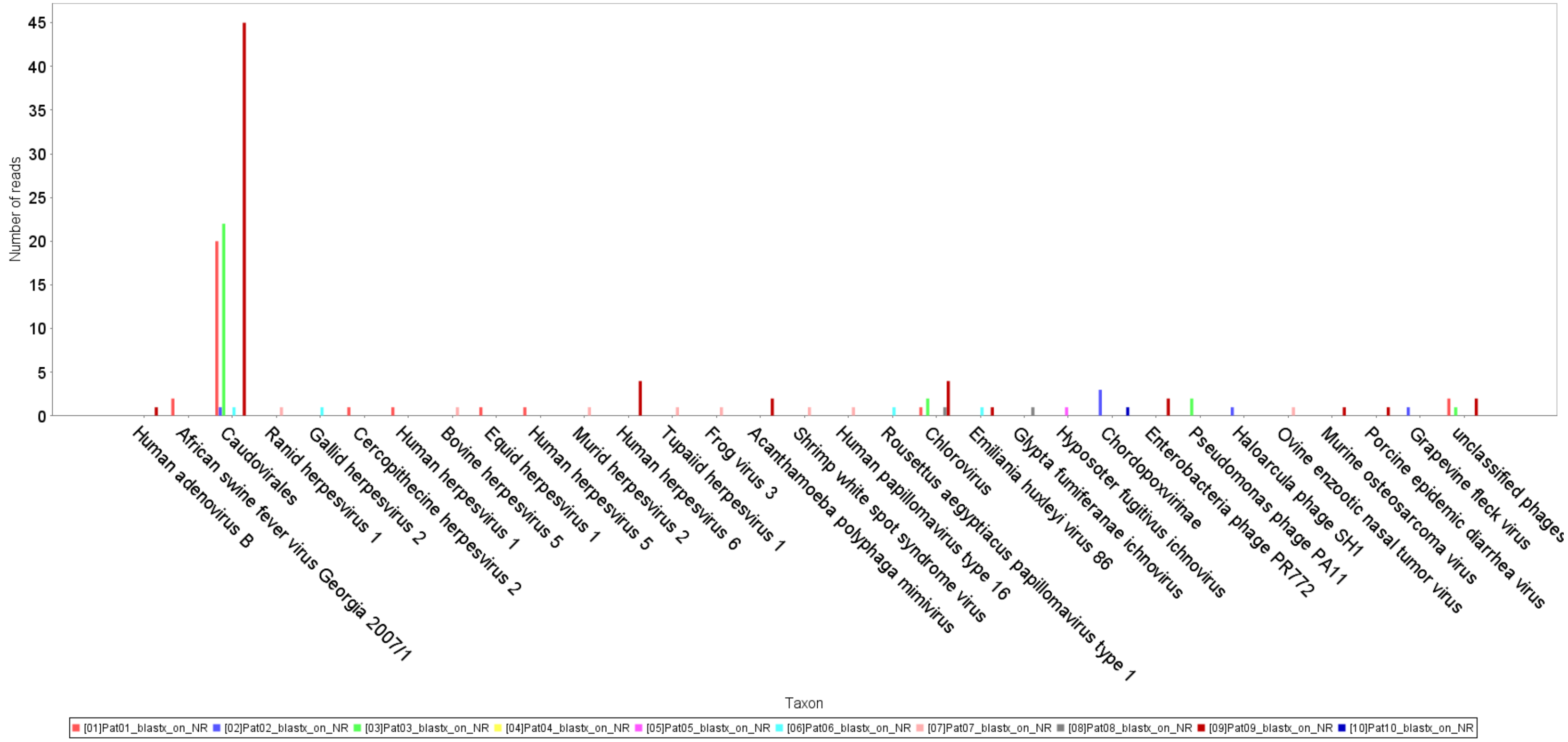


Fig. 7.1: Reads from 10 melanoma samples matching the viral RefSeq database were matched to non-redundant protein sequence collection (NR) using blastx and placed in a taxonomic tree using the tool MEGAN4. All hits on viral sequences are shown, reads with better matches on eukaryotes or bacteria are collapsed in the node "cellular organisms" for clarity. The number of reads that are assigned to each taxon is listed behind the taxon name. The hits assigned to Caudovirales are evenly distributed over about 40 different phages, the subtree is collapsed for clarity. The remaining hits seem to be randomly distributed with no viral group or species standing out, except for herpesviridae.

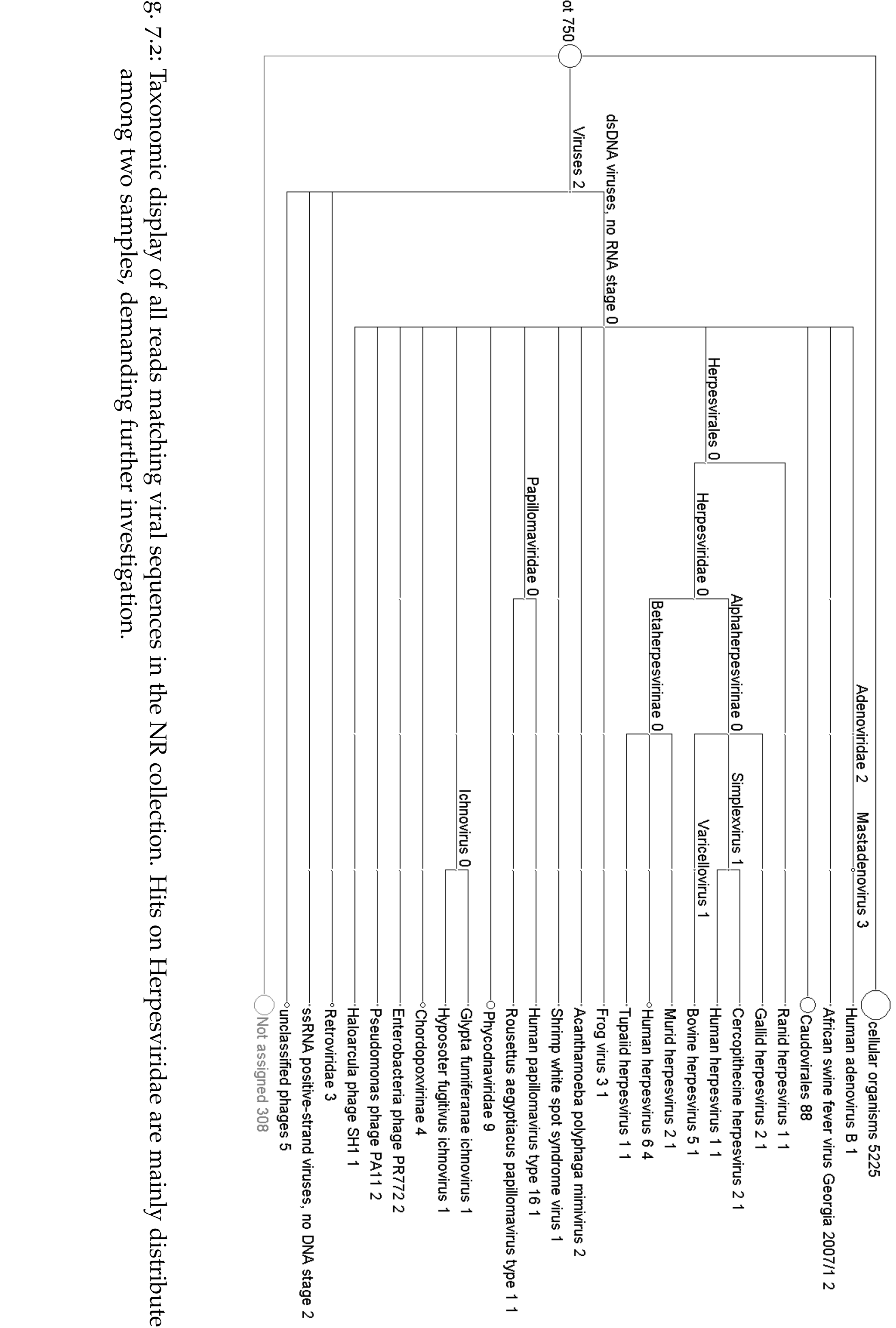
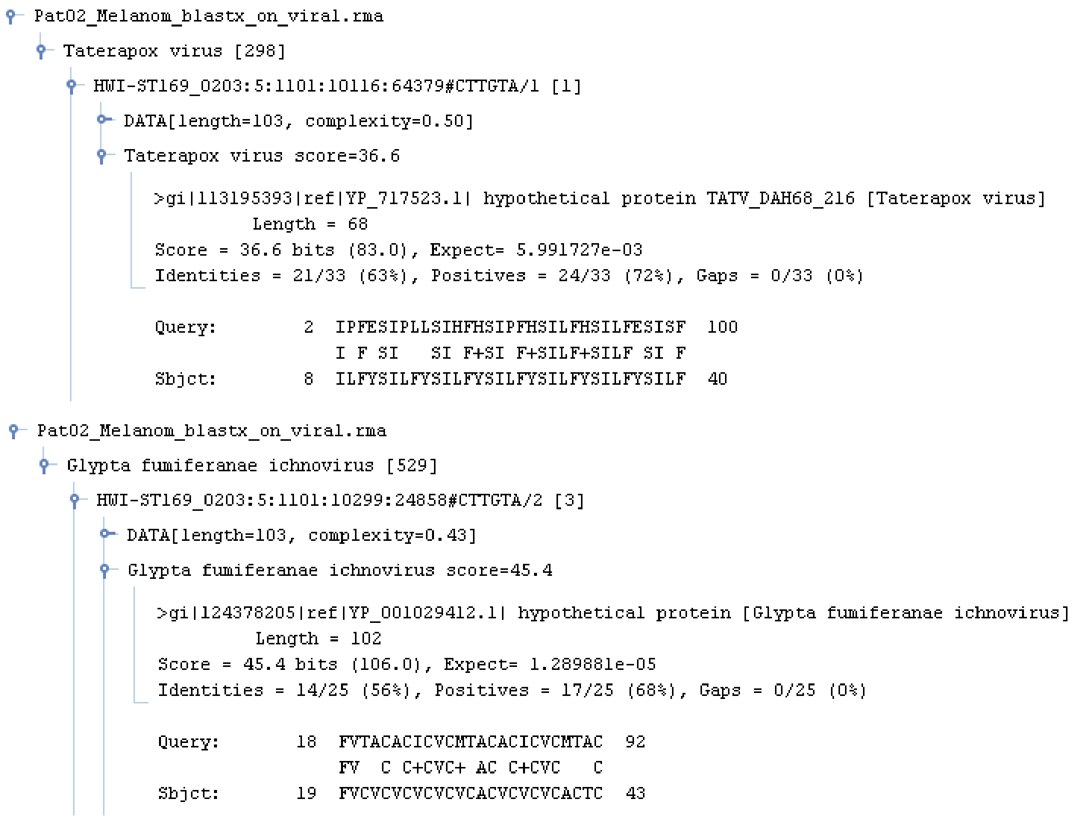


Fig. 7.2: Taxonomic display of all reads matching viral sequences in the NR collection. Hits on Herpesviridae are mainly distributed among two samples, demanding further investigation.



```
Pat02_Melanom_blastx_on_viral.rma
  Taterapox virus [298]
    HWI-ST169_0203:5:1101:10116:64379#CTTGTA/1 [1]
      DATA[length=103, complexity=0.50]
      Taterapox virus score=36.6
        >gi|113195393|ref|YP_717523.1| hypothetical protein TATV_DAH68_216 [Taterapox virus]
                Length = 68
        Score = 36.6 bits (83.0), Expect= 5.991727e-03
        Identities = 21/33 (63%), Positives = 24/33 (72%), Gaps = 0/33 (0%)

        Query:        2  IPFESIPLLSIHFHSIPFHSILFHSILFESISF  100
                         I F SI   SI F+SI F+SILF+SILF SI F
        Sbjct:        8  ILFYSILFYSILFYSILFYSILFYSILFYSILF  40

Pat02_Melanom_blastx_on_viral.rma
  Glypta fumiferanae ichnovirus [529]
    HWI-ST169_0203:5:1101:10299:24858#CTTGTA/2 [3]
      DATA[length=103, complexity=0.43]
      Glypta fumiferanae ichnovirus score=45.4
        >gi|124378205|ref|YP_001029412.1| hypothetical protein [Glypta fumiferanae ichnovirus]
                Length = 102
        Score = 45.4 bits (106.0), Expect= 1.289881e-05
        Identities = 14/25 (56%), Positives = 17/25 (68%), Gaps = 0/25 (0%)

        Query:       18  FVTACACICVCMTACACICVCMTAC  92
                         FV  C C+CVC+ AC C+CVC   C
        Sbjct:       19  FVCVCVCVCVCACVCVCVCACTC  43
```

Fig. 7.3: Exemple view of a hits on overrepresented viral sequences. The viral sequences are hypothetical proteins and are likely not to be real protein sequences.

assigned. We use the "Inspect" tool provided for taxonomic nodes in MEGAN4. The hits on these viral sequences are significant based on the blastx bit scores and E-values. However, a closer look at the viral protein sequences reveals that the sequences are suspect. The reads match on hypothetical viral sequences. The sequence for the *Glypta fumiferanae ichnovirus* hypothetical protein contains a large stretch of repeated "VC" (FVCVCVCVCVCACVCVCVCACTC). In addition, the host of *Glypta fumiferanae ichnovirus* is a wasp species. The hypothetical protein for *Taterapox* virus contains the subsequence "FYSIL" repeatedly. These observations, together with the facts that the alignments contain numerous gaps, make it highly unlikely that the observed hits represent real viral sequences that are present in the sample. We therefore ignore such hits.

However we noted that five samples contained one to four hits on viruses of the Herpesviridae family (Tab. 7.5). Different herpes viruses are known to cause cancer, such as HHV-8 causing Kaposi's sarcoma and Epstein-Barr virus (HHV-4) causing lymphomas [175]. As herpesviridae are potential candidates for causing cancer, we further investigated the samples containing reads that matched herpes virus sequences.

Tab. 7.5: Number of reads that match to a herpes sequence.

Sample ID	Species	Number of hits
Pat01	Cercopithecine Herpesvirus 2	1
	Human Herpesvirus 1	1
Pat06	Gallid Herpesvirus 2	1
Pat07	Ranid Herpesvirus 1	1
	Bovine Herpesvirus 5	1
	Murid Herpesvirus 2	1
	Tuapiid Herpesvirus 1	1
Pat09	Human Herpesvirus 6	4
Pat10	Simplexvirus	1

If there were transcripts of viral origin in our samples, it would be unlikely to find only single reads of these transcripts in our datasets. To exclude the possibility that we missed reads generated from the transcript of an unknown virus, we assembled all reads that were left after the first step of DTS (all reads that were not filtered out because of high similarity to human sequences) using Velvet [99]. No contigs longer than 250 bp were found and none of these contigs resembled herpes virus sequences. A real viral transcript should generate more reads that can be assembled into a longer contig. Thus, the hits on herpesviridae found in our samples are assumed to be artifacts.

Discussion The high number of short reads generated by Illumina sequencing leads also to a high number of potential viral hits, up to a point where manual analysis becomes impractical. We therefore analyzed all potential viral sequences with the metagenomics tool MEGAN4 that displays all reads in a taxonomic tree. We would expect to find any reads matching an unknown virus in one of the lower "virus" nodes and a set of subnodes, whilst artifacts are randomly distributed. One single read of a viral transcript is probably not sufficient to stand out in the taxonomic analysis. However, our positive controls (that consisted of less than 4 million reads) show that an actively transcribed virus will generate hundreds or thousands of hits that stand out in a taxonomic analysis (see Fig. 4.3). It should thus be easy to identify a viral infection in transcriptome datasets of 13 to 180 million reads even if we assume a level of transcription that is much lower than in HPV-infected cells.

It has to be kept in mind that DTS can only detect potential viral sequences under two conditions: a) The virus must share sequence similarities with known viruses. Our reads are compared with a database containing all known viral sequences, a viral read that has no similarity with a known virus would not be detected. However, we have assembled all non-human reads from our samples to find contigs that are long enough to be a transcript, potentially enabling us to find even an unknown virus. b) The virus has to be transcribed. An analysis of the transcriptome cannot detect a virus that is present in the sample cells but is not transcribed into mRNA. We have performed our study on transcriptome data because most known carcinogenic viruses are transcribed in tumor cells. There is however the possibility of a hit-and-run mechanism, where a viral infection initiates tumor growth, but viral transcripts are not needed for tumor maintenance [98]. In such a case it is even possible that the viral genome gets lost again, leaving no easily detectable evidence for a viral presence.

In conclusion, we found no viral transcripts in 10 melanoma metastases after deep-sequencing of the whole transcriptomes. It is thus unlikely that transcribed viruses play a role in the development of these melanomas.

A critical issue in DSS is the quality of the viral reference sequences used for comparison. As seen for the hits on *Glypta fumiferanae ichnovirus* or *Taterapox*, low complexity reference sequences can lead to false positive hits. Our pipeline includes a low-complexity filter for the reads, but not for the reference sequences. The choice of used reference sequences is critical and should be closely adapted to the question in mind. If one searches for a specific viral species, using all known strains of just this species would be a way to obtain a highly sensitive and specific comparison. In the application presented here, we want to be able to also identify unknown viruses that only remotely resemble a known virus. Using the full set of known viral sequences is then a suitable choice, although this increases the chance of irrelevant positive hits.

The pipeline we present still requires a manual or visual inspection of the potential viral reads. An even more automated method that does not require this manual step at the end would be desirable. Such fully automated methods can be developed if appropriate training data to train and fine-tune the filtering steps is available. The simulation of read data that contain different levels of different viral sequences can cover the theoretical aspects of such training data (influence of sequencing coverage, sequencing errors, viral diversity). However, there is no way to simulate the influence of experimental steps, for example which transcription level is required so that the viral sequences get actually captured by the sequencing. In addition, we do not know what level of similarity between an unknown virus and a known virus we can expect. A fully automated detection would require a definition and a quantitation of these parameters, and also a reliable set of reference sequences. With these issues in mind, the effort of a visual inspection of the remaining promising hits seems to be a good trade of between automated analysis and reliability of the results.

7.4 Analysis of SNVs found in the melanoma samples

Genetic variation was computed using the workflow described in Section 6.4.2. The settings and parameters used for the different steps in the workflow are listed in the following if they differ from the tool defaults:

- **SNPstore**: Only successful SNVs are listed in the output (-of 1), hide quality in output file (-hq), Illumina-quality encoding for reads is used (-sq0), pile-up correction is performed on merged lanes (-mmp).
- **annotate SNVs**: Gene annotation from RefSeq and SNP information from dbSNP132 are used for SNV annotation using ANNOVAR. *Tumor:* Tumor SNVs are annotated with information on the control tissue (SNVs in control tissue and coverage information for tumor SNV positions in control tissue). The original (unfiltered) control SNVs are used. *Control:* SNVs in the control tissue are filtered for coverage (>20) and quality (>20).
- **Filter somatic SNVs**: Tumor SNVs are filtered for non-synonymous SNVs. In addition, low-stringency filters are applied concerning coverage (< 10) and quality (< 10) in order to keep the number of candidates high. More stringent filters can be applied in later steps.
- **Predict SNV Epitopes**: HLA binding prediction is performed using the prediction method SYFPEITHI [116] with halfmax-values as binding thresholds. Predictions are performed for all of the patient's HLA alleles where a SYFPEITHI-prediction model is available (see Tab. 7.1). Predictions were only performed for peptides of length nine. The filtered SNV lists for the respective control tissues are used to generate patient- and tumor-specific protein sequences.

In the following sections we present an in-depth analysis for SNVs observed in nine melanoma patients (Pat02 was excluded from the analysis).

7.4.1 Somatic SNVs

The raw lists of SNVs in the tumor are analyzed further. As a first step we filter for mutations that lead to a tumor-specific variant, i.e. that, at a specific position, we observe a new allele compared to the respective control sample. This filter is applied by comparing the observed genotypes for tumor and control samples. For all tumor SNV positions the genotype is obtained directly as a result of SnpStore. To obtain the genotype for the control sample at a specific position we first check if a SNV was called in the control sample. If so, the called genotype is taken. If no SNV was called in the control sample, we

Tab. 7.6: Number of tumor mutations for different filter criteria. All filters that are listed in the column headers are applied. **TSA:** a Tumor-Specific Allele is observed. At least one of the bases observed in the tumor sample is not observed in the respective control sample. If no coverage is observed in the control, we assume that only the wildtype (hg19) is present in the control sample. **dbSNP:** SNVs that are annotated in dbSNP are excluded. **Coverage Control:** A non-zero coverage of the tumor SNV position in the control tissue is required.

Patient ID	TSA	TSA-dbSNP	TSA-CC	TSA-CC-dbSNP
Pat01	1,251	598	67	26
Pat03	3,689	2,800	689	318
Pat04	473	104	20	7
Pat05	1,712	1,129	59	21
Pat06	138	39	5	0
Pat07	7,956	7,125	140	42
Pat08	6,766	6,148	46	18
Pat09	23,225	21,891	176	53
Pat10	4,023	3,287	75	22

analyze the observed bases in the control obtained from the Position-Coverage-Lookup based on SnpStore. If the coverage of a tumor SNV position in the corresponding control sample is 0, we assume a homozygous genotype for the wildtype (the base that occurs at this position in the reference genome). To reliably decide that a mutation is present in the tumor but not in the control, a sufficient coverage in the control sample is needed. The coverage in the control tissue is the most restricting factor in this analysis as shown later in this section.

We additionally filter for SNVs that are not annotated in dbSNP and for a non-zero coverage in the control sample. The number of observed SNVs for the different combinations of these filters are summarized in Tab. 7.6.

We could not find a simple explanation for the large divergence between the number of mutation observed for the different patients. An extremely large number of mutations was observed for Pat09, the number of real candidates for somatic mutation for this patients however lies in the range of the numbers observed for the other patients.

For SNVs with no coverage in the control sample we cannot decide if the observed mutation is a real somatic mutation of if it is also present in healthy tissues of the same patient but not observable in the data. When applied to RNA-seq data, the coverage correlates with the expression of the respective genes. SNVs can therefore only be observed if they are sufficiently expressed. We use PBMCs samples as control for samples from melanoma metastases and thus expect the gene expression to differ between tumor and control samples. To assess the influence of gene expression or transcript abundance on the detection of somatic mutations we computed RPKM values (Reads Per Kilobase of exon model per Million mapped reads) as estimate for the transcript abundance. Computation of RPKMs for genes is based on Python code provided by Ramsköld *et al.* [176]. The tool was also integrated in our Galaxy server but for simplicity is not shown as part of our standard NGS analysis pipelines. We use the RefSeq genes as reference to compute RPKMs. As a basic estimate for the similarity in the gene expression between samples we use Pearson's correlation coefficient. The average correlation between the RPKMs for a tumor sample and the corresponding control sample is 0.55. The average correlation between all tumor samples is 0.6, and the average correlation between all control samples is 0.83. The highest average correlation is observed for the control samples. This is what one would expect, since these samples come from the same tissue and are supposed to have similar gene expression. The correlation between the tumor samples is lower. The samples come from melanoma metastases from different locations (skin, liver, lymph nodes) and are assigned to different disease states, so we expect these samples to be more divers than the PBMC samples. The correlation between the tumor and the corresponding PBMC samples is even lower. Since two very different tissues are compared we do not expect a high correlation in gene expression.

Additionally, we analyzed the number of expressed genes per samples. As proposed by Ramsköld *et al.* we use a threshold of 0.3 RPKM to detect expressed genes. The number of expressed genes in our samples varies between 56% and 70% of the RefSeq genes included in the RKPM analysis, which is in agreement with the 60-70% of expressed genes observed by Ramsköld *et al.* [176].

Due to the difference in the gene expression between tumor and control samples, the parts of the transcriptome that are available for the detection of genetic variants also differs between the samples. This leads to a high number of mutations observed in the tumors that cannot be confirmed or rejected as somatic mutations due to missing coverage in the control sample.

7.4.2 Genes affected by somatic SNVs

For the following analysis we only look at the genes in which a mutation has been found, disregarding of the exact position of the mutation in the gene. The number of genes with at least one mutation is summarized for all patients in Tab. 7.7. The number of genes in the group containing the most promising candidates for somatic mutations (non-zero coverage in the control sample and no annotation in dbSNP) varies between zero for patient Pat06 (zero candidates for somatic mutations for that patient were found) and 280 for Pat03.

Tab. 7.7: Number of genes with at least one mutation. All filters that are listed in the column headers are applied to select a set of SNVs for the analysis. **TSA:** a Tumor-Specific Allele is observed. At least one of the bases observed in the tumor sample is not observed in the respective control sample. If no coverage is observed in the control, we assume that only the wildtype (hg19) is present in the control sample. **dbSNP:** SNVs that are annotated in dbSNP are excluded. **CC:** A non-zero coverage of the tumor SNV position in the control tissue is required.

Patient ID	TSA	TSA-dbSNP	TSA-CC	TSA-CC-dbSNP
Pat01	737	445	49	12
Pat03	1,322	1,016	346	280
Pat04	278	83	17	4
Pat05	1,011	819	41	11
Pat06	83	29	5	0
Pat07	2,540	2,469	100	24
Pat08	2,525	2,461	27	6
Pat09	2,603	2,600	154	42
Pat10	1,629	1,532	51	13

For the data presented in Tab. 7.8 the SNV candidates from all patients are pooled. The number of genes with at least one mutation in at least one patient and the number of genes that have a mutation in more than one patient for each of the groups is listed. In addition, the maximal number of patients with a mutation in the same gene and the maximal number of total mutations observed in one gene are listed. The maximal number of mutations per gene is extremely high, especially for the groups where mutations are not filtered for coverage in the control tissue.

The gene that mainly contributes to these high numbers is PLEC. The protein encoded by PLEC, plectin, is expressed in nearly all mammalian cells and acts as a link between the main components of the cytoskeleton. PLEC has been previously reported in the context of cancer. Somatic mutations in PLEC have been reported for different cancer types according to the COSMIC database [80], however, only single somatic mutations are observed in the same tumor. PLEC has also been reported to be highly expressed in colorectal adenocarcinoma according to BioGPS [177]. We could, however, not find a clear association between PLEC and cancer in the literature. The very high number of mutations observed in one gene is suspicious and could also indicate problems with the reference sequence or with read mapping rather than real mutation hot-spots. Possible explanations are that the reads come from a similar but unknown gene or larger differences between the transcript sequences and the respective sequence stretch in the reference genome. If the transcript sequence diverges from the reference genome sequence, reads that are aligned to the transcripts and mapped back to the reference genome can lead to false positive mutations. Other possible explanation could be the

Tab. 7.8: Pooled SNV data for all patients. The same filters as for the data presented in Tab. 7.7 are used (**TSA-dbSNP-CC**). The number of genes with at least one mutation in at least one patient and the number of genes that have a mutation in more than one patient for each of the groups is listed. In addition, the maximal number of patients with a mutation in the same gene and the maximal number of total mutations observed in one gene are listed.

	TSA	**TSA-dbSNP**	**TSA-CC**	**TSA-CC-dbSNP**
Genes with ≥ 1 mutation	737	445	49	12
Genes mutated in > 1 patient	6,418	1,596	80	20
Max. number of patients	9	8	8	8
Max. mutations in one gene	1,399	1,389	51	47

expression of similar (pseudo) genes, or an erroneous amplification of PLEC. Most of mutations observed in PLEC cannot be validated as somatic mutations due to insufficient coverage in the normal tissues. We thus exclude PLEC from the in-depth analysis of the mutations in the following. The reason for the high number mutations in PLEC should however be investigated in a separate analysis. This could include the analysis of copy number variations, sequencing data from other cells using the same protocols to address the issue of problems with the reference sequence, and a more thorough literature search.

If we exclude PLEC from the analysis, the highest numbers of mutations found in one gene is 81 for the group **TSA** and 57 for the group **TSA,dbSNP**, in both for the gene HSPG2. These numbers are still relatively high, but the mutations are distributed over eight patients, leaving a more reasonable average number of 10 mutations per patient in the gene.

For the two groups that are filtered for coverage in the control tissue (**TSA-CC** and **TSA-CC-dbSNP**) the highest number of mutations are observed for the gene GLUD2 (51 and 47 mutations), mutations are found in eight of the nine patients. The number of mutations observed in gene GLUD2 is suspiciously high and mutations in GLUD2 are excluded from further analyses. In the group **TSA-CC** the gene PDE4DIP has 24 mutations in eight patients, these SNVs are however all annotated in dbSNP and therefore less likely to be real somatic mutations.

Recurrent SNVs

We also analyzed the SNVs for recurrent mutation. For the group **TSA-CC** we observe in total 879 mutated positions. 59 mutation occur in more than one patient. If we remove all SNVs that are annotated in dbSNP from this list (corresponding to **TSA-dbSNP-CC**) we obtain 446 mutated positions and 25 positions that are mutated in more than one patient. 10 of the 25 positions mutated in multiple samples (three to seven) are in the

Tab. 7.9: Mutations that are observed in more than one patient. The affected gene, the genomic positions, the observed bases, and the number of patients with that mutation are listed. All mutations are heterozygous. The wildtype ist given fist. The tumor variant is written in bold.

Gene	Position	Observed Bases	Number of Patients
ETFB	chr19:51857565	**G**/T	4
ETFB	chr19:51857567	**C**/A	4
TRAM1L1	chr4:118005979	**G**/C	4
TRAM1L1	chr4:118006021	**A**/G	4
RL28	chr19:55898064	**G**/C	3
SDF4	chr1:1153073	**G**/A	2
PIGP	chr21:38444840	**G**/C	2
PIGP	chr21:38444841	**T**/A	2
NEIL1	chr15:75646086	**A**/G	2
MPHOSPH6	chr16:82203743	**A**/C	2
TMEM161A	chr19:19243176	**G**/C	2
ZNF860	chr3:32030606	**A**/G	2
XPO5	chr6:43492579	**T**/A	2
SNX8	chr7:2297007	**A**/G	2

gene GLUD2. Mutations in GLUD2 are ignored, as described above.

The remaining 15 recurrent mutations are listed in Tab. 7.9. In all cases the base that is reported as the tumor-specific base (termed tumor variant in the following) is the wildtype at the respective position.

Gene ETFB has two mutations in four patients. ETFB is is an electron transfer flavoprotein and shuttles electrons between primary flavoprotein dehydrogenases involved in mitochondrial fatty acid and amino acid catabolism and the membrane-bound electron transfer flavoprotein ubiquinone oxidoreductase. Both mutations occur in the four patients (Pat01, Pat03, Pat09 and Pat10), the genotype for the SNV positions is heterozygous in all patients. The SNVs affect neighboring amino acids in the protein sequence, but do not affect an annotated protein domain. In both cases, the tumor-specific allele is the wildtype. In Pat01 and Pat03 the mutation is also observed in the control tissues but below the detection threshold (one of six reads for Pat01 and one of seven reads for Pat03). Patients Pat09 and Pat10 are homozygous for the non-wildtype with a coverage of four and five for the respective positions.

Translocation-associated membrane protein 1-like 1 (TRAM1L1) also has two mutations that occur in four patients (Pat01, Pat05, Pat07, Pat10). Both mutations also occur in the respective control tissues, the control tissues are homozygous for the mutation, the tumors are heterozygous and express mutation and the wildtype. Both mutations **F**177L and **Q**190E lie within an annotated protein domain (PROSITE accession PS50922).

Ribosomal protein L28 (RL28) is reported to be mutated in three patients (Pat01, Pat09, Pat10), the tumor-specific base is again the wildtype. A manual inspection of the coverage data for the control tissue reveals that in all control samples the wildtype is also observed but with a frequency below the detection threshold for a heterozygous genotype.

Stromal cell derived factor 4 (SDF4) is mutated in patients Pat03 and Pat08. In Pat03 the tumor variant is also observed in one of six reads. The coverage of the respective position in the control tissue for Pat08 is five, the tumor variant is not observed.

For PIGP the two mutations affect neighboring positions. The coverage in the control tissue is very low (<5) and in one of the patients the tumor variant is also observed with a low frequency.

Manual inspection for the mutation observed in gene NEIL1 shows that for one patient (Pat03) the tumor variant is also observed in the control sample (one of five reads). The second patient (Pat05) is homozygous for the mutation with a coverage of ten. The wildtype can therefore be considered as relatively reliable tumor-specific base in Pat05, however the mutation cannot be viewed as recurrent.

The tumor variant in the mutation observed in gene M-phase phosphoprotein 6 (MPHOSPH6) in two patients is the wildtype, as for all previous mutations analyzed in this section. The control samples for the two patients (Pat7 and Pat10) are homozygous for the mutation with a coverage of five and eight respectively.

For both patients with the mutation in gene TMEM161A (Pat07 and Pat09) the tumor variant is observed in one of four reads in the control sample.

The mutation in ZNF860 is observed in Pat08 and Pat09, the tumor variant is again the wildtype. For Pat08 the mutation is observed in eight of 18 reads and in the tumor and in nine of nine reads in the control. The tumor variant (wildtype) is observed in ten of 18 reads in the tumor but not in the control. In Pat09 the tumor variant is only observed in two of ten reads in the tumor and in none of the four reads in the control.

For both patients showing a mutation in XPO5 the tumor variant is also observed in the control samples but lies below the detection threshold. The same applies for the mutations observed in gene SNX8.

For most of the SNVs analyzed here, the tumor-specific allele is also detected in the control tissue with a frequency below the detection threshold, mostly due to low coverage in the control samples. This shows that the divergence in the coverage of the different samples makes the detection of somatic and recurrent mutations a hard problem. Thresholds for the detection of SNVs have to be chosen carefully, and, in case of low coverage, useful criteria to automatically decide if a mutation is tumor-specific are hard to define.

Mapping the mutations to biological pathways

The genes with somatic mutations are mapped onto known biological pathways. For all genes from the group **TSA-CC-dbSNP** we obtain all KEGG pathways that are linked to these genes. For all pathways that are affected by at least one mutation we count the number of overall mutations in all genes belonging to the pathway and the number of patients that have a mutation in one of the genes in the pathway. The KEGG API is used to map the genes to the pathways.

In total we obtain 157 different pathways with at least one mutation. 67 of these pathways are only affected by one mutation in one gene in one patient. Mapping mutations to pathways can help to reveal mutations that are relevant in a cancer. What this kind of simple analysis cannot do, however, is to reveal significantly affected pathways.

As for other analyses presented here the results depend on the quality and reliability of the underlying SNV data. In order to identify frequently affected pathways we need the information on the presence and absence of mutations in genes. With RNA-seq data we can identify SNVs, in case of no reported SNV at a position we however cannot be sure that this is a true negative. Statistical methods for the detection of significantly affected pathways exists for other types of omics data, for example Gene Set Enrichment methods (GSEA) [178]. These methods are mainly based on differentially expressed genes or proteins. Pathway enrichment methods have also been proposed for SNP data [179]. The SNP data is however generated from genome-wide association studies where SNP genotyping is available for a set of SNPs in large cohorts of individuals. We have a very low number of patients and in addition we only have a list of SNVs found in each patient and no information on the respective positions in the other patients. The data we have is very sparse and existing statistical analyses fail to identify significantly affected or enriched pathways on our data.

We excluded all pathways that contained mutations form HLA genes or from GLUD2. We also removed pathways that are associated with biosynthesis and diseases other than cancer. From the remaining 40 pathways, 10 are mutated in two patients. We present three of the pathways that are affected by a mutations. The 40 pathways, together with the number of mutations found are listed in Tab. A.1.

All pathway maps are generated using the KEGG API. By default, all genes that belong to a pathway in the organism under consideration are colored green. Genes in white are genes with similar function in different organisms. For all pathways presented here, the color for the mutated genes is based of a scoring scheme that combines the total number of mutations in the gene with the number of patients that have a mutation in the gene (number of mutations * number of patients with mutations in this gene). The scale is generated on the global maximum of the scores observed in any of the pathways, so the colors for different pathways are comparable. The color scale ranges from red (low mutation score) to yellow (high mutation score).

Only one pathway (Base excision repair, KEGG ID: hsa03410) is mutated in three patients. The pathway map is shown in Fig. 7.4. The mutated genes are NEIL1 and PARP4, mutations are observed in Pat03 and Pat05 for gene NEIL1 and for Pat09 for gene PARP4.

The pathways with the highest number of mutations is "Pathways in Cancer" hsa05200. The pathway is shown in Fig. 7.5. The mutated genes are FOXO1, BAX, APC, BRCA2, STK36, BCR, and CASP9.

Th pathway "Wnt signaling" (hsa04310) shows mutations in four genes, but all mutations occur in the same patient. The mutated genes are APC, NFATC3, VANGL1, and CSNK2A2. The pathway map is shown in Fig. 7.6.

The data we used for the pathway analysis is very sparse. We could identify cancer-related pathways that are mutated in single patients, however, no pathway that was affected in a larger percentage of the patients could be identified.

Mutations in BRAF

V600E in the gene BRAF is a common mutation in malignant melanoma [11]. In our nine melanoma samples we did not observe this or any other mutation in BRAF. As for all analyses presented here, missing coverage at a position of interest is a main issue with RNA-seq data. We therefore analyzed the coverage of BRAF in more detail. For all samples (tumors and controls) the RPKMs are above 0.3 and BRAF is therefore considered to be expressed in all samples (see Section 7.4.1 for details on the gene expression analysis). We analyzed the coverage of the respective genomic position (chr7:140,453,136) in detail for the nine tumor samples. A mutation from A to T in the genome and a T to A mutation at the position 1,796 in the respective transcript correspond to the V600E mutation. Results are summarized in Tab. 7.10.

The read counts observed in the nine melanoma samples of the position chr7:140,453,136 varies between zero and nine and is thus for all samples below the coverage threshold we require to accept a mutation. For two patients (Pat05 and Pat06) no reads were aligned to this genomic position. Reads with the A to T mutation at the respective position are only observed in one of the nine patients (Pat08). Five of the nine observed reads (56%) for Pat08 have a T instead of an A. The low coverage for position chr7:140,453,136 underlines the problems of low coverage for the detection of mutations in RNA-seq data.

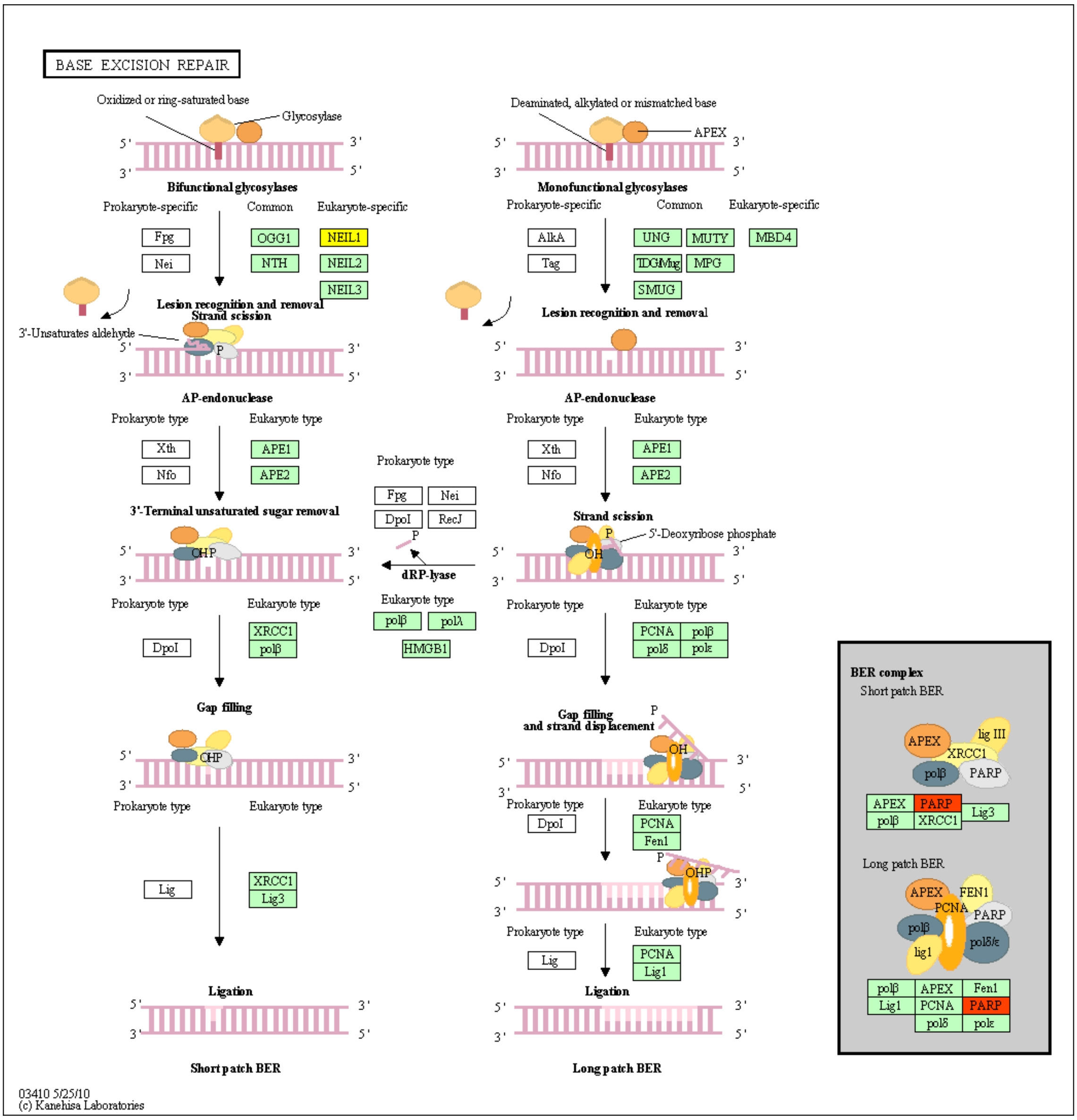


Fig. 7.4: Pathway map for the Pathway "Base excision repair" (hsa03410). The genes are colored by the number of mutations observed in the genes. Genes NEIL1 and PARP4 are affected by a mutation.

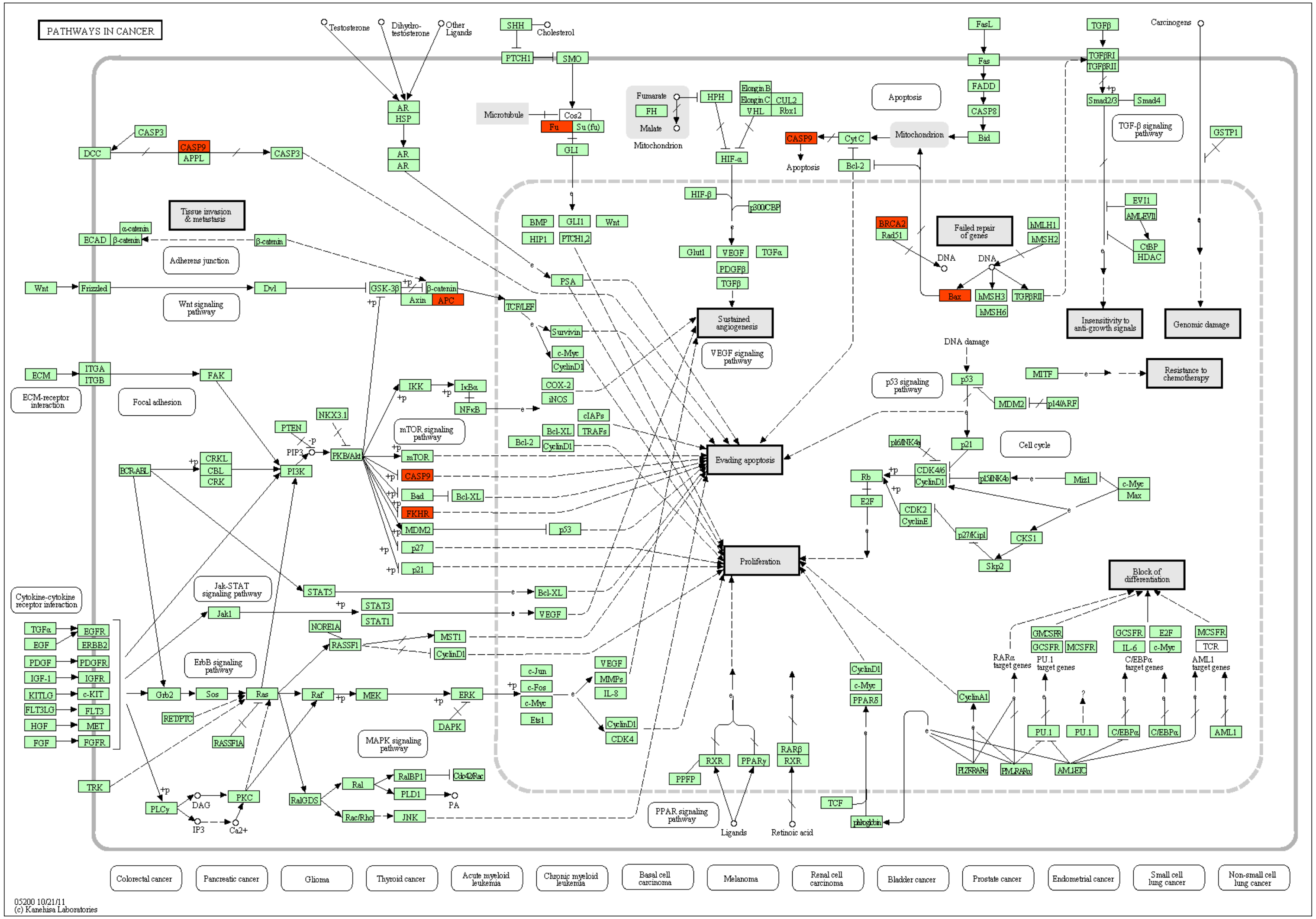


Fig. 7.5: Pathway map for the pathway "Pathways in Cancer" (hsa05200). The genes are colored by the number of mutations observed in the genes.

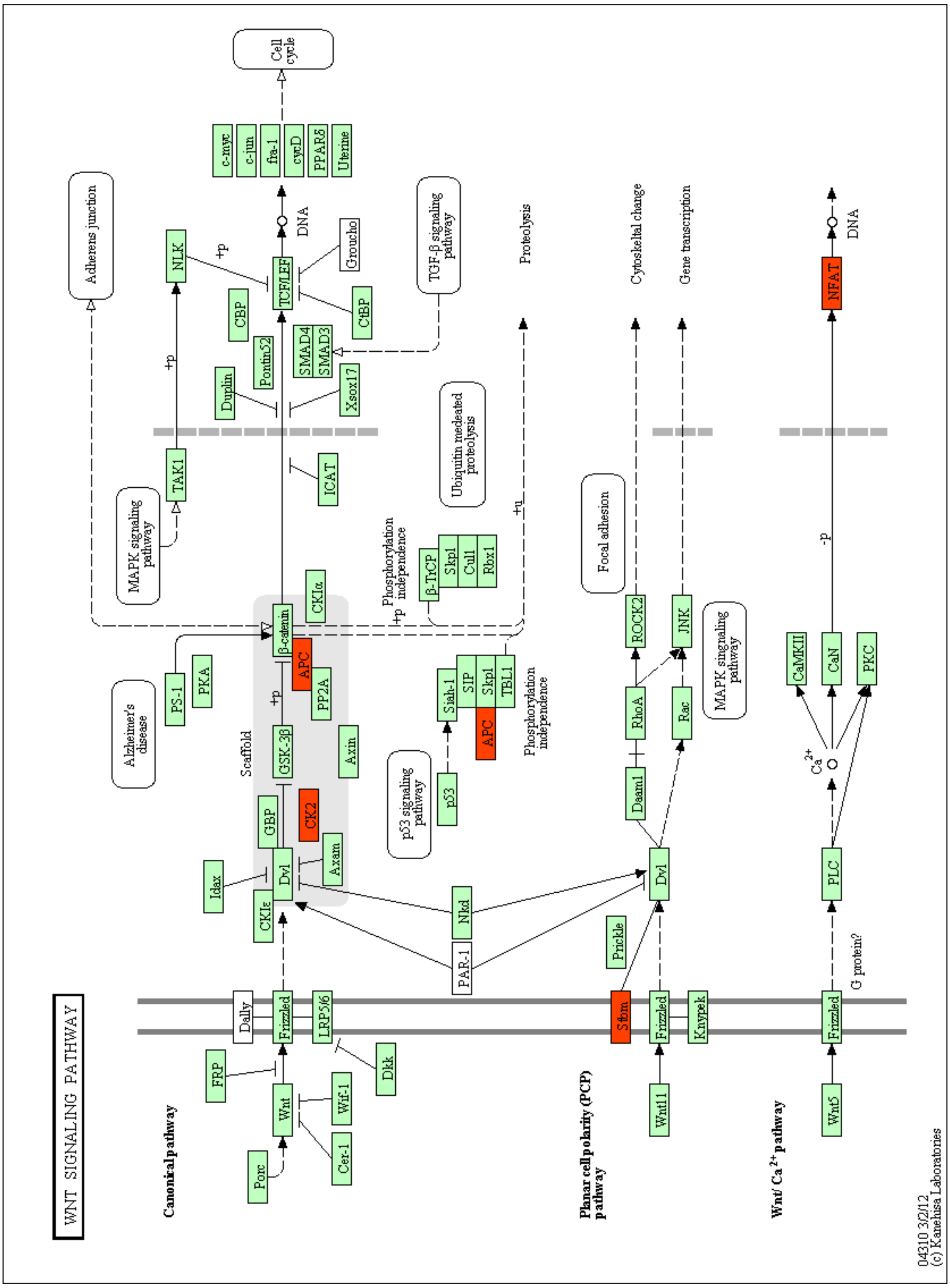


Fig. 7.6: Pathway map for the pathway "Wnt signaling" (hsa04310). The genes are colored by the number of mutations observed in the genes.

Tab. 7.10: Coverage analysis for genome position chr7:140,453,136 (corresponding to the common V600E mutation in BRAF) in the nine melanoma samples. Coverage of the respective position as well as the counts for the number of observed A to T mutations are listed.

Patient ID	Coverage	Count A→T
Pat01	2	0
Pat03	3	0
Pat04	5	0
Pat05	0	0
Pat06	0	0
Pat07	6	0
Pat08	9	5
Pat09	4	0
Pat10	3	0

7.4.3 Patient and tumor specific T-cell epitopes

We aim at finding candidates for tumor-specific epitopes for all nine patients. Experimental effort for the validation of such candidates is high. In a first step, the computationally detected mutations have to be validated by deep sequencing or Sanger sequencing. The next step is the synthesis of the respective peptides and the testing for immunogenicity. To restrict experimental effort to the most promising candidates we apply more stringent criteria than for the general analysis of the somatic mutation presented earlier in this section.

We exclude SNVs that are annotated in dbSNP. For the tumor SNVs we require a minimum coverage of 10 and that the tumor-specific base is at least observed in 3 reads. A minimum quality of the SNV is set to 20. The minimum coverage of 10 in the control tissue is required. Mutations in the control sample are expected to occur in no reads, in all reads (homozygous for wildtype or mutation) or in roughly 50% of the reads (heterozygous). For low coverage and RNA-seq we however expect the observed frequencies to differ from the expected values. If the coverage of a tumor SNV in the control lies below 20, we reject the SNV if the tumor-specific mutation is observed in at least one read. For a coverage over 20, we except up to two reads with the tumor-specific base. We also exclude SNVs in the genes PLEC and GLUD2, since these genes show suspiciously high numbers of SNVs in all patients (see Section 7.4.2). MHC binding prediction is performed for all peptides of length nine around the tumor-specific mutation using SYFPEITHI with halfmax-scores as binding thresholds.

A general observation form this analysis is that most candidate SNVs are rejected due to missing coverage in the control tissue. For SNVs with a coverage above ten in the control tissue we commonly observe the tumor-specific base in one or two reads.

These candidates are excluded as described above. Many of the remaining candidates get rejected because the frequency of the tumor-specific base is very low and occurs in less than three reads. Since tumors are heterogeneous and tumor samples also contain surrounding healthy tissue observing a mutation in only less than 20% is no reason to reject a SNV. A SNV that is however only observed in two reads is not very reliable and could result from sequencing error. We therefore reject SNVs where the tumor-specific base is observed in less than three reads.

In the following we present the results of the analysis for all nine patients individually. For all candidate epitopes the mutated position is written in bold. Only the tumor-specific peptides are reported.

Pat01 For Pat01 we find three SNVs that lead to predicted epitopes for the patient's HLA types. For one of these mutations the tumor-specific base is observed in 2 of 12 reads in the control sample and is therefore rejected. Two SNVs remain for HLA binding analysis.

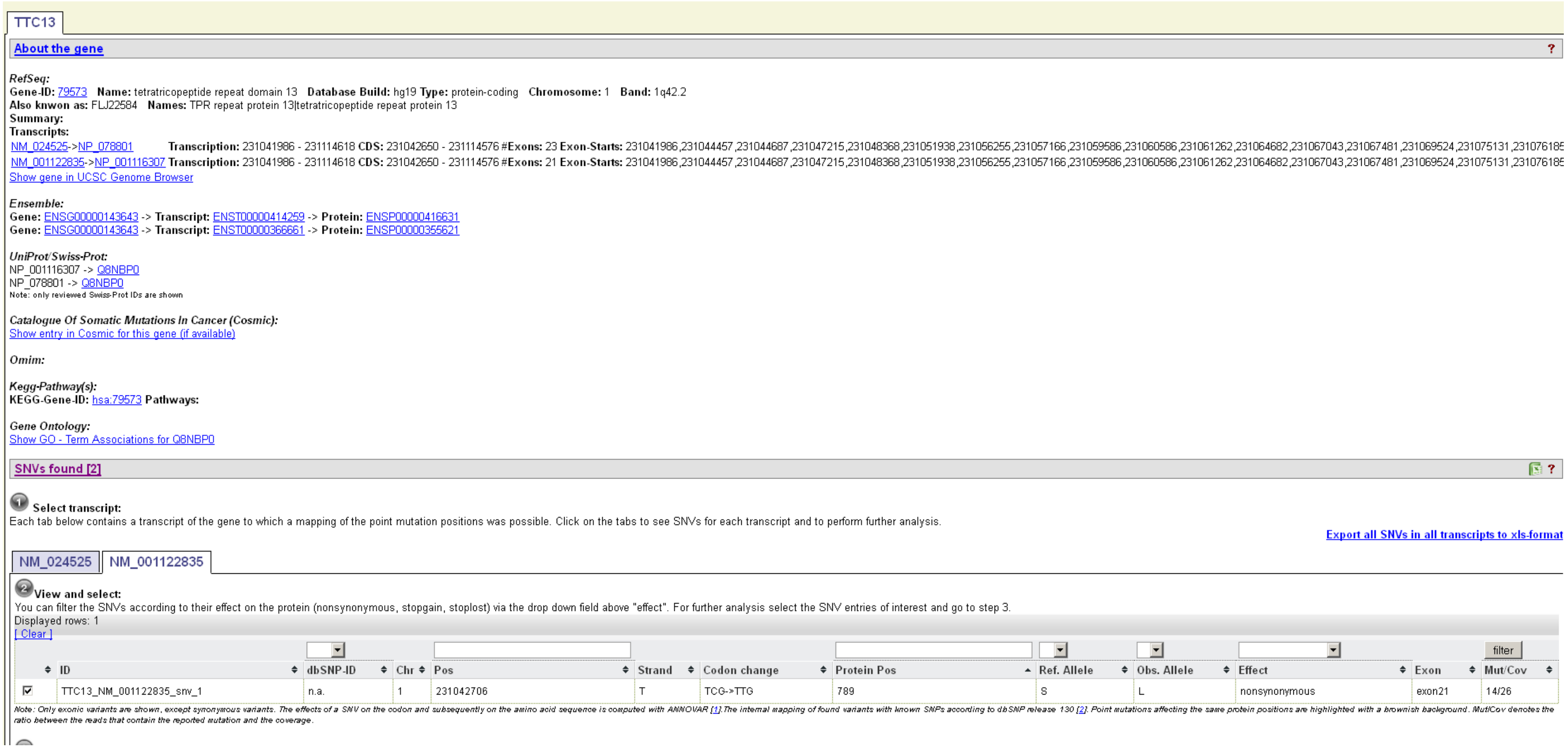
The first mutation is **E**70Q in the protein RPL28. The tumor-specific base is the wildtype, so the amino acid that is specific for the tumor is glutamic acid. This mutation leads to one peptide that is predicted to bind to B*08:01. The peptide is VIKRRSG**E**F and has a SYFPEITHI-score of 30.

The second mutation is **Q**191E in the protein TRAM1L1 with glutamine being the tumor-specific amino acid. This mutation leads to three predicted HLA binding peptides: FQKTKK**Q**DI is predicted to bind to B*08:01 with a SYFPEITHI-score of 27. KTKK**Q**DIPR is predicted to bind to A*68:01 with a SYFPEITHI score of 20. **Q**DIPRQLVY is predicted to bind to A*02:01 with a SYFPEITHI score of 20.

Pat03 Eight SNVs that lead to 14 predicted HLA binding peptide are found for patient Pat03. The mutations and the resulting MHC binding peptides are summarized in Tab. 7.11. Two of the patient's HLA alleles are covered by these peptides. One of the SNVs (L38**R** in gene UBE2L3) leads to peptides that are predicted to bind to different alleles. SNVs that are predicted to be presented by more than one HLA allele (in the sense that one of the peptides around the SNV are presented) are interesting candidates for peptide vaccines.

The mutation TTC13 is analyzed in VariartionDB. Fig. 7.7 and Fig. 7.8 show the general information on the gene and the mutation as well as the results for the HLA binding prediction for one of the peptides (TLR**L**MIEVL).

Pat04 No SNVs fulfill the quality criteria, except for one SNV in gene GLUD2. SNVs in GLUD2 were excluded from the analysis.



TTC13

About the gene

RefSeq:
Gene-ID: 79573 **Name:** tetratricopeptide repeat domain 13 **Database Build:** hg19 **Type:** protein-coding **Chromosome:** 1 **Band:** 1q42.2
Also knwon as: FLJ22584 **Names:** TPR repeat protein 13|tetratricopeptide repeat protein 13
Summary:
Transcripts:
NM_024525->NP_078801 **Transcription:** 231041986 - 231114618 **CDS:** 231042650 - 231114576 #**Exons:** 23 **Exon-Starts:** 231041986,231044457,231044687,231047215,231048368,231051938,231056255,231057166,231059586,231060586,231061262,231064682,231067043,231067481,231069524,231075131,231076185
NM_001122835->NP_001116307 **Transcription:** 231041986 - 231114618 **CDS:** 231042650 - 231114576 #**Exons:** 21 **Exon-Starts:** 231041986,231044457,231044687,231047215,231048368,231051938,231056255,231057166,231059586,231060586,231061262,231064682,231067043,231067481,231069524,231075131,231076185
Show gene in UCSC Genome Browser

Ensemble:
Gene: ENSG00000143643 -> **Transcript:** ENST00000414259 -> **Protein:** ENSP00000416631
Gene: ENSG00000143643 -> **Transcript:** ENST00000366661 -> **Protein:** ENSP00000355621

UniProt/Swiss-Prot:
NP_001116307 -> Q8NBP0
NP_078801 -> Q8NBP0
Note: only reviewed Swiss-Prot IDs are shown

Catalogue Of Somatic Mutations In Cancer (Cosmic):
Show entry in Cosmic for this gene (if available)

Omim:

Kegg-Pathway(s):
KEGG-Gene-ID: hsa:79573 **Pathways:**

Gene Ontology:
Show GO - Term Associations for Q8NBP0

SNVs found [2]

Select transcript:
Each tab below contains a transcript of the gene to which a mapping of the point mutation positions was possible. Click on the tabs to see SNVs for each transcript and to perform further analysis.

Export all SNVs in all transcripts to xls-format

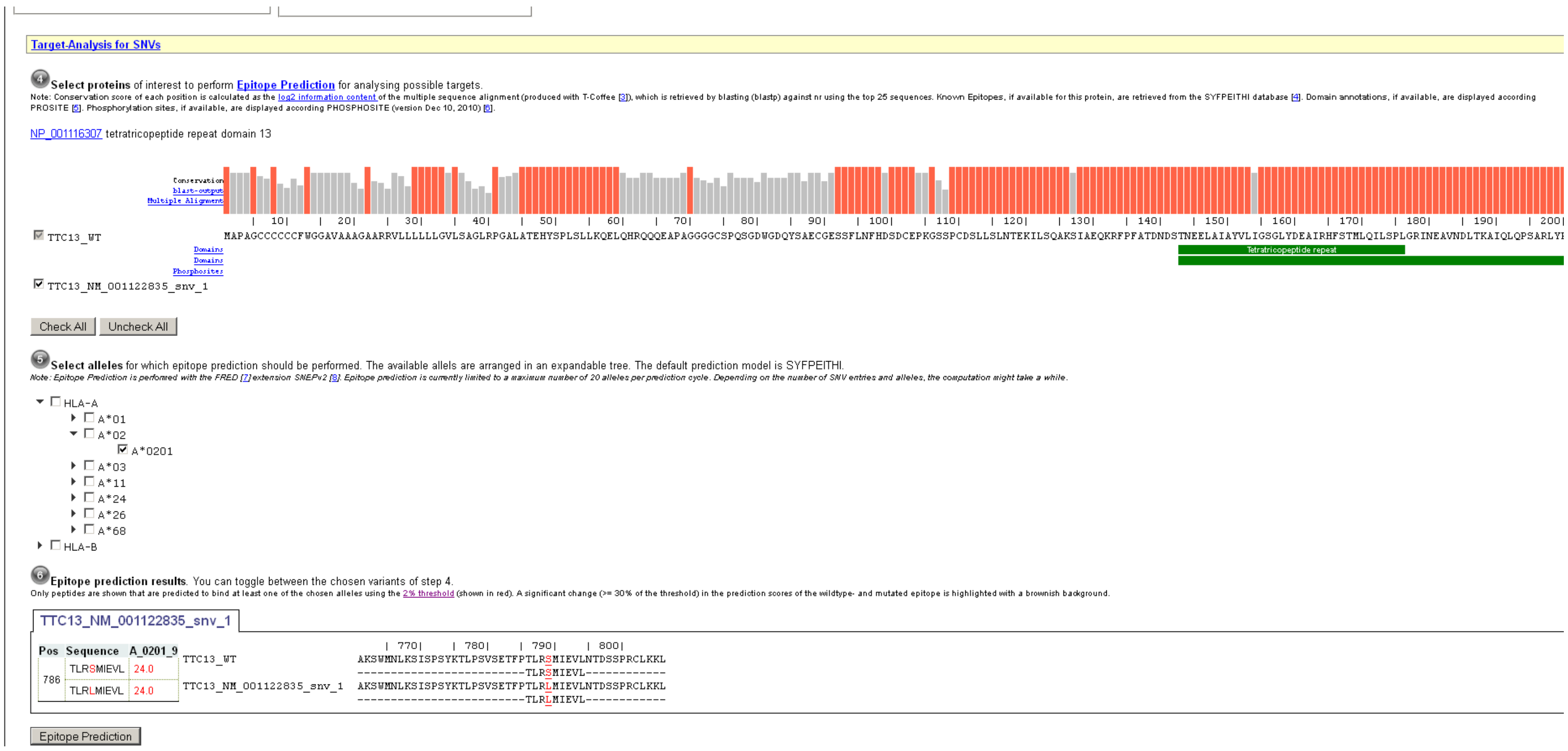
NM_024525 NM_001122835

View and select:
You can filter the SNVs according to their effect on the protein (nonsynonymous, stopgain, stoplost) via the drop down field above "effect". For further analysis select the SNV entries of interest and go to step 3.
Displayed rows: 1
[Clear]

	ID	dbSNP-ID	Chr	Pos	Strand	Codon change	Protein Pos	Ref. Allele	Obs. Allele	Effect	Exon	Mut/Cov
☑	TTC13_NM_001122835_snv_1	n.a.	1	231042706	T	TCG->TTG	789	S	L	nonsynonymous	exon21	14/26

Note: Only exonic variants are shown, except synonymous variants. The effects of a SNV on the codon and subsequently on the amino acid sequence is computed with ANNOVAR [1].The internal mapping of found variants with known SNPs according to dbSNP release 130 [2]. Point mutations affecting the same protein positions are highlighted with a brownish background. MutCov denotes the ratio between the reads that contain the reported mutation and the coverage.

Fig. 7.7: Analysis of mutation S789**L** in gene TTC13 in Pat03 with VariationDB. The general information on the gene and the information available for the mutation is displayed.



Target-Analysis for SNVs

4 **Select proteins** of interest to perform Epitope Prediction for analysing possible targets.
Note: Conservation score of each position is calculated as the log2 information content of the multiple sequence alignment (produced with T-Coffee [3]), which is retrieved by blasting (blastp) against nr using the top 25 sequences. Known Epitopes, if available for this protein, are retrieved from the SYFPEITHI database [4]. Domain annotations, if available, are displayed according PROSITE [5]. Phosphorylation sites, if available, are displayed according PHOSPHOSITE (version Dec 10, 2010) [6].

NP_001116307 tetratricopeptide repeat domain 13

```
                     |  10|    |  20|    |  30|    |  40|    |  50|    |  60|    |  70|    |  80|    |  90|    | 100|    | 110|    | 120|    | 130|    | 140|    | 150|    | 160|    | 170|    | 180|    | 190|    | 200|
☑ TTC13_WT           MAPAGCCCCCCFWGGAVAAAGAARRVLLLLLLGVLSAGLRPGALATEHYSPLSLLKQELQHRQQQEAPAGGGGCSPQSGDWGDQYSAECGESSFLNFHDSDCEPKGSSPCDSLLSLNTEKILSQAKSIAEQKRFPFATDNDSTNEELAIAYVLIGSGLYDEAIRHFSTMLQILSPLGRINEAVNDLTKAIQLQPSARLYI
```

☑ TTC13_NM_001122835_snv_1

Check All | Uncheck All

5 **Select alleles** for which epitope prediction should be performed. The available allels are arranged in an expandable tree. The default prediction model is SYFPEITHI.
Note: Epitope Prediction is performed with the FRED [7] extension SNEPv2 [8]. Epitope prediction is currently limited to a maximum number of 20 alleles per prediction cycle. Depending on the number of SNV entries and alleles, the computation might take a while.

- □ HLA-A
  - □ A*01
  - □ A*02
    - ☑ A*0201
  - □ A*03
  - □ A*11
  - □ A*24
  - □ A*26
  - □ A*68
- □ HLA-B

6 **Epitope prediction results**. You can toggle between the chosen variants of step 4.
Only peptides are shown that are predicted to bind at least one of the chosen alleles using the 2% threshold (shown in red). A significant change (>= 30% of the threshold) in the prediction scores of the wildtype- and mutated epitope is highlighted with a brownish background.

TTC13_NM_001122835_snv_1

Pos	Sequence	A_0201_9
786	TLRSMIEVL	24.0
	TLRLMIEVL	24.0

```
                                  | 770|     | 780|     | 790|     | 800|
TTC13_WT                  AKSWMNLKSISPSYKTLPSVSETFPTLRSMIEVLNTDSSPRCLKKL
                          ------------------------TLRSMIEVL------------
TTC13_NM_001122835_snv_1  AKSWMNLKSISPSYKTLPSVSETFPTLRLMIEVLNTDSSPRCLKKL
                          ------------------------TLRLMIEVL------------
```

Epitope Prediction

Fig. 7.8: Analysis of mutation S789**L** in gene TTC13 in Pat03 with VariationDB. The prediction HLA binding prediction results for one of the peptides is displayed.

Tab. 7.11: Somatic SNVs in Pat03 that lead to predicted HLA binding peptides. The name of the gene/protein is given together with the mutation relative to the protein sequence. The tumor-specific amino acid is written in bold. For each peptide the HLA alleles that are predicted to bind that peptide and the SYFPEITHI scores are listed.

Gene	Mutation	Peptides	HLA-Allele (score)
ARID1A	Q1334**P**	QQQ**P**RHDSY	B*15:01 (22)
TTC13	S789**L**	TLR**L**MIEVL	A*02:01 (24)
		R**L**MIEVLNT	A*02:01 (20)
		LMIEVLNTD	A*02:01 (21)
ZMIZ1	M200**V**	PMNPGGNP**V**	A*02:01 (21)
PDCD4	A293**G**	**G**ALDKATVL	A*02:01 (19)
NUP160	E146**K**	**K**TQNRVIIL	A*02:01 (19)
NFATC3	Q753**E**	SLICSIP**E**T	A*02:01 (25)
		LICSIP**E**TY	B*15:01 (19)
		SIP**E**TYASM	A*02:01 (20)
PRMT7	I284**V**	VLSWWD**V**EM	A*02:01 (19)
UBE2L3	L38**R**	NLLTWQG**R**I	A*02:01 (20)
		LLTWQG**R**IV	A*02:01 (21)
		RIVPDNPPY	B*15:01 (20)

Pat05 Five SNVs in four genes lead to predicted HLA binding peptides in Pat05 and are summarized in Tab. 7.12.

Two SNVs affect consecutive bases in the gene RNH1 (ribonuclease/angiogenin inhibitor 1). The affected positions are 500508 and 500509 on chromosome 11. Both SNVs affect the same codon and amino acid (P83). SNV calls that affect neighboring positions are less reliable than single SNV calls and can be caused by alignment errors. In this case, we observe a G to A mutation for both consecutive positions. The number of reads for which the mutation is observed is four and three, respectively and thus the data supporting the SNVs is rather slim. SNVs that affect the same protein position additionally bare the problem that from SNV calls alone we cannot decide if the SNVs occur together, i.e. that we have reads that contain both mutations and other that contain none. The read alignments shown that in this case the mutations occur separately as shown in Fig. 7.9. We therefore can assume that, in case that the SNV calls are true positive SNVs, the SNVs have to be introduced separately into the respective transcripts to generate the peptide sequences.

RHN1 is encoded on the negative strand. The observed changes on codon level are CCC to CTC and CCC to TCC and lead to the mutations P83L and P83S in the respective protein. Both mutations lead to predicted HLA binding peptides.

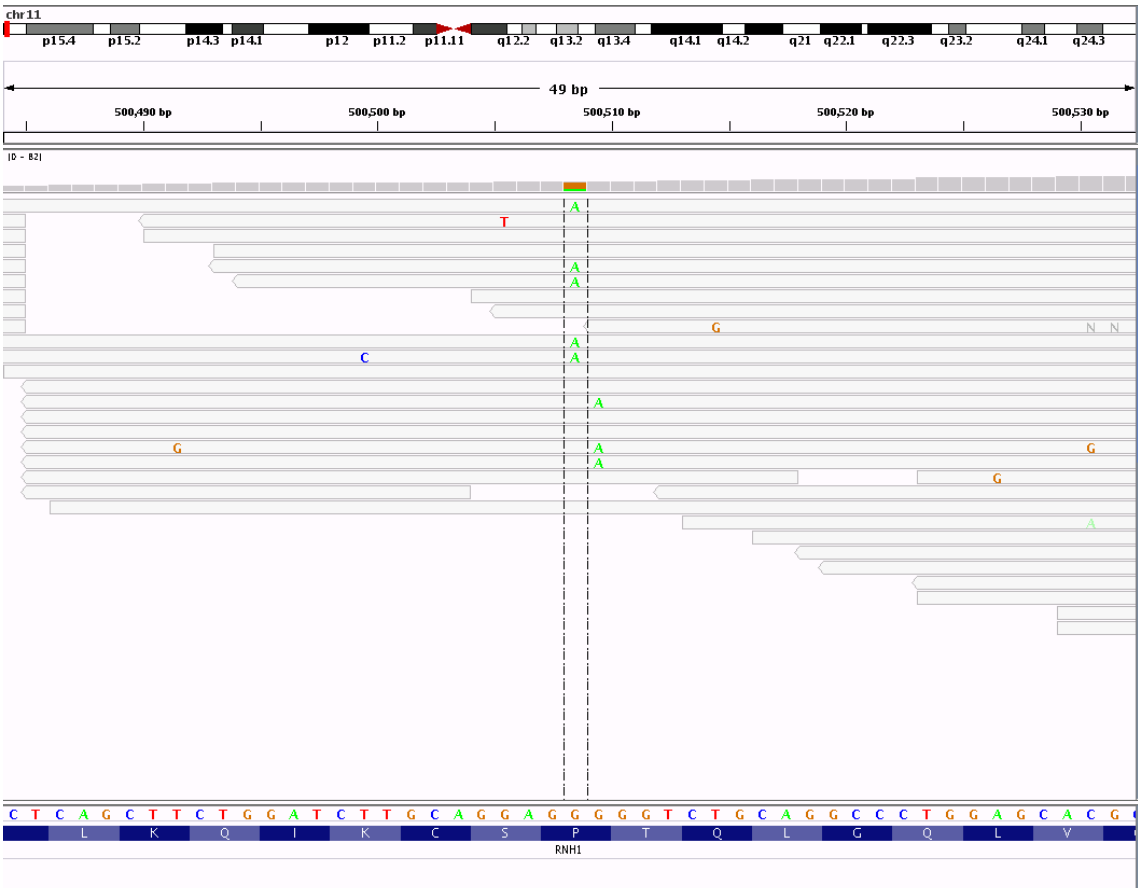


Fig. 7.9: Reads of Pat05 aligned to gene RNH1. Two consecutive SNVs are observed. The mutations do not occur in the same reads. The two mutations are therefore introduced separately into the respective transcripts to generate the tumor-specific peptide sequences.

Tab. 7.12: Somatic SNVs in Pat05 that lead to predicted HLA-binding peptides. The name of the gene/protein is given together with the mutation relative to the protein sequence. The tumor-specific amino acid is written in bold. For each peptide the HLA alleles that are predicted to bind that peptide and the SYFPEITHI scores are listed.

Gene	Mutation	Peptides	HLA-Allele (score)
KDM5A	P1592**L**	KVLDI**L**SKY	B*15:01 (21)
		REKKVLDI**L**	B*08:01 (20)
NEIL1	**K**242R	VVQLGG**K**GY	B*15:01 (20)
KDM5A	P1592**L**	LVKRR**V**RWY	B*15:01 (19) B*08:01 (20)
RNH1	P83**L**	CVLQGLQT**L**	A*02:01 (24)
		GLQT**L**SCKI	A*02:01 (22)
		T**L**SCKIQKL	A*02:01 (27) B*08:01 (27)
	P83**S**	GLQT**S**SCKI	A*02:01 (21)
		T**S**SCKIQKL	B*08:01 (21)

Pat06 No SNV for Pat06 fulfills the criteria.

Pat07 We observe two SNVs that lead to predicted HLA binding peptides in gene TRAM1L1, **F**177L and **Q**191E. For both SNV positions a homozygous mutation is observed in the control tissue while the tumor is heterozygous. The wildtype is the tumor specific base. **F**177L leads to the tumor-specific peptide QLAYW**F**HAF that is predicted to bind to A*03:01 (23) and B*15:01 (19). **Q**191E leads to the tumor-specific peptide **Q**DIPRQLVY that is predicted to bin to A*03:01. The tumor-specific amino acid is written in bold and the SYFPEITHI scores for the peptides are given in brackets.

Pat08 One SNV for Pat08 that fulfills all criteria and leads to a predicted HLA-binding peptide is observed in gene PABPC3. The mutation **A**313E, where the reference amino acid is specific for the tumor, leads to one predicted HLA-binding peptide: RLRK**A**FSPF is predicted to bind to B*15:01 with a SYFPEITHI score of 20.0.

Pat09 None of the SNVs for Pat09 fulfills all criteria.

Pat10 Only one SNV for Pat10 fulfills the criteria. This SNV is observed in gene HLA-DRB1, a gene that encodes for an HLA class II molecule. The HLA loci are known to be highly polymorphic, we therefore assume that the observed mutations can be attributed to the highly polymorphic character of the gene rather than to a somatic tumor mutation. This is supported by the fact that the mutation is also observed in the control tissue but below the detection threshold for polymorphisms in normal tissue (3 of 62 reads).

7.4.4 Discussion

RNA-seq is not the technology of choice for mutation analysis. Its advantage is that the found mutations are known to be expressed. Its dependency on transcript abundance, however, renders comparison between samples difficult. For regions that are not covered in a sample, we cannot decide whether a mutation is present or not. This applies to the identification of somatic mutations as well as for the analysis of recurrent mutations. For other types of analyses like the detection of gene fusions or the identification of alternative transcript isoforms RNA-seq however is the technology of choice. Gene fusions have been shown to be related to certain cancers (e.g., the Philadelphia chromosome in CML) and can lead to the identification of new targets and tumor specific T-cell epitopes. For the detection of somatic small mutations like SNVs and small INDELS Exome-seq combined with expression analysis using gene expression microarrays or RNA-seq would be more suitable. If mutation analysis is to be performed on RNA-seq data, the choice of the corresponding control samples is critical. DNA or Exome-seq of the control samples should be used, or at least RAN-seq should be performed on healthy cells from the same tissue type to increase the chance of similar expression. In the case of melanoma, the respective healthy cells, the melanocytes, are not available for analysis. It is not possible to isolate melanocytes from skin samples. Using Exome-seq on the PBMC samples as healthy samples would have been more appropriate.

The computational methods for detection of somatic mutations need to be improved. The biggest issue, especially for positions with low coverage, is the required frequency threshold for a mutation to be called. In the tumor samples, the observed frequency of a potential tumor mutation depends on the ratio of tumor to normal tissue in the sample as well as on the heterogeneity of the tumor. The amount of tumor tissue in a sample can be estimated by pathologists. However, a microscopic analysis of a tissue sample cannot reliably determine the percentage of tumor cells. The genetic heterogeneity of a tumor is unknown before sequencing since sequencing experiments are usually performed to assess the genetic landscape of a tumor. The fact that these important factors are unknown makes the statistical detection of mutations in the tumor a hard problem. Using RNA-seq data for SNV detection further complicates the problem due to the large variations in the coverage and its dependency of the transcript abundance.

The detection threshold for SNVs used for the control samples has a strong influence on the detection of somatic mutations. In healthy tissues we expect either homo- or heterozygous genotypes with an observation frequency for bases of roughly 100% for homozygous genotypes or of roughly 50% for heterozygous genotypes. It is unclear if this assumption also holds for RNA-seq data, since the sequencing depth depends on the transcript expression. To increase the chance of predicting true somatic mutations we use a very low detection threshold for heterozygous genotypes also in the control samples. This strategy however bears the risk of rejecting too many candidates, especially when the coverage in the control sample is low. Using Exome-Seq at least for the control samples would partly resolve this issue.

We were not able to identify reliable candidates for recurrent mutations. The 15 candidates that we analyzed manually in detail revealed that most of the candidates are false-positive somatic mutations, meaning that the tumor allele is also present in the control but below the detection threshold or that the coverage in the control is too low to reliably define the tumor variant as not present. Interestingly, for all of these candidate mutations the base detected as tumor variant is the wildtype. The analysis of recurrent mutations on RNA-seq data is problematic, because for the patients where the mutation is not observed, we cannot decide if this is because the mutation is not present or due to low or no coverage.

We did not identify genes or pathways with somatic mutations in a significant number of the patients. A pathway analysis could also be performed on the total set of observed mutations instead of somatic mutations. This kind of analysis could reveal SNPs that predispose for cancer development. More complete data would however be needed for such analyses in the sense that for all mutated positions included in the analysis we know if a mutations is present or absent. On our mutation data computed from RNA-seq, this information is however not available.

Studies like the one for melanoma samples presented here can reveal the problems that need to be addressed to make the detection of somatic mutations more reliable. Experimental validation of the proposed mutations by Sanger sequencing or deep sequencing can provide useful data for the improvement of variation detection.

Besides these drawback we were able to identify candidates for somatic mutations that lead to tumor-specific T-cell epitopes with respect to the patients alleles for five out of nine patients. The lack of predicted tumor-specific T-cell epitopes in four patients is due to the lack of appropriate SNVs. A higher number of reliable somatic SNVs increases the chance of finding patient- and tumor-specific T-cell epitopes. We used the prediction method SYFPEITHI in this analysis. As discussed in Section 5.2, the availability of appropriate allele-specific models for HLA binding prediction is an issue. Not for all of the patients HLA types a prediction model was included (see Tab. 7.1). Using the full set of the patients's HLA alleles for prediction by using pan-specific approaches (see Section 5.2) can increase the number of predicted epitopes.

We performed both, an analysis of the single mutations in single patients, and a comparative analysis for recurrent mutations. The size of our cohort (9) is very small. The analysis for recurrent mutations did not reveal any candidates that could be proposed as recurrent target for new therapies. The individual analysis however showed that we can still identify somatic mutations as targets for personalized immunotherapy. This demonstrates the potential of approaches that use truly individualized targets. However, the overall applicability and efficacy of an immunotherapy based on tumor-specific peptide vaccines still needs to be validated by clinical experiments.

We used VariationDB for the visual and interactive inspection of the results. This revealed some drawbacks and suggestions for improvements of the system. One useful feature would be a group view without the requirement for a second group. In addition, VariationDB does not directly account for tumor/control sample pairs. It is therefore not possible to directly use VariationDB for the identification of somatic mutations. A third issue is related to the fact that in some cases the tumor-specific base is the base present in the reference genome at this position. VariationDB interprets all mutations with respect to a reference genome. Cases where the wildtype is the one of interest will not be readily identifiable in VariationDB. An extension that accounts for these cases would facilitate the analysis of tumor-specific mutations.

The latter is a general problem in the analysis of somatic mutations. In the case that the tumor is homozygous for the wildtype but the control tissue has a mutation at a specific position, we would not identify this mutation as a tumor mutation. Such mutations would not be of interest for the identification of tumor-specific peptides, however a homozygous loss-of-function mutation in a tumor suppressor gene would be interesting for understanding the development of the tumor. However we do not expect these cases to occur often, because usually a tumor sample also contains surrounding normal tissue. A mutation with respect to the reference genome in the healthy tissue would then be detected in the tumor. This would lead to the detection of a somatic mutation with the wildtype as tumor variant (as observed in the analysis of recurrent mutations). Calling SNVs on tumor and control samples simultaneously could solve this problem.

The example of VariationDB shows that it is important to keep analysis systems flexible and extensible as long as the analyses are not fully standardized or state-of-the-art. Not all special cases and problems can be foreseen and handled in advance. The computational systems need to be able to evolve with the growing knowledge and experience in the identification and interpretations of genetic variations.

Chapter 8

Prediction of graft-versus-leukemia reaction after stem cell transplantation

SNP disparity between donors and recipients give rise to non-self antigens for donor and patient in HLA-matched allogenic stem cell transplantation (alloSCT). These antigens are called minor histocompatibility antigens (miHAs). Cytotoxic T cells specific to hematopoiesis-restricted SNPs eradicate circulating leukemic cells and leukemic progenitor cells *in vitro* and *in vivo*. Such T cells were isolated in the course of remission of donor lymphocyte infusion-treated patients after HLA-matched SCT. This beneficial side-effect of alloSCT is termed graft-versus-leukemia effect (GvL).

The use of donor-derived T cells specific for a patient's hematopoiesis-restricted SNPs is a new treatment modality for relapse in hematologic malignancies after alloSCT. The availability of this treatment is however restricted by the number and frequency of known hematopoiesis-restricted miHAs and by the frequency of the HLA alleles to which they are restricted. The goal of this project is to make miHA-based therapy available for more patients. In a collaboration project with the Department of Pediatric Hematology/Oncology at the Children's Hospital at the University of Tübingen, we developed a strategy for the large-scale identification, selection and validation of HLA class I-presented peptides that are derived from proteins that are physiologically expressed by hematopoietic tissue only and harbor a SNP variant.

The project is outlined in Fig. 8.1. The computational analysis is based on the pipeline presented in Section 6.2. In the first part of the project we aim at identifying a set of relevant SNPs and at designing a genotyping assay for SNPs that are promising to lead

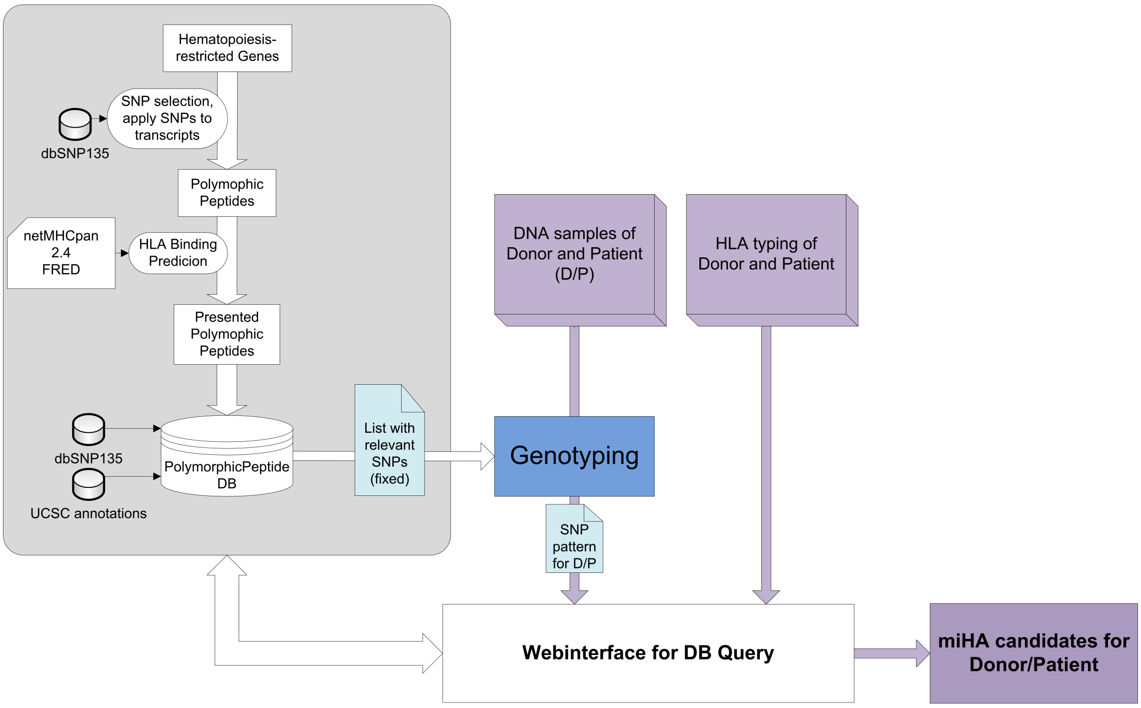

Hematopoiesis-restricted Genes
SNP selection, apply SNPs to transcripts
dbSNP135
Polymophic Peptides
netMHCpan 2.4 FRED
HLA Binding Predicion
Presented Polymophic Peptides
dbSNP135
UCSC annotations
PolymorphicPeptide DB
List with relevant SNPs (fixed)
DNA samples of Donor and Patient (D/P)
HLA typing of Donor and Patient
Genotyping
SNP pattern for D/P
Webinterface for DB Query
miHA candidates for Donor/Patient

Fig. 8.1: Large-scale identification of hematopoiesis-restricted miHAs. The first part (gray box) aims at the identification of relevant SNPs for genotyping. The second part proposes candidate miHAs based on genotyping information for donor-patients pairs.

to miHAs in a large number of patients. Using this approach we hope to be able to specifically, quickly, and cheaply identify hematopoiesis-restricted miHAs in a large group of donor-patient pairs, including patients with rare HLA alleles. In the following section we will describe the two parts in more detail, along with preliminary results.

8.1 Identification of relevant SNPs

The aim of this step is the identification of hematopoiesis-restricted SNPs that have the potential to lead to miHAs in a large number of patients. This step is based on the general large-scale screening setting described in Section 6.2. The most promising SNPs are selected and included in a customized genotyping assay. In order to keep the experimental effort for genotpying minimal we try to keep the number of selected SNPs low while at the same time maximizing the chance of detecting miHA candidates also in patients withe infrequent HLA types.

Generation of candidate polymorphic peptides and HLA binding prediction

As input we use a set of 67 genes that are specifically expressed in hematopoiesis. The selection was performed by our partners in the clinic based on gene expression data from BioGPS [177] and the Human Protein Atlas [180]. We retrieve all non-synonymous SNPs from dbSNP (release 135) for this set of genes.

In total, we find 2,161 non-synonymous SNPs for the 67 genes. The respective transcripts for all genes are retrieved from RefSeq. We map all SNPs to all of the corresponding transcripts. Peptides of length nine are generated around the SNP positions. If more than one SNP is included for one transcript, we consider all combinations. Duplicate peptides that stem from different transcripts for the same gene are removed. We generate 66,496 peptides from the 2,161 SNPs.

We use netMHCpan (version 2.4) [147] to predict HLA binding. The aim of this study is to include the largest possible number of HLA alleles, also HLA alleles with low frequencies. The identification of novel miHAs for rare HLA alleles is of major interest since those alleles are underrepresented in the set of currently known miHAs. We therefore use all HLA alleles that can be be predicted using netMHCpan. Models are available for 2,915 HLA-A, HLA-B and HLA-C alleles. We use the strong binder threshold (predicted $IC_{50} \leq 500$ nM).

1,928 of the SNPs lead to at least one peptide that is predicted to bind to at least one of the HLA alleles. For 402 of these genes no allele frequency data is available. 32 of the SNPs have a minimal minor allele frequency of 30%.

The total number of predicted HLA-binding peptides is 14,205. The number of predicted HLA-binding peptides per SNP ranges from 1 to 89. To obtain these numbers we count the number of different peptides that are generated by a SNP and are predicted binders. Each peptide is only counted once, irrespective of the number of HLA alleles the peptide is predicted to bind to. For a single SNP with two known alleles and therefore two different amino acids in the corresponding protein sequence the number of peptides that can be generated around the SNP position is 18 (9 peptides for the one allele or amino acid, 9 peptides for the other allele or amino acid). Not all of these peptides are expected to bind to an HLA allele. Nevertheless, the number of HLA-binding peptides that we observe for some SNPs exceeds 18. These cases can be explained by two facts. First, for some SNPs three alleles are known. If the codons corresponding to the three known variants all code for different amino acids the number of peptides is increased to 27. Second, if two SNPs are less than nine amino acids apart in the protein sequence we use all possible combinations of the two SNPs to generate the peptides. This explains the high number of predicted HLA-binding peptides for some SNPs.

An interesting criterion for including a SNP in the genotyping assay is the number of HLA alleles that are covered by a SNP. We consider an HLA to be covered by a SNP if at least one of the peptides around that SNP is predicted to bind to this HLA allele. We analyze the number of covered HLA alleles on the four-digit HLA-allele names (full resolution of HLA typing) and on a two-digit resolution. For the latter we cut the allele names after two digits, for example A*02:01 is reduced to A*02. We then count the unique reduced names for each SNP. This measure is used to better assess the number of different HLA alleles that are covered by a SNP. For some allele groups, a large number of single alleles is known, for example there are 247 alleles known for HLA-A*02. The alleles belonging to that group do not have the same peptide specificity, but the specificities are expected to be rather similar. Counting alleles only at the full resolution tends to overestimate large groups of similar alleles. We will therefore consider the number of single different alleles together with the number of different alleles on two-digit resolution (in brackets). The number of alleles covered by a SNP ranges from 718 (30) and 1 (1). SNPs that cover a large number of HLA alleles are interesting candidates to be included into the genotpying assay since they have the potential to be a miHA in patients with different HLA types.

Our analysis reveals the presence of promiscuous HLA binding peptides, i.e. peptides that are predicted to bind to more than one HLA allele. This is not surprising if we look at single alleles, since groups of similar alleles have similar binding specificities. We therefore focus on the number of different alleles on the 2-digit level. The largest number of different 2-digit alleles that is covered by one SNP is 33. The 33 alleles are covered by only 10 peptides. We define the promiscuity of a set of peptides as the number of covered alleles divided by the number of peptides that cover these alleles. The promiscuity for the 10 peptides covering 33 alleles is 3.3. We observe the highest promiscuity for the SNP rs150142878 in gene FAM65B. Two peptides cover 23 alleles from the A-,B-, and C-loci of HLA (A*02, A*23, A*24 , A*31, A*32, B*07 , B*15, B*27 , B*35, B*37, B*38, B*39 , B*40, B*41, B*42, B*48, C*02, C*03, C*06, C*08, C*12, C*14, C*16). These SNPs and the corresponding peptides are of great interest because the same peptides are candidate miHAs in a large number of patients.

Selection of SNPs for genotyping

The genotyping is performed by an external company (ATLAS Biolabs GmbH, Berlin) using the iPlex Gold Genotyping Assay from Sequenom (http://www.sequenom.com). It applies robust single-base primer extension to discriminate between the two alleles of a polymorphic DNA site. The assay is designed based on customer's SNP list. The assay and primer design are the most cost- and time-consuming step in the experimental part, the set of SNPs to be tested is therefore kept constant during the whole study. The selection of the final SNP list is based on the criteria presented above (allele coverage number of predicted HLA binding peptides, allele frequencies). Since the current setup is supposed to be used in a pilot study we make sure to cover the alleles of those individuals (donors and patients) that are already known to be included in the study. The selection has not yet been not completed, so the list cannot be presented here.

8.2 Detection of miHA candidates for donor-patient pairs

Patient and donor pairs are genotyped for the set of SNPs included in the SNP genotyping assay. For both, patient and donor, information on the zygosity for all SNP positions is generated. Once genotyping information is available, we can use the information to extract those miHA candidates that are relevant in the context of a donor-patient pair. The basic procedure is to filter for peptides that are uniquely present in the patient but not in the donor. We then select those peptides that are predicted to bind to one of the HLA alleles shared between donor and patient.

The input for this step is the genotpying information for patient and donor in the file format that we obtain from the company that performs the genotyping and a set of up to six different HLA class I alleles. Prediction data for all SNPs that are included in the genotyping assay are stored in a database. The database schema is depicted in Fig. 8.2. In this database the SNPs are linked to the respective genes and transcripts. Each SNP is also linked to all peptides of length nine that contain the SNP position. The peptides are linked to the results of the HLA binding prediction for all HLA alleles that can be predicted using netMHCpan.

In order to allow our collaboration partners to directly access the results we implemented a web interface for this step of the project. After login, the user can supply two files containing the genotyping information for donor and patient. The start page of the web-interface and the input form is shown in Fig. 8.3. In the next step all miHAs that are relevant in the context of the patient's and donor's genotypes and the shared HLA alleles are selected. The results are displayed in a way that allows a convenient selection of miHAs for this patient-donor pair.

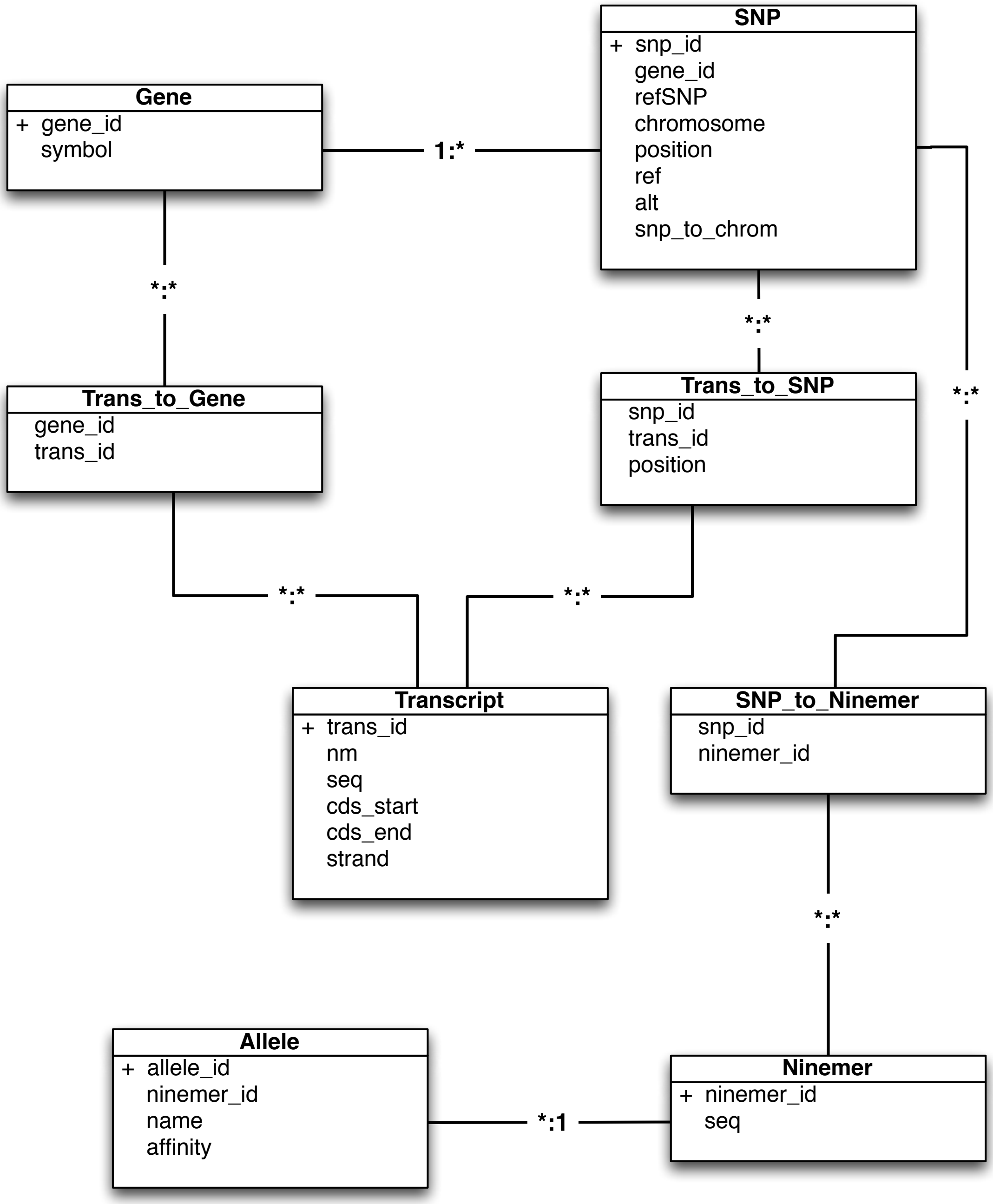
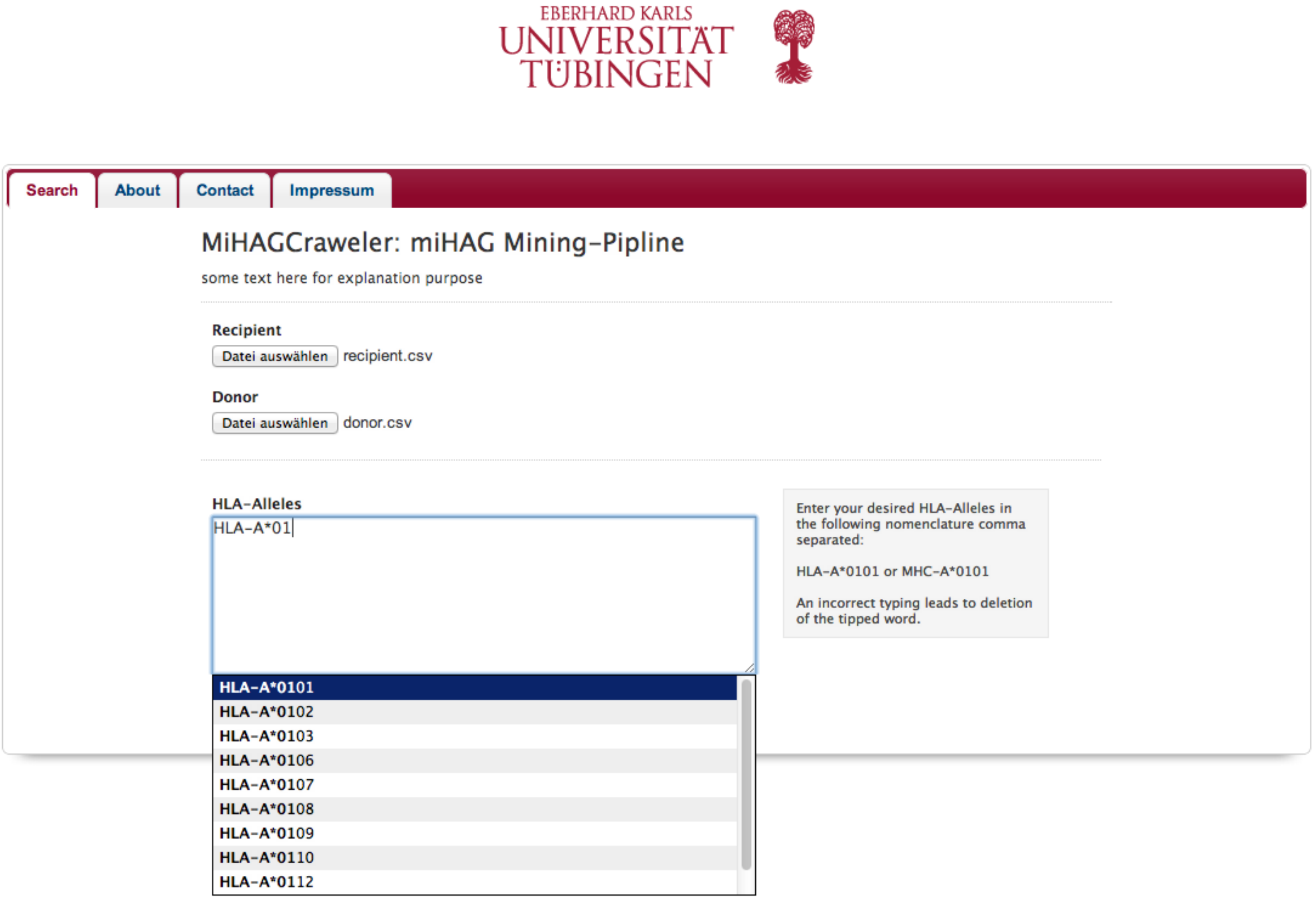


Fig. 8.2: Schema of the database containing all miHA candidates for the SNPs that are included in the genotyping assay. SNPs are linked to the respective genes and transcripts as well as to all related peptides of length nine. Results for HLA binding prediction for all netMHCpan alleles are stored for each peptide.



MiHAGCraweler: miHAG Mining–Pipline

some text here for explanation purpose

Recipient

Datei auswählen recipient.csv

Donor

Datei auswählen donor.csv

HLA–Alleles

HLA–A*01

Enter your desired HLA–Alleles in the following nomenclature comma separated:

HLA–A*0101 or MHC–A*0101

An incorrect typing leads to deletion of the tipped word.

HLA–A*0101
HLA–A*0102
HLA–A*0103
HLA–A*0106
HLA–A*0107
HLA–A*0108
HLA–A*0109
HLA–A*0110
HLA–A*0112

Fig. 8.3: Starting page of the webinterface for our automatic miHA detection pipeline. Inputs are two files containing the genotyping information for donor and patient in a predefined format and up to six different HLA alleles.

8.3 Discussion

We presented a computational pipeline for the identification of candidate miHAs for patients irrespective of the frequencies of the HLA types they expressed. The pipeline was developed in close collaboration with partners from Department of Hematology/Oncology at the University Children's Hospital, Tübingen. The pipeline is used in a pilot study that aims to make therapy based on hematopoietic miHAs available for 100% of the patient population. This new treatment modality of relapse in hematologic malignancies after alloSCT will significantly contribute to a better outcome of alloSCT in refractory and high-risk hematologic malignancies. Large-scale computational analyses are indispensable to identify a set of promising candidates for experimental testing. We propose a two-step procedure: 1) identification of candidate hematopoietic SNPs for genotyping, and 2) Identification of miHAs that are relevant for a patient-donor pair after genotyping of promising SNPs. We thereby minimize the number of SNPs that need to be genotyped but keep the number of HLA alleles that can be covered high.

The number of candidate SNPs after the general prediction step is promising with respect to HLA allele coverage. However, the clinical application also requires a suitable genotype combination for donor and patient for these SNPs. The selection of the SNPs for the genotyping assay was still in progress by the time this thesis was written, so the number of SNPs in the assay or first results of genotpying for patient-donor pairs cannot be presented. A clinical validation of the proposed miHAs also needs to be performed.

Despite the lack of clinical validation our results show that we can identify a significant number of potential miHAs in hematopoiesis-restricted genes with this two-step procedure.

Complete sequencing of the genomes or transcriptomes for all patient-donor pairs has become a realistic goal for the near future. However, the effort concerning time and money is currently to high to be broadly applied in alloSCT. For the identification of hematopoiesis-restricted miHAs, however, information for complete genomes and transcriptomes is not necessary since only genes that are specifically expressed in the hematopoietic system are of interest. In addition, we need high-confidence genotyping information, and a customized genotyping assay for SNPs is still more reliable than variation detection form sequencing data. Our two-step methods is a reasonable approach, since it allows to obtain high-quality data for all the regions of interest for a minimal price. If sequencing of whole genomes, transcriptomes, or exomes for patients and donors becomes a standard procedure in the clinic, the pipeline can be adapted and the first step can be skipped.

Some steps of the pipeline can be refined. Manual selection of the genes can be replaced by an automated procedure that identifies tissue-specific genes from publicly available gene expression data.

We now use a simplistic criterion to select SNPs that are likely to be disparate between two individuals. We base that selection on the minor alleles frequency (i.e. the frequency of the least frequent allele) reported for the SNPs in dbSNP. A statistical model that maximizes the chance of drawing a disparate and relevant SNP genotype for two individuals would be better. However, many factors that contribute to such a model are unknown. Haplotype frequencies have to be included instead of single allele frequencies. Donors in alloSCT are often related to the patients (in many cases siblings or parents of the patients are used as donor) and we do not know how the blood relationship influences the chance of observing disparate SNPs. We therefore decided to use simple minor allele frequencies in this pilot study. If the approach proves to be generally successful, the model for allele frequencies can be refined in a follow-up project based on the genotyping results from the pilot study.

We now use all HLA alleles for which a prediction is possible. This results in a very large number of HLA alleles (2,915 for netMHCpan 2.4). Prior knowledge on the HLA frequencies in the patient cohort can be used to reduce this number to those alleles that can be of interest in the respective cohort.

The two-digit representation is a very simplistic measure for the number of different alleles that are covered by a SNP. A measure based on the similarity of the peptide binding repertories of HLA alleles would be better, however, the similarity between the HLA alleles is hard to assess. For many of the alleles that are included in the pan-specific method netMHC no experimental binding data is available. The similarity can only be assessed on the prediction method, for example by computing the overlap between the predicted binding peptides for a large set of human proteins. This measure would inherently represent the similarity between prediction models rather than the similarity between the alleles. However, if the prediction method has a good performance on the alleles with no experimental binding data, we expect the prediction method to correctly represent the similarities between the alleles. Using groups of alleles obtained from clustering on the overlap of predicted peptide binding repertoires could thus be used to estimate the number of different alleles that is covered by a SNP.

The current implementation of the second part of our pipeline, where we include personalized genotyping information for donors and patients, is based on a database with prediction results for the selected SNPs. This is a suitable approach for the pilot study where the set of genes and SNPs is fixed. In order to be more flexible with respect to the set of genes and SNPs that need to be analyzed, the interface to the database could easily be replaced by an interface to a prediction step if necessary. The predictions can then be performed on the fly rather than be retrieved from a database.

Summarizing, this project shows that a close interaction between the clinics and experts for computational analyses in immunoinformatics offers new personalized treatment possibilities.

Part IV

Conclusion and perspectives

Chapter 9

Conclusion and perspectives

In this thesis we present computational methods for an integrated analysis of NGS data that allows the identification of tumor-specific mutations. A major problem is the detection of mutations in NGS data derived from tumor samples. The current approaches to detect short genetic variations from sequencing data aim at the identification of the most likely genotype in the sample. Sequencing data usually contains reads that are derived from different cells as single cell sequencing techniques are not standard. A heterogeneous mixture of cells as in cancer samples complicates the identification of mutations. Statistical models that account for the heterogeneity are not available today. Sequencing is used to determine the genetic profile of a tumor, so prior knowledge on the heterogeneity of a sequenced sample before data analysis to base a statistical model on is not at hand. A first step into the right direction is that tumor samples are histologically examined by a pathologist in order to estimate the percentage of tumor cells in a sample. However, histological examination cannot assess the genetic heterogeneity of a tumor sample. Several cancer sequencing projects are going on to date and results from these studies will provide data to improve variant detection in the future.

As observed in the analysis of transcriptome data from 10 melanoma samples presented in Chapter 7, experimental design and the choice of suitable control tissues are major issues in cancer sequencing projects. Choosing an inappropriate control tissue or sample can greatly reduce the significance of the results. A close collaboration between biomedical and computational researchers from the initial experimental design to the final data analysis can improve the outcome of a study. The computational parts depend on the experimental design not only with respect to the choice of appropriate control samples, but also with respect to the number of biological and technical replicates that are needed to obtain statistically significant results. As we have shown in Chapter 7, using RNA-seq data from a different tissue as control can lead to problems that are related to tissue-specific gene expression patterns. Healthy skin as control sample for melanoma, a skin cancer, seems more appropriate that PBMCs, however, melanocytes are very infrequent in in skin samples and are thus not appropriately represented by skin biopsies. A suitable control sample for RNA-seq is thus not available in this case. Using exome or genome sequencing for the control tissue would at least have solved some of the problems that arise from missing coverage in the control tissue. Additionally, samples are obtained under real-life conditions that sometimes largely diverge from ideal ones. For example, the major focus of a surgeon during a complicated liver carcinoma surgery is not to obtain a good tumor sample for sequencing but the removal of the tumor and the survival of the patient. Unquestioned usage of a sample obtained under such conditions might lead to sequencing of samples with very low percentages of tumor cells. This has to be taken into account during the data analysis in order to produce useful results. A close collaboration of scientists from all involved fields from the very beginning of an interdisciplinary project is indispensable for the success of the project.

Genetic mutation in cancer is just the beginning of the story. In order to assess the importance of mutations with respect to tumor development, additional levels have to be taken into account. The combined mapping to biological pathways of data on the genome (small mutations, methylation, copy number variations), transcriptome (expression levels and alternative transcripts), proteome (protein abundances, phosphorylation) and even the metabolome level is expected to allow a broader view on cancer than just looking at the genome does. Mutations can be mapped to biological pathways and, based on gene and protein expression data, the downstream effects of these mutation can be investigated. Two main challenges need to be solved before integrated, multi-omics analyses become feasible. First, the analysis of single omics levels is still far from being a solved problem, as shown in this thesis with respect to the detection of genetic variants from NGS data. Second, theoretical and computational models have to be developed to map the different levels to one. A one-to-one mapping of genes, transcripts and proteins is not possible due to overlapping genes, alternative transcripts or alternative protein isoforms. The picture gets even more complicated if metabolomics data is included. Research is going on to the address these challenges and new methods for data integration and pathway analysis will further promote the understanding of cancer development.

In order to develop new approaches to cancer treatment we need both, large-scale comparative studies and analyses of individual tumors. Large-scale studies on patient cohorts will allow to identify recurrent mutations or mechanisms. Studies on individualized tumors will allow identifying individual targets for personalized therapies. Personalized approaches to cancer therapies are a very promising strategy to develop new cancer treatments. Somatic mutations are a perfect goal for targeted and personalized treatments, however only a small fraction of these mutations are suitable as targets for chemotherapy. The immune system is a prime example for personalized and targeted reactions. Training the immune system by vaccines to attack cells harboring somatic cancer mutations is therefore a compelling approach.

In this thesis we present computational methods for the identification of tumor- and patient-specific T-cell epitopes. We develop and present methods for the prediction of HLA binding for a large number of alleles and a method to predict T-cell reactivity for HLA-binding peptides. We also show how these methods can be applied to clinical data to identify candidate peptides for epitope-based vaccines (as shown for the melanoma samples in Chapter 7) or for the identification for miHA candidates for exploiting graft-versus-leukemia reaction after stem cell transplantation (Chapter 8).

Some challenges still need to be addressed in this area. The first one concerns the prediction of T-cell epitopes. Today, we are able to reliably predict the HLA-binding for peptides. We have proposed a method to assess central tolerance as a factor that contributes to the immunogenicity of HLA-binding peptides. Many other factors that influence immunogenicity, for example peripheral tolerance and the influence of regulatory T cells or other immune-regulating mechanisms, are still not understood and the prediction is not possible. A second challenge is the selection of candidate T-cell epitopes for therapy. For the design of general epitope-based vaccines, strategies to select the optimal set of epitopes exist [157, 158]. These methods can in principle be adapted to select an optimal set of candidate epitopes in a personalized setting. However, we first need to define the selection criteria. The definition of such criteria is not trivial and clinical data to base them on is sparse. Clinical and immunological experiments will have to be performed to address questions concerning a suitable number of different epitopes, the alleles to be covered (e.g., is it beneficial to include T-helper cell epitopes to induce a cytotoxic T-cell reaction), and the form of administration (peptide cocktail, string-of-beads [181]) of the epitopes as vaccines. A somatic mutation that is present in the tumor on the genetic level, but is not expressed is no suitable target, so gene and protein expression should be taken into account. Targeting an epitope in a pathway that is down-regulated by a small-molecule drug (for example a kinase inhibitor) can impede efficacy of the treatment. Thus, the additional treatments that the patient obtains need to be considered. T-cell epitope-based personalized immunotherapy as cancer treatment is a promising approach. But cancers will be able to develop resistance and escape mechanisms, for example by downregulating MHC expression. Immunotherapies will thus not be the silver bullet in the war against cancer. But they can offer additional possibilities when combined with other treatments, especially since the expected side-effects are very mild.

The major challenge in personalized immunotherapy of cancer, however, is that a broad clinical validation of the concept remains elusive. In order to promote clinical application and first clinical trials, computational methods are indispensable. Computational methods need to be included in the design of new treatment approaches, despite the fact that some issues on the computational side, but also on the medical side, are still unresolved. An iterative process of computationally proposing candidate targets for immunotherapy, experimental testing and validation, and improvement of the computational methods is a promising procedure.

We integrated our computational methods into a flexible workflow system that allows the processing of clinical data in a timely manner and presents the results in a comprehensive way to our clinical partners. The application of these computational workflows led to first clinical tests of new T-cell epitope based therapy options. As results from experimental validation come at hand we can improve the computational methods and quickly adapt the workflows.

The conclusion that I personally draw from the work for this thesis is that computational methods can significantly promote the development of new anti-cancer treatment options. A close, constructive, and open-minded collaboration of experts from different fields, namely cancer immunology and biology, clinical cancer research, and computational biology is needed to close the gap between theoretical understanding of cancers and clinical application. Each field comes with a different view and understanding of cancer. Combining and reconciling these different perspectives and experiences rather than pursuing a single-disciplinary approach opens new roads in cancer research. Accurate computational methods are a valuable and indispensable part of such interdisciplinary approaches, however, without the application to clinical data these methods alone will not provide insight to cancer biology. In close collaboration with clinical partners, computational methods like the ones presented in this thesis, can be the basis of a major step towards personalized therapies.

Part V

Appendix

Appendix A

Pathways affected by somatic mutations in the melanoma samples

Tab. A.1: KEGG pathways that are mutated in at least one gene. The Pathways are selected manually. Metabolic pathways are excluded. We also excluded all pathways that contain HLA genes or the gene GLUD2.

KEGG PathwayID	Mutations	Patients	Genes	PatientIDs	Description
hsa05200	7	2	FOXO1 BAX APC BRCA2 STK36 BCR CASP9	Pat03,Pat09	Pathways in cancer
hsa03410	3	3	NEIL1 PARP4	Pat03, Pat05, Pat09	Base excision repair
hsa04210	3	2	BAX CFLAR CASP9	Pat03, Pat09	Apoptosis
hsa05210	3	2	BAX APC CASP9	Pat03, Pat09	Colorectal cancer
hsa04520	2	2	PTPN6 CSNK2A2	Pat03, Pat09	Adherens junction
hsa04110	2	2	PTTG2 ORC3	Pat03, Pat09	Cell cycle
hsa04115	2	2	BAX CASP9	Pat03, Pat09	p53 signaling pathway
hsa04630	2	2	IL10RA PTPN6	Pat03, Pat09	Jak-STAT signaling pathway
hsa03015	2	2	MAGOHB PABPC3	Pat07, Pat08	mRNA surveillance pathway
hsa04623	2	2	DDX58 TREX1	Pat03, Pat10	Cytosolic DNA-sensing pathway
hsa04310	4	1	APC NFATC3 VANGL1 CSNK2A2	Pat03	Wnt signaling pathway
hsa05202	4	1	FOXO1 BMP2K UTY NCOR1	Pat03	Transcriptional misregulation in cancer
hsa04510	3	1	FLNB TLN2 VAV1	Pat03	Focal adhesion
hsa04810	3	1	APC VAV1 MSN	Pat03	Regulation of actin cytoskeleton
hsa04010	3	1	NF1 FLNB DUSP7	Pat03	MAPK signaling pathway
hsa04060	3	1	IL10RA ACVR2A CCR5	Pat03	Cytokine-cytokine receptor interaction
hsa04120	2	1	UBE2L3 UBOX5	Pat03	Ubiquitin mediated proteolysis
hsa03420	2	1	RFC4 ERCC2	Pat09	Nucleotide excision repair
hsa04370	2	1	NFATC3 CASP9	Pat03	VEGF signaling pathway
hsa05217	2	1	APC STK36	Pat03	Basal cell carcinoma
hsa05215	2	1	FOXO1 CASP9	Pat03	Prostate cancer
hsa05212	2	1	BRCA2 CASP9	Pat03	Pancreatic cancer
hsa05213	2	1	APC CASP9	Pat03	Endometrial cancer
hsa05223	1	1	CASP9	Pat03	Non-small cell lung cancer
hsa05222	1	1	CASP9	Pat03	Small cell lung cancer
hsa05220	1	1	BCR	Pat03	Chronic myeloid leukemia
hsa04350	1	1	ACVR2A	Pat03	TGF-beta signaling pathway
hsa03022	1	1	ERCC2	Pat09	Basal transcription factors
hsa04620	1	1	TLR4	Pat03	Toll-like receptor signaling pathway
hsa04966	1	1	SLC12A7	Pat09	Collecting duct acid secretion
hsa04622	1	1	DDX58	Pat03	RIG-I-like receptor signaling pathway
hsa03430	1	1	RFC4	Pat09	Mismatch repair
hsa03030	1	1	RFC4	Pat09	DNA replication
hsa04974	1	1	KCNQ1	Pat03	Protein digestion and absorption
hsa04975	1	1	GOT2	Pat04	Fat digestion and absorption
hsa04976	1	1	SLC4A5	Pat03	Bile secretion
hsa04977	1	1	BTD	Pat03	Vitamin digestion and absorption
hsa04971	1	1	KCNQ1	Pat03	Gastric acid secretion
hsa04972	1	1	KCNQ1	Pat03	Pancreatic secretion
hsa03320	1	1	ACSL1	Pat09	PPAR signaling pathway

Appendix B

Contributions

Viral Integration

Magdalena Feldhahn (MF), Nico Weber (NW), Moritz Menzel (MM), Diana Meckbach (DM), Oliver Kohlbacher (OK), Daniel Huson (DH), and Jürgen Bauer (JB) contributed to this project.
JB and OK conceived the projects. MF and OK designed the project. MF and NW performed the bioinformatics experiments. MF, NW, MM and JB analyzed the results. OK, MF, JB, DH, MM, and DM contributed to the discussion. Parts of this section have been published in [92].

UniTope

Magdalena Feldhahn (MF), Nora C. Toussaint (NCT), Matthias Ziehm (MZ), and Oliver Kohlbacer (OK) contributed to this project.
OK, NCT, and MF designed the original study. NCT, MF and MZ performed the experiments for the original UniTope approach presented in this thesis. For the improved regression version of UniTope, NCT performed the experiments, MZ retrieved the 3D-structures and determined the pocket profiles. NCT, OK, and MF contributed to the discussion. Parts of this section are taken from an unpublished manuscript by Magdalena Feldhahn, Nora C. Toussaint, Matthias Ziehm, and Oliver Kohlbacher.

Prediction on T cell-reactivity

Magdalena Feldhahn (MF), Nora C. Toussaint (NCT), Sebastian Briesemeister (SB), Matthias Ziehm (MZ), Gunnar Rätsch (GR), Stefan Stevanović (SS), and Oliver Kohlbacher contributed to this project. NCT and OK designed the experiments, NCT performed the experiments. SS provided experimental data. MZ determined the thymus proteomes. Ok and MF implemented the distance tries, MF performed the distance-to-self calculations and the MHC binding predictions. NCT, MF, SB, GR and OK contributed to the discussion. Parts of this section were presented at the *Second Immunoinfomratics and Computational Immunology Workshop (ICIW 2011)* and are included in the workshop proceedings [149].

FRED

Magdalena Feldhahn (MF), Oliver Kohlbacher (OK), Pierre Dönnes (PD), Philipp Thiel (PT), and Mathias Walzer (WZ) contributed to this project. OK, PD, and MF conceived the project. MF and PD designed the project, MF implemented the main framework, PT and MW contributed to the implementation of polymorphism handling. Parts of this section have been published in [159].

EpiToolKit

Magdalena Feldhahn (MF), Oliver Kohlbacher (OK), Philipp Thiel (PT), Mathias M. Schuler (MMS), Nina Hillen (NH), Stefan Stevanović (SS), and Hans-Georg Rammensee (HGR) contributed to this project. OK and MF conceived the project. MF, OK and PT designed the project, PT and MF contributed to the implementation. MF, OK, PT, MMS, NH, SS, HGR contributed to the testing and the discussion. Parts of this section have been published in [163].

VariationDB

Sebastian Bögel (SB), Magdalena Feldhahn (MF), Oliver Kohlbacher (OK), Stefan Stevaniović (SS), Jürgen Bauer (JB), Moritz Menzel (MM), and Benjamin Schubert (BS) contributed to this project.

OK and MF conceived the project. MF and SB designed the project. MF, SB, and BS contributed to the implementation. MF, SB, OK, SS, JB, MM contributed to the testing and the discussion.

High-throughput detection of minor histocompatibility antigens

Magdalena Feldhahn (MF), Karin Schilbach (KS), Pierre Dönnes (PD), Benjamin Schubert (BS), Oliver Kohlbacher (OK) and Hans-Georg Rammensee (HGR) contributed to this project.
MF, KS, OK and HGR conceived and designed the project. MF and BS implemented and tested the pipeline and the user interface. MF, KS, OK, PD and HGR contributed to the discussion. Parts of this section have been published in [182].

Workflows in Galaxy

Magdalena Feldhahn (MF) and Nico Weber (NW). NW designed and implemented the Galaxy server. MF integrated and tested the tools and designed and implemented the workflows.

Appendix C

Publications

Published Manuscripts

- **Feldhahn, M**, Dönnes, P, Schubert, B, Schilbach, K, Rammensee, HG, Kohlbacher, O. miHA-Match: computational detection of tissue-specific minor histocompatibility antigens. *Journal of Immunological Methods*, 386:94-100, 2012. [182]
  - Text and figures from this manuscript appear in Section 6.2 of this thesis.
- Toussaint, NC, **Feldhahn, M**, Ziehm, M, Stevanović, S, and Kohlbacher, O. T-Cell Epitope Prediction Based on Self-Tolerance. In: *Proceedings of the Second Immunoinformatics and Computational Immunology Workshop 2011*, 2011. [149]
  - Text and figures from this manuscript appear in Section 5.3 of this thesis.
- **Feldhahn, M**, Menzel, M, Weide, B, Bauer, P, Meckbach, D, Garbe, C, Kohlbacher, O, and Bauer, J. No evidence of viral genomes in whole-transcriptome sequencings of three melanoma metastases. *Experimental Dermatology*, 20(9):766-768, 2011. [92]
  - Text and figures from this manuscript appear in Sections 4.2 and 7.3 of this thesis.
- **Feldhahn, M**, Dönnes, P, Thiel, P, and Kohlbacher, O. FRED - A Framework for T-cell Epitope Detection. *Bioinformatics*, 25(20):2758-9, 2009. [159]
  - Text and figures from this manuscript appear in Sections 6.1.1 of this thesis.
- **Feldhahn, M**, Thiel, P, Schuler, M, Hillen, N, Stevanović, S, Rammensee, H, and Kohlbacher, O. EpiToolKit - A web server for computational immunomics. *Nucleic Acids Res.*, 36:W519-22, 2008. [163]
  - Text and figures appear in Section 6.1.2 of this thesis.

Bibliography

[1] The National Cancer Act, Senate Bill 1828 - Enacted December 23, 1971 (P.L. 92-218), 1971.

[2] American Association for Cancer Research (AACR). AACR Cancer Progress Report 2011. www.cancerprogresreprot.org, 2011.

[3] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[4] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113, 2007.

[5] American Cancer Association. Cancer Facts & Figures 2011. Atlanta:American Cancer Association, 2011.

[6] American Cancer Society. The Global Economic Cost of Cancer. http://www.cancer.org/acs/groups/content/@internationalaffairs/documents/document/acspc-026203.pdf, 2010.

[7] J. V. Melo. The molecular biology of chronic myeloid leukaemia. *Leukemia*, 10(5):751–756, 1996.

[8] P. S. Steeg. Perspective: The right trials. *Nature*, 485(7400):S58–S59, 2012.

[9] C. Gouttefangeas, A. Stenzl, S. Stevanović, and H.-G. Rammensee. Immunotherapy of renal cell carcinoma. *Cancer Immunol Immunother*, 56(1):117–128, 2007.

[10] M. D. Vesely, M. H. Kershaw, R. D. Schreiber, and M. J. Smyth. Natural innate and adaptive immunity to cancer. *Annu Rev Immunol*, 29:235–271, 2011.

[11] H. Davies, G. R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M. J. Garnett, W. Bottomley, N. Davis, E. Dicks, R. Ewing, Y. Floyd, K. Gray, S. Hall, R. Hawes, J. Hughes, V. Kosmidou, A. Menzies, C. Mould, A. Parker, C. Stevens, S. Watt, S. Hooper, R. Wilson, H. Jayatilake, B. A. Gusterson, C. Cooper, J. Shipley, D. Hargrave, K. Pritchard-Jones, N. Maitland, G. Chenevix-Trench, G. J. Riggins, D. D. Bigner, G. Palmieri, A. Cossu, A. Flanagan, A. Nicholson, J. W. C. Ho, S. Y. Leung, S. T. Yuen, B. L. Weber, H. F. Seigler, T. L. Darrow, H. Paterson, R. Marais, C. J. Marshall, R. Wooster, M. R. Stratton, and P. A. Futreal. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954, 2002.

[12] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[13] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[14] T. Boveri. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J Cell Sci*, 121 Suppl 1:1–84, 2008.

[15] L. E. Macconaill and L. A. Garraway. Clinical implications of the cancer genome. *J Clin Oncol*, 28(35):5219–5228, 2010.

[16] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, 2011.

[17] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, Wellcome Trust Case Control Consortium, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010.

[18] http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml.

[19] C. A. Pratilas, B. S. Taylor, Q. Ye, A. Viale, C. Sander, D. B. Solit, and N. Rosen. (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. *Proc Natl Acad Sci USA*, 106(11):4519–4524, 2009.

[20] F. Mitelman, B. Johansson, and F. Mertens. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*, 7(4):233–245, 2007.

[21] H. Zur Hausen. The search for infectious causes of human cancers: where and why. *Virology*, 392(1):1–10, 2009.

[22] H. Feng, M. Shuda, Y. Chang, and P. S. Moore. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science*, 319(5866):1096–1100, 2008.

[23] C. A. Janeway, P. Travers, M. Walport, and M. J. Shlomshik. *Immunobiology - the immune system in health and disease*. Garland Science Publishing, 6th edition, 2005.

[24] A. K. Abbas and A. H. Lichtman. *Basic Immunology*. W.B. Saunders Company, 2nd edition, 2004.

[25] J. Neefjes, M. L. M. Jongsma, P. Paul, and O. Bakke. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol*, 11(12):823–836, 2011.

[26] IMGT/HLA Database Statistics. http://www.ebi.ac.uk/imgt/hla/stats.html, 2012.

[27] G. P. Dunn, A. T. Bruce, H. Ikeda, L. J. Old, and R. D. Schreiber. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol*, 3(11):991–998, 2002.

[28] T. Fojo and D. R. Parkinson. Biologically targeted cancer therapy and marginal benefits: are we making too much of too little or are we achieving too little by giving too much? *Clin Cancer Res*, 16(24):5972–5980, 2010.

[29] D. Dornan and J. Settleman. Dissecting cancer heterogeneity. *Nat Biotechnol*, 29(12):1095–1096, 2011.

[30] K. A. Chianese-Bullock, W. P. Irvin, Jr, G. R. Petroni, C. Murphy, M. Smolkin, W. C. Olson, E. Coleman, S. A. Boerner, C. J. Nail, P. Y. Neese, A. Yuan, K. T. Hogan, and C. L. Slingluff, Jr. A multipeptide vaccine is safe and elicits T-cell responses in participants with advanced stage ovarian cancer. *J Immunother*, 31(4):420–430, 2008.

[31] G. G. Kenter, M. J. P. Welters, A. R. P. M. Valentijn, M. J. G. Lowik, D. M. A. Berends-van der Meer, A. P. G. Vloon, J. W. Drijfhout, A. R. Wafelman, J. Oostendorp, G. J. Fleuren, R. Offringa, S. H. van der Burg, and C. J. M. Melief. Phase I immunotherapeutic trial with long peptides spanning the E6 and E7 sequences of high-risk human papillomavirus 16 in end-stage cervical cancer patients shows low toxicity and robust immunogenicity. *Clin Cancer Res*, 14(1):169–177, 2008.

[32] C. L. Slingluff, Jr, G. R. Petroni, K. A. Chianese-Bullock, M. E. Smolkin, S. Hibbitts, C. Murphy, N. Johansen, W. W. Grosh, G. V. Yamshchikov, P. Y. Neese, J. W. Patterson, R. Fink, and P. K. Rehm. Immunologic and clinical outcomes of a randomized phase II trial of two multipeptide vaccines for melanoma in the adjuvant setting. *Clin Cancer Res*, 13(21):6386–6395, 2007.

[33] N. Yajima, R. Yamanaka, T. Mine, N. Tsuchiya, J. Homma, M. Sano, T. Kuramoto, Y. Obata, N. Komatsu, Y. Arima, A. Yamada, M. Shigemori, K. Itoh, and R. Tanaka. Immunologic evaluation of personalized peptide vaccination for patients with advanced malignant glioma. *Clin Cancer Res*, 11(16):5900–5911, 2005.

[34] C. L. Slingluff, Jr, G. R. Petroni, G. V. Yamshchikov, D. L. Barnd, S. Eastham, H. Galavotti, J. W. Patterson, D. H. Deacon, S. Hibbitts, D. Teates, P. Y. Neese, W. W. Grosh, K. A. Chianese-Bullock, E. M. H. Woodson, C. J. Wiernasz, P. Merrill, J. Gibson, M. Ross, and V. H. Engelhard. Clinical and immunologic results of a randomized phase II trial of vaccination using four melanoma peptides either administered in granulocyte-macrophage colony-stimulating factor in adjuvant or pulsed on dendritic cells. *J Clin Oncol*, 21(21):4016–4026, 2003.

[35] A. Sette and J. Fikes. Epitope-based vaccines: an update on epitope identification, vaccine design and delivery. *Curr Opin Immunol*, 15(4):461–470, 2003.

[36] H.-I. Cho and E. Celis. Optimized peptide vaccines eliciting extensive CD8 T-cell responses with therapeutic antitumor effects. *Cancer Res*, 69(23):9012–9019, 2009.

[37] S. A. Perez, E. von Hofe, N. L. Kallinteris, A. D. Gritzapis, G. E. Peoples, M. Papamichail, and C. N. Baxevanis. A new era in anticancer peptide vaccines. *Cancer*, 116(9):2071–2080, 2010.

[38] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–5467, 1977.

[39] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995.

[40] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, R. D. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. Hutchison, 3rd, and

J. C. Venter. The minimal gene complement of mycoplasma genitalium. *Science*, 270(5235):397–403, 1995.

[41] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen,

M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[42] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V.

Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and I. H. G. S. C. . Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[43] C. A. Hutchison, 3rd. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res*, 35(18):6227–6237, 2007.

[44] M. L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, 2010.

[45] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[46] G. E. Moore. Cramming more components onto integrated circiuts. *Electronics*, 38:4–7, 1965.

[47] C. Walter. Kryders's law. *Scienticic American*, 293:32–33, 2005.

[48] R. Tehrani. As we may communicate. *TMCnet*, 2000.

[49] L. D. Stein. The case for cloud computing in genome informatics. *Genome Biol*, 11(5):207, 2010.

[50] M. Meyerson, S. Gabriel, and G. Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11(10):685–696, 2010.

[51] Fastqc. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

[52] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[53] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[54] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, 2008.

[55] D. Weese, A.-K. Emde, T. Rausch, A. Döring, and K. Reinert. RazerS–fast read mapping with sensitivity control. *Genome Res*, 19(9):1646–1654, 2009.

[56] G. Lunter and M. Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*, 21(6):936–939, 2011.

[57] A.-K. Emde, M. H. Schulz, D. Weese, R. Sun, M. Vingron, V. M. Kalscheuer, S. A. Haas, and K. Reinert. Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using splazers. *Bioinformatics*, 2012.

[58] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[59] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–i180, 2008.

[60] G. Jean, A. Kahles, V. T. Sreedharan, F. De Bona, and G. Rätsch. RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics*, Chapter 11:Unit 11.6, 2010.

[61] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498, 2011.

[62] R. Goya, M. G. F. Sun, R. D. Morin, G. Leung, G. Ha, K. C. Wiegand, J. Senz, A. Crisan, M. A. Marra, M. Hirst, D. Huntsman, K. P. Murphy, S. Aparicio, and S. P. Shah. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–736, 2010.

[63] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 2009.

[64] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009.

[65] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 2012.

[66] A. Roth, R. Morin, J. Ding, A. Crisan, G. Ha, R. Giuliany, A. Bashashati, M. Hirst, G. Turashvili, A. Oloumi, M. A. Marra, S. Aparicio, and S. P. Shah. JointSNVMix: A Probabilistic Model For Accurate Detection Of Somatic Mutations In Normal/-Tumour Paired Next Generation Sequencing Data. *Bioinformatics*, 2012.

[67] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. W. Edwards, G. R. Bignell, M. R. Stratton, and P. A. Futreal. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–729, 2008.

[68] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, 2009.

[69] C. A. Maher, N. Palanisamy, J. C. Brenner, X. Cao, S. Kalyana-Sundaram, S. Luo, I. Khrebtukova, T. R. Barrette, C. Grasso, J. Yu, R. J. Lonigro, G. Schroth, C. Kumar-Sinha, and A. M. Chinnaiyan. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A*, 106(30):12353–12358, 2009.

[70] A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M. G. F. Sun, M. Griffith, A. Heravi Moussavi, J. Senz, N. Melnyk, M. Pacheco, M. A. Marra, M. Hirst, T. O. Nielsen, S. C. Sahinalp, D. Huntsman, and S. P. Shah. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol*, 7(5):e1001138, 2011.

[71] D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. T. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, 6(1):99–103, 2009.

[72] J. F. Sathirapongsasuti, H. Lee, B. A. J. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush, and S. F. Nelson. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, 27(19):2648–2654, 2011.

[73] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–628, 2008.

[74] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[75] P. C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814, 2003.

[76] C. Schaefer, A. Meier, B. Rost, and Y. Bromberg. SNPdbe: Constructing an nsSNP functional impacts database. *Bioinformatics*, 2011.

[77] B. Reva, Y. Antipin, and C. Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*, 39(17):e118, 2011.

[78] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311, 2001.

[79] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–D517, 2005.

[80] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, and R. Wooster. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*, 91(2):355–358, 2004.

[81] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*, 39(Database issue):D945–D950, 2011.

[82] J. Küntzer, D. Maisel, H.-P. Lenhof, S. Klostermann, and H. Burtscher. The Roche Cancer Genome Database 2.0. *BMC Med Genomics*, 4:43, 2011.

[83] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[84] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114, 2012.

[85] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[86] M. J. Heller. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, 4:129–153, 2002.

[87] T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res*, 37(13):4181–4193, 2009.

[88] P. van den Ijssel, M. Tijssen, S.-F. Chin, P. Eijk, B. Carvalho, E. Hopmans, H. Holstege, D. K. Bangarusamy, J. Jonkers, G. A. Meijer, C. Caldas, and B. Ylstra. Human and mouse oligonucleotide-based array CGH. *Nucleic Acids Res*, 33(22):e192, 2005.

[89] A. Döring, D. Weese, T. Rausch, and K. Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, 2008.

[90] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*, 37(Database issue):D32–D36, 2009.

[91] P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39(Database issue):D876–D882, 2011.

[92] M. Feldhahn, M. Menzel, B. Weide, P. Bauer, D. Meckbach, C. Garbe, O. Kohlbacher, and J. Bauer. No evidence of viral genomes in whole-transcriptome sequencing of three melanoma metastases. *Exp Dermatol*, 2011.

[93] H. Feng, J. L. Taylor, P. V. Benos, R. Newton, K. Waddell, S. B. Lucas, Y. Chang, and P. S. Moore. Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. *J Virol*, 81(20):11332–11340, 2007.

[94] SeqClean. `http://compbio.dfci.harvard.edu/tgi/software/`.

[95] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4):462–467, 2005.

[96] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Res*, 21(9):1552–1560, 2011.

[97] S. T. Arron, J. G. Ruby, E. Dybbro, D. Ganem, and J. L. Derisi. Transcriptome sequencing demonstrates that human papillomavirus is not active in cutaneous squamous cell carcinoma. *J Invest Dermatol*, 131(8):1745–1753, 2011.

[98] H. H. Niller, H. Wolf, and J. Minarovits. Viral hit and run-oncogenesis: genetic and epigenetic scenarios. *Cancer Lett*, 305(2):200–217, 2011.

[99] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829, 2008.

[100] E. L. Anson and E. W. Myers. ReAligner: a program for refining DNA sequence multi-alignments. *J Comput Biol*, 4(3):369–383, 1997.

[101] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, 2010.

[102] P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek, and B. Zhang. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–1561, 2004.

[103] C. J. A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3(3):265–274, 2002.

[104] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, 2002.

[105] J.-S. Boegel. VariationDB - An Interface for in silico Analysis of Genetic Variation in Cancer Reasearch. Master's thesis, University of Tuebingen, 2011.

[106] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

[107] V. Vapnik. The Nature of Statistical Learning Theory [M]. *NY: Springer-Verlag*, 1995.

[108] P. Cascio, C. Hilton, A. F. Kisselev, K. L. Rock, and A. L. Goldberg. 26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide. *EMBO J*, 20(10):2357–2366, 2001.

[109] K. L. Rock, I. A. York, and A. L. Goldberg. Post-proteasomal antigen processing for major histocompatibility complex class I presentation. *Nat Immunol*, 5(7):670–677, 2004.

[110] P. Saxova, S. Buus, S. Brunak, and C. Kesmir. Predicting proteasomal cleavage sites: a comparison of available methods. *Int Immunol*, 15(7):781–787, 2003.

[111] H. G. Holzhütter, C. Frömmel, and P. M. Kloetzel. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. *J Mol Biol*, 286(4):1251–1265, 1999.

[112] M. Nielsen, C. Lundegaard, O. Lund, and C. K. cmir. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, 57(1-2):33–41, 2005.

[113] P. Dönnes and O. Kohlbacher. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci*, 14(8):2132–2140, 2005.

[114] M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, S. Brunak, O. Lund, and M. Nielsen. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol*, 35(8):2295–2303, 2005.

[115] B. Peters, S. Bulik, R. Tampe, P. M. V. Endert, and H.-G. Holzhütter. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J Immunol*, 171(4):1741–1749, 2003.

[116] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, 1999.

[117] T. Sturniolo, E. Bono, J. Ding, L. Raddrizzani, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M. P. Protti, F. Sinigaglia, and J. Hammer. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol*, 17(6):555–561, 1999.

[118] P. A. Reche, J.-P. Glutting, and E. L. Reinherz. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol*, 63(9):701–709, 2002.

[119] M. Nielsen, C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, 20(9):1388–1397, 2004.

[120] V. Brusic, G. Rudy, and L. C. Harrison. Prediction of MHC Binding Peptides Using Artificial Neural Networks. *Complexity International*, 02, 1995.

[121] K. Gulukota, J. Sidney, A. Sette, and C. DeLisi. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol*, 267(5):1258–1267, 1997.

[122] M. C. Honeyman, V. Brusic, N. L. Stone, and L. C. Harrison. Neural network-based prediction of candidate T-cell epitopes. *Nat Biotechnol*, 16(10):966–969, 1998.

[123] S. Buus, S. L. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, and S. Brunak. Sensitive quantitative predictions of peptide-mhc binding by a 'query by committee' artificial neural network approach. *Tissue Antigens*, 62(5):378–384, 2003.

[124] P. Dönnes and O. Kohlbacher. SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res*, 34(Web Server issue):W194–W197, 2006.

[125] N. C. Toussaint and O. Kohlbacher. Towards in silico design of epitope-based vaccines. *Expert Opinion on Drug Discovery*, 4(10):1047–1060, 2009.

[126] D. Rognan, S. L. Lauemoller, A. Holm, S. Buus, and V. Tschinke. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem*, 42(22):4650–4658, 1999.

[127] O. Schueler-Furman, Y. Altuvia, A. Sette, and H. Margalit. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci*, 9(9):1838–1846, 2000.

[128] N. Jojic, M. Reyes-Gomez, D. Heckerman, C. Kadie, and O. Schueler-Furman. Learning MHC I–peptide binding. *Bioinformatics*, 22(14):e227–e235, 2006.

[129] A. Sette and J. Sidney. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, 50(3-4):201–212, 1999.

[130] O. Lund, M. Nielsen, C. Kesmir, A. G. Petersen, C. Lundegaard, P. Worning, C. Sylvester-Hvid, K. Lamberth, G. Roder, S. Justesen, S. Buus, and S. Brunak. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, 55(12):797–810, 2004.

[131] J. Sidney, B. Peters, N. Frahm, C. Brander, and A. Sette. HLA class I supertypes: a revised and updated classification. *BMC Immunol*, 9(1):1, 2008.

[132] D. S. DeLuca, B. Khattab, and R. Blasczyk. A modular concept of HLA for comprehensive peptide binding prediction. *Immunogenetics*, 59(1):25–35, 2007.

[133] G. Chelvanayagam. A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics*, 45(1):15–26, 1996.

[134] B. Peters, J. Sidney, P. Bourne, H.-H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J. V. Ponomarenko, M. Sathiamurthy, S. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, and A. Sette. The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol*, 3(3):e91, 2005. URL: http://www.immuneepitope.org/ Release: release 2007 1 31 23 16.

[135] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, 28(1):45–48, 2000.

[136] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

[137] O. Kohlbacher and H. P. Lenhof. BALL–rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. *Bioinformatics*, 16(9):815–824, 2000.

[138] J. Robinson, M. J. Waller, P. Parham, N. de Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, and S. G. E. Marsh. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res*, 31(1):311–314, 2003.

[139] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, 1994.

[140] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transaction on Neural Networks.*, 12(2):181–201, 2001.

[141] M. S. Venkatarajan and W. Braun. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J Mol Model*, 7:445–453, 2001.

[142] S. Kawashima, H. Ogata, and M. Kanehisa. AAindex: Amino Acid Index Database. *Nucleic Acids Res*, 27(1):368–369, 1999.

[143] P. Dönnes and A. Elofsson. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, 3:25, 2002.

[144] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm/, 2001.

[145] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[146] N. C. Toussaint. *New Approaches to in silico Deisgn of Epitope-Based Vaccines*. PhD thesis, University of Tübingen, 2011.

[147] M. Nielsen, C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Roder, B. Peters, A. Sette, O. Lund, and S. Buus. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE*, 2(8):e796, 2007.

[148] C.-W. Tung and S.-Y. Ho. POPI: Predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*, 2007.

[149] N. C. Toussaint, M. Feldhahn, M. Ziehm, S. Stevanović, and O. Kohlbacher. T-cell epitope prediction based on self-tolerance. *Proc. ICIW*, 2011.

[150] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988, 2004.

[151] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue):D885–D890, 2009.

[152] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868–D872, 2009.

[153] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, 1992.

[154] A. Sette, A. Vitiello, B. Reherman, P. Fowler, R. Nayersina, W. M. Kast, C. J. Melief, C. Oseroff, L. Yuan, J. Ruppert, J. Sidney, M. F. del Guercio, S. Southwood, R. T. Kubo, R. W. Chesnut, H. M. Grey, and F. V. Chisari. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J Immunol*, 153(12):5586–5592, 1994.

[155] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN Machine Learning Toolbox. *Journal of Machine Learning Research*, 11:1799–1802, 2010.

[156] D. L. Mueller. Mechanisms maintaining peripheral tolerance. *Nat Immunol*, 11(1):21–27, 2010.

[157] N. C. Toussaint, P. Dönnes, and O. Kohlbacher. A mathematical framework for the selection of an optimal set of peptides for epitope-based vaccines. *PLoS Comput Biol*, 4(12):e1000246, 2008.

[158] N. C. Toussaint and O. Kohlbacher. OptiTope–a web server for the selection of an optimal set of peptides for epitope-based vaccines. *Nucleic Acids Res*, 37(Web Server issue):W617–W622, 2009.

[159] M. Feldhahn, P. Dönnes, P. Thiel, and O. Kohlbacher. FRED - A Framework for T-cell Epitope Detection. *Bioinformatics*, 2009.

[160] K. C. Parker, M. A. Bednarek, and J. E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 152(1):163–175, 1994.

[161] I. Doytchinova, S. Hemsley, and D. R. Flower. Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. *J Immunol*, 173(11):6813–6819, 2004.

[162] M. Nielsen, S. Justesen, O. Lund, C. Lundegaard, and S. Buus. NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res*, 6:9, 2010.

[163] M. Feldhahn, P. Thiel, M. M. Schuler, N. Hillen, S. Stevanović, H.-G. Rammensee, and O. Kohlbacher. EpiToolKit–a web server for computational immunomics. *Nucleic Acids Res*, 36(Web Server issue):W519–W522, 2008.

[164] M. Bleakley and S. R. Riddell. Molecules and mechanisms of the graft-versus-leukaemia effect. *Nat Rev Cancer*, 4(5):371–380, 2004.

[165] J. L. M. Ferrara, J. E. Levine, P. Reddy, and E. Holler. Graft-versus-host disease. *Lancet*, 373(9674):1550–1561, 2009.

[166] E. Goulmy. Minor histocompatibility antigens: from transplantation problems to therapy of cancer. *Hum Immunol*, 67(6):433–438, 2006.

[167] P. Hombrink, S. R. Hadrup, A. Bakker, M. G. D. Kester, J. H. F. Falkenburg, P. A. von dem Borne, T. N. M. Schumacher, and M. H. M. Heemskerk. High-throughput identification of potential minor histocompatibility antigens by MHC tetramer-based screening: feasibility and limitations. *PLoS One*, 6(8):e22523, 2011.

[168] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 39(Database issue):D38–D51, 2011.

[169] C. A. M. Van Bergen, C. E. Rutten, E. D. Van Der Meijden, S. A. P. Van Luxemburg-Heijs, E. G. A. Lurvink, J. J. Houwing-Duistermaat, M. G. D. Kester, A. Mulder, R. Willemze, J. H. F. Falkenburg, and M. Griffioen. High-throughput characterization of 10 new minor histocompatibility antigens by whole genome association scanning. *Cancer Res*, 70(22):9073–9083, 2010.

[170] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

[171] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, 2000.

[172] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S. Riat, D. Rios, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Herrero, T. J. P.

Hubbard, A. Parker, G. Proctor, J. Vogel, and S. M. J. Searle. Ensembl 2011. *Nucleic Acids Res*, 39(Database issue):D800–D806, 2011.

[173] UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, 38(Database issue):D142–D148, 2010.

[174] J. Taylor, I. Schenck, D. Blankenberg, and A. Nekrutenko. Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics*, Chapter 10:Unit 10.5, 2007.

[175] X. F. Zhao, M. Reitz, Q. C. Chen, and S. Stass. Pathogenesis of early leukemia and lymphoma. *Cancer Biomark*, 9(1-6):341–374, 2011.

[176] D. Ramsköld, E. T. Wang, C. B. Burge, and R. Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, 5(12):e1000598, 2009.

[177] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, J. W. Huss, 3rd, and A. I. Su. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*, 10(11):R130, 2009.

[178] A. Keller, N. Ludwig, C. Backes, B. F. M. Romeike, N. Comtesse, W. Henn, W.-I. Steudel, C. Mawrin, H.-P. Lenhof, and E. Meese. Genome wide expression profiling identifies specific deregulated pathways in meningioma. *Int J Cancer*, 124(2):346–351, 2009.

[179] L. Weng, F. Macciardi, A. Subramanian, G. Guffanti, S. G. Potkin, Z. Yu, and X. Xie. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, 12:99, 2011.

[180] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Ponten. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*, 28(12):1248–1250, 2010.

[181] N. C. Toussaint, Y. Maman, O. Kohlbacher, and Y. Louzoun. Universal peptide vaccines - optimal peptide vaccine design based on viral sequence conservation. *Vaccine*, 29(47):8745–8753, 2011.

[182] M. Feldhahn, P. Dönnes, B. Schubert, K. Schilbach, H.-G. Rammensee, and O. Kohlbacher. miHA-Match: Computational detection of tissue-specific minor histocompatibility antigens. *J Immunol Methods*, 386(1-2):94–100, 2012.