TECHNISCHE UNIVERSITÄT MÜNCHEN

Fachgebiet für Bioinformatik

Clinical Application of Next Generation Sequencing Data in Virology and Oncology

Kerstin Haase

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Michael Schemann
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Dimitri Frischmann
	2. apl.Prof. Dr. Christine Leib-Mösch (Ruprecht-Karls-Universität Heidelberg)

Die Dissertation wurde am 11.08.2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 07.12.2015 angenommen.

Kerstin Haase: *Clinical Application of Next Generation Sequencing Data in Virology and Oncology*, Dissertation, München, August 2015

The saddest aspect of life right now is that science gathers
knowledge faster than society gathers wisdom.
— Isaac Asimov

What we know is a drop, what we don’t know is an ocean.
— Isaac Newton

ABSTRACT

With the invention of traditional Sanger sequencing, genome analysis opened a wide field of research aiming at unveiling and understanding the genomic sequence of many organisms. Since then the sequencing methods have been improved, they have become automatised and are now capable of producing a large number of long reads at comparably low cost. In this thesis we wanted to harness the technology of next generation sequencing (NGS) to examine virus evolution, detect expression of viral remnants integrated in the human genome and provide an expression database to aid T cell engineering.

In the first part of this work we analysed the evolution of noroviral genomes within chronically infected hosts. By using longitudinal data, we could determine mutation rates and find traces of positive selection. The high resolution of next generation sequencing enabled us to reconstruct the different viral variants present in our samples and track their dynamics over time.

The second and third part of this thesis analyse expression of human endogenous retroviruses (HERVs) in different cell types. By extracting HERVs from cancer and matched healthy tissues and subjecting them to NGS, we wanted to identify all loci that were expressed in each sample and determine differences between cancerous and normal cells. We find however, that the expressed loci seem to depend more on the individual patient than the disease state. Additionally, to reveal the background expression level of HERVs in various tissues and cell types, we analysed RNA sequencing data from the ENCODE project with regard to HERV expression and performed differential expression analysis to reveal cell specific patterns. The study revealed two samples that show very uncharacteristic expression profiles which imply a potential change in the state of the cell.

The last part of this work introduces Expitope, a web server that provides in silico prediction of potential off-target effects of engineered T cell receptors used in immunotherapy. It enables users to search for a peptide of interest in a protein database and obtain all exact and approximate hits sorted by an adapted scoring function. We can show that known cases of fatal off-target reactions can be identified with Expitope's functionalities.

Overall this thesis presents four studies which are only feasible through the availability of next generation sequencing technologies. NGS makes it possible to examine mutational dynamics even in low frequency viral variants, to detect transcripts that only comprise a small fraction of the complete amount of cellular RNA and to identify mRNAs which are only scarcely expressed but still vital to a cell. With the ongoing improvement of sequencing technologies, further advancements to the fields of virology and oncology are approaching swiftly.

KURZZUSAMMENFASSUNG

Die Erfindung der traditionellen Sanger-Sequenzierung hat der Genomanalyse ein weites Feld an Möglichkeiten eröffnet, um Genome verschiedenster Organismen aufzuklären und zu verstehen. Seither sind die Sequenzierungsmethoden stetig verbessert worden; inzwischen laufen sie automatisiert ab und sind fähig, eine große Anzahl langer Reads zu vergleichbar geringen Kosten zu produzieren. In dieser Arbeit haben wir die Technologie der Next-Generation Sequenzierung (NGS) genutzt, um Evolution in Viren zu untersuchen, die Expression von viralen Überbleibseln im menschlichen Genom aufzudecken und eine Expressions-Datenbank zur Unterstützung von T-Zell Therapien zu entwerfen.

Im ersten Teil dieser Arbeit haben wir die Evolution von Norovirus Genomen innerhalb chronisch infizierter Patienten analysiert. Durch den Einsatz von longitudinalen Daten konnten wir die Mutationsrate bestimmen und Anzeichen für positive Selektion entdecken. Die hohe Auflösung von Next-Generation Sequenzierung hat es uns ermöglicht, die verschiedenen Varianten des Virus in unseren Proben zu rekonstruieren und ihre Veränderungen im Laufe der Zeit zu verfolgen.

Der zweite und dritte Teil dieser Dissertation analysieren die Expression von humanen endogenen Retroviren (HERVs) in verschiedenen Zelltypen. Indem HERVs aus Krebs- und zugehörigem gesundem Gewebe extrahiert und NGS unterzogen wurden, wollten wir alle Loci identifizieren, die in den verschiedenen Proben expremiert waren und Unterschiede zwischen Tumor- und normalen Zellen feststellen. Allerdings kamen wir zu dem Ergebnis, dass die expremierten Loci mehr von dem jeweiligen Patienten, als vom Krankheitszustand abzuhängen scheinen. Zusätzlich haben wir RNA Sequenzierungsdaten aus dem ENCODE Projekt im Bezug auf HERV Expression analysiert, um die grundlegenden Expressions-Level von HERVs in verschiedenen Geweben und Zelltypen festzustellen und eine Differential Expression Analysis durchgeführt, um zellspezifische Expressionsmuster aufzudecken. Die Studie hat gezeigt, dass zwei Proben sehr uncharakteristische Expressionsprofile haben, was auf eine vermeintliche Änderung im Zustand der Zelle hindeutet.

Der letzte Teil dieser Arbeit stellt Expitope vor, ein Webserver der eine in-silico Vorhersage von potentiellen Nebeneffekten von zur Immuntherapie künstlich erzeugten T-Zell-Rezeptoren zur Verfügung stellt. Er ermöglicht es dem Benutzer, nach einem bestimmten Peptid in einer Protein-Datenbank zu suchen und alle exakten und Näherungstreffer, sortiert nach einer angepassten Scoring-Funktion, zu erhalten. Wir konnten zeigen, dass bereits bekannte, tödlich verlaufende Nebeneffekte mit Expitopes Fähigkeiten identifiziert werden können.

Zusammengefasst präsentiert diese Dissertation vier Studien, die ohne die Verfügbarkeit von Next-Generation Sequenzierungs-Technologien nicht möglich wären. NGS ermöglicht es, Mutations-Verläufe selbst in Virus Varianten mit geringer Häufigkeit zu untersuchen, Transkripte zu finden, die nur einen geringen Anteil der gesamten zellulären RNA ausmachen, und mRNAs zu identifizieren, die nur in geringer Menge expremiert werden, aber dennoch entscheidend für die Zelle sind. Durch die andauernde Verbesserung der Sequenzierungstechnologien ist damit zu rechnen, dass es bald noch weitere bedeutende Fortschritte in den Feldern der Virologie und Onkologie geben wird.

PUBLICATIONS

Some of the presented concepts and figures have previously been published in scientific journals or have been presented as talks and posters at conferences:

POSTERS:

- Haase K, Raffegerst S, Schendel DJ, Frishman D: Expitope: Web server for epitope expression, presented at the *European Conference on Computational Biology 2014*, Track: Bioinformatics of Health and Disease, Straßburg, France, 2014
- Haase K, Frishman D: Intraindividual noroviral evolution analysed by next generation sequencing, presented at the *RECESS Retreat 2013*, Venice, Italy, 2013

PRESENTATIONS:

- Haase K: Expression of human endogenous retroviruses in cancer tissue, presented at the *International Conference on Molecular and Evolutionary Oncology 2014*, St Petersburg, Russia, 2014 (invited speaker)
- Haase K: NGS technology opens new insights into virus evolution, presented at the *RECESS Retreat 2013*, Venice, Italy, 2013

PUBLICATIONS:

- Haase K, Mösch A, Frishman D: Differential expression analysis of human endogenous retroviruses based on ENCODE RNA-seq data. *BMC Medical Genomics*, 8(71), 2015
- Haase K, Raffegerst S, Schendel DJ, Frishman D: Expitope: a Web server for epitope expression. *Bioinformatics*, 31(11):1854–1856, 2015

ACKNOWLEDGEMENTS

I am grateful to many people who have guided or accompanied me during the development of this thesis.

First and foremost I would like to thank Prof. Dmitrij Frishman for giving me the opportunity to work in his group and for introducing me to academic research and teaching. I really enjoyed having the freedom to pursue my own ideas while always knowing where to find an open door for discussions.

I would also like to thank Prof. Christine Leib-Mösch for introducing me to the interesting topic of HERVs and sharing her expertise. I am grateful that I learned how to address problems from a different perspective than a purely bioinformatic one.

I would like to thank Prof. Michael Schemann for agreeing to become head of my thesis committee.

I wish to express my appreciation to everybody I had the honour to collaborate with during the course of my dissertation. Dr. Dieter Hoffmann, Prof. Dolores Schendel, Dr. Silke Raffegerst and Ingmar Göttesdorfer helped me to see and understand challenges in virology and immunology and gave me the opportunity to enjoy many fruitful cooperations.

I would like to thank all students I was given the opportunity to supervise and teach, especially Malena Bestehorn and Valentina Klaus, from whose questions I gained a lot.

I am grateful to everybody in our group in Weihenstephan to make the long commute worthwhile and the lunch breaks entertaining, but I am especially indebted to Stefanie Kaufmann and Florian Goebels for making the best out of the precious spare time. Wether we had nice dinner parties, went on exhausting day-long hikes or spent the evenings and nights with entertaining and slightly geeky board games, with you I could talk about work when discussions were needed or talk about everything else when work needed to rest for a few hours.

I wish to thank Jonathan Hoser who taught me a lot about programming and always had an immediate cure for all the hiccups of our system. Without his help I would not have completed this thesis in time.

I would also like to thank Claudia Luksch for helping with all the bureaucracy hurdles attached to a dissertation and Leonie Corry for being the heart of the group and always having an ear for everybody.

I am thankful for the research scholarship I was granted by the Universität Bayern e.V. and the Elite Network of Bavaria to finance my dissertation work.

My deepest gratitude goes to my parents and grandparents for always supporting me morally and financially and never loosing their patience, although most of my work is still double Dutch to them.

Finally, I would like to express my gratitude to Florian and Lilly who stoically endured all my moods, were always forgiving when I could not spend as much time with them as they deserved and still never failed to cheer me up.

Thanks to all of you!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

d_N/d_S	Ratio of nonsynonymous substitutions per nonsynonymous site d_N to synonymous substitutions per synonymous site d_S
A549	Epithelium cancer cell line
B-cells CD20+	Blood cell line
bp	Base pair(s)
Caltech	California Institute of Technology
CSHL	Cold Spring Harbor Laboratory
CTL	Cytotoxic T cell
DNA	Deoxyribonucleic acid
DRiPs	Defective ribosomal products
EM	Expectation Maximisation
ENCODE	Encyclopedia of DNA Elements
ER	Endoplasmic reticulum
ERV	Endogenous retrovirus
FDR	False discovery rate
FPKM	Fragments per kilobase per million mapped reads
GEO	Gene Expression Omnibus
GM12878	Blood cell line
H1-hESC	Embryonic stem cell line
HBGA	Histo-blood group antigen
HeLa-S3	Cervix cancer cell line
HepG2	Liver cancer cell line
HUVEC	Blood vessel cell line
ID	Identifier
IMR90	Lung cell line
K562	Blood cancer cell line
lncRNA	Long non-coding RNA
LTR	Long terminal repeat
MaLR	Mammalian-apparent long terminal repeat retrotransposon
Mb	Megabases
MCF-7	Breast cancer cell line
MHC	Major histocompatibility complex

miRNA	Micro RNA
Monocytes CD14+	Monocyte cell line
mRNA	Messenger ribonucleic acid
MS	Multiple sclerosis
MSA	Multiple Sequence Alignment
NHGRI	National Human Genome Research Institute
NPCDR1	Nasopharyngeal carcinoma down-regulated gene protein
NPSR1	Neuropeptide S receptor
nt	Nucleotide(s)
ORF	Open reading frame
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLC	Peptide loading complex
PYDC2	Pyrin domain-containing protein 2
QS	Quasispecies
RefSeq	Reference sequence database
RNA	Ribonucleic acid
RPKM	Reads per kilobase per million mapped reads
rRNA	Ribosomal RNA
SK-N-SH	Brain cancer cell line
SMRT	Single-Molecule Real-Time (Sequencing)
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
STAR	Spliced Transcript Alignment and Reconstruction
TAP	Transporter associated with antigen processing
TCR	T cell receptor
tRNA	Transfer ribonucleic acid
UCSC	University of California, Santa Cruz
VP1	Viral Protein 1

Part I

INTRAINDIVIDUAL NOROVIRUS EVOLUTION

In this part of the thesis we analysed the evolution of norovirus genomes. Noroviruses are the most prevalent non-bacterial cause of gastroenteritis and the reason for many costly outbreaks every year. By means of next generation sequencing and the acquisition of viral samples from chronically infected patients, we were able to trace the changes within the norovirus genome over time in the same host. This enabled us to perform many interesting analyses which are impossible or very challenging to do in acute infections.

This work is the result of a collaboration with the Institute of Virology at the Klinikum rechts der Isar, Technische Universität München.

Excerpts and figures from this part of the thesis have been published previously in the following forms:

> POSTER Haase K, Frishman D: Intraindividual noroviral evolution analysed by next generation sequencing, presented at the *RECESS Retreat 2013*, Venice, Italy, 2013
>
> PRESENTATION Haase K: NGS technology opens new insights into virus evolution, presented at the *RECESS Retreat 2013*, Venice, Italy, 2013

1 INTRODUCTION

1.1 NOROVIRUS

Norovirus is the most common cause of gastroenteritis worldwide [62] and is, among Rotavirus, the most often found to be causative of cases in children. Estimates place the number of infections with fatal consequences at over 200,000 a year in children under five from developing nations [136]. Deaths caused by norovirus are usually due to dehydration and occur either in very young or very old patients. Healthy individuals normally recover after 3 to 5 days of showing symptoms [171].

A norovirus infection in healthy individuals usually persists for less than a week

The strong medical relevance of noroviruses is based on the complete absence of effective anti-viral agents or of a satisfying vaccination strategy. In fact, not even a cleared previous infection confers lasting immunity, thus it is possible to acquire the virus multiple years in a row. The causes for the severe infectivity of this virus family are its high survival rate of viral particles outside the host and the low number of particles needed to established an infection. Studies showed that norovirus particles can survive at least 12 days on surfaces and that as few as 18 particles are sufficient to set up an infection [181]. The very high mutation rate, even compared to other ribonucleic acid (RNA) viruses [192], enables the virus genome to undergo so many changes within the patient or between outbreak waves that lasting immunity is nearly impossible to create.

All these combined features make it possible for noroviruses to cause large outbreaks every year around the globe, resulting in large efforts to contain and clear up the infection [88]. Due to the high infectivity, epidemic waves usually start in confined spaces such as care facilities, schools or cruise ships where the virus can quickly spread from person to person via the fecal-oral route [131]. As noroviruses usually spread most rapidly in the colder months of the year, it is often termed by the reporting media as "winter vomiting disease".

1.1.1 *Discovery and Structure*

The first isolate of Norovirus was obtained in 1972 by Kapikian et al. [91] from an outbreak in an elementary school in Norwalk, Ohio. The original name of the isolated particle was hence "Norwalk virus".

Noroviruses are also commonly known as "Norwalk-Virus"

Norovirus is a member of the caliciviridae family. The particle is non-enveloped and icosahedral with a diameter of ≈ 38 nanometres. The genome is organised in an approximately 7.5 kb long RNA fragment and belongs to group IV of the Baltimore system [7], meaning it is a single-stranded positive sense molecule that does not use a DNA intermediate when producing mRNA. The norovirus genome contains three open reading frames (ORFs) that encode all proteins the virus requires to replicate its genome and to assemble new particles. The first ORF ($\approx$5 kb) is translated into a polyprotein consisting of six non-structural proteins needed for genome replication. These are an RNA-dependent RNA-Polymerase, p48, nucleoside triphosphatase, p22, viral genome-linked protein and the protease to auto-cleave the polyprotein.

ORF2 is 1.8 kb in length and encodes the major structural capsid protein, viral protein 1 (VP1). VP1 is divided into two domains, the shell domain and the protruding domain, whereas the latter is further divided into two subdomains known as P1 and P2. ORF3 is $\approx$ 0.6 kb in length and encodes a minor basic structural protein, VP2, that is integrated in smaller numbers during capsid formation. [44]

The P2 domain of VP1 provides the receptor-binding region and is hypervariable

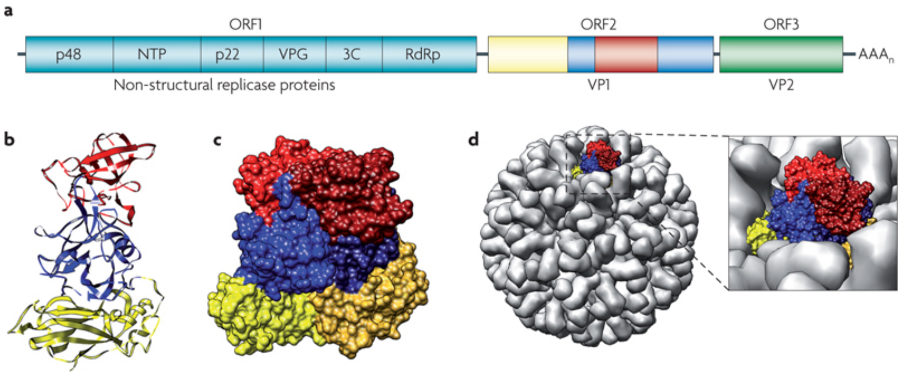
The virus particle is formed by 180 copies of VP1 which first dimerises and then forms a closed casing with icosahedral symmetry. As P2 is the part of VP1 extending the farthest away from the particle body, it provides the receptor-binding region and the sites of antigenic variation. Hence, VP1 is the target of the host's immune system and responsible for the survival of the virus which is the reason why it shows the highest mutation rate within the viral genome. Especially the P2 subdomain of the protruding region is known to be hypervariable [44]. For a depiction of the genome organisation please see figure 1.

1.1.2 *Classification*

The most prevalent outbreak causing cluster in humans is GII.4

Based on the sequences of the major capsid protein, noroviruses can be divided into genogroups and genotypes. At the moment there are five genogroups defined, labeled with the roman numbers I-V. Each group contains a different amount of known genotypes, distinguished with an arabic number. The most prevalent cluster that causes outbreaks in humans is GII.4. In order to categorise newly identified sequences, their distance to existing strains is assessed. Two sequences belonging to the same genotype should conform to each other in at least 80% of positions. [209]

So far, only members of the genogroups I, II and IV could be shown to infect humans [209]. Genogroup III contains bovine virus strains [132] and genogroup V the murine noroviruses which, until very recently, were the only subgroup that could be cultured in vitro [202].



Nature Reviews | Microbiology

Figure 1: Genome organisation and capsid structure.

a The norovirus genome is composed of three open reading frames. ORF1 (≈ 5 kb) is located in the first two-thirds of the genome and encodes a polyprotein that is auto-processed. ORF2 is 1.8 kb in length and encodes the 57 kDa major structural capsid protein, viral protein 1 (VP1). ORF3 is ≈ 0.6 kb in length and encodes a 22 kDa minor basic structural protein, VP2.

b The structure of the VP1 monomer is shown. Yellow: shell domain; blue: P1 domain; red: P2 domain.

c Two capsid protein monomers form the A-B dimer (indicated with the A monomer in lighter shades and the B monomer in darker shades), which allows the P2 domain to protrude from the viral particle.

d The virus-like particle is formed of 180 monomers of the capsid protein that assemble through different dimers.

Taken from Donaldson et al. [44].

1.1.3 *Tropism*

The receptors used by noroviruses to enter cells remained elusive for a long time. More recent studies revealed that histo-blood group antigens (HBGAs) expressed in epithelium cells of the small intestines are a very likely candidate [83]. However, there seemed to be yet unknown co-receptors, as no cell culture could be established from HBGAs alone.

In December of 2014 a publication by Jones et al. [89] finally introduced a possibility to culture noroviruses in vitro. They showed that the virus could in fact infect B cells, as long as HBGAs were present, either free or in form of HBGA-expressing bacteria. Additionally to a new infection model, this study also opens the possibility of impairing a successful norovirus infection by depleting intestinal microbiota instead of targeting the virus directly. [89]

The histo-blood group antigens can be grouped into ABO-, Lewis- and secretory antigens of which all three can be bound by noroviruses, although they exhibit certain preferences [177]. The high incidence of genotype GII.4 seems to be founded in its capability to bind all different HBGAs [79].

1.1.4 *Quasispecies*

When an organism becomes infected by a virus, it is usually not just one particle initiating the infection, but multiple of them. Thus, not one single genome starts to replicate within the host but many simultaneously. These founding sequences are very rarely identical, but differ due to the high mutation rate of viruses, especially the ones carrying an RNA genome. After multiplication within the host, the genomes start to accumulate even more changes, leading to a whole population of coexisting viral sequences, commonly referred to as quasispecies [16].

Quasispecies populations are extremely dynamic, as different variants can have varying fitness with regard to immune escape, leading to some species' depletion while other variants might flourish [41]. Certain events in a virus replication cycle pose bottleneck events for quasispecies evolution, such as the infection of a new organism. Only variants with a genome sequence that is adapted to the encountered immune system will be able to replicate, hence the population is strongly diminished and the consensus sequence over all genomes might change [42]. For a schematic example of a viral quasispecies population, see figure 2.

With traditional Sanger sequencing, usually only a small number of clones was created from a given viral sample. As most clones revealed different sequences, every single one was regarded as a quasispecies from the population pool. However, exact frequencies could not be determined in this manner and only the higher frequency variants were likely to be picked up this way. With the rise of next generation sequencing and ultra deep coverage, it has become much easier to get an insight into a nearly complete viral population, including low frequency species [144].

Next generation sequencing technologies enable us to reconstruct nearly complete viral populations of quasispecies

An important step when reconstructing the underlying quasispecies distribution from a set of sequencing reads is the differentiation between bona fide variant separating mutations and sequencing errors introduced by the applied methodology. This distinction is a crucial step to understanding quasispecies evolution dynamics and hence, multiple tools have been implemented in recent years to solve the problem for different sequencing technologies and viruses [207, 145, 6, 143].

1.2 SEQUENCING

At the time of its invention the costs of traditional Sanger sequencing were very high and the process took a long time and manual labour, even for short sequences. Additionally, the nucleotides needed for the technique had to be radioactively labelled demanding high care in handling and laboratory safety [162]. So improvements were made

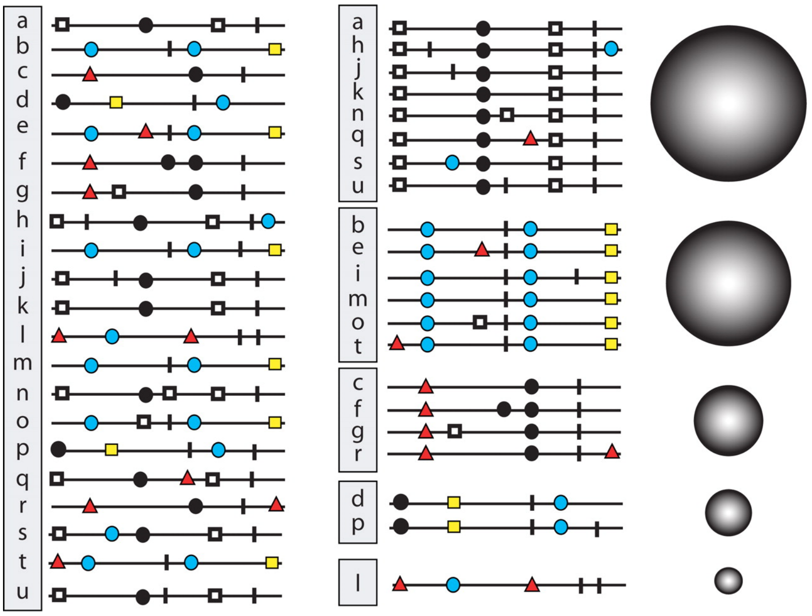


Figure 2: Schematic quasispecies distribution.
The left panel shows a schematic of a complete viral population, with symbols representing mutations.
The middel panel reveals the quasispecies distribution, after sorting all sequences by their shared, species defining, mutations. Some variants do have additional, individual mutations or sequencing errors.
The right side visualises the proportion each species takes up from the overall population.
Taken from Domingo et al. [42].

to conquer this drawbacks. This resulted in three major methods that were introduced by different companies in the beginning of the 21st century.

One of these technologies, advancing the sequencing of DNA, is the so called pyrosequencing. It was first described in 1996 by Ronaghi et al. [157] and marketed in 2003 by the company 454 Life Sciences which now belongs to Roche. Pyrosequencing does not use dideoxynucleotides to determine a sequence, but instead applies "sequencing by synthesis". It starts by fragmenting the sequence of interest into pieces of 300 to 800 base pair length. These are rendered single stranded and to both ends specific adaptors are ligated. With one of these adaptors the strand is bound to a bead, which is located inside an enclosed reaction space, called a "microreactor". This microreactor contains all reagents that are needed to amplify the captured template. After this process, an identical fragment of DNA is bound to the bead surface millions of times. Afterwards all sequence coated beads are loaded onto a "PicoTiterPlate" by placing one bead in each well. The plate is then exposed to a sequential flow of nucleotides, one base at a time. A mixture from sulfurylase and luciferase that is included in the wells emits a light signal whenever a nucleotide is incorporated in a well. With a special camera this emission can be recorded and thus the sequence for each well reconstructed [157, 159].

The great advantage of this method is its speed. It manages to sequence about one million reads per run, while taking approximately six hours. Thereby it is able to yield good results for reads with a mean length of 500 nt. A disadvantage of the method is its problem with homonucleotide repeats. If a certain base is incorporated more than twice or trice subsequently, the detected light emission can not be correlated to the correct count.

Due to its very long read length 454 is often chosen for de novo-sequencing, as it alleviates the assembly afterwards and its speed is also highly helpful for this appliance. In 2014, Roche ceased the support for pyrosequencing, as no further improvements could be achieved with regard to quality or cost reduction and the second next generation of sequencing technologies is slowly taking over [107].

1.2.1 *Amplicon Sequencing*

Most modern sequencing technologies follow the goal to reveal complete new genomes or to re-sequence them to find variants specific to individuals or conditions. For these aims an even distribution of reads among the genome is desirable in order to obtain a uniform whole genome coverage. The primers used in this scenario are usually random hexamers to achieve unbiased binding. Although there are still notable biases, in most cases a sufficient coverage along the complete sequence is reached [69].

In other cases, however, only a small excerpt of a genome is of interest, for example, when a single gene is in the focus of a study. For these studies, the specific region needs to be targeted effectively and can then be analysed very deeply with next generation sequencing. The only requirement for this approach is that the surrounding regions of the fragment of interest are known, so that two primers framing the amplicon region can be designed [33].

An amplicon is a small region of interest that can be amplified using a known forward and reverse primer

The accomplished high coverage over the region of interest makes it possible to even detect low frequency variants in a mixture of sequences, with a sensitivity that can not be reached by whole genome approaches. This makes amplicon sequencing especially valuable for viral analysis studies, as there are usually multiple quasispecies present in single sample which can be analysed in great detail with an amplicon sequencing approach. The only obstacle that has to be overcome, is to define primers that are strongly conserved, which can be hard to find in a viral genome, especially in the highly variable virus families. A solution is the choice of slightly degenerative primer sequences or very lenient PCR settings, although the amplification of off-target sequences should at the same time be kept to a minimum [45].

1.3 MOTIVATION

Noroviruses are a severe hardship for public health with multiple outbreaks every year. While fatal cases can mostly be prevented in the developed world, there are still reported deaths, mostly among very old, young or immunocompromised patients. However, there are still no effective anti-viral drugs or vaccination strategies available that could cure or prevent infection. Furthermore, not even the complete tropism of noroviruses could be revealed yet.

We wanted to contribute to the ongoing task of understanding more about the evolutionary dynamics of noroviruses and shed light on its interaction with the immune system. As a longitudinal study is the best way to trace mutational patterns in viral genomes over time, we needed to obtain temporally spaced samples which are very hard to get from a normally only week-long infection. Fortunately, the Klinikum rechts der Isar had collected multiple samples from chronically infected patients with varying interim times. These pose a great opportunity to not only reveal the selection pressure acting on the virus or the phylogeny between agents infecting different hosts, but also to apply NGS techniques to analyse viral sequences.

The ultra high coverage that can be achieved by applying next generation sequencing to a comparably short region of interest makes it possible to identify viral variants that make up only a small fraction of the overall virus population. This way, we can trace changes in the quasispecies dynamics and might gain insights into the fitnesses associated to different variants.

With this study we want to integrate approaches to norovirus analysis that have not been presented in this way before. By using data collected from chronically infected patients, we hope to obtain a more detailed evolutionary time line as opposed to using only major outbreak strains. At the same time, applying NGS instead of traditional Sanger sequencing will enable us to reconstruct the complete virus population within one sample and not restrict ourselves to the consensus sequence or few major variants.

2 METHODS

2.1 NOROVIRUS SAMPLES

2.1.1 *Cohort 2012*

In our first cohort from 2012, norovirus sequences were extracted from stool samples of three patients. All three individuals had previously received a bone marrow transplant and were treated with immune suppressants to prevent a rejection of said transplant. During the course of their treatment the patients also acquired a norovirus infection. Due to their suppressed immune system, a normally only week-long infection could persist in their organism for up to 15 month.

A region of interest in the norovirus genome was amplified with two primers, one forward and one reverse, which targeted an amplicon of 760 nt length. These region is located in the second noroviral ORF and contains the complete hypervariable P2 domain.

The patients provided four or five samples over the course of their infection with interim times between 54 and 206 days. After extraction and amplification of the region of interest, all fragment libraries were sequenced on a GS FLX+ pyrosequencer. To facilitate multiplexing, the sequences originating from one time point were marked by a four nucleotide long barcode for identification. Patient 1 has six associated barcodes although only five samples were collected from this subject. The stool from the last time point has been treated with two different extraction methods, so that our downstream analysis can identify, if these different approaches yield varying results.

Study subjects are bone marrow transplant recipients who acquired a norovirus infection while being treated with immune suppressants

All patients in the study have undergone HLA-genotyping, meaning their MHC alleles are known. Furthermore, the previous and current blood type of all subjects is recorded, as a conversion occurs due to the bone marrow transplant from the recipient's to the donor's type. All data available for the three study subjects can be found in table 2.

2.1.2 *Cohort 2013*

In a second study from the year 2013, we obtained sequences from five patients, although this time they provide less longitudinal data as only two, three or, in one case, four time points have been sampled. These study subjects are also bone marrow transplant recipients who have acquired a norovirus infection while undergoing treatment with immune suppressants.

Patient	Blood Group	MHC Alleles	Tag	Date
1	Donor: A Recipient: O	HLA-A 0101/0201 HLA-B 5101/1302 HLA-Cw 1502/0602 HLA-DRB1 0402/1303 HLA-DQB1 0302/0301	ACAC AGAG ATAT AGCT ATGA TCAG	08.01.2008 11.03.2008 05.05.2008 04.10.2008 05.02.2009* 05.02.2009*
2	Donor: B Recipient: A	HLA-A 2601/6801 HLA-B 5101/3501 HLA-Cw 0401/1502 HLA-DRB1 1601/1001 HLA-DQB1 0502/0501	CGCG CTCT CACA CTAG	15.01.2008 19.05.2008 11.12.2008 16.04.2009
3	Donor: A Recipient: O	HLA-A 0201/3301 HLA-B 07/3503 HLA-Cw 0401/0702 HLA-DRB1 1101/1501 HLA-DQB1 0301/0602	CATG TATA TCTC TGTG	20.05.2008 15.09.2008 04.12.2008 25.02.2009

Table 2: Norovirus samples from 2012.
The table shows all available information of the patients and their corresponding samples. The last column gives the date of stool sample collection, the second to last column the corresponding barcode used in the sequencing run.
For all patients the five HLA alleles A, B, Cw, DRB1 and DQB1 are known.
*: Same sample treated in two different ways.

Samples were obtained from the patients with interim times between 48 and 470 days and marked with the same 14 four nucleotide long barcodes as in the study from 2012. Unlike for the first cohort, we do not have any information on blood type or MHC genotypes for this group of patients. All data available for the five patients involved in the 2013 cohort can be found in table 3.

2.1.3 *Second Cohort 2013*

A third batch of reads, consisting of two or three samples each collected from five infected individuals, had been prepared and sent away for sequencing. Unfortunately, the sequencing run resulted in very few reads and the following mapping step could in some cases not even place 2% of the remaining ones onto a norovirus reference. Upon further inspection it was revealed that the library had been contaminated with sequences of prokaryotic origin and hardly any

Patient	Tag	Date
1	ACAC	08.04.2010
	AGAG	06.01.2011
	ATAT	25.05.2011
	AGCT	12.07.2011
2	ATGA	03.02.2010
	CGCG	26.05.2010
	CTCT	21.07.2010
3	CTAG	23.04.2010
	CACA	24.06.2010
4	TATA	09.04.2010
	CATG	23.07.2011
5	TCTC	24.06.2010
	TCAG	30.09.2010
	TGTG	18.08.2011

Table 3: Norovirus samples from 2013.
The table shows all available information of the patients and their corresponding samples. The last column gives the date of stool sample collection, the second to last column the corresponding barcode used in the sequencing run.

noroviral material was included. Thus, we omitted these data from downstream analysis.

2.2 SEQUENCING READS

2.2.1 *Quality Control*

All samples were prepared for sequencing at the Klinikum rechts der Isar and handed over to an external company that performed the pyrosequencing. As a result we obtained three files for every barcode, one containing the sequenced reads, another the corresponding qualities in phred encoding and the last file contained the flow information. We converted all flowgrams to fastq files with *sff2fastq* version 0.8.0.

We used an in-house processing pipeline developed by Jonathan Hoser [78] and extended for the purpose of our analysis to trim away the read ends until the mean quality in a window of 100 nt reached a quality of above 20. We furthermore discarded all reads that had a mean quality below 20, contained Ns or were shorter than 50 nt.

2.2.2 *Read mapping*

As reference norovirus strain for an initial mapping, we chose GenBank entry *AY502023.1*. It was submitted in 2003, contains the complete noroviral genome and belongs to genogroup II, which is the most prevalent in human infections and thus likely to be the causative agent in our study subjects.

Because RNA viruses have very high mutation rates, we expect many differences between our isolates and every already known sequence. Hence, we planned to do an initial mapping against an existing strain, then creating a consensus sequence from the resulting alignment, which can afterwards be used as reference for a second round of mapping. Therefore, the filtered reads were first mapped to *AY502023.1* with *Mosaik* [106] using the unique alignment method (reads with multiple possible hits are only assigned to one). We set the maximum allowed mismatch percentage to the very lenient value of 15% to be able to align reads over regions of high variability.

We use a two-step mapping approach to compensate for the extremely high variability of RNA viruses

To obtain the consensus sequences for every sample which would serve as references for the second mapping, we used an adjusted *samtools* [109] pipeline. At first, a pileup is created to list the read coverage of the reference sequence position wise. At this step it is important to set the read depth to a high enough value, as per default the depth used to analyse variants is rather small to save computing time. But because our coverage is very high, we want to use all aligned reads for the examination. Thus, the maximum read depth is dynamically set to the overall number of reads in the corresponding sample. As output the bcf format is chosen, a binary file that contains information about the variance of every alignment position. The samtools functionality *bcftools* is applied to call SNPs from the initial alignment. We used a likelihood based SNP calling which suppresses genotype information, as it per default assumes a diploid organism. We used a flat bayesian prior for this step, as we do not have any initial information on the mutation distribution. In a last step we used the created vcf file to get a consensus fastq output with the tool *vcf2fq* which also belongs to the samtools framework. We modified its function, so that in case of two possible nucleotides at a position not the ambiguity code is used, but the major variant.

The above described mapping steps are then repeated with the consensus fasta for each sample instead of *AY502023.1*.

2.3 PHYLOGENY

Every consensus sequence was translated into an amino acid sequence as well. From both sets of sequences, the nucleotide and amino acid version, we constructed multiple sequence alignments (MSAs) with ClustalW2 (version 2.1) and calculated phylogenetic trees to reconstruct the sequence inherent relationships between the samples. For this purpose we used the phylogeny package *PHYLIP* [52] (version 3.69), specifically its tool *seqboot*, to create 100 bootstrap samples from our initial MSAs and then used the programs *DnaML* and *ProML* to create maximum likelihood trees for all replicates. In a last step the tool *consense* merged all created trees into a consensus bootstrap tree. The resulting trees were visualised via *Newick utilities* [90].

2.4 MUTATION ANALYSIS

2.4.1 *Evolutionary distance*

For most virus families the common mutation rates are known and for noroviruses it has been estimated to lie between 1.21×10^{-2} to 1.41×10^{-2} changes per site per year which is a high amount even for an RNA virus [192]. We compared these values to the ones observed in our datasets by counting the substitutions occurring between two consecutive time points and normalising by sequence length and interim time.

2.4.2 *Trajectories*

From the previously prepared MSAs on amino acid level, we created minimal versions, that only contained those positions per patient which changed at least once over the course of our study. We implemented a colour encoding of these mutation trajectories based on how often a certain position mutates or reverts back to the initial variant observed at a previous analysed time point.

A publication of Bok et al. [21] examined evolutionary dynamics of noroviruses, especially GII.4. They list multiple sites in the genome that either show great variability or conservation and name positions that are responsible for receptor interaction or are connected to epidemic waves. We compared all these given sites against our set of positions which showed variability in our samples.

We compare the intraindividual mutations against a list of sites associated with important interactions and outbreak waves

2.4.3 *Selection pressure*

We analysed all reconstructed consensus sequence with regard to occurrence of selection pressure. For this purpose we used the tool *paml* [203](version 4.4b). As input for its selection analysis the program needs a codon alignment which can be created by the included perl script *pal2nal*. The needed input files for the construction of a codon alignment are the already created MSA of amino acid sequences and all corresponding nucleotide sequences in fasta format. The program *codeml* was used to calculate d_N/d_S ratios between all samples in our data set from the codon alignment and the maximum likelihood tree of amino acid sequences. The number of non-synonymous substitutions per non-synonymous site compared against the number of synonymous substitutions per synonymous sites is supposed to be an indicator of the overall selective direction of a sequence. Is the resulting value large, i.e. greater than one, the sequence is assumed to be under positive selection. A result less than one implies purifying selection and a result close to one is a pointer at neutral selection. [98]

A d_N/d_S ratio > 1: positive selection; < 1: purifying selection; ≈ 1: neutral selection

Especially for our viral sequences an overall d_N/d_S calculation is not a necessarily a sufficient measurement for positive selection as only certain regions are usually subject to selection due to their interaction with the host's immune system. Thus, we additionally implemented a more local analysis. It uses the same tools as mentioned above but applies two selection models to the complete set of input sequences, one checks the assumption of neutrality, the other for positive selection. Instead of taking into account every nucleotide or amino acid position, it calculates the rates for every codon. With this approach the tool can statistically identify the form of evolution corresponding to every codon.

2.5 QUASISPECIES RECONSTRUCTION

A great challenge when analysing deep sequencing data is the reconstruction of viral quasispecies from the sets of reads. Due to the high mutation rate and improper error correction of the RNA-dependant polymerase used in norovirus replication, a patient is usually not infected by agents with identical sequences, but rather with a population of viruses that differ from each other.

Obviously these population diversity is ignored when constructing the overall representative consensus. In traditional sequencing approaches the identification of different quasispecies was hardly possible and limited to the types with the highest prevalences. Each individual clone was usually regarded as one species and thus only between ten and twenty of them were obtained in one study. With the rise of the next generation technologies and their inherent high coverage of amplicon regions, it becomes possible to find even low frequency quasispecies.

Next generation sequencing enables the identification of low frequency quasispecies

2.5.1 *Existing tools*

To assess the reconstruction problem, multiple tools have been published since 2011. They have been compared and the problem of quasispecies reconstruction has been described by Beerenwinkel and Zagordi [11]. The publication lists six different approaches for the reconstruction, four of which have already been implemented. Of this subset, three are able to perform a global assembly over the full-length sequence, so we tried ShoRAH [207], ViSpA [6] and PredictHaplo [143]. One of the methods that had not been implemented at the time of the review has been published as usable tool in the meantime, thus we added QuRe [145, 146] to the list.

All these tools use different approaches to assemble and align the sequencing reads as well as correct sequencing errors and estimate relative frequencies of each species.

2.5.1.1 *ShoRAH*

The first tool we applied to our data to reconstruct quasispecies was *ShoRAH*. Although it was only first published in 2011, it is often referred to as "state-of-the-art" by other publications addressing the same issue.

ShoRAH allows the user to reconstruct quasispecies within a certain window that should have the length of the average read length or to do a global analysis which mainly concatenates multiple local analysis. For the global analysis which we wanted to achieve, the software provides a wrapper class that starts with aligning the sequencing reads to a given reference and corrects all of the included sequencing errors with a clustering process. While using a parsimony principle, the program then constructs a minimal set of quasispecies that is able to explain all reads. Afterwards the frequency of these reconstructions is estimated with an Expectation-Maximisation-Algorithm (EM).

2.5.1.2 *PredictHaplo*

The next tool we applied to our data was *PredictHaplo*. It also implements an approach to explain all sequencing reads with a minimal set of haplotypes. It does not correct the input reads beforehand but instead extracts prior probabilities for local windows on which the following global assembly is based. The authors state that reads that were produced by PCR lead to a high rate of false positives that are reconstructed by PredictHaplo.

2.5.1.3 *ViSpA*

The last reviewed tool for the reconstruction of quasispecies is *ViSpA*. It is a java implementation that relies on the alignment tool *segemehl* for the mapping step. The authors state in their publication that *ViSpA* is able to outperform *ShoRAH* when applied to already corrected reads. The tool iteratively updates the initial used reference sequence with aligned reads to get an as correct as possible alignment. Then, it constructs a read graph from the alignment which is afterwards reduced and the frequencies are estimated with an EM-algorithm.

2.5.2 *In-house implementation*

After the unsuccessful utilisation of the published tools, we decided to reconstruct quasispecies ourselves. Therefore we used the read pileup we had constructed to find the consensus sequences and extracted all variable positions. These are genome regions at which at least two different nucleotides with a frequency above a certain threshold are observed. We used a threshold of 1.5% for further analysis. Additionally, we only took positions into account that were covered by at least a given number of reads. This cutoff was set to ten. It should be noted that the variable positions are not based on all mapped reads but, due to the construction of the pileup, only on the ones that have a certain quality.

Afterwards our program executes *samtools* to list every read sorted by its start alignment position. Because our reads were created by an amplicon sequencing with a forward and a reverse primer, all aligned reads either start at the amplicon beginning or stop at the amplicon end. For both classes of reads we created a tree in which we stored the different kinds of reads which were beforehand corrected with help of their cigar string.

The cigar string is a concatenation of numbers followed by a character as for example `12M5I3M1D2M`. This string describes a 22 nucleotide long read with respect to its alignment position. It states that the first twelve positions of the read match the reference. The `M` does not only describe exact matches but rather serves as a distinction from insertions and deletions, hence it also can contain substitutions. Then the example read contains five nucleotides which have no counterpart in the reference and are thus insertions. After three following matches the read lacks one base which is present in the reference and hence noted as deletion. In our pipeline deletions are corrected with Ns in the reads because we know that the reference is completely coding and viable, thus we can deduce that frameshifts are highly unlikely.

The created trees are rooted at the amplicon start and end and every previously determined variable position is a potential node. The first read added to our tree creates a path from the root to a leaf (represented by $) which denotes the end of a read. The following reads create branches in the existing paths if they differ at one of the variable positions. In this way we create two trees in which every $-leaf represents one quasispecies. These trees are not balanced due to the very varying length of the input reads. Furthermore is it not possible to combine forward and reverse tree on their leafs as their overlap is far to small. Because we only allow branching at the previously defined high quality variable positions, we do not include sequencing errors as species separating events.

In our tree construction, every branch from root to leaf represents one distinct quasispecies

Each reconstructed species is assigned a frequency which shows how many of the reads that are assigned to the corresponding root reach the quasispecies-representing leaf.

2.5.3 *QuRe*

The last published tool we utilised was *QuRe*. The underlying statistical methods had been published previously by Prosperi et al. [146] before the group implemented the tool. It also expands local analyses to an overall reconstruction with help of a heuristic algorithm. It constructs multinomial distributions for defined windows and then tries to combine adjacent ones. The error correction method of the program is not specific to 454 reads but instead based on a Poisson distribution.

We executed *QuRe* on all our read files and used the sequences of GenBank entry *AY502023.1* as a reference. As the tool is optimised on read length of over 100 bp, it has default error probabilities for different sequencing technologies producing long reads. Thus, we used the values for 454 which are specialised for correcting errors in homopolymer stretches.

To visualise the results, we took the most prevalent twenty quasispecies of every time point and created a phylogenetic tree per patient from these.

3 RESULTS

3.1 Reads have a sufficient length and quality

After we filtered all sequencing reads by length, mean quality and unresolved sequence positions as well as trimmed away low quality read ends, we were left with a mean length per read of around 390 nt for the 2012 run and 440 nt for 2013. The mean quality is approximately 33 for the earlier cohort and 35 for the later. This improvement in length and quality at the same time is probably caused by a change in reagents used during the sequencing procedure. These are often updated to facilitate the production of higher read length. Although this is very helpful for our downstream analysis, it has not been communicated by the company conducting the sequencing which exact changes lead to the quality improvement.

The distribution of the reads among the barcodes is in both cohorts very uneven. In 2012 overall 96,352 reads were left after filtering, of which more than 15% belong to barcode ATGA, while only 1.9% originate from the sample tagged TCAG. In 2013 overall only 77,529 reads survived the filtering process, the most (20%) belonging to barcode TCTC, the least (1.3%) originate from the CATG sample. For an overview of the read statistics post filtering see tables 4 and 5.

Reads are distributed very uneven among barcodes

In both cohorts most reads were lost in the filtering process due to containing Ns. A mean quality below a phred score of 20 was only observed in 17 reads from the 2012 run and in three reads from 2013. As can be expected when looking at the average length of the two cohorts, we lose much more reads in the filtering step which disposes of reads shorter than 50 nt in the 2012 than in the 2013 cohort. The statistics of the trimming and filtering process can be seen in table A1.

The remaining reads have a more than sufficient quantity, quality and length to perform all of our intended downstream analyses.

3.2 Two-step mapping approach can position over 92% of reads

In the first mapping round of all filtered reads against reference strain *AY502023.1* only about 80% of the reads from the 2012 run and 40% of the reads from 2013 could be mapped. This result shows what a big difference the slightly increased read length in the second cohort can make, as the larger amount of sequence information makes it much harder to place fragments onto the reference genome below the

			Read length (nt)					Read quality (Phred score)				
Patient	Sample	# Reads	mean	median	min	max	σ	mean	median	min	max	σ
1	ACAC	4051.0	388.01	420.00	53.00	453.00	85.77	33.83	34.24	22.99	39.84	2.57
	AGAG	6319.0	395.68	423.00	50.00	458.00	87.38	33.49	33.79	22.94	39.89	2.67
	ATAT	8208.0	384.75	421.00	50.00	463.00	98.26	33.66	33.99	21.63	39.83	2.67
	AGCT	8580.0	387.34	424.00	50.00	482.00	99.62	33.29	33.57	22.15	39.94	2.94
	ATGA	14895.0	421.37	425.00	52.00	466.00	23.80	32.95	33.34	22.52	39.85	2.31
	TCAG	1842.0	423.02	426.00	51.00	446.00	25.13	32.93	33.29	22.47	39.82	2.31
2	CGCG	4752.0	371.38	423.00	50.00	478.00	112.25	33.71	33.85	24.41	39.88	2.85
	CTCT	10893.0	374.75	420.00	50.00	459.00	107.43	33.62	33.89	21.09	39.86	2.86
	CACA	8881.0	392.02	421.00	50.00	526.00	88.48	33.55	33.92	21.93	39.92	2.85
	CTAG	4047.0	401.64	416.00	55.00	455.00	64.12	33.68	34.09	23.48	39.89	2.61
3	CATG	7803.0	389.96	412.00	50.00	469.00	77.78	33.79	34.21	23.94	39.91	2.57
	TATA	5608.0	394.29	418.00	50.00	455.00	74.28	33.72	34.03	23.17	39.85	2.70
	TCTC	5834.0	377.00	415.00	50.00	454.00	96.88	33.93	34.28	22.13	39.85	2.62
	TGTG	4639.0	386.57	413.00	50.00	470.00	84.11	34.08	34.50	24.21	39.81	2.50

Table 4: Statistics of 2012 sequencing reads after filtering.

			Read length (nt)					Read quality (Phred score)				
Patient	Sample	# Reads	mean	median	min	max	σ	mean	median	min	max	σ
1	ACAC	6377.0	444.78	444.00	50.00	495.00	24.33	35.78	36.20	21.74	38.99	1.86
	AGAG	4956.0	433.13	435.00	50.00	477.00	30.03	36.16	36.70	22.38	39.11	1.94
	ATAT	5234.0	432.82	435.00	50.00	512.00	30.95	36.31	36.82	23.74	39.23	1.90
	AGCT	4719.0	430.97	433.00	51.00	474.00	32.99	36.39	36.91	23.79	39.49	1.86
2	ATGA	6891.0	429.94	431.00	52.00	500.00	31.91	36.07	36.48	24.27	38.86	1.84
	CGCG	1509.0	440.03	442.00	77.00	486.00	24.38	35.78	36.21	27.09	38.88	1.99
	CTCT	6570.0	429.45	429.00	52.00	534.00	30.71	36.21	36.71	23.32	39.17	1.96
3	CTAG	7897.0	444.49	445.00	50.00	516.00	26.12	35.88	36.35	22.86	39.02	1.89
	CACA	3715.0	443.60	444.00	53.00	476.00	22.46	35.90	36.31	24.32	39.02	1.79
4	TATA	1418.0	441.19	441.00	79.00	496.00	19.04	36.09	36.57	26.92	38.68	1.83
	CATG	1007.0	436.85	437.00	75.00	467.00	19.27	35.77	36.23	28.13	38.72	1.91
5	TCTC	15561.0	438.82	440.00	50.00	494.00	32.38	35.70	36.10	23.33	39.96	1.87
	TCAG	8038.0	437.70	439.00	52.00	527.00	37.30	35.43	35.88	23.29	39.02	2.13
	TGTG	3637.0	441.17	448.00	51.00	502.00	50.70	35.40	35.86	22.24	39.96	2.10

Table 5: Statistics of 2013 sequencing reads after filtering.

given mismatch threshold. The location among the norovirus genome was exactly as intended, located between position 5727 and 6486 of sequence *AY502023.1* which contains the complete hypervariable P2 domain. We plotted the coverages for all first-round mappings per sample and confirmed not only the covered positions but could also observe the typical pattern of coverage caused by the amplicon sequencing. The forward and reverse reads start at the above mentioned positions and overlap only for 100 - 150 bp in the middle of the region of interest. For an exemplary coverage plot, see figure A1. The forward and reverse primers produced a varying fraction of the overall reads in a sample, usually with a slight preference for the forward one.

Read distribution among the two amplicon primer is biased towards the forward primer

After we reconstructed a preliminary consensus sequence from the first mapping, we used the same mapping method to map all reads per sample against the individual references. In this round we were able to place more than 92% of reads in every sample and thus have a large amount of positional sequence data for further mutation analysis. For all mapping statistics refer to table 6.

Patient	Tag	% Aligned Reads	
		initial	cons.
1	ACAC	84.15	92.15
	AGAG	87.31	95.62
	ATAT	84.25	94.42
	AGCT	78.11	94.88
	ATGA	96.08	99.77
	TCAG	96.09	99.78
2	CGCG	88.07	96.89
	CTCT	83.81	94.69
	CACA	90.79	96.30
	CTAG	93.35	97.95
3	CATG	74.70	96.77
	TATA	74.05	97.40
	TCTC	52.86	93.23
	TGTG	44.60	96.90

(a) Run from 2012

Patient	Tag	% Aligned Reads	
		initial	cons.
1	ACAC	47.33	99.59
	AGAG	39.06	99.17
	ATAT	42.87	99.29
	AGCT	34.75	99.07
2	ATGA	30.36	99.23
	CGCG	60.97	99.54
	CTCT	54.09	98.83
3	CTAG	51.16	99.66
	CACA	58.47	99.70
4	TATA	47.74	99.58
	CATG	40.71	98.71
5	TCTC	50.65	99.24
	TCAG	41.29	98.91
	TGTG	45.45	98.02

(b) Run from 2013

Table 6: Percentage of mappable reads.
The alignment mode used was *unique* and the maximum mismatch percentage per read was set to 15%. Values for both, the initial mapping against GenBank entry *AY502023.1* and the second mapping against the preliminary consensus, are shown in the last two columns.

3.3 SEQUENCE RELATIONSHIP IS ONLY DISTINCT ON NUCLEOTIDE LEVEL

When constructing maximum likelihood trees from the MSAs of all samples' consensus sequences, we obtained very high bootstrap values for nearly all branches of the nucleotide trees, representing a very stable topology. Especially some pairwise relationships between consecutive samples reach extremely high values.

When examining the amino acid sequence based trees of the same samples, the bootstrap values are considerably lower. Furthermore, we can still find pairs of sequences which are clustered together similar to the nucleotide tree but larger subtrees containing samples with continuous numbers do not occur, except for patient 1 from 2012 and patient 5 from 2013 (compare figure 3).

This phylogenetic make-up shows that although the nucleotide sequences are pretty diverse when compared between patients, the sequence space of the translated amino acids is actually limited and thus similar among samples from different patients.

3.4 OBSERVED MUTATION RATE IS HIGHER THAN PREVIOUSLY REPORTED

We calculated the evolutionary distances between two consecutive samples by counting all occurring substitutions and normalising by alignment length and interim time between sample collections. The obtained value then represents the rate of substitutions per site and day and is thus very small. To increase the number and at the same time make it comparable to the known noroviral mutation rate, we multiplied it by 365 to calculate the evolutionary rate per year.

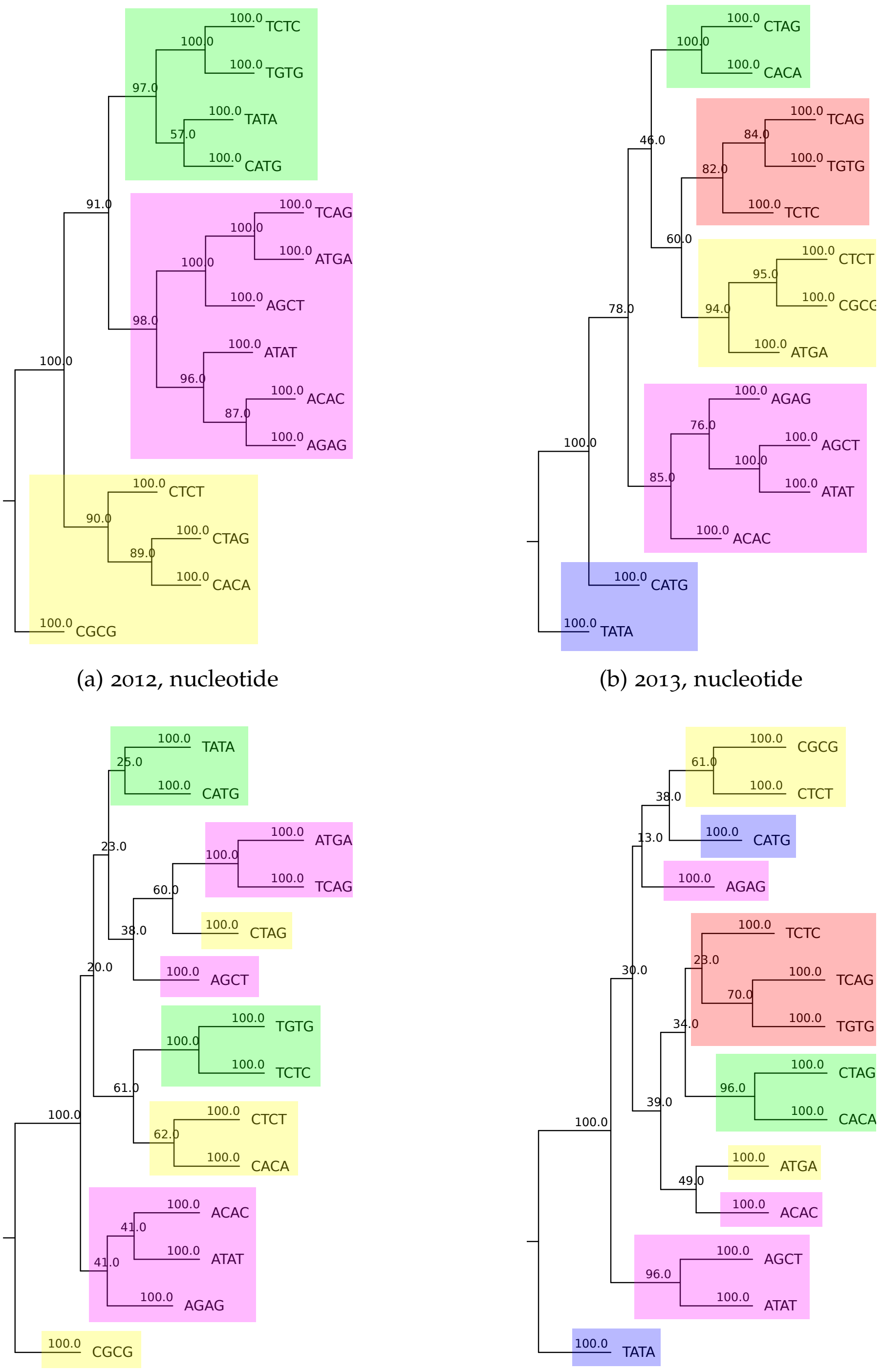
It can be observed that in nearly all cases the distance on protein level is higher than on nucleotide level which could be a first sign for existing positive selection. Furthermore, many of the transitions have mutation rates on nucleotide level of 3×10^{-2} which are much higher than the previously reported rates between 1.21×10^{-2} and 1.41×10^{-2} [192]. It is possible that the discrepancy results from the fact that the mentioned publication derived the mutation rates from multiple sequences in different outbreaks, whereas we are looking at intraindividual dynamics.

We can observe mutation rates of over 10^{-1} per site per year

Only in two cases do the consensus sequences of two consecutive time points show no mutation at all, underlining the strong variation of noroviruses. For all mutation rates compare table 7.

3.5 MUTATION TRACING SHOWS SPECIFIC HOT SPOTS

We visualised all occurring mutations in the consensus sequences of the different time points by minimising the alignments to only those



(a) 2012, nucleotide

(b) 2013, nucleotide

(c) 2012, protein

(d) 2013, protein

Figure 3: Phylogenetic trees constructed from consensus MSAs.
100 bootstrap replicates were calculated to define the values at branch points.
Colors are assigned to patient numbers as follows: 1 - magenta, 2 - yellow, 3 - green, 4 - blue, 5 - red.

Patient	Interval	Distance on level	
	[days]	Nucleotide	Protein
1	63	0.0153	0.0458
	54	0.0591	0.1525
	152	0.0690	0.0380
	124	0.0464	0.1047 *
	124	0.0464	0.1047 *
2	124	0.0340	0.0568
	206	0.0284	0.0489
	125	0.0805	0.1731
3	118	0.0122	0.0367
	80	0.1019	0.1803
	83	0.0404	0.0869

(a) Run from 2012

Patient	Interval	Distance on level	
	[days]	Nucleotide	Protein
1	273	0.0123	0.0264
	138	0.0487	0.0627
	48	0	0
2	111	0.0259	0.0650
	56	0.0086	0.0258
3	62	0	0
4	470	0.0276	0.0522
5	98	0.0597	0.1379
	322	0.0400	0.0761

(b) Run from 2013

Table 7: Evolutionary distances along the time scale.
The values in the cells show the number of base substitutions per site per year (365 days) in the corresponding sequence pairs.
*: Distances were calculated to the sequence of sample AGCT, not to each other.

positions which change at least once during the sample collections from each patient. We regarded the sequence from the first available time point as a founder sequence and indicated mutations away from it by a darker colour. In the case of a back mutation, the colours change back to the starting hue. As a reference for the residue numbers, we chose the associated capsid protein sequence of Genbank entry *AY502023.1*, *AAR97663.1*.

For patient 3 of the 2013 cohort no trajectory was created, as no mutations between the two consensus sequences were observed. The only other step at which no changes occurred was the last in patient 1 of the 2013 group. The last two samples in patient 1 of 2012 are identical, showing that the change in the sequence extraction protocol did not have any influence on the consensus sequence.

Certain positions exhibit variation in multiple, even all patients of the 2012 cohort, such as 297 and 393. Furthermore can clear hot spots be identified, meaning regions that have around four consecutive position that all show mutations throughout the time course. The mentioned positions are both located within two of those mutational centres.

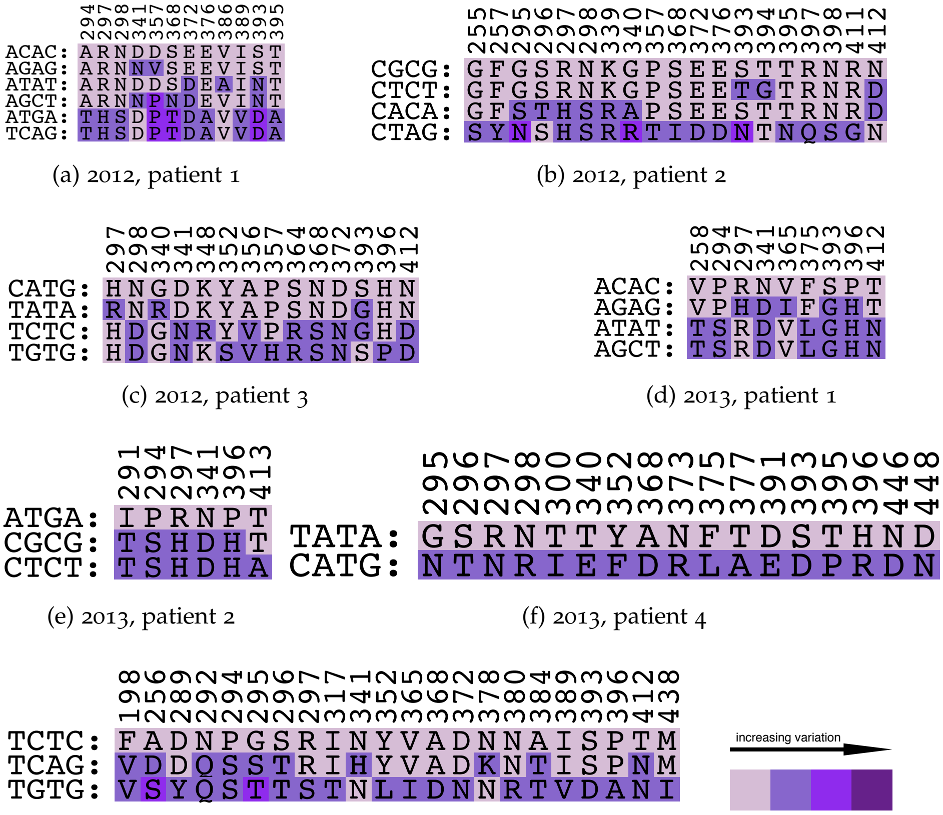
The overall number of residues with variation is very diverse among the different patients. While only six positions show changes over three time points in patient 2 of 2013, patient 5 of the same cohort has 22 mutating amino acids. The most different residues at one position we observed in our study are three. However, we can very often see back mutations to a previous variant, or even "flip-flop" mutations, where one position switches constantly between two amino acids. For these cases it is important to remember that we are working on consensus sequences, meaning instead of a constant mutation we could in fact be looking at a change in the underlying quasispecies dynamics. As we use a majority rule for consensus sequence construction, only a small change in frequency of a specific species could change the overall fraction of a variant.

"Flip-flop" mutations could be caused by a change in the underlying quasispecies composition

For all created mutation trajectories see figure 4.

Many of the residues shown by our trajectories to be variable over the course of the study have also been mentioned in the publication by Bok et al. [21] naming important capsid protein positions. The first important site mentioned in the paper is 292-295 which showed the conserved amino acids RVGI in sequences obtained before 1970. Later it was replaced with the four residues HIVG and additional changes at position 294. This is not consistent with our sequences, as although positions 292 and 293 are conserved (except for patient 5 from 2013), both 294 and 295 are variable in our samples.

When Bok et al. [21] searched for residues with positive selection, they identified one site that is included in our samples, 395. The positive selection of this residue has been identified before by another group [113]. In our samples this site belongs to the variable positions



(a) 2012, patient 1

(b) 2012, patient 2

(c) 2012, patient 3

(d) 2013, patient 1

(e) 2013, patient 2

(f) 2013, patient 4

(g) 2013, patient 5

(h) legend

Figure 4: Mutation trajectories for all patients.
The alignments only contain the variable positions in the consensus sequences for each patient and are sorted by the chronological progression. The background colour of a residue gets darker the more variation occurred at the corresponding position.
Residue numbering was taken from the associated capsid protein sequence *AAR97663.1* of norovirus strain *AY502023.1*.

of three patients but does not show up as significant in our analysis of selection pressure.

The paper reports which part of the P2 region potentially interacts with HGBA trisaccharides to mediate viral entry that is conserved since the 1970s. In agreement with this finding, our samples do not show any variance at the corresponding positions. But the authors also describe a site that stabilises the interaction which is located at positions 390-393 and 395. Interestingly, Bok et al. [21] claim that these region also has been stable over time, except for positions 393-395. Our sequences underline this statement, as although positions 393-395 are all variable in our samples, positions 390-392 are conserved in all patients but one (patient 4 from 2013).

Our sequences show mutations in a region associated with mediating viral entry

Tan et al. [179] list additional residues that are sterically close to the interaction site and hence can influence the binding to HBGA. One finding is that changes in the positions 331, 346, 348 and 389 alter the binding pattern [180]. At position 389 either isoleucine or valine could be found, depending on the outbreak cluster the corresponding strain belonged to. Contrary to this statement, we can see the change between isoleucine to valine at this position over the course of our study within two patients.

The last mentioned sites in the report are two *hot spots* identified by another group which impact the biochemical properties of the P2 domain and apparently change between epidemic waves [2]. These regions are positions 296-298 and 393-395. At least one of these sites and surrounding regions changes between samples in every patient, underlining the importance of the two hot spots with regards to virus evolution. But our finding also contradicts that the residues change only between epidemic waves. Furthermore, the *hot spots* both seem to be slightly longer than reported, as in our samples we also see variation in sites in the direct vicinity.

Two hot spots responsible for the biochemical properties of the P2 domain are highly variable in our samples

A very recent publication by Zakikhany et al. [208] also found positions 296-298 and 393-395 to be highly variable and traced their changes within next generation sequencing data obtained in London in the last ten years and samples deposited in GenBank. In their overall 1,312 P2 domain sequences they found 82 different strains, when only considering differences in the six positions of the hot spots. The authors also state that these patterns change over time and one major occurring motif is replaced between outbreaks. Years that showed an overall high activity of norovirus were correlated to switches in the surface type motif of the first hot spot or an increasing variability in the second site.

3.6 EVERY PATIENT SHOWS POSITIVE SELECTION

At least at one time point transition per patient, the determined d_N/d_S ratio infers positive selection by being larger than one. In patients 2, 4 and 5 from the 2013 cohort, this is even true for all analysed steps. In three cases, exclusively non-synonymous mutations were observed, with no occurring synonymous mutation.

In the patients in which also purifying selection can be seen, the mutation pattern are different. In patients 1 and 3 from 2012, the d_N/d_S ratio starts high, decreases at the middle time points, only to increase again at the last transition. In patient 2 of 2012 however, the rate gradually increases while it decreases over time in patient 1 of 2013.

The different stages at which the d_N/d_S ratios inferring positive selection can be observed in the patients could reflect, when the virus has to undergo the most changes in order to escape the immune response.

For all calculated d_N/d_S ratios compare table 8.

Patient	Interval [days]	d_N/d_S
1	63	> 1*
	54	2.0857
	152	0.3459
	124	1.4359
	124	1.4359
2	124	0.8173
	206	0.8867
	125	1.2740
3	118	> 1*
	80	0.6980
	83	0.9502

(a) Run from 2012

Patient	Interval [days]	d_N/d_S
1	273	2.3689
	138	0.4845
	48	†
2	111	2.3798
	56	> 1*
3	62	†
4	470	2.4611
5	98	1.7228
	322	1.9455

(b) Run from April 2013

Table 8: Pairwise d_N/d_S ratios.
The pairwise ratio between non-synonymous mutations per non-synonymous site and synonymous mutations per synonymous site is shown for sequence pairs that were drawn consecutively.
*: At these transitions all observed mutations were non-synonymous and thus would call for a division by zero. As this is not possible, we can only denote that the value would infer positive selection.
†: Here, no mutations were observed at all, thus no selection analysis could be performed.

When analysing the consensus sequences codon wise to identify the direct sites of acting selection, only three residues in patient 1 from 2012 reached a significant p-Value of > 0.95. These three sites are 341, 357 and 393. That only results for this one patient could be obtained is mostly due to a lack of data. The models can only be applied to a set containing at least three sequences and the calculation becomes more statistically powerful, the more data is provided. *Codeml's* manual even states that its minimum number of sequences for meaningful results is four to five.

3.7 Our experimental design is incompatible with most reconstruction algorithms

When we tried to apply the various published reconstruction algorithms for viral quasispecies, we encountered multiple errors. The tools *ShoRAH, PredictHaplo* and *Vispa* are all incompatible with the amplicon sequencing design. Ideally, these tools expect the input reads to originate from a sequencing approach in which the reads are randomly distributed among the viral genome or region of interest. This method would produce a more or less equal coverage over all bases. However, as discussed in section 3.2, due to our amplicon approach, we have a high coverage in primer regions, nearly a doubled amount in the overlap region and lower values in between. This setup causes *ShoRAH* to reconstruct two different sets of species for both halves of the amplicon, *PredictHaplo* to enter an infinity loop in regions of decreased coverage and *ViSpA* to simply exit with an error message. After correspondence with the authors, the group behind *ShoRAH* has subsequently published an amplicon specific method, but unfortunately it requires the complete reconstructed region to be covered by reads of both orientations. Thus, in our case it only creates quasispecies in the approximately 150 bp long overlap region in the middle of the amplicon.

Established tools expect a sequencing design with randomly distributed reads

The same outcome applies to our selfmade implementation. While the construction of the trees rooted at both primers works well, a joining of the leaves is not feasible as to many possibilities exist. We tried to collapse the branches by elongating the very short ones. Now, if such an early $-leaf has a non-$ sibling, we merge these two branches, resulting in only one path. The frequency belonging to the $-leaf will be added to its sibling. If the short branch has two non-$ siblings, its frequency will be divided between the two of them in the same ratio the two already have to each other. Thus, the pre-existing ratio is preserved.

The problem of combining the forward and reverse tree unfortunately still remains. The two possibilities we are left with would be to either create all possible combinations of leaf overlaps or to apply a maximum parsimony approach and construct a complete tree with a minimum number of branches explaining all reads. However, the former approach would very likely create a large number of false positives that were not present in our initial sample, while the latter could disregard true positive species.

Many factors make the quasispecies reconstruction a vague undertaking. Obviously, if not all or at least most reads cover the complete region of interest, one has always to use statistical measure to assign reads which only overlap in short regions to different or the same species. If they differ in their overlapping part, it is clear that they stem from distinct variants, but if they are equal in the overlap part and no information is available for the regions that are unique to one of the reads, the assignment is an educated guess at best.

Quasispecies reconstruction relies heavily on good error correction and solid statistics to combine local, read-sized analyses

Even in the regions that are observed to be different the problem of differentiating between real differences and sequencing errors arises. Although 454 mostly shows problems with inserations and deletions in polynucleotide repeats, substitution errors can never be ruled out completely. Thus, one needs a good error correction method that can distinguish between these cases.

While the above mentioned problems are at least addressable in the resulting sequencing data, there are additional problems which can not be observed. It is known that reverse transcription and PCR can introduce errors in the sequences as well, two methods which are both used to obtain the amplicon region. Furthermore can PCR create so called chimera by combining two different strains into one sequence. These in vitro created mixtures can afterwards not be told apart from in vivo cross overs.

The PCR process can also distort the frequency analysis which is applied to determine the ratio of the different quasispecies to each other. Some sequences are biased to be multiplied more often than others in the replication cycles, leading to in vitro changed ratios.

When considering all of the listed problems it becomes clear why the implemented tools were all tested mainly on error free, engineered reads rather than on real data comparable to ours.

3.8 MANY MAJOR QUASISPECIES EMERGE FROM MINOR VARIANTS AT PREVIOUS SAMPLES

QuRe reconstructed between 1 (ACAC from 2012) and 109 (TCTC from 2012) quasispecies for each sample. Nearly all possible developments with regard to changes in the population size can be observed: Patient 3 from 2013 has nearly the same amount of species at both measured time points, while the number progressively increases in patient 5 of the same cohort. Other patients show a more wave-like pattern. In patient 1 and 3 of 2012, as well as patient 2 of 2013, the number of quasispecies increases at first and then drops back at the later time points to numbers similar to the ones in the initial sample. The opposite is true in patients 2 of 2012 and patient 1 of 2013, where the amount of variants decreases at first, only to increase again at the last samples.

Overall we can see that the amount of reconstructed quasispecies strongly depends on the number of used sequencing reads. This becomes especially clear when comparing the samples barcoded ATGA and TCAG of patient 1 from 2012 which originate from the same sample. In all of the mutation analyses these two samples behave identical or strongly similar, as we would expect given their origin, but in the reconstruction, ATGA has twice the amount of associated species than TCAG. The only difference between the two sets of reads we found previously is their size, thus we can assume that the higher number of input reads in ATGA lead to the larger amount of reconstructed quasispecies (compare table 9).

Number of input reads strongly influences the amount of reconstructed quasispecies

When analysing the phylogeny of the reconstructed quasispecies by patient, we could observe different dynamics. In patient 1 from 2012, only one species is present in the initial sample and many of the species in the second sample, including the major variant seem

Patient	Tag	# QS
1	ACAC	1
	AGAG	53
	ATAT	95
	AGCT	60
	ATGA	8
	TCAG	4
2	CGCG	63
	CTCT	44
	CACA	5
	CTAG	23
3	CATG	62
	TATA	60
	TCTC	109
	TGTG	55

(a) 2012 cohort

Patient	Tag	# QS
1	ACAC	30
	AGAG	4
	ATAT	10
	AGCT	34
2	ATGA	9
	CGCG	26
	CTCT	8
3	CTAG	14
	CACA	17
4	TATA	45
	CATG	36
5	TCTC	14
	TCAG	52
	TGTG	68

(b) 2013 cohort

Table 9: Number of reconstructed quasispecies per sample.
Shown is the amount of quasispecies reconstructed by *QuRe* for each sample in both cohorts.

to directly derive from it. The third sample however, seems to have evolved from it and is most closely related to a variant that only made up 0.58% of the population in the second sample. The fourth and fifth major variants have again evolved further and seem to origin from a variant that was only present at 3% in the third sample. Additionally to this observable timeline, there are a few branches that contain minor variants from all time points, showing that some species seem to neither increase nor decrease their fitness over time.

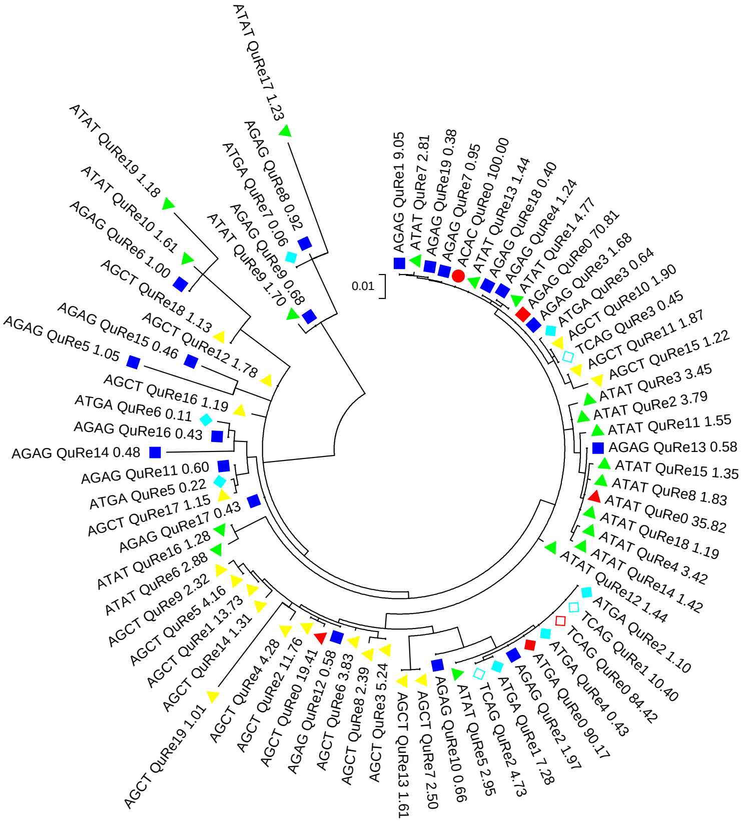
In patient 2 of 2012 we can see some minor outlier species that have a large mutational distance to most other sequences, that could either be briefly existing real variants occurring at the time of sample collection, or false positives erroneously constructed by *QuRe*. Additionally, it is hard to tell which quasispecies of the second time point lead to the population in the third sample, as it is very distinctly clustered in the tree. Patient 3 from the 2012 study shows a very straightforward phylogeny with only very few changes separating the different species and only two minor outlier cluster. For the phylogenetic trees of the 2012 cohort see figure 4.

The quasispecies reconstructed for the first time point in patient 1 from 2013 show a very strong variation among themselves. The closest variant to the major species at the second sample collection is a very minor one that only accounted for 0.46% of the initial population. Between the first and last two time points a major break occurs that makes it hard to determine which early variant established the later populations.

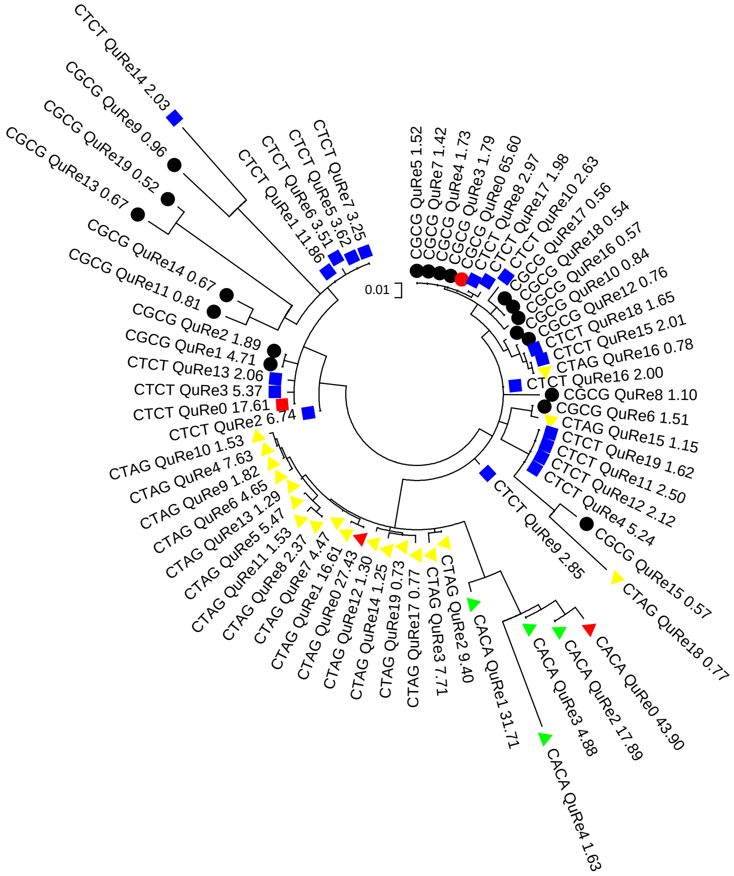
While the major variant of the second sample of patient 2 in 2013 directly emerged from the initial major species, the population of the third obtained sample seems to have mostly arisen from very minor variants.

Patients 3 and 4 from 2013, while both having only two time points, show very diverse quasispecies dynamics. In Patient 3 the major variant of the first time point accounts for over 93% of that population and is extremely closely related to the major species of the later time point. That one, however, merely represents half of its population, meaning the initial strong haplotype is in the process of being replaced by other upcoming species. In patient 4 all species from the first time point are very uniform and show minor divergence from each other and it is unclear which variant is most closely related to the population of the second time point. Only smaller clusters show a clear relationship to two minor variants of the first sample. The species from the second sampling are also much more variable when compared among each other.

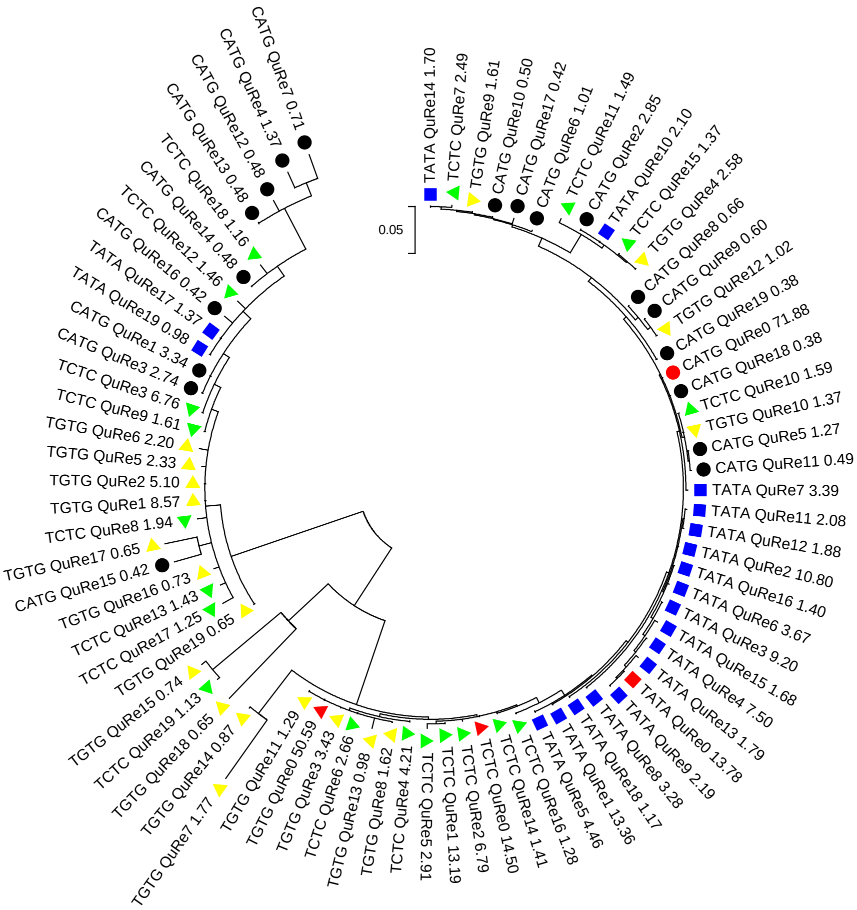
Patient 5 also follows a very straightforward relationship between the populations of the different time points. Most notably is only the increasing internal variance of the quasispecies belonging to one sample. For all trees representing the relationship of the 2013's cohort quasispecies refer to figure 5.



(a) 2012, patient 1

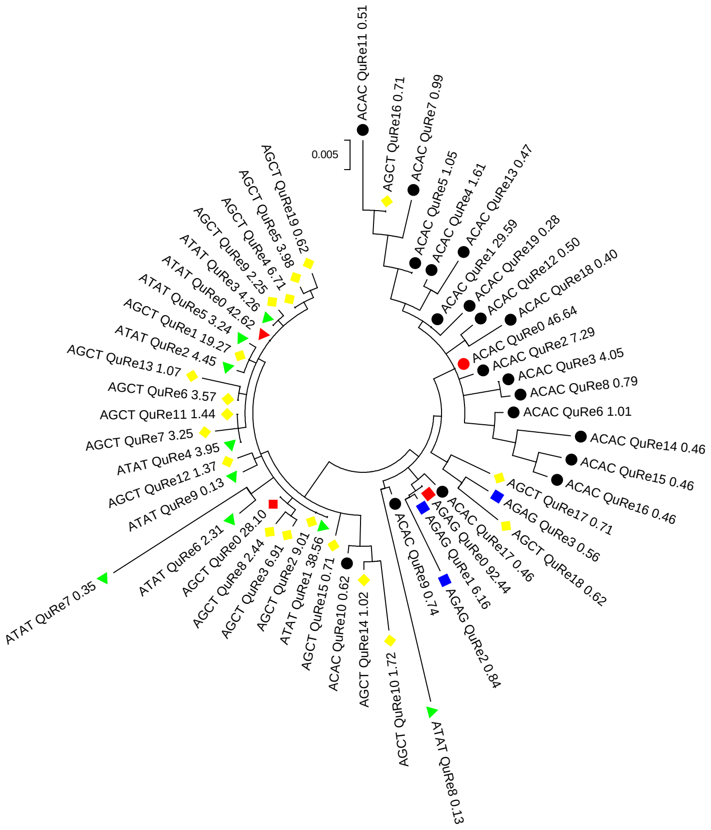


(b) 2012, patient 2

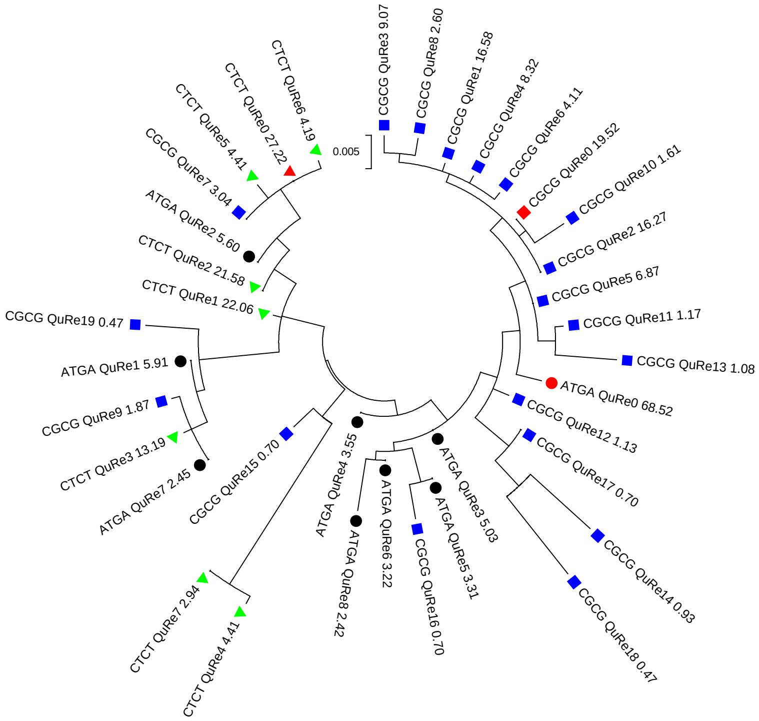


(c) 2012, patient 3

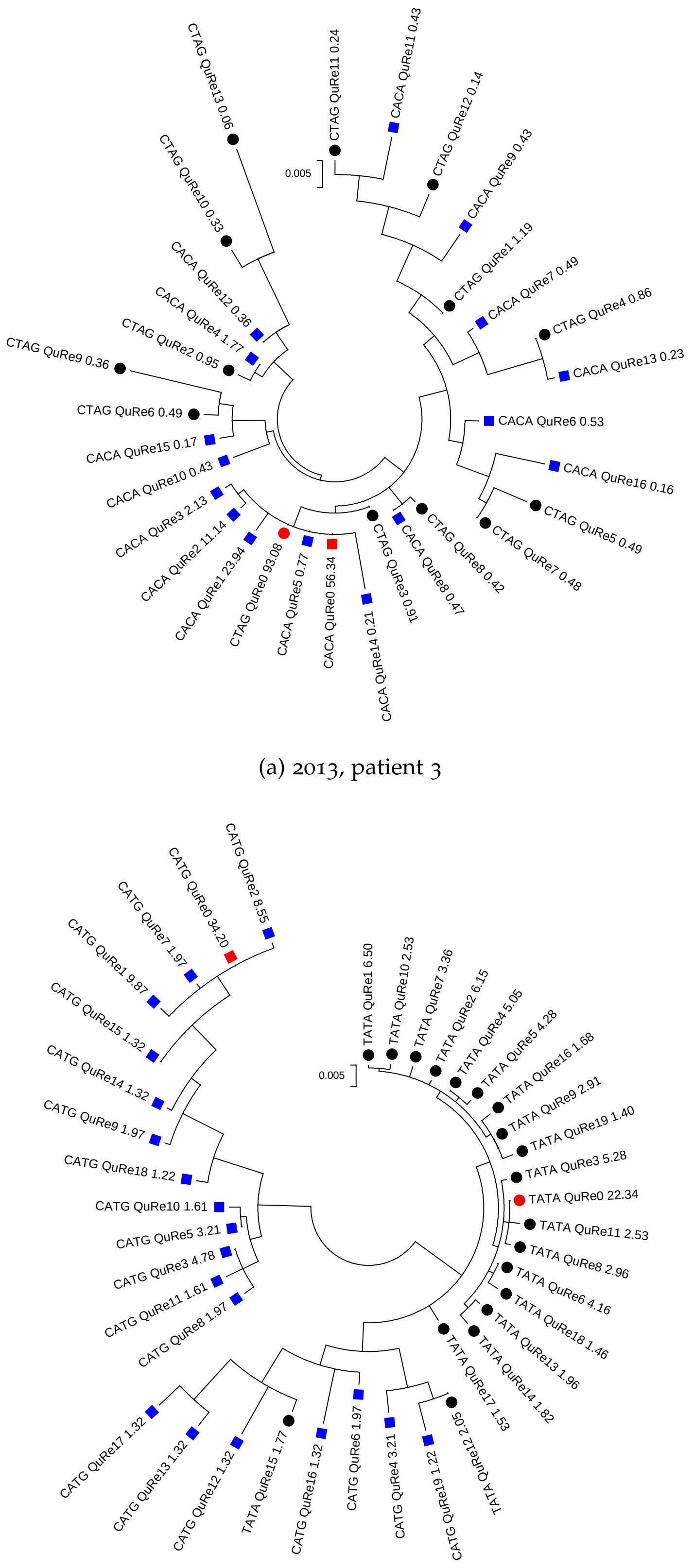
Figure 4: Quasispecies maximum likelihood tree for 2012 cohort. Quasispecies were reconstructed using *QuRe*. The most abundant 20 species from every time point were used for tree construction. The labels give the sequencing barcode, followed by a ranking from zero (most abundant) to 19 (least abundant used species) and the fraction the species makes up in the overall population in percent. Symbols are a visual aid to recognise the time progression: black circles - first sample, blue squares - second, green downward triangle - third, yellow upwards triangle - fourth, cyan diamond - fifth. Red symbols belong to the sample their shape indicates but represent the major variant of the associated time point.



(a) 2013, patient 1

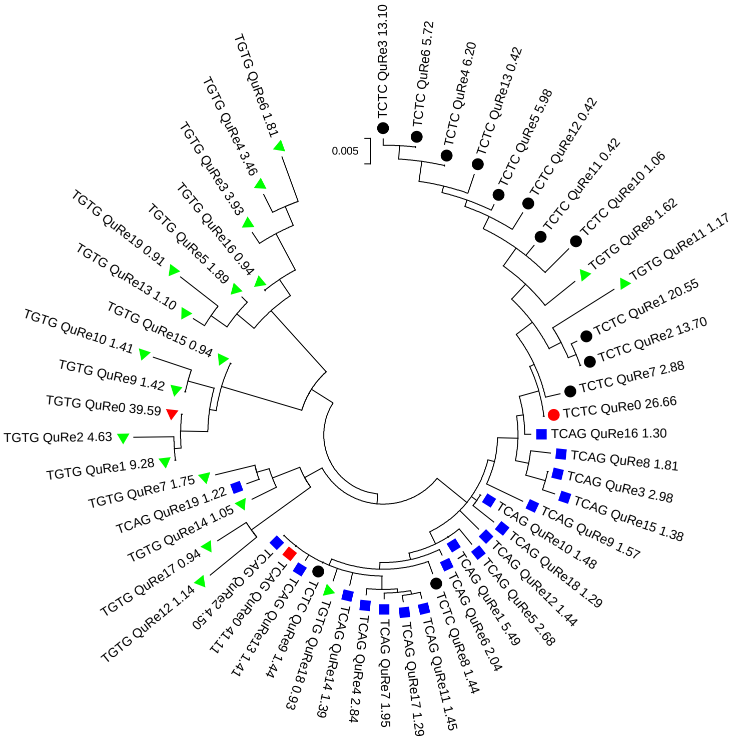


(b) 2013, patient 2



(a) 2013, patient 3

(b) 2013, patient 4



(c) 2013, patient 5

Figure 5: Quasispecies maximum likelihood tree for 2013 cohort. Quasispecies were reconstructed using *QuRe*. The most abundant 20 species from every time point were used for tree construction. The labels give the sequencing barcode, followed by a ranking from zero (most abundant) to 19 (least abundant used species) and the fraction the species makes up in the overall population in percent. Symbols are a visual aid to recognise the time progression: black circles - first sample, blue squares - second, green triangle - third, yellow diamond - fourth. Red symbols belong to the sample their shape indicates but represent the major variant of the associated time point.

4 CONCLUSION

Our longitudinal study of norovirus sequences obtained from chronically infected patients showed the great advantage of applying next generation sequencing to viral samples. The large number of reads covering a fairly small region of the norovirus genome leads to extremely high coverage and thus a large amount of data per nucleotide to base the downstream analysis on. Furthermore, the quality of the sequenced reads is very high, facilitating a very statistically stable mutation analysis.

Although the mutation rate of noroviruses is very high and our amplicon even especially targeted the hypervariable domain of its capsid protein, we were able to map more than 92% of the reads in every sample with a two-stage alignment approach.

The consensus sequences which we derived from the read mapping clustered in a phylogenetic tree according to their corresponding patients when working on their nucleotide sequences. On amino acid level, however, the patient relationship can not be reconstructed, showing a limited sequence space of the translated protein.

The mutation rates we could observe in our data were in most cases at least twice as high as the previously reported values for noroviruses, in some cases up to ten times as much. This could either result from the fact that we are only focusing on the hypervariable region of the capsid protein as opposed to the complete genome, or that we are looking at intraindividual dynamics, not comparing major outbreak strains.

When comparing the non-synonymous mutations in all three patients, clear mutational "hot-spots" could be identified. These are changes that occur in all patients along the course of the examination and involve often multiple consecutive amino acids. Furthermore do most of the observed variable regions correspond to important protein residues which have been previously described by other publications. These residues are mostly involved in the binding of HBGA trisaccharides and the stabilisation of said interaction.

The calculated d_N/d_S ratio between successive samples resulted in values greater one for multiple steps, hence representing positive selection. A more detailed analysis for single codons only resulted in three significantly positively selected positions in patient 1.

Reconstruction of viral quasispecies was extremely difficult given our experimental design. As most established programs expect a random distribution of reads, they were incompatible with input of amplicon origin. Even our own implementation, while working fine for each orientation of the amplicon, had trouble combining both sides for a full-length quasispecies reconstruction.

We finally were able to use a recent update of a published tool to obtain a set of complete species per sample. When analysing the evolution of the most prevalent quasispecies within one patient, we could observe very often, how a minor variant at a previous sample rose to being the major species at a subsequent time point, hence showing a fast paced evolution in the ongoing infection. However, as discussed in section 3.7, quasispecies reconstruction is a hard to solve problem and relies a lot on sound statistics. Thus, our current analysis might not contain all true positive variants present in the samples and might be influenced by library size, sequencing errors or a replication bias.

5 OUTLOOK

5.1 STUDY IMPROVEMENTS

There are different ways in which our current study design could be improved, if we were to repeat it. In the following we will touch briefly on changes in the patient cohorts or used technologies which would give additional insights into the intraindividual mutation dynamics.

5.1.1 *Sequencing technology*

In our herein presented study, we sequenced the libraries extracted from the patient samples with 454 pyrosequencing which was the method producing the longest read length at the time of realisation. However, in the mean time PacBio has officially introduced their Single-Molecule Real-Time (SMRT) sequencing technology [47] that can produce reads of up to 15,000 nt long reads for prices comparable to our 454 expenses. As each of the produced reads represents a single input fragment with up to 99% accuracy, we would be able to obtain quasispecies directly from the machine without any error correction or sophisticated quasispecies reconstruction algorithm.

SMRT sequencing makes it possible to obtain quasispecies sequences directly from the machine without further reconstruction

The only remaining difficulty to be overcome would be to distinguish between real representatives of the same species and PCR duplicates in the calculation of frequencies.

SMRT sequencing has already been applied to viral genomes and could successfully identify and help in assembly of full-length strains [186, 185].

5.1.2 *Genotype associations*

It has been established that the cellular receptors for noroviruses are most likely histo-blood group antigens that are expressed in the epithelium cells of the small intestine [83, 112]. While there seem to be additional co-receptors involved that have yet to be identified, there have been hypothesis that varying blood types convey different levels of susceptibility to a norovirus infection [156, 178]. Furthermore, it could be shown for Hepatitis C virus that patients carrying specific HLA alleles were either more likely to clear the virus or were especially slow and inefficient in mounting an immune response [204, 60].

It would have been very interesting to analyse both of these associations between HLA- and blood type to the mutational changes in our samples in order to identify the genomic regions responsible for interaction. It might further be possible to see how variation in receptors or MHC structures leads to different mutations in the infecting virus population. While we had MHC and blood type information for our earlier cohort, three patients is a far too low number to perform a meaningful correlation analysis. Thus, it would be of particular interest to carry out our complete examination on a much larger patient cohort.

A larger patient cohort would enable association analyses between mutations and HLA- and blood type information

However, this is a difficult undertaking as a chronic norovirus infection is hard to acquire and thus, a rare event. It might be interesting to instead use patients with acute infection, which are more numerous. With infection duration of approximately one week, the longitudinal samples would need extremely short interim times. But we could also observe a full-blown immune answer, not one that is strongly reduced by transplant accompanying medication. Thus, it might be possible to observe immune evading variants with respect to blood and HLA type.

5.2 CULTURING

Until the end of last year, norovirus research was limited by the fact that all sequences for research had to be extracted from infected individuals. Because there was no working cell culture system available, studies involving genetic engineering of sequences or vaccine trials could not be conducted.

In December of 2014 Jones et al. [89] published their study showing that noroviruses can be grown in vitro in B cells, as long as there are free HBGAs or HGBA-expressing bacteria added to the culture. They are also formulating the assumption, that noroviruses are capable of entering B cells in vivo based on their observations. This breakthrough discovery in norovirus research makes it finally possible to culture norovirus strains. Thus, certain hypothesis regarding norovirus mutational behaviour can now potentially be tested in a new infection model and are not limited to collected samples anymore.

Part II

EXPRESSION OF HUMAN ENDOGENOUS RETROVIRUSES IN CANCER TISSUE

In this part of the thesis we compared the expression of selected HERV families in mammary and urothelium cancer. In tumour and matched normal samples of the same patients, we analysed if there was a difference in the major expressed loci between the two conditions. In order to find annotated endogenous viral elements in the human genome, we used the most comprehensive HERV database, HERVd, which was adapted and extended to reference human genome assembly hg19.

This work is the result of a collaboration with the Institute of Virology at the Helmholtz Zentrum Munich.

Excerpts and figures from this part of the thesis have been published previously in the following form:

> PRESENTATION Haase K: Expression of human endogenous retroviruses in cancer tissue, presented at the *International Conference on Molecular and Evolutionary Oncology 2014*, St Petersburg, Russia, 2014

6 INTRODUCTION

6.1 RETROVIRUSES

6.1.1 *Structure and Replication*

The family of *Retroviridae* are enveloped viruses that carry their genome in RNA form while encapsidated, but rely on DNA in order to replicate and produce new virus particles. In contrast to most virus families which members enter a host organism and replicate by highjacking the host transcription and translation machinery to work with their packaged genetic material, retroviruses need to perform the additional step of integrating their genome into that of the infected organism.

While the complexity of retroviruses can be very diverse among the different families, the basic structure is shared among all. The unintegrated virus genome consists of three major classes of proteins which are surrounded by repetitive sequences on both ends. When the RNA gets reverse transcribed into DNA, these ends get copied so that the $5'$ and $3'$ boundary of the HERV sequence become identical. These end regions of the integrated retroviral genome, referred to as provirus , are called long terminal repeats (LTR) [31].

Retrovirus sequences that have integrated into the host's DNA are called provirus

The enzymes which are needed for reverse transcription and integration are encoded in the *pol* part of the retroviral genome, together with a protease which is responsible for cleavage of the resulting polyprotein. The *gag* portion of the viral genome (short for group-specific antigen) encodes multiple structural proteins which are needed for assembly of the viral capsid. The last group of proteins that are present in most retroviruses are *env* elements that consist of glycoproteins which are integral in the virus envelope after assembly [31].

To be able to transcribe all of the described proteins from the retrovirus genome, stop codon skipping, splicing and proteasomal cleavage has to be used. Most commonly expressed are the *gag* proteins which are needed in a higher amount than the enzymes because they are structural proteins. In a fraction of translation processes, the termination signal following the *gag* sequence is suppressed and a *gag-pol* precursor is transcribed. From this elongated sequence, the enzymes are cleaved. In order to create the envelope proteins, the aforementioned *gag-pol* region has to be spliced from the viral mRNA and different glycoproteins can be cleaved from the resulting translated polyprotein [31]. For a visualisation of the retrovirus structure and replication, compare figure 6.

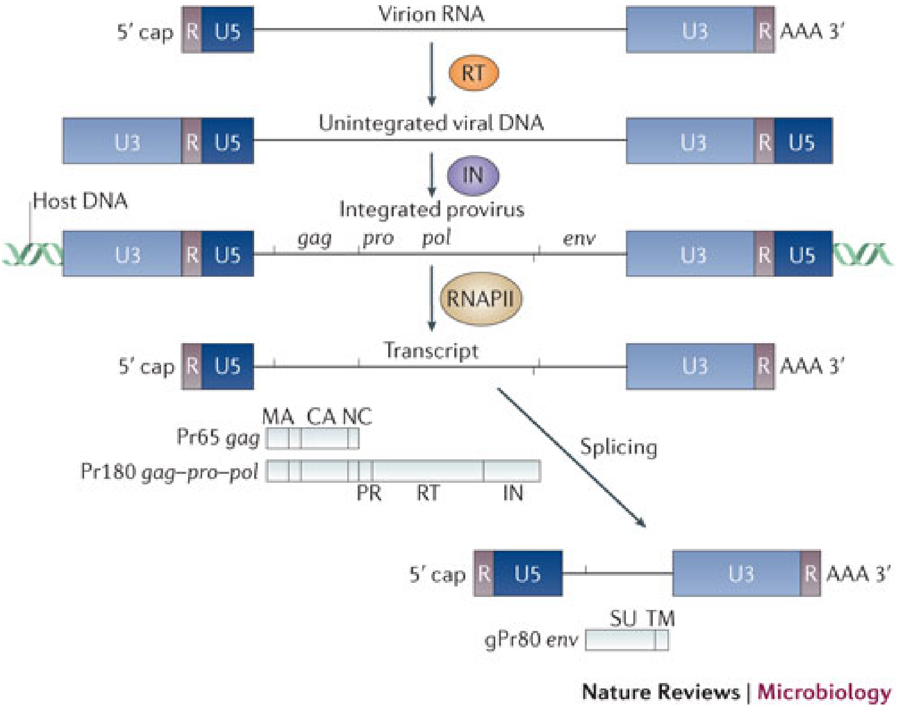


Figure 6: Schematic structure of a retrovirus genome.
Shown are the steps reverse transcription, integration and the translation of the three retroviral genes *gag*, *pol* and *env*.
RNAPII: RNA polymerase II, MA: matrix protein, CA: capsid protein, NC: nucleocapsid protein, PR: protease, RT: reverse transcriptase, IN: integrase, SU: surface proteins, TM: transmembrane proteins.
Taken from Stoye [175].

6.1.2 *Endogenisation*

Unlike most other virus families, retroviruses have three means of propagation. The first method is the horizontal transmission which they share with other infectious agents. For this way of transmission they need to be able to produce all parts to create new virus particles and to transcribe complete new copies of the viral genome. Due to the provirus containing a functional reverse transcriptase, retroviruses are also capable of propagating further within the same host. Through retrotransposition a copy of the integrated viral genome can be inserted in other locations throughout the host’s DNA. For this process, which can lead to a massive increase in copy number of the provirus in question, no proteins needed for the virus envelope are required.

The third way of propagation is vertical transmission. Retrovirus integrations happening in somatic cells are limited to the infected individual unless external transmission occurs. However, if the provirus is produced in a germ line cell, the retroviral genome gets added to the inheritable DNA of the organism. By this mechanism retroviruses are added to the overall genomic material of their host and can persist in them over the course of evolution.

After a provirus has become fixed in the genome of the species it used to infect, it is called endogenous retrovirus (ERV). These endogenised elements have become a fossil record, as they are extremely helpful in phylogenetic analysis. On the one hand, comparing which species share the same HERV families and loci can help to place the integration event on a time scale with regard to speciation, on the other hand, the LTRs can give a pretty accurate measurement of the age of the ERV [64, 120]. This is due to the aforementioned fact that at the time the retrovirus becomes a provirus in the host's genome, the LTRs on both ends are completely identical after reverse transcription. Because the provirus is not under purifying selection, as its functionality is not necessary for the organism, the viral sequence decays over time with neutral selection rate. Given this mutation rate and the current difference of the 5′ and 3′ LTR, a good estimate of the integration time point can be determined. [87]

Directly after reverse transcription the 5′ and 3′ LTR are identical in sequence

Due to the deterioration of the integrated viral sequences, most instances found in published genome sequences are incomplete, e.g. do not consist of the 8-9kb long ERV sequence. Usually, the *env* gene is the first to accumulate mutations and deletions, followed by the other internal open reading frames of the provirus. Throughout ERV carrying genomes, solitary LTRs are the most numerous instances of viral origin [12]. As long as the ERVs still retain their enzymatic activity, they are capable of further proliferating in the organism.

The many solitary LTRs are of particular interest, as they contain a strong promoter which originally is supposed to facilitate a satisfying expression of the provirus, but in case of isolated elements, can increase the expression of the genomic region in the direct vicinity.

ERVs have been found in all sequenced vertebrate genomes to date and make up between four and ten percent of the overall genomic sequence [142, 80]. The human endogenous retroviruses comprise 8% of the human genome, which has been shown after the first draft of the genome assembly has been finished. Furthermore, retroviruses are counted towards the repetitive element content of the human sequence, which accounts for more than half of the size of the human genetic material. [82]

About 8% of the human genome sequence consists of endogenous retrovirus elements

6.2 ENDOGENOUS RETROVIRUSES

6.2.1 *Evolution and Expression*

The overall age of ERVs could be estimated, as mentioned, by comparing integration sites and LTR deviation. Although for integration events dating back further than 200 million years ago an estimation is unfeasible, the age of many ERV loci could be determined. It could be seen that there are still very recent integrations in the human genome sequence [61]. By comparing the human genome sequence to that of Old World monkeys it was observed that most ERV families were present across all species, dating their fixation at least 30 million years ago [170].

The most recent integration into the human genome, HERV-K, has shown to be replicationally active for an extended time period, although it has accumulated random mutations [189]. The majority of HERV proviruses does in fact not encode functional proteins anymore, however, many of the loci are still being expressed. This expression was seen first in germline cells and placental tissue [153], but has since been shown in different cell lines, as well as cancer tissue. A specific microarray, containing HERV *pol* genes, has been designed to identify retroviral sequences present in an analysed transcriptome [168]. While these chips could successfully identify different HERV families in a follow-up study [169], it is limited to transcripts which still contain an undisrupted *pol* region. Furthermore, this approach only gives an overview of different families present in a transcript mixture and does not define the specific loci from which they originated.

6.2.2 *Classification*

HERVs are generally classified into three Genera, analogous to the classification of exogenous retroviruses. The different families are distinguished based on sequence comparisons of the retroviral genes, specifically the *pol* gene and the reverse transcriptase. HERVs with the highest similarity to γ-retroviruses belong to class I, β-retrovirus related HERVs are grouped into class II and HERVs most similar to spuma-viruses are sorted into class III [85]. For a visualisation of the retrovirus classes, see figure 7. It can be seen that the endogenous viral sequences from primate genomes are in some cases more closely related to those of bird and rodent genomes than to currently circulating exogenous retroviruses infecting primates, e.g. HIV-1.

HERV families are named after the tRNA which is used to initiate reverse transcription

While HERV nomenclature is not standardised, most families are named after the tRNA that is used to initiate reverse transcription (e.g. HERV-W uses a tryptophan tRNA) [121].

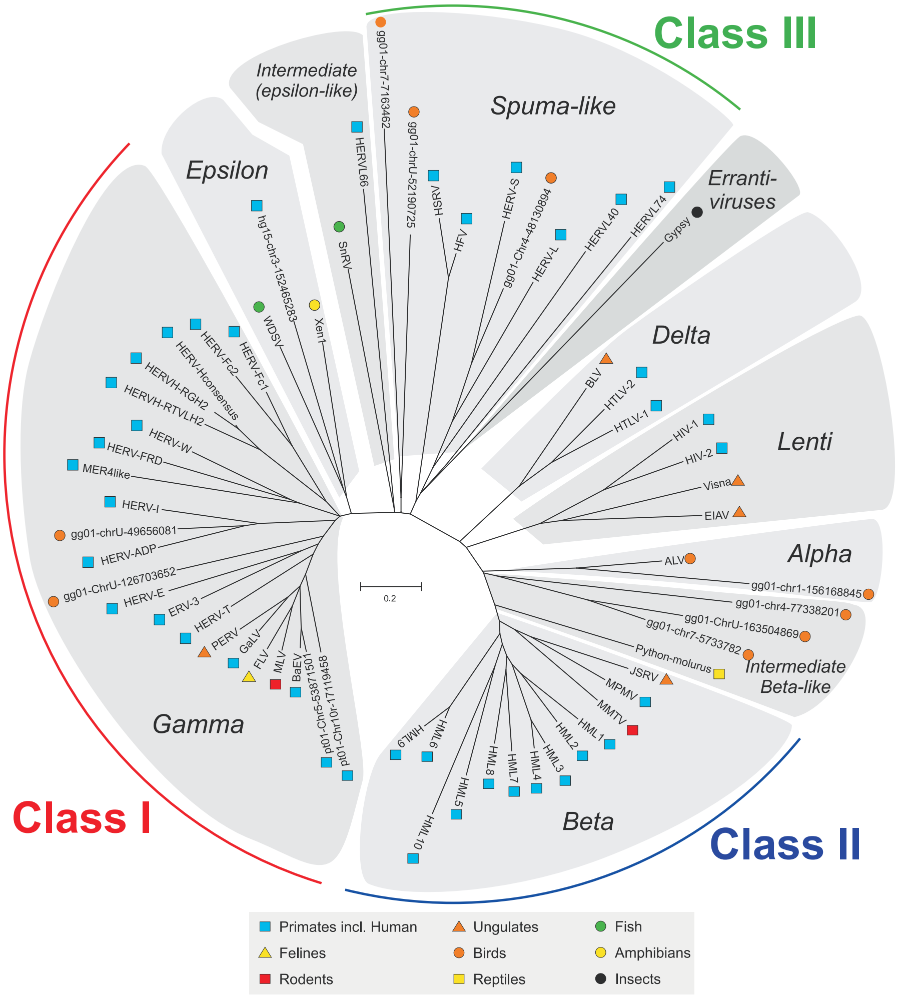


Figure 7: Neighbour joining dendrogram of retroviral *pol* sequences.
Shown is an unrooted neighbour joining dendrogram based on *pol* sequences (500 bootstraps consensus) of seven retroviral genera. The endogenous retroviral classes are indicated in the periphery. The host species are indicated with symbols next to each taxonomic unit. The novel sequences are named according to their chromosomal positions within respective genomes.
hg15 and 16: Human genome; gg01: Chicken genome and pt01: chimpanzee genome.
Taken from Jern et al. [85].

6.2.3 *Impact*

The integration of the provirus into the host's genome can lead to a variety of consequences with regard to the surrounding genetic loci. Due to promoter and enhancer sequences located in the LTR regions, the expression patterns of neighbouring genes can be altered. This could either be an increase in expression [32] or a change in tissue specificity if the LTR provides a transcription factor binding side atypical for the normal localisation of the gene in question [184]. An expression analysis of the youngest HERV family, HML2, showed that at least half of the analysed LTRs do have strong promoter activity [27]. Another way in which proviruses can affect host genes is via creation of antisense transcripts. By expressing short transcripts that are complementary to known genes, the mRNAs can be bound and silenced [63].

HERVs not only introduce new promoter signals into the host's genome, they also carry polyadenylation and splicing signals which can have influences on genes in close proximity. There are known cases in which the splice signal of an integrated HERV creates a shorter version of one of the host's genes [81].

Besides these effects on the infected genome, some functions of HERVs seem to be beneficial for their host. One proposed role of HERVs is that they convey a certain resistance to other viral infections. It could be shown in mice that a protein encoded by an otherwise disrupted endogenous retrovirus can prevent an infection of murine leukaemia virus, an exogenous retrovirus [14].

Furthermore, the family of HERVW retroviruses, which became fixed in the genome after the separation of New and Old World monkeys [121], is essential for the placenta development in its hosts. Most of the existing loci have deteriorated over time so that their protein coding genes have become disrupted. One locus, however, positioned one chromosome 7q21.2, still contains a full-length *env* gene from which syncytin-1 can be transcribed. During placental morphogenesis syncytin is thought to facilitate membrane fusion [123].

One locus of the HERVW family is essential for placental morphogenesis

Another HERV family with a known function that is beneficial to the host organism is HERVH. It is important for stem cell development, as it is suggested that HERVH can recruit pluripotency-associated transcription factors [163, 116].

6.2.3.1 *Potential involvement in carcinogenesis*

HERVs have been associated with multiple diseases by a variety of studies. The causation for these involvement has been named either production of viral proteins, expression of viral transcripts, changes in expression patterns through the mechanisms listed in section 6.2.3 or simply the disruption of a host gene by a retrotransposition into the coding region.

While one locus of the HERVW family has proven beneficial to the host, other viral sequences originating from HERVW proviruses have been associated with multiple sclerosis (MS) [139]. Although the exact mechanism in pathogenesis still needs to be revealed, the finding that HERVW sequences are expressed in MS-patients could be reproduced [102, 5]. HERV expression has also been linked to bipolar disorder and schizophrenia when transcribed HML2 and HERVW sequences could be isolated in brain tissue from affected patients [92, 54].

Probably the most commonly reported association between HERV expression and disease is for cancer. Different HERV families have been reported as transcribed in various tumour types. A direct causative effect in carcinogenesis has so far only been shown in mice and sheep [174, 172], while in the human organism studies aim at identifying similar processes. Possible mechanisms how HERVs can influence tumour formation are via *env* proteins that facilitate cell fusion, as seen in the case of syncytin in placental morphogenesis [176], or via regulation of tumour suppressor genes, as over 1,500 LTRs in the human genome have a nearly perfect p53 binding site [196].

For the aforementioned endogenous retroviruses involved in cancer pathogenesis in mice and sheep, exogenous counterparts still exist. As this is not the case for any of the known ERV elements in the human genome, the most likely active candidate is HML2, as it is the youngest integration which still has many full-length proviruses and can even produce viral particles [22]. HML2 particles could be seen in teratocarcinomas, although they were found to be defective and having no infectious potential [17]. Yet, studies found expressed RNA, as well as *gag* and *env* proteins and in some patients increased antibody levels for said proteins were measured [99].

HERV sequences are more similar to endogenous viruses of other species than to currently circulating exogenous viruses

The *env* genes of HERVK were also identified as expressed in melanoma cell lines and tissue [26] and multiple HERV families could be identified in mammary tumour tissue with a microarray approach [55].

6.3 MOTIVATION

While the molecular mechanism of retroviral replication and integration into host's genomes is well studied and mostly understood, all associations with diseases, especially cancer, are still spurious and not yet linked to causation. Previous attempts in identifying viral sequences in a complete transcriptome were made by using PCR with family specific primer sequences. While this approach can identify the contained HERV families, it does not provide an overview of the exact loci that contribute to the transcriptome.

We wanted to take a closer look at the different loci that still express endogenous viral sequences within the human genome and compare this activity between healthy and cancer tissue. By applying this approach to samples from multiple patients, we hope to identify any systematic changes in loci usage that could potentially be liked to tumour formation. Even if no change in transcript levels can be reported, by using highly sensitive next generation sequencing techniques, we will still be able to gain a detailed overview of all the transcribed loci from the analysed HERV families.

7 METHODS

7.1 amplicon reads

7.1.1 *Library Construction*

7.1.1.1 *Urothelium cancer*

Tissue from healthy and malignant urothelium were provided from the university hospital of the TU Munich, Klinikum rechts der Isar. The healthy samples were extracted from the periphery of the tumour. Overall 63 tissue samples were obtained, divided in 32 cancerous and 31 matched healthy probes from overall 32 patients. RNA extraction was performed on all samples and reverse transcribed into cDNA.

Based on initial analysis with microarrays analysing differential expression, five patients and five HERV families were chosen for amplicon sequencing. Amplicons were generated using primers specific for the five different HERV families ERV-9, HERV-E, HERV-W, HML2 and HML6. ERV-9 was chosen because it showed increased expression in about 30% of tumour samples, HERV-E and HERV-W had previously been shown to be expressed in urothelium tissue (HERV-E [66]) and T cell lymphoma (HERV-W [119]) with potential carcinogenic function. The two members of the HERV-K family, HML2 and HML6, were chosen because there are multiple cases reported in the literature showing increased HERV-K expression in tumours [205, 55]. For HML2 there was additional evidence in the microarray data hinting at increased expression in the urothelial cancer samples.

The five HERV families chosen for the expression study are ERV-9, HERV-E, HERV-W, HML2 and HML6

The polymerase used for amplicon replication had a specific proof reading function in order to prevent introduction of sequence errors during polymerase chain reaction (PCR). During PCR the temperature was set lower than calculated as optimal in order to create less stringent annealing conditions. With this approach we hoped to amplify sequences that do not have an exact match to the chosen primers. All described wet lab tasks were performed by Ingmar Göttesdorfer at the Institute of Virology at the Helmholtz Zentrum Munich and are described in more detail in his dissertation [67].

The created library was sequenced on a GS FLX+ System by an external company. To allow for the reads to be sequenced multiplexed in one run, they were tagged with four nucleotide long barcodes to distinguish the healthy and tumour samples per patient from each other.

7.1.1.2 *Mammary cancer*

The mammary tumour data comes from a study analysing 125 samples from 52 different female patients comprised of matched normal, cancer and lymph tissue. Four subjects (patients 1, 11, 21 and 33) were chosen for in-depth analysis with next generation sequencing. To obtain an even higher resolution than in the urothelium study, this time only two HERV families were selected for amplicon sequencing. HML2 and HML6 were chosen, due to their often reported involvement in breast cancer [198, 55, 34].

The experimental preparations, barcoding and library construction were done analogously to the urothelial samples (compare section 7.1.1.1). Afterwards, the library was also sequenced on a GS FLX+ System.

7.1.2 *Sample separation*

The downloadable sequencing files provided by the sequencing company contained the data separated by barcodes. For every barcode a fasta file (.fna), a quality file (.qual) and a file containing the standard flowgram format (.sff) was provided. The latter contains all information that are recorded during the sequencing run.

We converted the sff-files with the tool *sff2fastq* (version 0.8.0) into fastq files which contain the sequences of all reads and their corresponding qualities in phred encoding [50]. Phred defines the quality as a logarithmic function depending on the error probability P in a way that phred-score $= -10 \cdot \log_{10}(P)$. Thus, a phred-score of 30, for example, describes an error probability of 1 in 1,000 and hence an accuracy of 99.9%. Quality values between 0 and 40 are represented by ASCII characters, thereby the quality of a read can by displayed by a string of the same length as the nucleotide sequence.

The phred-score is a logarithmic function of error probability P*:* $-10 \cdot \log_{10}(P)$

We analysed the number of reads, their length and qualities before sorting the reads into their corresponding samples and filtered and trimmed them accordingly.

Although initially all samples from one patient and status (healthy/-tumour for urothelium and healthy/lymph/tumour for breast) should be distinguishable by a barcode added during library preparation, this could not be done in all cases. Some of the HERV-family specific primers would have formed a secondary structure with some of the prepared tags. Thus, some patients had more than one associated barcode and some tags were used for multiple samples. To differentiate the latter into their sets of origin, we had to additionally use the primer sequencing for sorting. For a list showing the used primers and barcodes in the urothelium study, refer to table A2. In the breast cancer analysis only one barcode (GTGT) has been used for multiple samples and needed to be separated by primer comparison (see table A3).

We were expecting that we would not to able to find the complete primer sequences at the beginning of the reads, on the one hand due to sequencing errors, on the other hand due to mutations which are present in the genomic HERV loci. To be less stringent in our search for the sequences, we took the first 21 bases (length of the shortest primer) of every read in a barcode set and clustered them on 90% sequence identity with *cd-hit-est* (version 4.5.4) [110]. In an ideal setup this would result in as many clusters as samples are associated with the analysed barcode, but in most cases we obtained more clusters than expected and thus lost a few reads when splitting the reads into the sample sets.

The separation steps resulted in one fastq files for each patient and condition.

7.1.3 *Quality analysis*

In all resulting fastq files we afterwards removed the first 21 bases for the urothelium samples and the first 24 bases for the breast samples to dispose of the primer sequences. We furthermore filtered all reads that contained Ns, meaning unresolved sequence positions and shortened the read ends until the quality reached a value of at least 20. In this way we counteract the normal quality decline of sequencing accuracy towards read ends. After the filtering and trimming we removed all reads which were now shorter than 100 bp. All of the trimming and filtering steps were carried out by the *RawReadManipulator* implemented by Jonathan Hoser [78].

All reads containing Ns and shorter 100 bp were discarded, low quality read ends and primer sequences were trimmed

The stringent quality control steps are necessary as our analysis is dependent on the exact mapping of reads to their locus of origin. Because most HERV loci belonging to the same family share a high similarity, especially among the youngest integrations, it is important to decrease potential sequencing errors which could be mistaken for sequence variation in the downstream analysis.

7.2 HERVD

In order to map our obtained reads back to the HERV loci on the human genome, we used the currently most comprehensive collection of annotated HERVs, the HERVd database [137, 138]. This database contains 98,008 entries describing 139 different HERV families, from full-length proviral elements to singular long terminal repeats (LTRs). Since the HERVd annotation is based on the hg17 assembly of the human genome we transferred all genomic coordinates to hg19 using the *liftOver* tool [74]. A number of HERVd entries did not survive the lifting process: 2,342 entries are completely or partially deleted and another 24 entries are split in the latest hg19 assembly. We nevertheless attempted to identify the location of these entries in hg19 by sequence similarity searches using BLAT [96]. Similarity hits were accepted as the origin of a given HERV if the corresponding alignments were gap free, covered the complete query sequence, and had a minimum sequence identity of at least 98%. In the same fashion we identified additional viral elements in hg19 by using all known HERV sequences as query and accepting new origins when they met the identity cutoff. The original HERVd database was obtained and extended by Anja Mösch in the scope of her bachelor's thesis [127].

Our extended and updated version of HERVd contains 100,495 HERV loci on the human genome assembly 19

The initial HERV data set contained 100,495 locations in hg19. We created basic statistics on the HERVd data to see, which fraction of each chromosome is taken up by endogenous viral sequences. Furthermore, we analysed how many loci of the five HERV families of interest are annotated in HERVd.

7.3 MAPPING

7.3.1 *Mosaik*

We mapped all reads against the human genome assembly (hg19) with the *Mosaik aligner* (version 2.1.73) [106] which is optimised to handle pyrosequencing reads. Instead of only mapping against the HERV sequences stored in our HERVd database, we decided to perform an alignment against the complete human genome, so we could identify potential HERV loci which are so far not annotated.

During the mapping we allowed a maximum mismatch percentage of 2% relative to read length and chose the unique mapping mode. Both settings are supposed to ensure the exact mapping of each read to its genomic origin although many other loci share a high sequence similarity.

7.3.1.1 *Expression per chromosome bands*

We compared the alignments of every sample against a list of our HERVd entries and counted the number of mapped reads that overlapped every sample with *bedtools coverage* (v2.17.0) [148]. Because the different samples were created using enrichment by PCR we can not compare them quantitively. Thus, we acquired the cytoBand table (for hg19) from UCSC table browser [93] and assigned all HERVd entries their corresponding chromosome band using *bedtools closest*. Afterwards, we calculated the fraction of reads in each sample that is allotted to a certain band and compared theses values between the different tissues of a patient based on HERV families. With this approach we avoid performing a quantitative analysis and instead evaluate which locus creates the most transcripts per family and can investigate changes in the activity based on disease condition.

7.3.2 *Blat*

Because the traditional mapping approach was only able to map an unsatisfying fraction of the reads, we attempted to improve the amount by using Blat [95]. This method has been reported to be successful for HERV sequences [165, 166]. Hence, we converted the filtered and trimmed reads to fasta format and mapped them to hg19 with Blat. As this results in many matches per read, we afterwards examined the output and assigned every read one of the classes *ambiguously aligned*, *unambiguously aligned* or *not aligned*. We only considered hits in this analysis if they covered more than a required fraction of the read and if they were located on one of the canonical chromosomes. If no match had less than a user defined number of mismatches the read is considered unaligned. If it can be optimally aligned, it is labelled *unambiguously* if the second best hit has at least one mismatch more, otherwise it is ambiguously aligned.

Reads with too many mismatches are not aligned, reads that could be mapped equally well to more than one locus are ambiguously aligned and all others are unambiguously aligned

Thus, two parameters have to be considered when analysing the Blat output: covered fraction of the read and allowed number of mismatches. The former parameter is of importance as Blat does not necessarily map the complete read to the reference but sometimes only aligns fractions. To identify an ideal combination of parameters, we calculated the fraction of all reads that would be unaligned, ambiguously and unambiguously aligned for 88 coverage and mismatch threshold combinations (the former ranging from 0.0 to 1.0, the latter from 3 to 10).

8 RESULTS

8.1 Reads have sufficient length and quality but unequal distribution

8.1.1 *Urothelium study*

The urothelium sequencing run yielded 14,987 reads. Their distribution among the different used barcodes is very uneven, ranging from 2,292 reads for ATGC to only 9 for barcode AATT. However, this can be explained with the different amount of samples that the barcodes were used for. A total of 843 reads could not be assigned to a tag-group because their first four bases did not exactly correspond to one of the used barcodes. The obtained length and qualities of the reads were very satisfying, with the former having its mean at 505 nucleotides (nt) and the quality having a mean phred score of about 32, thus an accuracy of over 99.9%. Table A4 shows the basic statistics for all urothelium reads.

As some barcodes have been used for more than one patient and condition (see table A2), we needed to separate those while referring to the used primer sequences. As these are of 20 nt length and longer, their probability of containing a sequencing error is higher than for the short barcodes, making it harder to assign them without ambiguity. Thus, we clustered the leading nucleotides of the reads starting with GAGA, GCGC and GTGT barcodes. Every resulting cluster was assigned a representative sequence which we then compared against all the known primers and linked the corresponding cluster with a patient and condition. All reads were then rearranged into files that belonged to a certain patient and condition rather than a given barcode. Additionally, we filtered the reads for length and quality criteria as described in section 7.1.3.

Ideally, when having nearly 15,000 from 10 samples, we should see 1,500 reads per data set. However, the read distribution deviates from this, by favouring certain conditions over others. Overall, the most reads are being allotted to patient 1 and the least to patient 4. The strongest difference between two conditions of the same patient are seen in patient 5. The reason for this varying read number could be an amplification bias due to the different used primers, or even unforeseen secondary structure formation, leading to a low yield of some primer-barcode combinations.

The mean length of the reads has decreased only slightly due to the removal of the primer sequences and the end trimming. Although we filtered low quality reads, the mean quality has stayed the same, because we also removed the leading bases of all reads which usually have the highest qualities. Table 10 shows the same statistics as table A4, but this time for the filtered and to patients and conditions assigned reads.

8.1.2 *Breast study*

In the mammary tumour analysis the sequencing run yielded 85,543 reads with lengths ranging between 41-744 nt and a peek at about 500 nt. The reached qualities were very satisfying as well, with a mean phred score of about 33.

Unfortunately, when clustering the read starts of all sequences beginning with the barcode GTGT, instead of three clusters, one representing every corresponding sample, we obtained 21. But three of these include the majority of reads, namely of all 4,800 reads starting with GTGT, 4,643 are covered by clusters of size 505, 2,277 and 1,861 sequences. When comparing the cluster representative sequences to the used primers, however, only the second and third cluster's representative can be found in the list of used primers. Although the representative of the first cluster does also match a primer used in the study (HML2 reverse primer 1, see table A3), according to the sample list, it was not used in combination with the barcode GTGT. Thus, we could only assign the two larger clusters to a patient and condition. As non of the smaller clusters matched the missing barcode either, we further investigated the already sorted reads and noticed, that the primer in question, HML2 forward primer 2, can not be found in any read. Hence, only three of the four primers used for HML2 loci targeting were effective.

One of the four primers used to amplify HML2 sequences did not produce any reads

From the initial 85,543 reads we were able to assign 68,413 to a specified sample, the remaining 16,214 did not show a known barcode. After filtering the reads have a mean length of 465 nt and a mean quality of about 34. The amount of reads is very unevenly distributed among the different patients and conditions. The highest fraction is attributed to patient 1's normal tissue with the second largest sample (patient 11's tumour) having less than half the number of reads. Patient 33 has the overall lowest read counts compared to the other three patients (compare table 11).

When looking only at the read statistics from the mammary tumour read statistics, the results seem very satisfying with regard to mean length and quality. However, when we start to compare the results to the previous urothelium study, especially because a similar library preparation and the same sequencing method was used, two major differences are striking. First, the length statistics differ drastically. While the mean is only about 35 nt shorter in the mammary study, the minimum length is 100 which means it is defined by our length filter-

Patient	Tissue	# Reads	Read length (nt)					Read quality (Phred score)				
			mean	median	min	max	σ	mean	median	min	max	σ
1	normal	1529.0	507.86	480.00	360.00	619.00	67.60	32.63	32.98	22.76	37.74	2.49
	tumour	2306.0	508.32	482.00	358.00	619.00	68.01	32.90	33.30	22.55	37.89	2.60
2	normal	1727.0	509.64	480.00	363.00	621.00	68.24	32.71	33.00	22.45	38.45	2.53
	tumour	1175.0	490.59	473.00	360.00	628.00	63.50	31.19	31.35	20.93	37.06	2.50
3	normal	1035.0	468.99	468.00	357.00	616.00	41.56	32.03	32.32	23.12	38.14	2.42
	tumour	1422.0	500.59	476.00	362.00	615.00	66.00	32.48	32.71	22.48	37.52	2.45
4	normal	889.0	493.92	477.00	358.00	620.00	60.82	32.45	32.83	23.27	38.37	2.59
	tumour	722.0	492.14	471.00	367.00	654.00	62.76	32.82	33.19	23.71	38.37	2.33
5	normal	778.0	473.10	465.00	364.00	613.00	48.67	31.87	31.98	24.23	37.51	2.42
	tumour	2247.0	477.68	470.00	363.00	617.00	53.51	32.01	32.22	21.94	37.87	2.52

Table 10: Statistic of urothelial reads filtered and divided into samples.

ing criteria. In the urothelium study this filter, while applied, was not needed, as the shortest occurring reads were still over 350 nt in length. Second, although the existence of the unusual short reads seems to show that problems occurred in the mammary sequencing run, the mean quality is still on average two units larger than in the urothelium study. We can not say what exactly caused these performance differences between our two cancer studies, as even troubleshooting in cooperation with the sequencing company did not yield helpful hints.

8.2 HERVD COVERS APPROXIMATELY 3% OF THE GENOME

We wanted to analyse how our extended HERVd annotations are distributed among the human genome. For this purpose we counted all bases which were covered by a HERV locus and divided the number through the length of the respective chromosome. Overall, approximately 2.9% of the genome are made up of HERVs. Chromosomes which exceed this value by more than one standard deviation are chr4, chr19 and the two gonosomes. The chromosomes 15-17 and chr22 show the least coverage by HERVs (compare figure 8).

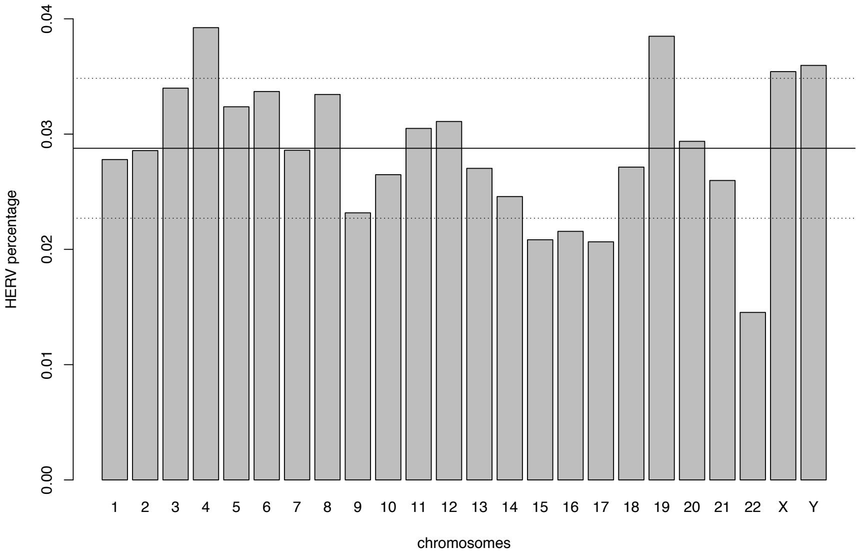


Figure 8: HERVd coverage of the human genome.
Shown are the number of bases covered by a HERVd annotation divided by chromosome length. The continuous line marks the overall mean, the dashed lines the standard deviation.

It is striking that the overall fraction of 2.9% lies a lot lower than the usual cited amount of 8% which is often given as percentage of the human genome arising from viral elements. The 8% quote stems from the publication accompanying the initial draft of the human

Patient	Tissue	# Reads	Read length (nt)					Read quality (Phred score)				
			mean	median	min	max	σ	mean	median	min	max	σ
1	lymph	6358.0	465.66	476.00	100.00	519.00	54.60	34.17	34.51	24.04	39.34	2.25
	normal	16482.0	463.25	475.00	100.00	527.00	55.58	34.54	35.00	21.23	39.75	2.23
	tumor	2555.0	465.56	476.00	101.00	504.00	52.59	34.19	34.50	23.74	39.17	2.24
11	lymph	4438.0	464.64	475.00	100.00	515.00	52.09	34.03	34.37	21.45	39.44	2.29
	normal	5463.0	464.63	475.00	101.00	515.00	50.05	34.53	34.97	21.06	39.50	2.09
	tumor	8030.0	466.19	475.00	105.00	517.00	44.60	35.01	35.51	20.88	39.39	2.16
21	normal	3753.0	468.97	476.00	102.00	516.00	52.05	33.99	34.19	21.62	39.40	2.25
	tumor	5153.0	462.80	478.00	100.00	527.00	62.79	33.88	34.13	23.63	39.27	2.31
33	lymph	2217.0	464.64	475.00	107.00	503.00	48.72	35.14	35.59	24.54	39.67	2.12
	normal	1163.0	467.62	477.00	111.00	521.00	56.47	33.89	34.13	25.61	39.16	2.28
	tumor	1266.0	470.13	481.00	100.00	509.00	54.45	33.77	33.94	23.47	39.39	2.39

Table 11: Statistic of mammary reads filtered and divided into samples.

genome sequence from 2001 [82], but they count all LTR-containing sequences towards this number. Our annotations from HERVd do not contain any mammalian apparent long terminal repeat retrotransposon (MaLR) sequences, as these are non-autonomous and do not contain internal viral sequences [137]. Furthermore, some prominent HERV sequences, e.g. HML2 K111, are located in centromeric regions [35] which are not completely reconstructed in the human genome assembly due to their highly repetitive nature and are thus not contained in HERVd.

HERVd contains annotations for 2.8% of the human genome because MaLRs and HERVs in highly repetitive regions are not included

To examine how many loci are known in HERVd for our five analysed HERV families, we counted them chromosome wise. The previously analysed percentage is reflected in this values, as chromosomes containing overall low amounts of HERVs, also only carry a smaller number from the families of interest. An exception make the gonosomes, as chrX does not contain a strikingly high amount of loci, but chrY has nearly twice as many ERV9 and HML6 loci than any other chromosome. When compared across all chromosomes, the most annotated loci in HERVd belong to ERV9 (442) and the least for HML6 (131). For the chromosome-wise family distribution see figure 9.

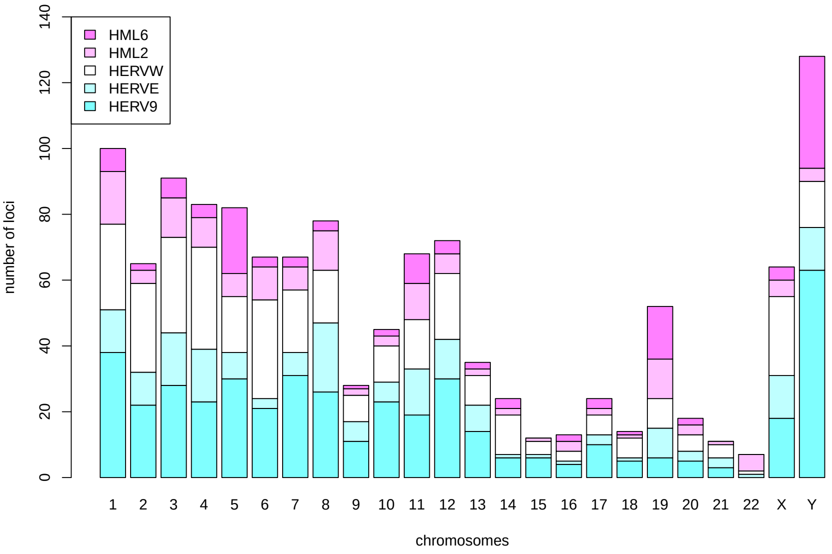


Figure 9: Number of loci belonging to five analysed HERV families per chromosome.

Patient	Condition	
	healthy	tumour
1	1070 (69.98)	1670 (72.42)
2	1226 (70.99)	741 (63.06)
3	684 (66.09)	1022 (71.87)
4	581 (65.35)	509 (70.50)
5	508 (65.30)	1467 (65.29)

Table 12: Mapped urothelium reads per sample.
Values in brackets show the percentage of mapped reads relative to all reads entering the mapping process.

8.3 only a low fraction of reads can be mapped to the genome

8.3.1 *Urothelium study*

With our initial mapping strategy, using a program specialised for pyrosequencing reads, an allowed mismatch percentage of 2% and the placement of unique reads only, we were able to map on average 65% of all reads. Table 12 shows the number of mapped reads in every sample.
That only a comparably low amount of reads can be mapped can mostly be due to two reasons. First, it is possible that our allowed mismatch percentage of 2% is too strict and reads can not be placed because they show a greater variance within the HERV loci compared to the hg19 reference. Second, the reads could be matched equally well to two or more loci, because members of the same family are still very similar to each other and thus the mapping would not be unique. Our analysis of the cause can be found in section 8.5.

The unambiguously mapped reads were then compared against our HERVd annotations and we could assign all reads to one of the five families of interest. We also checked for reads that had been mapped onto hg19 but were not covered by a HERVd annotation and found that only seven reads mapped in a region without a known HERV locus. Because non of these seven reads were mapped to the same region, we regarded them as too weak evidence to define a new locus. Although all other reads could be assigned to a HERV family, some of them mapped to loci that were not targeted in the library construction. We found reads mapping to HML9, HERVH, HERV3 and HERVFb. When analysing the distribution among the five HERV families of interest, we noticed that despite equally sized libraries previous to sequencing, the produced reads were extremely unequally distributed. HERVEa is hardly contained in any sample, taking up

Additionally to the targeted HERV families we also found reads mapping to HML9, HERVH, HERV3 and HERVFb

between 1 and 4% of all reads, while HML2 is favoured in replication and being the source of up to 50% of all mapped reads. The other three families show variations between samples, but not as strong as the other two (compare table 13).
The strong differences in the number of mapped reads can either be due to a preference in amplification during the sequencing process, or the reads we were not able to map belong primarily to family HERVE.

8.3.2 *Breast study*

When mapping the reads originating from the breast cancer samples with Mosaik to hg19, an on average even lower percentage than in the urothelium study could be placed. While in one case (patient 11 tumour) more than 65% of all reads can be mapped under our applied constraints, in two cases (patient 33 lymph and patient 11 normal) not even half of the total set can be aligned (compare table 14).

Instead of continuing with the analysis analogous to the urothelium samples, we decided to use a Blat mapping to assign the reads to their loci, hoping to identify more fitting parameters and to understand why on average 45% of the reads could not be aligned. We executed Blat runs with multiple combinations of the two parameters *read coverage*, meaning fraction of the read involved in a calculated alignment, and *allowed mismatches*. For each parameter pair we could then count the number of reads that could not be mapped due to too many mismatches, the ones that could be aligned to more than one origin and the reads which could be unambiguously aligned. With this approach we can determine, if the reads we lost in our initial mapping were filtered out by the *unique* setting, or the maximum of 2% mismatches.

With a requested read fraction of 50% up to 90%, the read classification solely depends on the number of allowed mismatches. It can also be seen that if we request at least 50% of a read to be involved in a match, even if we allow up to ten mismatches, we can not unambiguously align more than 75% of the reads (compare figure 10). This is in line with the previous Mosaik mapping, as ten mismatches in an approximately 500 bp long read equates to 2%. At least 40% of a read should be part of the match to keep the number of ambiguously aligned reads considerably low. But even then, an amount of mismatches between 8 and 16 per read would still lead to about 15% of ambiguously mapped reads. This high percentage is probably due to the young age of the two HERV families of interest and hence, their high sequence similarity between loci.
To create alignments comparable to the Mosaik mapping, we chose the parameter combination 10 mismatches and 80% read coverage to carry out Blat runs. With this setting we were able to unambiguously

Patient	Assigned reads									
	Normal					Tumour				
	ERV9	HERVEa	HERVW	HML2	HML6	ERV9	HERVEa	HERVW	HML2	HML6
1	0.16	0.01	0.20	0.47	0.13	0.12	0.01	0.21	0.46	0.18
2	0.19	0.01	0.14	0.50	0.13	0.20	0.01	0.21	0.33	0.19
3	0.28	0.02	0.38	0.10	0.17	0.17	0.02	0.20	0.41	0.17
4	0.19	0.02	0.28	0.36	0.14	0.24	0.01	0.17	0.31	0.26
5	0.41	0.04	0.07	0.17	0.19	0.27	0.03	0.19	0.21	0.17

Table 13: Urothelium reads assigned to HERV families.
Values show the number of reads assigned to the corresponding family divided by all unambiguously aligned reads for that sample.

Patient	Condition		
	lymph	normal	tumour
1	3447 (54.22)	9123 (55.35)	1354 (52.99)
11	2511 (56.58)	2434 (44.55)	5400 (67.25)
21	–	2333 (62.16)	3239 (62.86)
33	1048 (47.27)	678 (58.30)	784 (61.93)

Table 14: Mapped mammary reads per sample.
Values in brackets show the percentage of mapped reads relative to all reads entering the mapping process.

map the majority of reads per sample, while keeping the ambiguous mappings low.

The second most common class for all samples, except patient 21 tumour, are reads with more than 20 mismatches and the minority of reads could only be mapped ambiguously. For the exact read classification compare table 15.

When trying to determine what caused so many of our reads to be un-mappable to the human genome, at least when requiring more than half of the read to be involved, we found that our sets of reads contained many chimaera. Through a Blast analysis of a selected subset of reads, which were classified as *not aligned* in our Blat mapping, with the chosen parameters and an alignment examination by eye, we found that more than half of the tested reads showed two good local alignments, each involving one half of the sequence. Thus, the multiple rounds of amplification that were needed during library preparation and sequencing seem to facilitate the formation of chimaera.

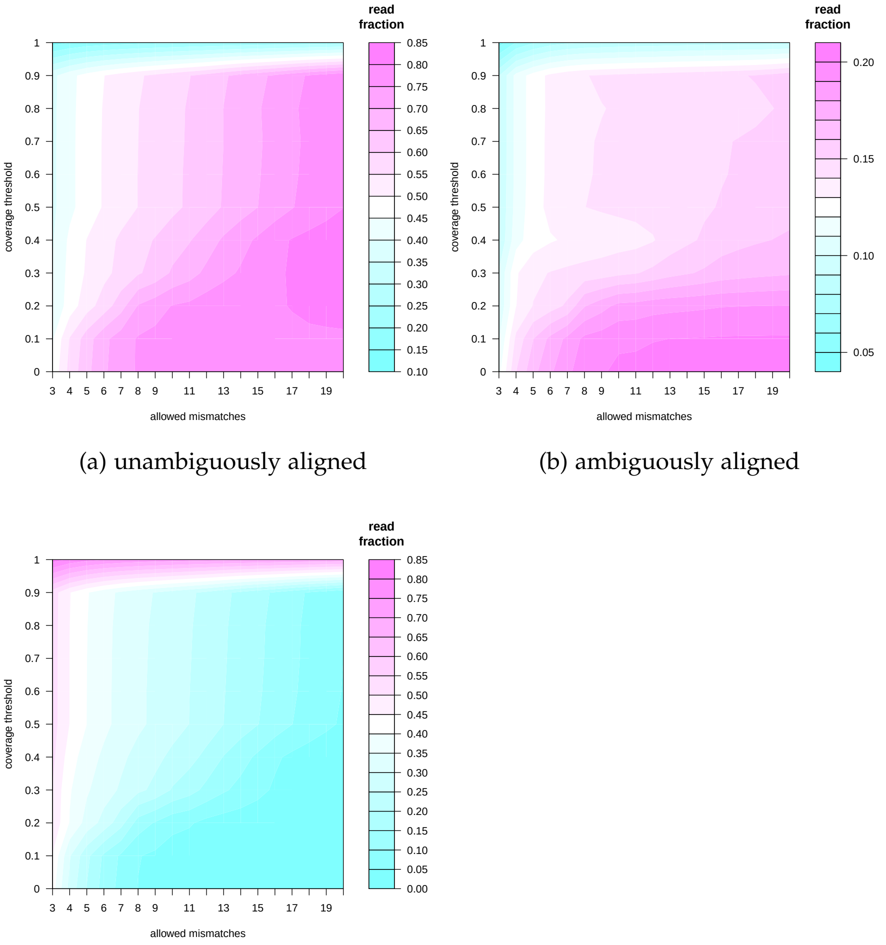
The high amount of un-mappable reads is probably due to chimaera

8.4 LOCI USAGE MORE DEPENDANT ON INDIVIDUAL PATIENT THAN CONDITION

8.4.1 *Urothelium study*

When analysing the differential usage of loci between the two tissue conditions of each patient, it was striking that most variances in expression seemed to be specific for a particular patient, not for the tumour state.

Overall we could identify 66 expressed HERV9 loci, 26 HML2 loci, 24 HERVW loci, 15 HML6 loci and although only very few reads were remaining from the HERVE set, we still found 18 expressed loci assigned to this family. For HERV9 we can observe that the three most strongly expressed loci (2q33.2, 6p22.3 and 11q21) are the same for all five patients and both disease states. While in the other four patients, these three loci each mostly make up between 10 and 25% of



(a) unambiguously aligned

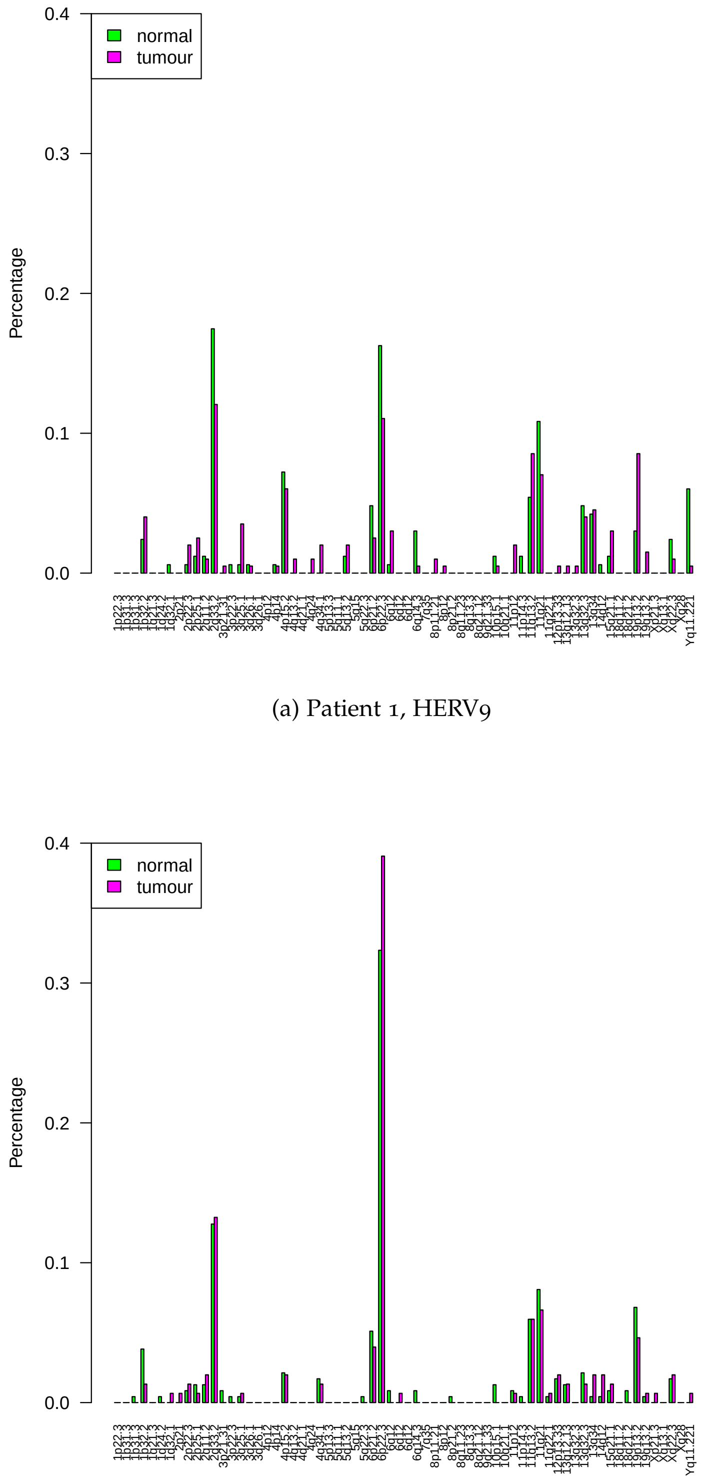
(b) ambiguously aligned

(c) not aligned

Figure 10: Fraction of mammary reads being unambiguously, ambiguously or not mapped.
The three classifications are assigned depending on the allowed number of mismatches and the minimum fraction of a read involved in a match.

the HERV9 reads, in patient 2 locus 6p22.3 is strongly favoured and accounts for 32% in healthy tissue and 39% in the tumour sample. While there are minor differences in loci usage when comparing the two tissue types with each other, a significant switch can not be found (compare figure 11 and A2).

In the case of HERVW the results are very similar than for HERV9, as we can mostly see three prominently expressed loci in all patients. Only for HERVW even the ranking of the three major loci is consistent among the study subjects, except for one case. The locus from which most reads are transcribed is Xq22.3, followed by 14q21.2 and 7q21.2. The exception to this rule are the tumour samples from patients 2 and 4, as well as both tissues from patient 5. In the former two locus 7q36.1 is slightly stronger expressed than 7q21.2. In patient 3, how-



(a) Patient 1, HERV9

(b) Patient 2, HERV9

Figure 11: Distribution of expressed HERV9 loci in patients 1 and 2.
The x-axis contains all HERV9 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.
Only patients 1 and 2 are shown here as representatives, the plots corresponding to the remaining three patients can be found in figure A2.

Patient	Class	Assigned reads		
		lymph	normal	tumour
1	unambiguous	3503 (55.1)	11315 (68.7)	1388 (54.3)
	ambiguous	986 (15.5)	518 (3.1)	404 (15.8)
	not aligned	1868 (29.4)	4638 (28.2)	763 (29.9)
11	unambiguous	2488 (56.1)	3105 (56.9)	5906 (73.6)
	ambiguous	752 (16.9)	286 (5.2)	579 (7.2)
	not aligned	1197 (27.0)	2069 (37.9)	1543 (19.2)
21	unambiguous	–	2141 (57.1)	2669 (51.8)
	ambiguous	–	760 (20.3)	1370 (26.6)
	not aligned	–	849 (22.6)	1113 (21.6)
33	unambiguous	1416 (63.9)	636 (54.7)	668 (52.8)
	ambiguous	43 (1.9)	231 (19.9)	280 (22.1)
	not aligned	758 (34.2)	296 (25.5)	317 (25.1)

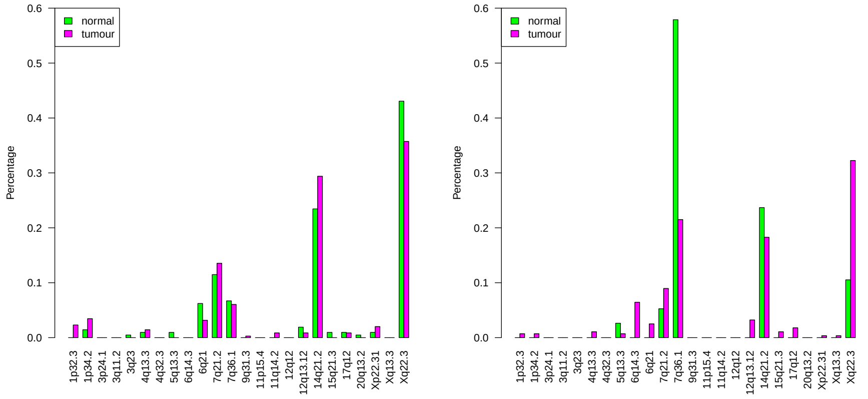
Table 15: Mammary reads assigned to classes after Blat mapping.
Shown are the classification results, when requiring 80% of the read to be involved in the match and allowing at most 10 mismatches.
Values in brackets show the percentages.

ever, the same locus is so highly expressed in healthy tissue, that it provides 58% of all reads. Consistent with the results for HERV9, no significant change in loci usage can be seen when comparing healthy and tumour tissue (compare figure 12 and A3).

From the 26 identified HML2 loci showing expression in our data, 25 are expressed in varying degrees in all samples and make up between 0 and 20% of all HML2 reads. One locus is the major transcript producer in all patients: 3q21.2. Although in patient 1 this transcription origin produces nearly twice the proportion in tumour than in healthy tissue, this pattern is reversed in patient 5. Hence, also for HML2 there is no clear cancer-associated locus (compare figure 13 and A4).

3q21.2 is the major HML2 locus in all five patients

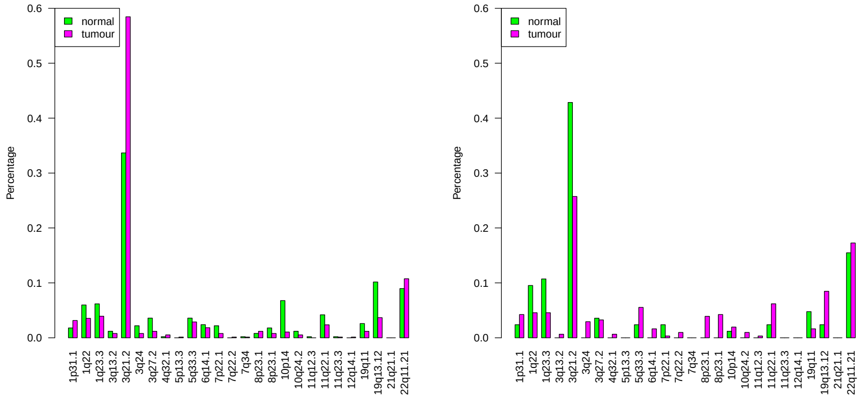
Our analysis of expressed HML6 loci revealed that in patients 3, 4, 5 as well as for the healthy tissue of patient 1 and the tumour sample of patient 2 the strongest expressed locus is Xp11.21. In patient 4's tumour tissue this locus produces so many reads that nearly no other expressed loci can be found (Xp11.21 accounts for 89%). When only examining patient 1, a clear locus shift between healthy and tumour tissue could be observed, with 14q24.2 being nearly exclusively expressed in the cancer sample and Xp11.21 and 19q13.41 being nearly exclusively expressed in the healthy tissue. But when looking at the



(a) Patient 1, HERVW

(b) Patient 5, HERVW

Figure 12: Distribution of expressed HERVW loci in patients 1 and 5. The x-axis contains all HERVW loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus. Only patients 1 and 5 are shown here as representatives, the plots corresponding to the remaining three patients can be found in figure A3.

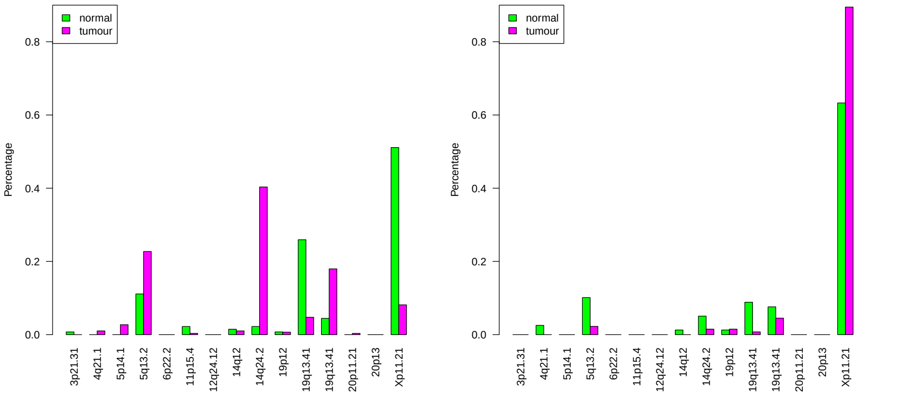


(a) Patient 1, HML2

(b) Patient 5, HML2

Figure 13: Distribution of expressed HML2 loci in patients 1 and 5. The x-axis contains all HML2 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus. Only patients 1 and 5 are shown here as representatives, the plots corresponding to the remaining three patients can be found in figure A4.

other four patients, these findings are revealed to be limited to patient 1 and not a general disease specificity (compare figure 14 and A5).



(a) Patient 1, HML6

(b) Patient 4, HML6

Figure 14: Distribution of expressed HML6 loci in patients 1 and 4. The x-axis contains all HML6 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus. Only patients 1 and 4 are shown here as representatives, the plots corresponding to the remaining three patients can be found in figure A5.

Due to the small amount of reads which could be mapped to HERVE loci, we decided that their distribution would not be representative of the underlying mechanisms and did not analyse them further. For reasons of completeness, the corresponding barplots can be found in figure A6.

HERVE is omitted from further analysis because of sparse data

8.4.2 *Breast study*

All reads which were unambiguously mapped by Blat were then overlapped with the HERVd hg19 annotation to obtain coverage values for every HERV locus. We also ran *bedtools coverage* in the opposite direction, with the reads as query, to control if every mapped read was positioned inside a HERV locus. If reads were aligned to the genome outside of HERVs, we compared their positions with *bedtools intersect* against the GENCODE V17 annotation to see if the read can be explained by a known transcript. For most patients and conditions only between 0 and 6 reads fall outside the HERVd annotations, all of which can be explained by GENCODE transcripts. An exception are patient 21's tumour and patient 1's normal sample. In the former, 48 reads lie outside of HERV loci, 47 of which are contained in GENCODE V17. The normal sample of patient 1, however, contains 125 reads that map to hg19 but not within a known HERV locus. Only 113 of these transcripts can be explained by GENCODE. It might be that the extreme amount of reads in patient 1's normal sample orig-

Patient 1's normal sample contains reads that can not be mapped to a known transcript

Patient	Assigned reads					
	Lymph		Normal		Tumour	
	HML2	HML6	HML2	HML6	HML2	HML6
1	1525 (43.53)	1954 (55.78)	1638 (14.48)	9523 (84.16)	523 (37.68)	850 (61.24)
11	1073 (43.13)	1396 (56.11)	450 (14.49)	2645 (85.19)	745 (12.61)	5146 (87.13)
21	–	–	807 (37.69)	1316 (61.47)	1236 (46.31)	1363 (51.07)
33	49 (3.46)	1366 (96.47)	299 (47.01)	330 (51.89)	403 (60.33)	263 (39.37)

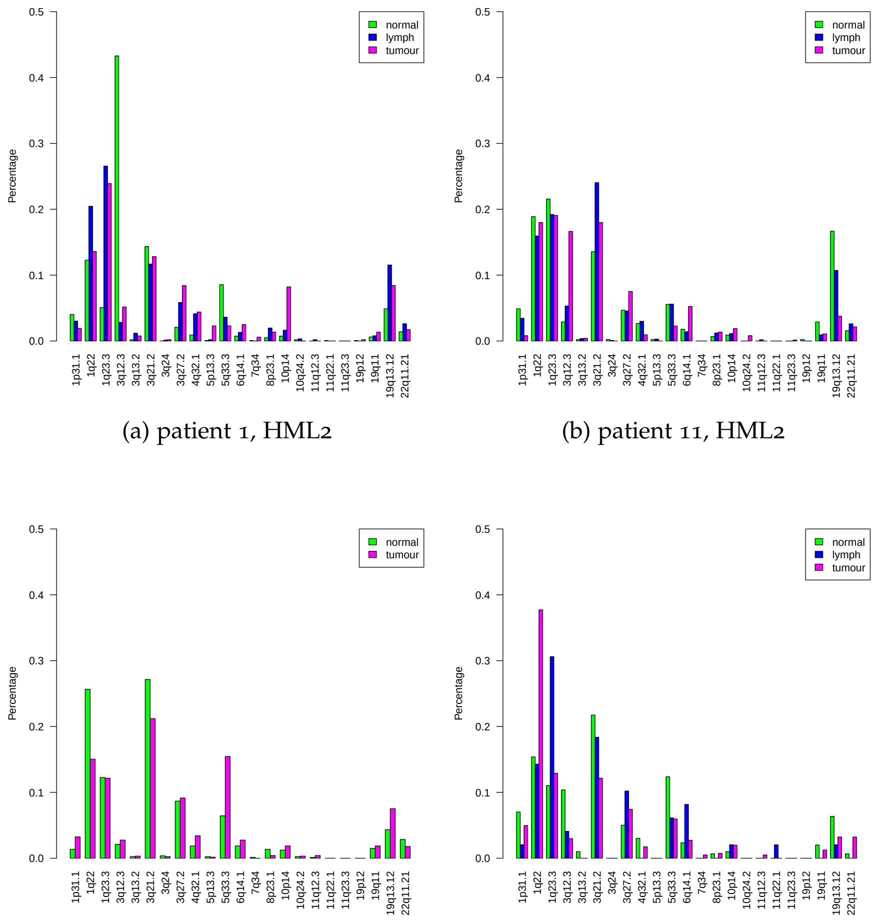
Table 16: Mammary reads assigned to HERV families.
Values in brackets show the percentage of reads assigned to the corresponding family divided by all unambiguously aligned reads for that sample.

inates from a less specific amplification during library construction than in the other sample preparations.

Most reads that were unambiguously mapped to a HERV locus, belonged to one of the families associated with either HML2 (HML2 or HERV-K) or HML6 (HML6, HERVK3 or HERVK3I). Only eight other loci were covered: two HML9, one HERVK(II), one ERVL, one HERVL, one MER9, one HERVK14Cl and one LTR5_Hs entry. The distribution of reads between HML2 and HML6 loci is very biased towards HML6, as in all samples more reads are assigned to HML6 than to HML2. The strongest case can be seen in patient 33's lymph sample, where nearly all reads (96.47%) originate from HML6 loci. The overall assignment of reads to the HERV families HML2 and HML6 can be seen in table 16.

When analysing the loci usage among the different samples of the same patient, the observations are very similar to the results of the urothelial study. In the overall 23 expressed HML2 loci no clear expression pattern can be found. While in some patients single loci produce the majority of reads in a condition (e.g. 3q12.3 in patient 1 normal), these events are limited to a patient and do not occur as patterns (compare figure 15).

Of the 22 expressed HML6 loci only 3 to 5 make up a notable proportion (>10%) of reads per patient. Besides 19p12 and 11p15.4 these loci vary between the four analysed subjects. A strong difference in expression between conditions can only be seen in patient 1, where locus 11p15.4 is responsible for twice as much percentage of reads in the normal than the tumour tissue and in the overall expression of patient 11. Here, we can nearly find one locus associated with each of the three analysed tissue types: 3p21.31 for normal, 19p12 for cancer and Xp11.21 for lymph tissue. In all cases the denoted locus produces



(a) patient 1, HML2

(b) patient 11, HML2

(c) patient 21, HML2

(d) patient 33, HML2

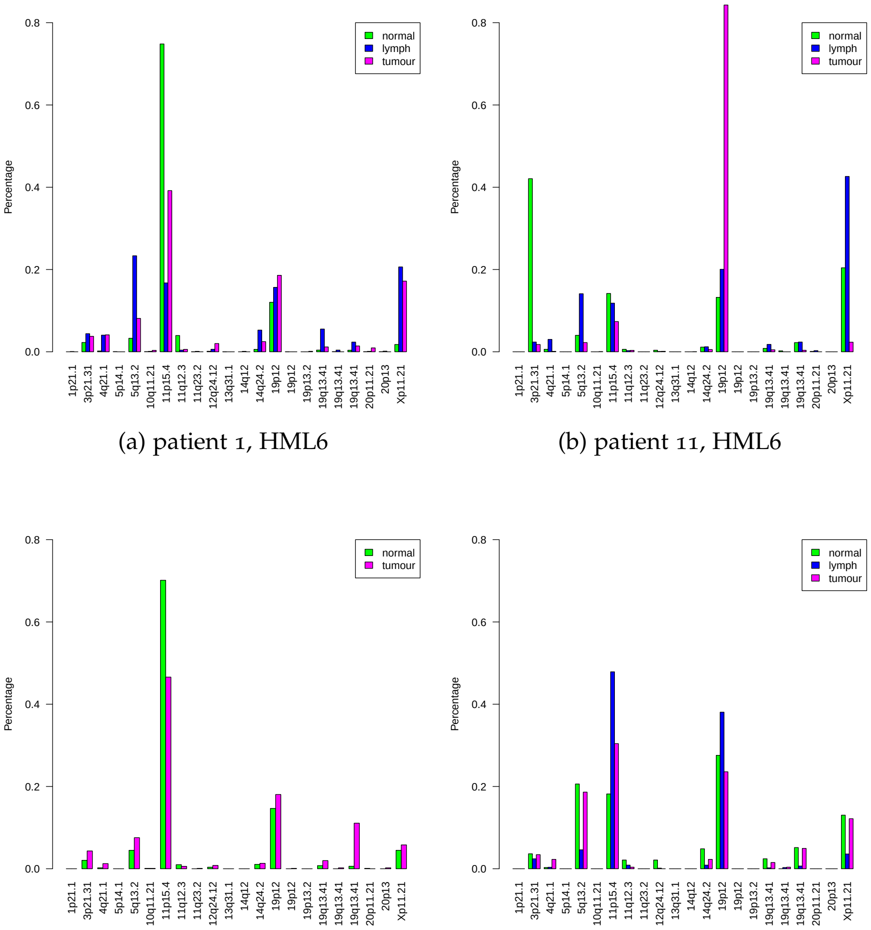
Figure 15: Distribution of expressed HML2 loci across chromosome bands for BLAT alignments.
The x-axis contains all HML2 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.

the majority of reads in the respective tissue and is much less involved in the other two tissue types. However, this association can only be observed in patient 11 and thus, does not seem to be an universal pattern.

In patient 11 each analysed tissue has a preferred HML6 locus

8.5 UNMAPPED UROTHELIAL READS HAVE TOO MANY MISMATCHES

We repeated the Blat approach, used to map the mammary tumour reads to the genome, in order to identify why we also lost approximately 35% of the urothelium reads. We found, unsurprisingly, that the number of unambiguously aligned reads increases with decreasing fraction of the read involved in the match and increasing allowed mismatches. But it is striking that with a requested read coverage of 40% up to 90%, the read classification solely depends on the number of allowed mismatches. It can also be seen that if we require a minimum of 40% of a read to be involved in the match, no more than 5% of reads are aligned ambiguously. Thus, the reads we are not able to map with our initial approach seem to have too many mismatches compared to the reference rather than having to many possible origins (compare figure 17).



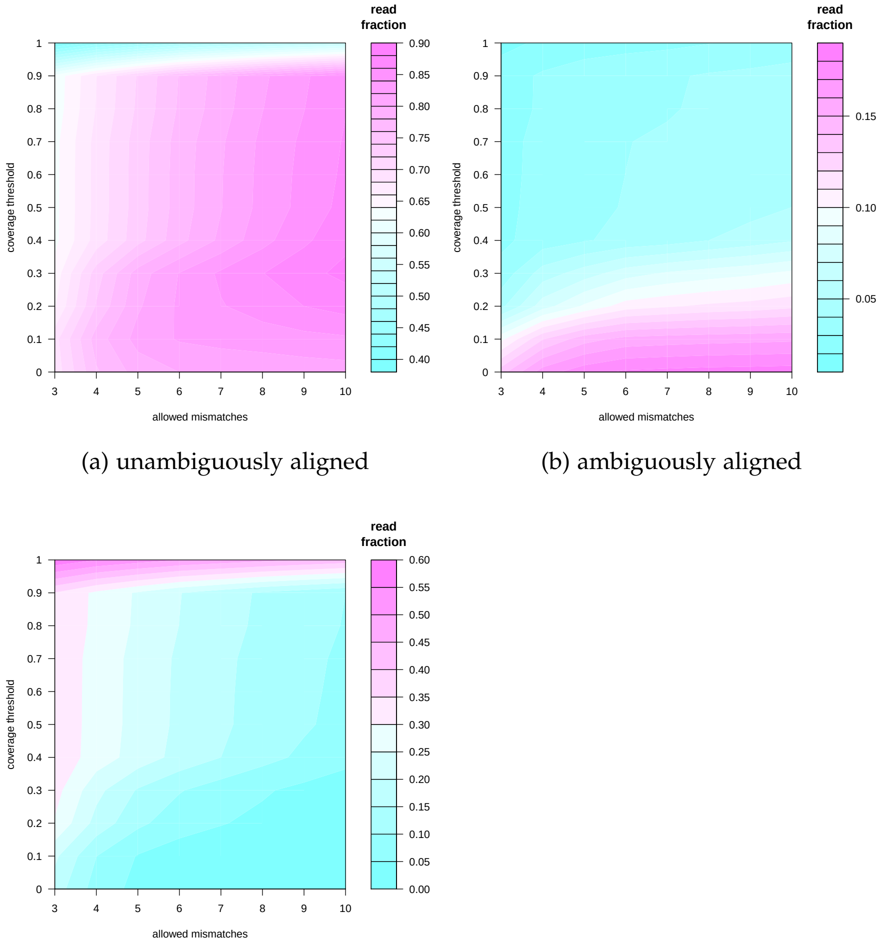
(a) patient 1, HML6

(b) patient 11, HML6

(c) patient 21, HML6

(d) patient 33, HML6

Figure 16: Distribution of expressed HML6 loci across chromosome bands for BLAT alignments.
The x-axis contains all HML6 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.



(a) unambiguously aligned

(b) ambiguously aligned

(c) not aligned

Figure 17: Fraction of urothelium reads being unambiguously, ambiguously or not mapped.
The three classifications are assigned depending on the allowed number of mismatches and the minimum fraction of a read involved in a match.

9 CONCLUSION

We analysed HERV expression in urothelium and mammary cancer tissue, as well as in matched normal and, for the breast study, lymphoid samples. In both cases specific HERV families, which had previously been shown to play a potential role in tumourigenesis were extracted, amplified and sequenced.

We observed certain difficulties with the laboratory processes, as barcoding was complicated by the capability of the long primers to form secondary structures, which could hinder them from binding to the target sequences efficiently. Furthermore, the multiple amplification steps in PCR and the sequencing itself introduced biases into the read distribution among patients, conditions and HERV families, although their initial concentrations were deemed to be identical. Another problem arising from amplification is the high percentage of chimeric reads, found especially in the breast cancer analysis. These are probably produced by template switching due to the high similarity of the analysed HERV loci.

We could also observe that the chosen sequencing technique yielded extremely different results for the mammary and urothelial library. Given that the samples were treated equally before they were send off to the sequencing company, we can only assume that different settings or reagents were used on their end that caused the much worse quality in the breast cancer study.

To identify the genetic origins of the produced reads, we obtained and extended the HERVd annotations, containing all known HERV loci on the current human genome assembly. We found this resource to be very comprehensive and only lacking some LTR sequences which do not contain internal HERV elements (MaLRs).

When trying to map our obtained reads to the genome, we found that only an unsatisfying low amount could be placed when enforcing a unique positioning and a maximum of 2% or 10 mismatches per read. After analysing the reason behind this, we found that mostly too strong divergences of our reads from the hg19 reference were causative and only a small percentage was rejected because it could be mapped equally well to multiple origins. One of the reasons for the mismatches were the discussed formation of chimaera during library construction.

In our final comparison of loci usage within the same HERV family between patients and conditions, we found that while in some cases we could see a switch in the most commonly used loci in a patient between healthy and tumour sample, most evidence pointed at a very individual expression pattern of HERVs. Usually the variation in expressed HERV loci depended more on the currently considered patient than the disease state.

10 OUTLOOK

10.1 IMPROVE SEQUENCING PROCESS

There are plans on repeating the described project, but applying a newer sequencing technology. We plan to use the real-time sequencing approach developed by pacific biosciences [47] which has the great advantage over pyrosequencing that it produces much longer reads (on average 10,000 to 15,000 nt as of 10/2014 press release) and that no additional replication step is needed in the sequencing process, thus hopefully decreasing the amount of produced chimaera. All reads are created as circular molecules, merging the fragments in forward and reverse orientation by artificial connecting pieces. Through this design the RNA polymerase, once attached, keeps iterating over the same read multiple times, on both strands. Although the error rate for each base is still around 15%, by calculating a consensus sequence from the multiple readings of the same position, an overall accuracy of over 99% can be reached [29].

We will probably be able to map much more reads to the genome when using this technique. The rejected reads due to mismatches will be reduced through the increased accuracy and lower amount of chimaera and rejected reads due to ambiguity will be reduced by increased read length. A higher number of mapped reads could help us to identify potential expression patterns on a more solid statistical basis.

10.2 HERV EXPRESSION MAP ACROSS VARIOUS CELL TYPES

As many of the studies linking HERVs to carcinogenesis do so by reporting strong expression levels of viral elements in cancer tissue but not necessarily comparing them to normal counterparts, we wanted to get an overview of the background expression of HERVs in various cell types. We are only able to define over- or under-expression of a certain locus, when we have a clear picture what its normal level of transcription is.

Although there are many whole RNA-Seq studies deposited in public databases which we could in theory use easily for these analysis, we were unsure whether HERV expression would provide a strong enough signal for detection when admixed with the rest of the human transcriptome. In an initial test we checked for eight whole RNA-Seq cancer studies from the gene expression omnibus (GEO)[46], at what number of reads in a dataset a saturation of detected HERVs would be reached. Therefore we counted the number of HERVs found in all complete sets and then artificially lowered the coverages in each sample with the tool *downsampleSam* from the picard tools package [141]. It iterates over all reads in the original file and keeps every one with a predefined chance (read pairs are discarded together). After we decreased the initial amount in nine steps down to 10% we could observe that no saturation with regard to HERV detection was reached, but instead followed a nearly linear increase (compare figure 18). Thus, we decided to work on the largest available whole transcriptome data sets as they would result in the most comprehensive analysis of HERV loci.

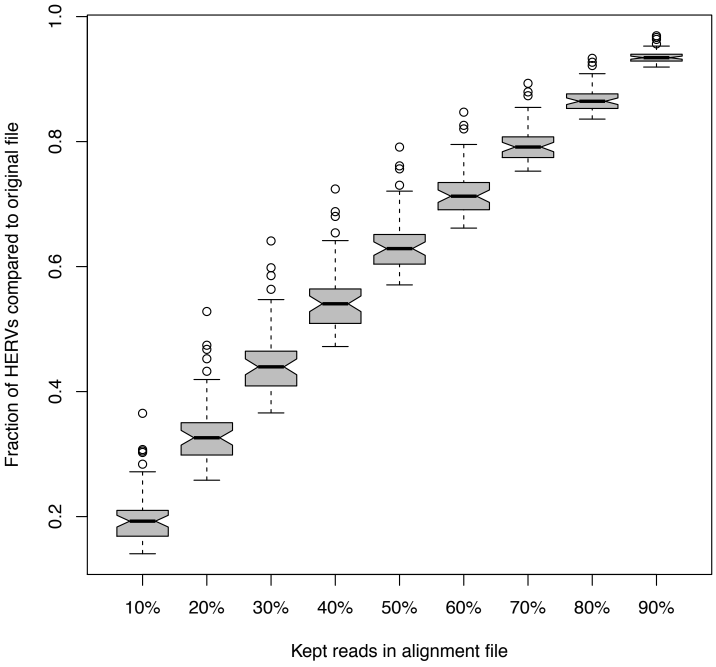


Figure 18: Identified expressed HERVs in down-sampled cancer studies.

Based on this preliminary examination, we decided to analyse HERV expression in the different cell lines available in the Encyclopedia of DNA Elements (ENCODE), as it is the largest whole transcriptome RNA-Seq compilation across various cell types to date. The description of this study can be found in part iii of this thesis.

Part III

DIFFERENTIAL EXPRESSION OF ENDOGENOUS VIRUS SEQUENCES IN ENCODE RNA-SEQ DATA

In this part of the thesis we describe a systematic study of HERV expression patterns in a multitude of healthy and cancerous cell types. We present a comprehensive differential analysis of HERV expression based on ENCODE Tier 1 and Tier 2 RNA-Seq data produced by Cold Spring Harbor Laboratories and the California Institute of Technology. Our study focuses on analysing the comparability of the different laboratories contributing to ENCODE and the expression patterns of HERV elements among tissues.

Excerpts and figures from this part of the thesis have been published previously in the following form:

> PAPER Haase K, Mösch A, Frishman D: Differential expression analysis of human endogenous retroviruses based on ENCODE RNA-seq data. *BMC Medical Genomics*, 8(71), 2015

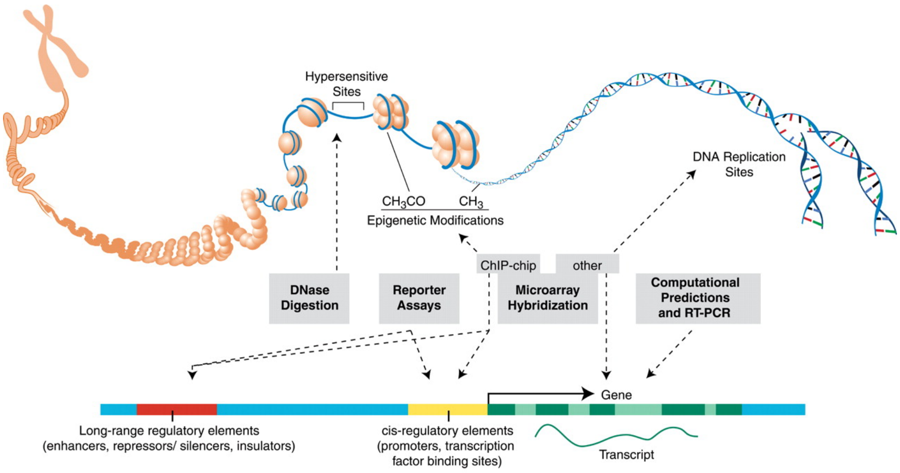
11 INTRODUCTION

11.1 ENCODE

In September of 2003 the National Human Genome Research Institute (NHGRI) launched a research consortium named Encyclopedia Of DNA Elements (ENCODE) [182] with the declared goal of identifying all functional elements in the human genome.

The Encyclopedia Of DNA Elements (ENCODE) wants to identify all functional elements in the human genome

The project started with a pilot phase that only focused on a small portion of the genomic sequence (30 Mb ≈ 1%) in order to find optimal suited methods to analyse multiple aspects of the genome. The different tasks were distributed internationally among laboratories and computational experts to combine various expertise [182]. At the end of the pilot phase a solid toolbox was developed that makes it possible to comprehensively analyse genomic data and search for its functional elements in a high-throughput manner. A great public benefit of the ENCODE project is that all produced data that adheres to the quality standards, is released into freely available databases and can thus be accessed by the interested research community. Among the experiments of the pilot phase were microarray hybridisation and reporter assays which helped to identify regulatory elements, epigenetic modifications and replication sites. For an overview over the different assays and the targeted functional elements see figure 19, taken from the pilot phase's publication.



Hypersensitive Sites
CH_3CO
CH_3
Epigenetic Modifications
DNA Replication Sites
ChIP-chip
other
DNase Digestion
Reporter Assays
Microarray Hybridization
Computational Predictions and RT-PCR
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

Figure 19: Functional genomic elements being identified by the ENCODE pilot phase.
Taken from The ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636-640, 2004. Reprinted with permission from AAAS.

The pilot phase alone helped to uncover many previously unknown functionalities in the human genome. One of the major findings was how much of the genome is really associated with a transcript which, contrary to former beliefs, was the majority of all bases. Furthermore, many non-coding transcripts could be identified, showing the strong regulation between genomic elements. Another finding was that histone-modification status was directly linked to the transcription activity or silencing. [18]

With the introduction of more powerful sequencing technologies, the possibilities for identification of functional elements increased and new experiments were added to the ENCODE repertoire. The second major ENCODE publication already describes results based on 1,640 datasets from 147 different cell types. Because the more recent studies did not focus on a small subset of the genome anymore, statistics for the complete genome could be presented. They found that 99% of the genome has a distance of less than 1.7kb to the closest biochemical event analysed by ENCODE [183].

ENCODE's second phase showed that 99% of the genome is in $<$ 1.7kb vicinity of a biochemical event

The second phase has added multiple new methods to the ENCODE experiment list, most notably sequencing approaches that take advantage of the next generation technologies, such as RNA-Seq and ChIP-Seq. The overall 25 kinds of experiments are applied to a defined set of cell types in order to facilitate comparisons and integration of the results from different contributing laboratories. The Tiers have associated levels of priority, starting with the most important Tier 1 which only contains three cell types: K562, erythroleukemia cells, GM12878, a B-lymphoblastoid cell line and the H1 embryonic stem cell (H1-hESC) line. In order for a group to participate in the ENCODE project, submission guide lines have to be fulfilled. For the sequencing experiments these state, among other things, that at least two biological replicates have to be provided which enables a more statistically robust downstream analysis.

The launch of the modENCODE project [30] which focuses on the identification of functional elements in the prominent model organisms C. elegans and D. melanogaster makes it possible to apply very detailed comparisons of the distantly related species human, fly and worm. So far, large scale studies looking into the different transcriptomes [58] and chromatin organisations [75] have been published. Furthermore, in 2012 a parallel project to the successful human ENCODE was introduced, mouse ENCODE [173]. Its goal was to provide the same level of high sensitivity, integrated genome analysis as in the human consortium to decode all functional elements in the commonly used biomedical research model organism mus musculus. The first results of mouse ENCODE were presented in the end of 2014 and showed that many mouse orthologs to human genes showed considerable distinct expression profiles while the chromatin states were highly similar [206]. Additionally to the studies conducted by the EN-

CODE associated consortia, there have also been external groups that used the provided resources to integrate or compare them to their own samples or conducting meta-analyses across ENCODE cell lines [39, 71, 133].

With the most recent data release, made available in September 2014, ENCODE incorporates 27 different kinds of experiments, conducted by 26 research laboratories. There are 437 RNA-Seq studies for human available, which in contrast to previous phases, now have also been performed on tissue samples, in addition to various cell lines. When we began our study there were a total of 151 RNA-Seq experiments available, with 87 of them using small RNA-Seq (as of September 2014).

11.2 HERV DIFFERENTIAL EXPRESSION ANALYSIS

As mentioned in section 6.2, multiple studies have linked HERVs to diseases, most importantly cancer, by identifying transcripts of retroviral origin in affected patients. However, so far we are lacking a comprehensive analysis of HERV expression in healthy tissues in order to know the base line of expressed HERV transcripts in an organism. Even very early studies that only focused on a single HERV family found a strong variation in expression between individuals and tissue types and thus suggesting regulation of endogenous retrovirus activity [4].

Previous attempts to create an overview of HERV expression patterns in different tissues relied on a specifically designed microarray with 52 representative captured retroviral *pol* sequences from 20 major HERV families [168] and were thus limited to the subset of HERV family members that still contain intact *pol* sequences. Nevertheless, the study could show HERV expression in all of the 19 analysed healthy human tissues. Some of the captured families were identified in nearly all tissue while others are restricted to a small subset. [169]

HERV expression has already been shown to exist across multiple tissues, different disease states and seems to be regulated

In order to increase the number of covered retroviral elements, a more comprehensive approach, capable of identifying the full-length sequences would be required. RNA-Seq has become a method of choice for addressing such problems [197] as it provides precise measurements of transcript levels in the cell and thus makes it possible to map all retroviral elements, both structurally intact and partial, back to their genomic loci. However, such an expression analysis is complicated by the fact that HERVs are repetitive elements spread over the entire genome, which makes mapping of their transcripts to genomic loci particularly challenging and requires an as exact as possible method for read alignment.

ENCODE data has been used in multiple computational studies on gene expression, but as of now transcriptome analyses spanning multiple cell types aim at protein-coding genes or functional regulatory RNAs [8, 70] while overlooking everything classified as repetitive element. Examination of ENCODE RNA-Seq data with regard to HERV expression has either been limited to single cell types [116] or covers HERVs only as a very small subset of the overall analysis, usually categorising them together with other transposable elements [37].

11.3 MOTIVATION

In part ii of this thesis we saw that expression of loci from certain HERV families could vary quite a lot when compared between different individuals. However, unlike many publications linking HERVs to cancer were proposing, no clear association of HERV expression and disease state could be observed.

Our findings were in line with a publication from Flockerzi et al. [54] who reported that when comparing expressed sequence tags originating from loci of the HML2 family, some contributed more to the overall amount of transcripts than others and their distribution among tissues was different. But the authors also conclude that expressed sequence tags are not sufficient to capture all of the transcriptional activities of HERVs and that "A specialised (H)ERV Transcriptome Project is needed".

Need for a HERV Transcriptome Project has already been formulated

Thus, we attempted to comprehensively analyse RNA-Seq data covering all annotated HERV loci in a broad variety of cell lines, disease and developmental stages. We chose ENCODE to provide the used RNA-Seq data as it contains various cell types and its studies all adhere to quality guide lines that enable optimal comparisons. We sought to gain an insight into the overall expression patterns of HERV elements and to examine on a large scale if there are measurable differences in HERV activity between cancer and normal cells, as already reported for individual tumor types. Furthermore, a major goal of our study was to assess the consistency of different ENCODE-contributing laboratories with regard to expression values from the same cell lines.

12 METHODS

12.1 RNA-SEQ DATA

When deciding which subset of the vast amount of data deposited within ENCODE to use, we followed the importance ranking of cell types into tiers. The first two Tiers, regarded as most important by the project, contain only three and 14 cell lines, respectively but have been used in a multitude of experiments. The third Tier contains a large number of cell types, but for most of them only one to a handful of experiments are available. Thus, we decided to limit our approach to the RNA-Seq data from Tier 1 and Tier 2.

While it is possible to download the raw reads produced in each sequencing run, we decided to obtain the ready-made alignments in bam-format. The advantage of this format is that we do not need to execute the mapping step ourselves and thus save a lot of computing time. Furthermore, we can presume that the applied quality measures and mapping processes are already very sophisticated due to ENCODEs quality standards. The files contain the alignments of all reads mapped against the latest human genome assembly (hg19) and were obtained using the UCSC track download portal [158, 152].

Only samples from the ENCODE category long RNA extracts (>200 bp) were considered, as short RNA extractions aim at identifying small non-coding RNAs while our proviral remnants of interest are considerably longer (mean length of 928 nt in our HERVd compilation). We further restricted the considered tracks to whole cell extracts, as we are interested in the overall analysis of HERV expression in entire human cells rather than in individual compartments. We focused on data produced by either the Cold Spring Harbor Laboratories (CSHL) or the California Institute of Technology (Caltech) (for complete list see table 17) because these provide the most comprehensive coverage of (mostly) the same cell types, facilitating a direct comparison of the results produced by these two groups.

We chose RNA-Seq data in bam-format, extracted from the whole cell with a length > 200 bp and produced by CSHL or Caltech for our analysis

Moreover, due to the fact that Caltech performed both single-end and paired-end sequencing on a subset of cell types, expression analysis results can also be compared between different library preparations.

12.2 HERV ANNOTATION

As basis for the HERV loci annotation in the human genome we used our extended HERVd database as described in section 7.2 of this thesis. From the initial 100,495 locations on hg19 we subtracted all entries

Cell Type	Tissue	Cond.	CSHL	Caltech	
				single-end	paired-end
A549	epithelium	cancer	2	-	-
B-cells CD20+	blood	normal	2	-	-
GM12878	blood	normal	2	2	2
H1-hESC	ESC	normal	2	2	4
HUVEC	blood vessel	normal	2	2	2
HeLa-S3	cervix	cancer	2	2	2
HepG2	liver	cancer	2	2	2
IMR90	lung	normal	2	-	-
K562	blood	cancer	2	2	2
MCF-7	breast	cancer	2	-	3
Monocytes CD14+	monocytes	normal	2	-	-
SK-N-SH	brain	cancer	2	-	-

Table 17: ENCODE cell lines used in analysis.
Shown are the ENCODE RNA-seq tracks that remain when filtering for the features *view, alignment format, cell sample, RNA extract, cellular compartment* and *producing laboratory*.
The numbers in the last three columns denote the amount of replicates available for the corresponding sample.

positioned on chromosome Y, as this chromosome is not covered by all ENCODE datasets used in our study and would lead to artificial differential expression results when comparing Y containing sets against the others. This filtering step left us with a total of 98,998 annotated HERV loci for which we could obtain read counts.

12.3 HERV EXPRESSION IN ENCODE RNA-SEQ

We calculated the read coverage over the HERVd entries for every RNA-Seq experiment using the *featureCount* tool of the subread package [111]. For every annotated viral element it returns the number of reads mapped to it in every analysed sequencing run. The program was executed with the *primary* option, which forces featureCount to only take primary alignments into account, thus avoiding biased expression values through non-uniquely mapped reads.

12.3.1 *Differential expression analysis*

The coverage depth of HERV loci between the 25 ENCODE samples (compare table 17) was compared using the R bioconductor [149, 57] package *DESeq* [3] which is specially designed for differential expression analysis. To achieve a better comparability between samples we normalised their count data by library size and carried out a variance stabilising transformation based on the inherent biological variability between the replicates of the same condition. We then performed a principal component analysis (PCA) and calculated the Euclidean distances between the transformed expression values to detect overall differences between the samples.

The following analyses were limited to the paired-end RNA-Seq data as PCA revealed extensive differences between single- and paired-end library preparations. Hence, to avoid introducing a bias in the differential expression analysis, we excluded single-end data. The read count value of every condition, normalised by the library size, was compared in a pairwise fashion against every other condition, resulting in 171 differential expression analyses and the corresponding fold changes. The DESeq implementation of the negative binomial test was than used to find significant differences in the calculated expression values. The initial p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure [13].

In order to identify significantly differentially expressed HERVs we filtered for loci whose absolute logarithmic fold change was at least one and whose adjusted p-Value did not exceed 0.001 (which is equivalent to a false discovery rate of 0.1%). For every analysed cell type, we compiled a list of HERV loci up-regulated in at least one of the pairwise comparisons. By considering the corresponding families of loci, we sought to identify HERVs that are particularly active in certain cell types and under certain conditions.

We regarded HERVs as significantly differentially expressed when their logarithmic fold change > 1 *and the adjusted p-Value* < 0.001

12.3.2 *Validation on housekeeping genes*

In order to ascertain that the differences in HERV expression between different conditions and library preparations reported in this study are not due to computational or experimental biases specific to endogenous viral elements, we repeated our analysis with a set of housekeeping genes.

For this purpose we used the list of 3,804 genes compiled by Eisenberg and Levanon [48]. This list was created based on RNA-seq data from 16 different human tissues by first identifying housekeeping exons, i.e. those exons expressed in all tissues, displaying low variance between tissues, and showing no exceptional expression in any single data set. Housekeeping genes were then defined as those genes, for which at least one annotated RefSeq [147] transcript has more than half of its exons classified as housekeeping.

When acquiring the annotation file for the housekeeping genes from the UCSC genome browser, only 3,801 entries could be retrieved, as three identifiers (NM_032937, NM_003926, NM_032560) had been removed from the RefSeq database. The assessment of coverage for every housekeeping gene in all our selected ENCODE data sets was carried out exactly as described above for HERVd (section 12.3.1), including normalisation and principal component analysis.

13 RESULTS

13.1 Differential expression analysis shows strong differences between sequencing technologies

The principal component analysis of the transformed expression values of all 25 samples shows that the strongest differences in HERV expression result from variations of the sequencing technology used.

The first principal component (which captures 19.06% of the total variance) clearly subdivides the Caltech samples into those analysed by single- and paired-end sequencing. The second and third principal components, accounting for 10.30% and 5.70% of the variance, respectively, separate certain cell types from the rest of the datasets. While the second component clusters together all eight embryonic stem cell samples, the third component lets the six K562 samples stand out. Finally, the fourth component (5.01% of the variance) mostly reflects the differences between normal and cancer cell types, with the exception of the MCF-7 sample from Caltech where one replicate is clearly separated from the other two and overlaps with a normal tissue (HUVEC). For a visualisation of the PCA results, see figure 20.

Given that single-end datasets lead to very different results compared to their paired-end counterparts, we decided to exclude them from the further analyses to prevent them from introducing a bias into the expression data. The paired-end data seems to give a better overview of the expression rates. Note that in the GEO summary of the ENCODE Caltech RNA-Seq data, the single-end protocol which is also strand-specific, is described as less reliable for quantification [28].

Single-end sequenced datasets are not reliable for quantification and are excluded from further analysis

13.2 Housekeeping genes confirm strong differences in ENCODE expression data

Differences in expression of housekeeping genes were also mostly due to the used library protocols. While the first principal component (accounting for 30.67% of the variance) primarily divides the CSHL cell lines into cancerous and normal ones (with the exception of CSHL's GM12878 and a small overlap involving SK-N-SH), the samples provided by Caltech are neatly separated into paired- and single-end protocols (compare figure 21).

The observed differences thus do not depend on the chosen transcript family, but are rather indeed an inherent pattern in the ENCODE datasets.

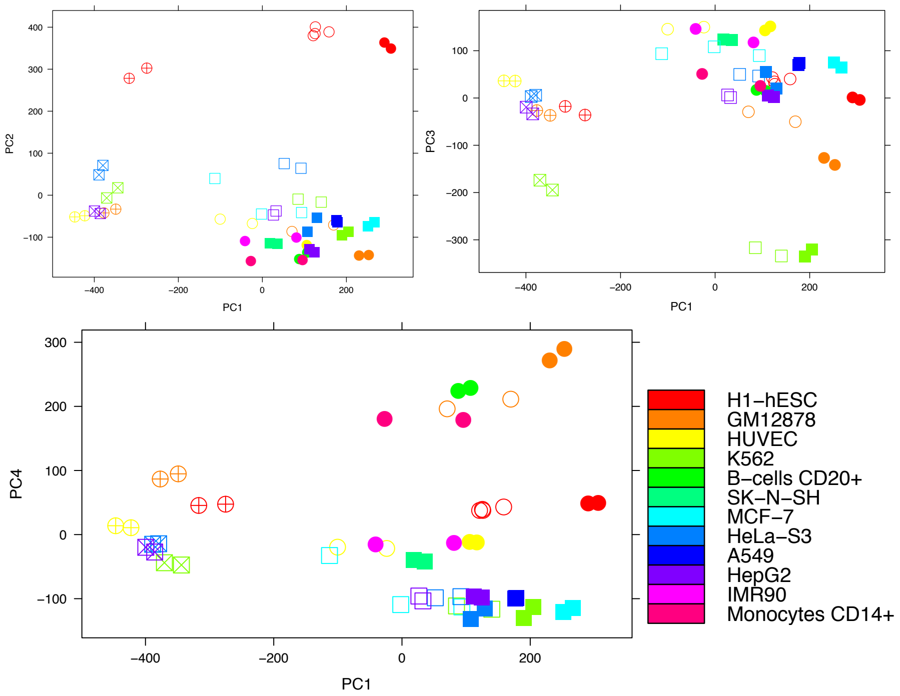


Figure 20: Principal component analysis of normalised HERV expression values in 25 ENCODE RNA-Seq data sets.
The first, second, third, and fourth principal components account for 19.06%, 10.30%; 5.70%, and 5.01% of the variance, respectively. Each data point represents one replicate.
Circles: normal cell lines; squares: cancer cells; filled symbols: CSHL samples; empty symbols: Caltech; crosses in the symbols: single-end sequencing.
Upper left panel: The first component separates single- and paired-end libraries, while the second separates ESC samples from all the others.
Upper right panel: The third component separates K562 from the other cell types.
Bottom panel: The fourth component divides samples into normal and cancerous cell types.

13.3 TWO CELL LINES SHOW BEHAVIOUR ATYPICAL FOR THEIR DISEASE AND DEVELOPMENTAL STAGE

After excluding the single-end Caltech samples, we calculated the Euclidean distances between the transformed expression vectors for all remaining paired-end datasets and performed a hierarchical clustering. As seen in figure 22, most replicates are highly similar to each other with regard to HERV expression, with the exception of IMR90 and the second replicate of Caltech's MCF7 sequencing. ESCs are clearly the most diverse among the differentiated cell types, serving as an out-group. Particularly striking is the clustering of CSHL's GM12878 replicates. While the same cell line, analysed by Caltech, is branched together with two other healthy blood cell types, GM12878's

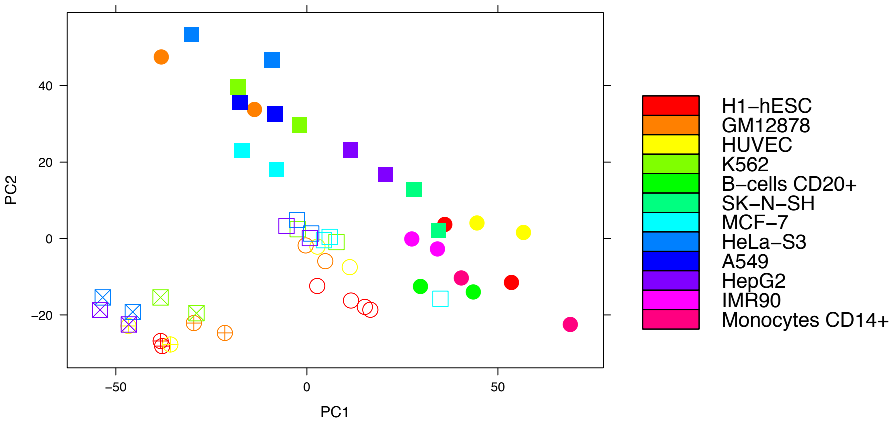


Figure 21: Principal component analysis of normalised expression values of housekeeping genes in 25 ENCODE RNA-Seq data sets.
The first and the second principal components account for 30.67% and 21.33% of variance, respectively.
Circles: normal cell lines; Squares; cancer data sets; Filled symbols: CSHL; Empty symbols: Caltech; crosses in the symbols: single-end sequencing.

expression vectors determined for CSHL cluster with all four examined blood cancer samples (K562).

CSHL's GM12878 cell line clusters with blood cancer samples instead of other datasets analysing the same cell type

In addition to the global clustering presented above we also compared every condition against every other condition in a pairwise fashion, leading to 171 differential expression analyses. Following the common practice [56, 115] we defined significantly differentially expressed HERV loci as those with an absolute value of the logarithmic fold change of at least one and the adjusted p-value smaller than 0.001 (FDR of 0.1%).

The number of identified loci varied depending on the compared samples. Pairwise analyses involving embryonic stem cells led to the largest number of differences with up to 956 significant loci (Caltech H1-hESC vs. CSHL HepG2). The smallest number of significantly differentially expressed loci between two different cell types is three and occurs three times (Caltech and CSHL HUVEC vs. IMR90 and Caltech MCF-7 vs. CSHL HeLa-S3).

While the largest numbers of differentially expressed HERVs are seen when ESCs are compared to the other cell types, it is remarkable that Caltech's HeLa-S3 line shows very few loci, which are significant in comparison with both H1-hESC samples (compare figure 23).

13.4 CALTECH'S HELA-S3 CELL LINE SHOWS A STRONG UP-REGULATION OF HERVH

Upon extracting all significantly over-expressed loci per cell type from the pairwise comparisons and grouping them by their family affiliation, we observed characteristic patterns for different sam-

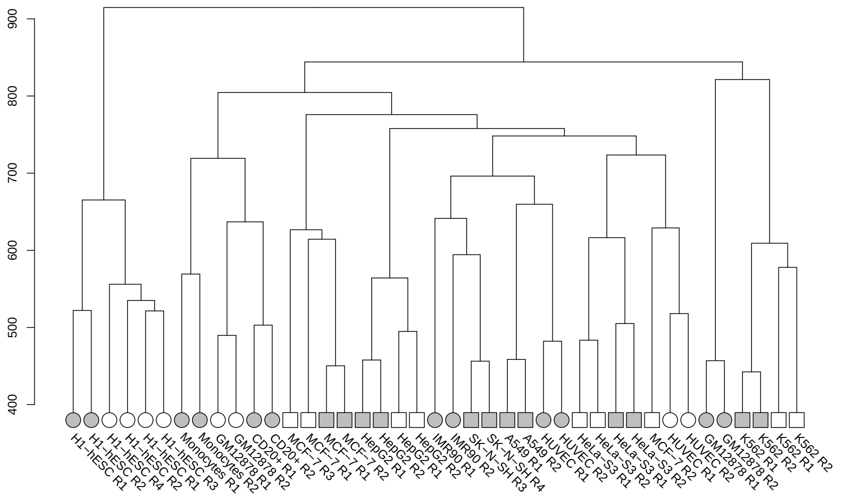


Figure 22: Hierarchical clustering of Euclidean distances between the transformed expression values in paired-end ENCODE data sets. *Rn* in the dataset identifier represents the replicate number n. Circles show normal, squares cancer cell lines. Samples analysed by CSHL are drawn with filled symbols, those from Caltech are drawn with blank symbols.

ples. The eight most often up-regulated HERV families are four internal and four LTR sequences. Internal regions that are most overexpressed in the pairwise comparisons are ERVL, a very old endogenous element that is found even outside of primates and shows high similarity to foamy retroviruses [36] and its younger relative HERVL.

The most commonly up-regulated family is ERVL

Additionally, two LTR sequences, LTR16C and LTR33, are also among the eight most often over-expressed families, which belong to the ERVL superfamily. The other two internal families are the pluripotency marker HERVH and the young human endogenous viral element, which is also the most active in terms of expression and transposition, HERVK [187]. The two additional often up-regulated LTR sequences are LTR7, which is a long terminal repeat sequence of HERVH, and LTR12, belonging to the HERV superfamily ERV1.

Overall, the same cell types analysed by any two laboratories show a similar composition of HERV families, with the exception of GM12878 from CSHL, which exhibits a nearly double amount of significant HERVs compared to its Caltech counterpart. Especially the large number of ERVL members in the CSHL sample is unmatched in the corresponding Caltech cell line. The only other cell types with a similar large number of active HERVL loci is CSHL's K562 sample, which is also a blood cell type but, contrary to GM12878, cancerous.

Another cell line that exhibits an extremely deviant behaviour in the laboratory comparison is HeLa-S3 from Caltech. It appears to

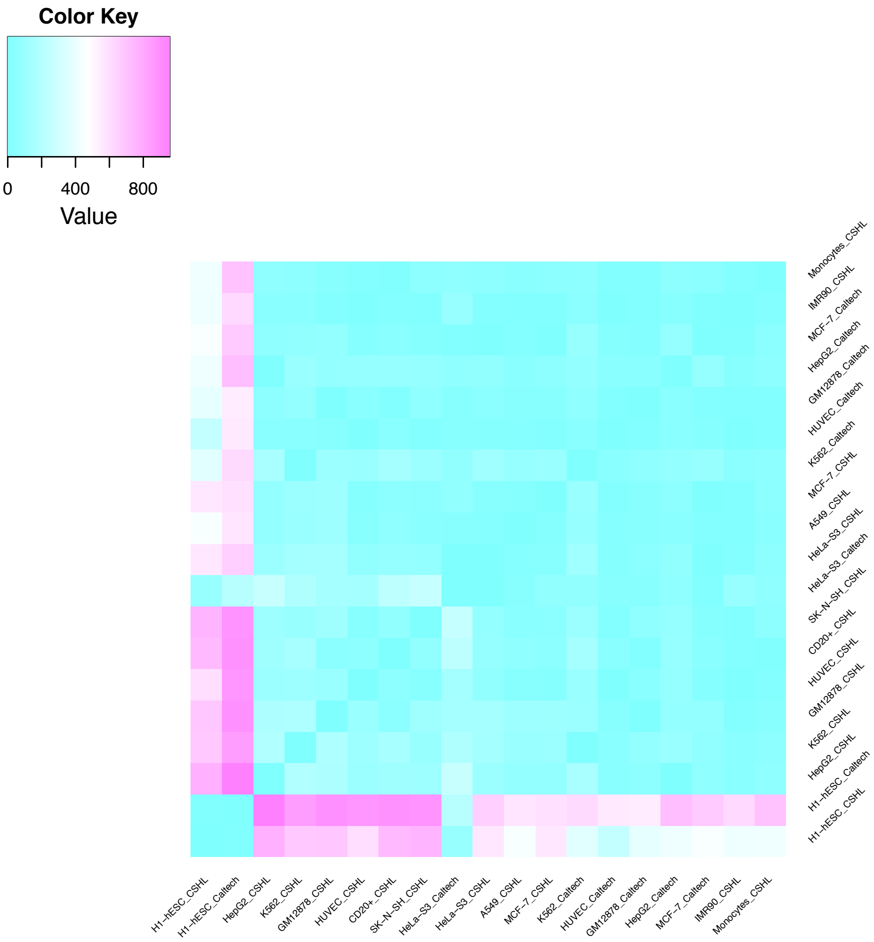


Figure 23: Heatmap of significantly differentially expressed HERV loci. Number of significantly (|$\log_2$ fold change| at least one, adjusted p-value < 0.001) differentially expressed HERV loci in pairwise comparisons.
Ice blue fields show comparisons yielding none or very few differentially expressed loci, hot pink represents large numbers (up to 956).

over-express an immense amount of HERVH family members (290 loci), which are only found in low numbers in all other specialised cells. The over-expression is not as strong as in the embryonic stem cells, but is higher than any other number of a single HERV family in all other cell lines (compare figure 24 and table A5).

Caltech's HeLa-S3 cell line over-expresses an unusual amount of HERVH loci

The difference between cancerous and normal cell lines revealed by the principal component analysis could not be linked to a particular over-expressed HERV family. We were not able to identify any expression patterns separating the six normal from the six cancer cell lines on the basis of individual HERV families.

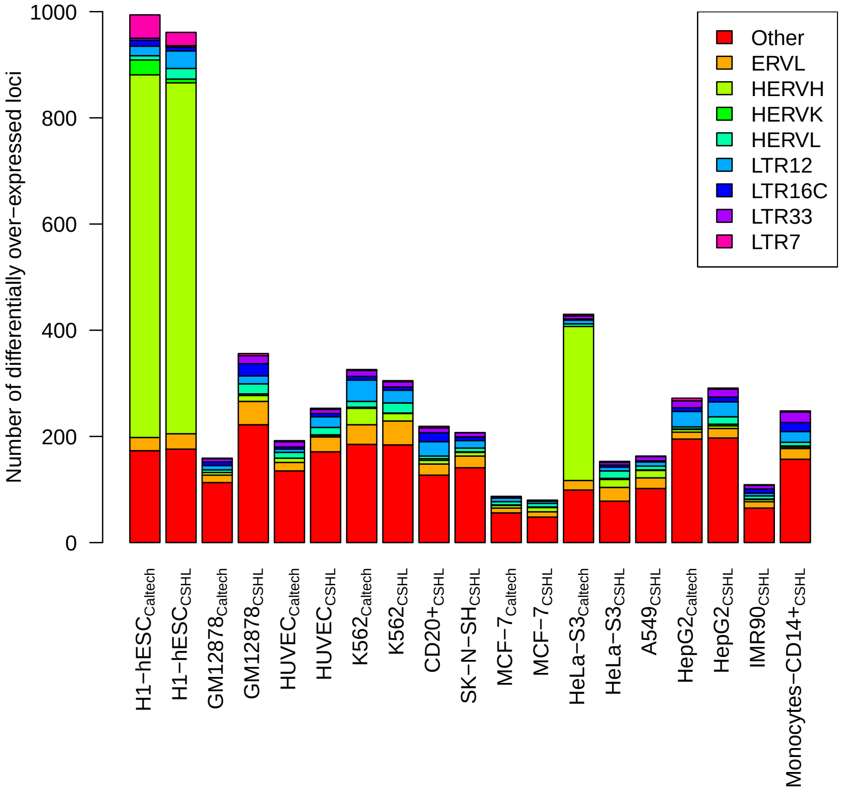


Figure 24: Up-regulated HERV families in different cell types.
The plot shows HERV families that are significantly stronger expressed when comparing the indicated cell type against all others. All families that have fewer than 20 members significantly over-expressed in all samples are grouped together in the *other* class. The exact numbers can be found in table A5.

14 CONCLUSION

In this study we analysed the expression of known human endogenous retroviral elements in RNA-Seq samples from the ENCODE project. It is the first comprehensive examination of sequencing data from a multitude of cell types and laboratories with regard to HERV expression patterns.

Based on the analysis of 25 RNA-Seq samples from the ENCODE project with regard to their HERV expression we find that datasets created with different sequencing library methods (paired- vs single-end) are not very easily comparable, because single-end samples achieve less coverage. This is expected, as the sequencing technique used by Caltech is strand specific and as such trades quantification against a qualitative analysis.

Although a principal component analysis of the HERV expression patterns in different cell types revealed the possibility to distinguish cancerous from healthy samples based on HERV activity, we could not link this difference to a specific HERV family. However, our study revealed unexpected results regarding GM12878 from CSHL, which showed hints of being a tumorous cell line in two different analyses. First, a hierarchical clustering of all HERV loci expression grouped this cell type with all four K562 (blood cancer) replicates instead of the GM12878 samples from CSHL. Second, the composition of upregulated HERV families in this sample, when compared to all others, is much more similar to that of K562, especially regarding the strong activity of ERVL. A possible reason for this behaviour could be the transformation of an initially normal cell line to a tumorous one prior to experimental measurements. However, this explanation does not seem to be particularly plausible given that ENCODE imposes strict data quality requirements, especially with regard to Tier 1 cell lines to which GM12878 belongs.

CSHL's GM12878 cell line shows signs of being cancerous

The respective Caltech GM12878 RNA-Seq track has been accessible through the UCSC genome browser [74, 96] since August 2012 and so far no unusual features of this dataset, including a possible progression towards a tumor line, have been reported. It is conceivable that the change in HERV expression detected in our study, which is the first comprehensive investigation of HERV expression in ENCODE samples, occurs very early in the transition from a normal to a cancer cell type and hence remained undetected in studies focusing on protein-coding gene expression, although we were able to detect aberrant behaviour hinting at this change when performing PCA on housekeeping genes. Further research is needed to verify this hypothesis, as it implies that unusual HERV expression could serve as an early indication of oncogenic transformation and thus represent a valuable diagnostic lead.

Another striking finding is the low amount of differentially expressed HERVs when comparing Caltech's HeLa-S3 sample to the ESCs. The strongest difference between HERV expression in ESCs compared against specialised cell types is the very strong up-regulation of HERVH family members. Because HERVH activity is also high in Caltech's HeLa cells, unmatched in any of the other differentiated cell types, the difference in expression pattern to ESCs is understandably small.

The HERVH family is known to play a vital role in embryonic stem cells. In particular, since they can serve as a marker for pluripotency due to their strong association with binding sites for the pluripotency transcription factors NANOG, OCT4 and SOX2 [163]. Furthermore, it has been suggested that HERVH and its LTR7 can recruit the transcription factors p300 and OCT4 to regulate the transcription of pluripotency-associated transcripts [116]. Intriguingly, Santoni et al. [163] also used ENCODE RNA-Seq data from Caltech to analyse HERVs in hESCs, although they relied on the 2010 data release whereas in this study we utilised the most recent data published in 2012.

Unusual high number of expressed HERVH loci in Caltech's HeLa-S3 cell line could be a sign for pluripotency

For comparison with differentiated cells, Santoni et al. [163] also obtained the 2010 data on corresponding HeLa-S3 cells and found that "HERV-H expression is barely detectable in HeLa", although it was identified when using transient-transfection assays [167]. It is thus apparent that there has been a significant change between the 2010 and 2012 HERV expression data submitted to the ENCODE project by Caltech.

Santoni et al. [163] observed that the HERV expression strength in ESCs diminishes during differentiation. Expression is highest at the undifferentiated N0 stage, still observable during N1 (early initiation), and only barely measurable during N2 (neural progenitor). Thus, a conceivable explanation for the behaviour of Caltech's HeLa-S3 cells would be reprogramming towards pluripotency, although an underlying mechanism for this process remains enigmatic.

Part IV

EXPITOPE: A WEB SERVER FOR EPITOPE EXPRESSION

In this part of the thesis we will describe the implementation of Expitope, a web server that provides a first in silico prediction of potential off-target effects of engineered T cell receptors used in immunotherapy. Our web server enables the user to search for a peptide of interest in all annotated protein sequences and returns exact and approximate matches together with their expression values in a multitude of healthy human tissues which have been calculated from RNA-Seq data. All results are scored based on the probability of the corresponding peptide being created by the different steps of epitope presentation. We expect our web service to aid the exclusion of potential targets with catastrophic side effects before entering the much more expensive experimental phase of target selection.

This work is the result of a collaboration with the Immune Monitoring Group at the Helmholtz Zentrum Munich and Medigene Immunotherapies GmbH a subsidiary of Medigene AG.

Excerpts and figures from this part of the thesis have been published previously in the following forms:

POSTER Haase K, Raffegerst S, Schendel DJ, Frishman D: Expitope: Web server for epitope expression, presented at the *European Conference on Computational Biology 2014*, Track: Bioinformatics of Health and Disease, Straßburg, France, 2014

PAPER Haase K, Raffegerst S, Schendel DJ, Frishman D: Expitope: a Web server for epitope expression. *Bioinformatics*, 31(11):1854–1856, 2015

15 INTRODUCTION

15.1 IMMUNE RESPONSE

The human immune system consists, like that of other vertebrates, of two major components: the innate and adaptive immune response. The innate immune system, also termed nonspecific immune response, is the first line of defence against intruding pathogens. Its function is mostly the recruitment of immune cells to the location of infection and triggering of signalling pathways through cytokines. The innate immune system is evolutionary much older than its adaptive counterpart, as it exists throughout multiple kingdoms [84]. While the innate response itself does not confer long lasting immunity or can invoke a memory response to reoccurring infectious agents, one of its functions in vertebrates is the activation of the adaptive immune response which can fulfil this tasks.

The adaptive immune system, also termed specific immune system, is able to produce memory cells which can induce a faster and stronger immune response upon any subsequent infection with the same pathogen, building the molecular basis for vaccination strategies. Alongside the increased speed of reaction time comes a high specificity tailored to the infectious agent. But the adaptive immune system is also at fault in diseases like asthma or allergies, when molecules introduced into the organism are wrongfully identified as precarious antigens. Furthermore, despite a training stage in the thymus where the cells of the adaptive immune response are trained to distinguish self- from foreign-antigens, disruptions of this process lead to autoimmune disorders, when the immune system starts to attack self-antigens. [84]

The major agents of the adaptive immune system are a class of white blood cells, called lymphocytes. One subclass of them, B-lymphocytes or B cells, activate immunoglobulin proteins which are also known as antibodies. These proteins are distributed in the organism via the bloodstream and are able to bind their specific antigen when encountering it, thus rendering it ineffective for binding to host cells. The second subclass of lymphocytes are T cells, the cell mediated adaptive immune response. T cells can directly react to antigens which are presented on host cell surfaces [1]. To facilitate the recognition, it is of importance that peptides specific for the invading pathogen are cleaved from their respective protein and processed by the cell machinery, so that they can be presented on the outside of the cell. The steps facilitating this presentation will be discussed in the following section.

15.1.1 *Epitope Processing*

While all T cells recognise cellular proteolysis products, there are differences regarding the origin of the proteins. CD8$^+$ T cells, also known as cytotoxic T cells, interact with the major histocompatibility complex I (MHC I) which presents foremost self-derived peptides, meaning cleaved cellular proteins. CD4$^+$ T cells, on the other hand, recognise the major histocompatibility complex II (MHC II) which mostly presents foreign epitopes derived from organism-invading pathogens.

CD8 and CD4 are surface glycoproteins expressed on the respective T cells and serve as co-receptor in MHC binding

The human genome encodes three genes for each complex, HLA-A, -B and -C for MHC I and HLA-DR, -DQ and -DP for MHC II. Striking is the extreme variance existing in all MHC loci, leading to a broad range of peptides which can be bound by different HLA alleles [84].

The MHC molecules form a so called binding groove, in which the presented peptide is confined by a β-sheet and two α-helices. While the MHC II structure allows the ends of the bound peptide to be unconstrained, in the MHC I complex the peptide is restricted on both ends, leading to mostly 8-10 amino acid long fragments being presented [20]. As for our Expitope web server the focus is on cellular antigens, in the following we will concentrate on the creation of MHC I epitopes.

In order for a T cell receptor (TCR) of a cytotoxic T cell to recognise a peptide, it has to be cleaved from its protein by proteasomal proteolysis, translocated from the cytosol into the endoplasmic reticulum (ER) and mounted onto a MHC I complex to be presented on the cell surface. Because in all these steps certain boundaries have to be observed, not every part of a random protein can serve as epitope. In rare cases exogenous peptides derived from pathogens can enter this pathway, when they get into the cell by means of endo- or phagocytosis. [20]

15.1.1.1 *Proteasomal cleavage*

Proteins not currently needed by the cell have to be degraded in order to be able to control their function and to reuse their components. Mainly two pathways are responsible for this degradation. Lysosomes and endosomes, on the one hand, decompose mostly extracellular and membrane proteins, thus this pathway produces peptides usually presented by the MHC II complex [122]. The other degradation pathway is facilitated by the proteasomal complex, which marks proteins to be degraded with ubiquitin and hydrolyses them via the 26S proteasome [73]. The proteasome is made up from three catalytic β subunits which each contains cleavage specificities. Nevertheless could it be shown that substrates can be cleaved after virtually every amino acid [200]. Most common substrates for proteasomal cleavage are defective ribosomal products (DRiPs) which are destroyed in order to prevent faulty protein functions that could lead to a diseased phenotype. The C-terminus is usually cleaved after a hydrophobic or positively charged side chain to fulfil the preferences of the downstream processing via the transporter associated with antigen processing (TAP) and MHC I [190]. While the C-terminus is in most cases defined by the proteasome, the N-terminus is left unconstrained by the cleavage and is only cut inside the ER [100].

15.1.1.2 *Transporter associated with antigen processing*

Because the previous degradation of proteins by the proteasome takes place in the cytosol, the produced peptides have to be transported into the ER where they are mounted onto the MHC I molecules. The responsible transporter to introduce the peptides into the ER lumen across the membrane is TAP. It belongs to the class of ATP binding cassette transporter which require ATP as energy source in order to fulfil their function. TAP binds the peptide with hydrophobic transmembrane domains and relocates it across the membrane through a change in conformation for which ATP is consumed. [72]

Certain peptides are more likely to be transported into the ER than others due to preferences of TAP for length and charge. Substrates between 8 and 16 amino acids in length and with a hydrophobic or positively charged C-terminus [124]. Most peptides are transported into the ER in an N-terminal elongated form and only cleaved to final length by ER aminopeptidases known as ERAP1 and ERAP2 after crossing the membrane [103],

Not all peptides which are presented by MHC I have to enter the ER via TAP. Some ligands, mostly signal sequences reach the MHC biosynthesis through a signal receptor pore [199].

15.1.1.3 *Major histocompatibility complex synthesis*

The major histocompatibility complex class I is a transmembrane glycoprotein heterodimer which consists of a heavy α chain and β2 microglobuline. The heavy chain forms three extracellular domains, α1, α2 and α3, in which α1 and α2 bind the peptide. The function of α3 is to interact with T cell co-receptors to facilitate recognition and ensure specificity [19].

While the N- and C-terminus of the binding groove are highly conserved among different MHC genes, the internal part is highly polymorphic which enables the representation of a broad range of peptides [151]. The binding of a peptide to the MHC is realised in the peptide loading complex (PLC) which associates with the TAP transporter. Through an interaction of the MHC I heavy chain and tapasin, a part of the peptide loading complex, the empty MHC complex which is inherently unstable is stabilised and the binding groove is brought in an optimal conformation for loading. In the case that a suboptimal peptide finds its way onto the MHC, the whole complex becomes a substrate for UGT1 which reglucosylates the heavy chain and thus marking it for re-entry of the PLC. There the suboptimal ligand is exchanged for a high-affinity peptide. [20]

After successful loading of the MHC, it is transported by exocytosis via the Golgi apparatus to the cell surface where it can be recognised by specific T lymphocytes. These can then in turn activate an immune reaction cascade that most likely results in the induction of apoptosis of the presenting cell. [160]

For a visualisation of all steps in the MHC class I synthesis see figure 25 which is taken from Blum et al. [20].

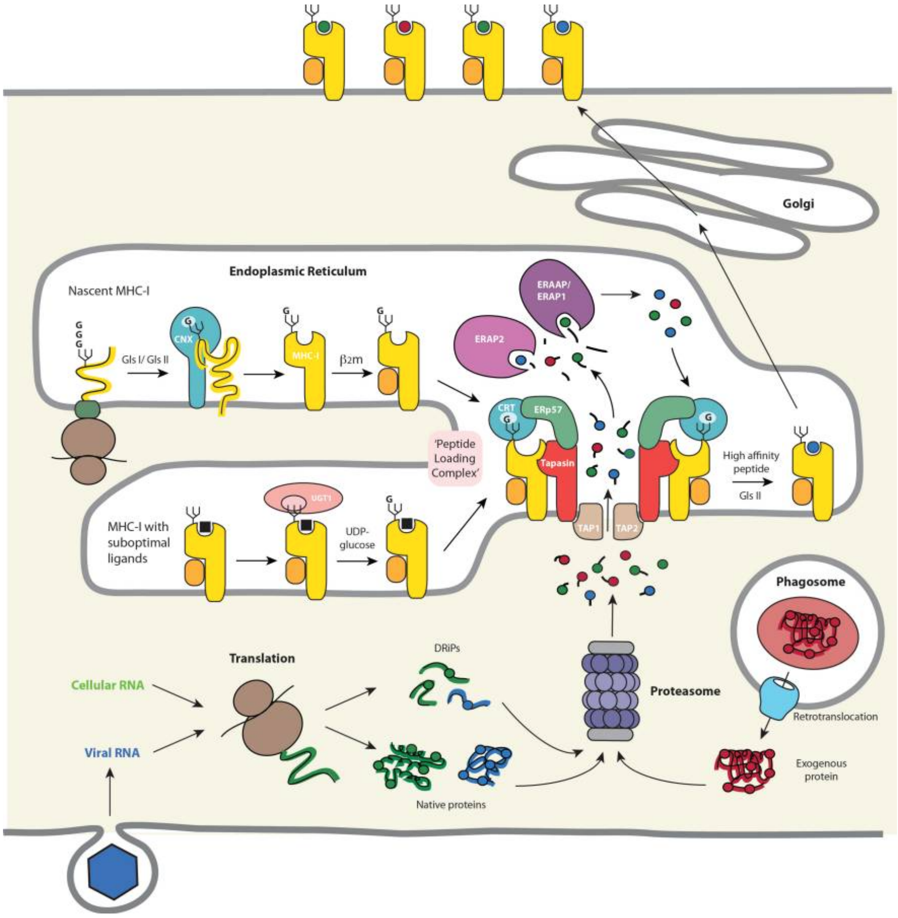


Figure 25: MHC I biosynthesis and antigenic peptide binding.
Shown are the different pathways how antigens presented by MHC I molecules are created and bound. The more common processing of endogenous proteins as well as the cross-presentation of exogenous peptides are depicted.
Taken from Blum et al. [20].

15.1.2 *Prediction Programs*

Due to the fact that many steps in the aforementioned peptide processing and MHC complex loading pathways have specific preferences regarding which parts of a protein are considered as epitope, multiple prediction programs have been implemented for each step. Most of them use a supervised learning approach, meaning the algorithms are trained on the data gathered so far to derive rules for yet unknown proteins.

Most of the prediction tools for different steps of epitope processing apply supervised learning algorithms

15.1.2.1 *Proteasomal cleavage*

The freely available tools for proteasomal cleavage prediction are MAPPP (MHC I Antigenic Peptide Processing Prediction; based on FRAGPREDICT) [77, 76], NetChop [94, 129] and PAProC (Prediction Algorithm for Proteasomal Cleavages) [105, 130]. In addition to proteasomal cleavage prediction, MAPPP also provides a binding prediction to the MHC I complex. The module of MAPPP responsible for calculating proteasomal cleavage probabilities is the older program FRAGPREDICT. It starts with a statistical analysis to identify cleavage-determining amino acid motifs. Afterwards, given the previously determined motifs, major proteolytic fragments are generated. The algorithm uses a kinetic model of the 20S proteasome describing time-depending degradation of substrates. [77, 76]

The core algorithm of netChop is implemented as a neural network that has been trained on verified MHC class I ligands. As not all proteasomal cleavage products are good binders to MHC, this approach is biased towards the immune-relevant peptides. [94, 129]

PAProC provides prediction of proteasomes in humans and yeast. The cleavage probabilities are calculated using the amino acids and their sequence position with a stochastic hill-climbing algorithm. The functionality has been trained on a set of experimentally determined cleavage and non-cleavage sites. [105, 130]

Out of this three available methods, netChop is the only one that, additionally to a web interface which exists for all presented tools, also provides a command line version that makes it easier to integrate its functionalities into other works. Furthermore, in a comparison review by Saxová et al. [164] which compared the performance of all three programs, netChop was identified as the most reliable prediction tool in determining the C-terminus of an epitope. This could be due to the fact that netChop's neural network design is better in modeling the complex specificity of the proteasome or that it uses a larger training set to initialise the network. [164]

15.1.2.2 *TAP affinity*

While there have been studies showing that the motifs of the transporter associated with antigen processing can be sufficiently modelled [25], not many approaches have been made available. The only web service that predicts TAP binding affinity from an input sequence is TAPPred [15]. The algorithm uses a support vector machine that has been trained on the physicochemical properties of amino acids and sequences of peptides which affinities were determined by binding essays. When validating the prediction results of TAPPred to experimentally defined binding affinities, a correlation coefficient of 0.88 could be reached [15]. Unfortunately, TAPPred provides no standalone version, only a web interface which does not lend itself to integration into other pipelines.

A method that can easier be adapted into other frameworks has been published by Peters et al. [140]. They have assembled a 20×9 consensus matrix for nonamer epitopes, giving for every peptide position and amino acid the positional $\log(IC_{50})$ value. The consensus matrix is compiled from three input matrices, two being created for peptide libraries of defined amino acid compositions, one is created with the stabilised matrix method (SMM) which minimises the distance of predicted and experimentally determined IC_{50} scores. This method yields the best results, when only taking the C-terminal position and the three N-terminal positions into account while also considering N-terminal elongated peptides. With these adaptions, the matrix method can reach a correlation coefficient of 0.79 when comparing experimental to predicted values. [140]

15.1.2.3 *MHC affinity*

Of all parts of the epitope processing pathway, predicting binding affinity to the MHC complex is the most commonly attempted. Hence, there are many different approaches available, for varying organisms and both MHC classes. The most cited web service with integrated database of more than 7000 MHC class I and II epitopes is SYFPEITHI [150]. The underlying data is compiled exclusively from published experimental data and used to derive binding motifs. Predictions are made by scoring the giving sequence based on how often the residues are found at the same position in the database entries.

The name SYFPEITHI acknowledges the first MHC-eluted peptide that was directly sequenced by Falk et al. [51] in 1991

Most published MHC affinity prediction tools function in a similar way, they derive position-specific rules combined with amino acid properties and define a scoring function which can be applied to new input sequences (e.g. BIMAS [134]). One of the most widely used tools implementing this approach is netMHC [128, 117, 118]. It uses an artificial neural network that has been trained on 78 human (HLA) alleles representing all 12 HLA-A and B supertypes. Of all published methods so far, netMHC uses the most comprehensive training set, as it combines the SYFPEITHI database with additional sources. Additionally to its web service, netMHC is available as stand alone software that can be used in other pipelines.

15.2 T CELL THERAPY

As described in the previous section, T cells are able to recognise invading pathogens. But research shows that they are also capable of identifying cancer cells which present tumour-specific antigens. This functionality has been used in the development of immunotherapies which use T cells to specifically target cancer cells.

The most straightforward approach to immunotherapy is the use of the host's own T cells which have already shown to be tumour-infiltrating lymphocytes. Another method is the genetical engineering of tumour-specific T cells. The T cells are in both cases expanded ex vivo and then re-introduced into the patient, a treatment that is called adoptive immunotherapy. In order to be effective, these therapies are usually accompanied by lymphodepletion, to prevent the tumour from mounting excessive anti-immune defences.

Lymphodepletion describes the temporary depletion of the immunesystem, usually achieved through chemotherapy or total body irradiation

The most critical step in designing an effective immunotherapy is the choice of TCR and thus the recognised target antigen. The targeted epitope should be available in a large enough fraction of tumour cells in order to sufficiently decrease the mass of the lesions, but on the other hand needs to be highly specific to the tumour, so that no healthy tissue is targeted. In order to fulfil both limitations, most immunotherapies so far tested target antigens that belong to one of five classes: The first group are antigens derived from proteins which play a role in tissue-differentiation. After their transformation, most cells continue to express antigens which are specific for the tissue they originated from. If this tissue of origin is non-essential, the specific antigens are ideal targets for immunotherapy.

Another group are neo-antigens that are derived from proteins which have obtained non-synonymous mutations in cancer cells and thus provide new epitopes. Due to their mutation, these antigens are not found in healthy cells, not even in the same tissue as the tumour site of origin.

The third group of antigens which can be targeted with immunotherapy are viral antigens. While these epitopes are only found in infected cells and hence make the therapy highly specific, they also limit treatment to cancer histologies associated with virus infections like Epstein-Barr or human papillomavirus.

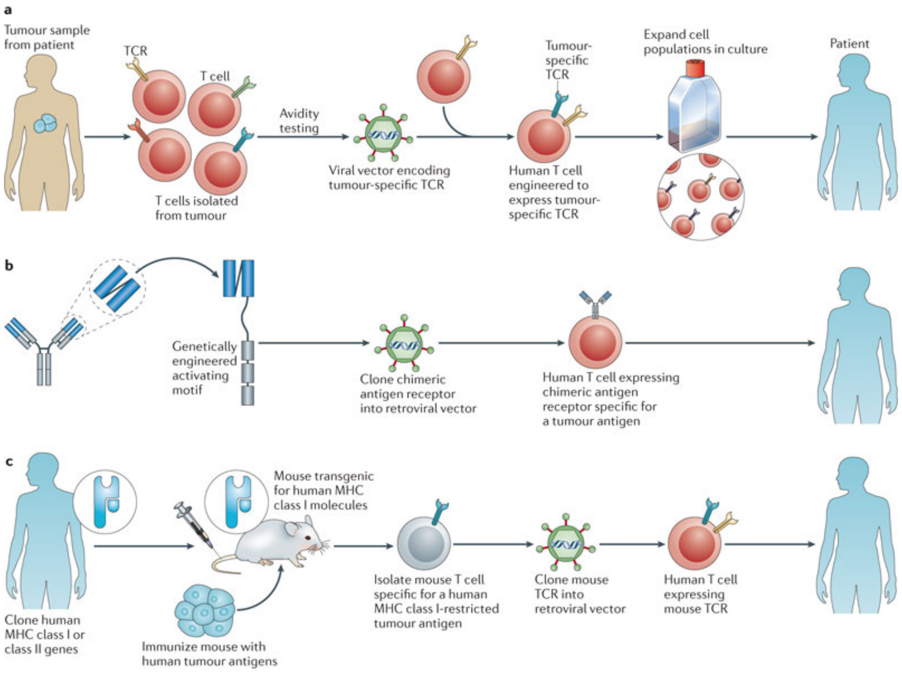
The fourth group of promising antigens are peptides derived from the tumour micro-environment instead of the tumour itself. It has been shown so far that a successful tumour formation also depends on a tumour supporting cell environment where vasculature and stroma have to aid the neoplasm formation. Thus, it has been proposed that

targeting these cells might be as effective in tumour elimination as attacking tumour cells directly [104].
The last group of target antigens for immunotherapy arise from epigenetic changes. Some tumours start expression of cancer germline antigens, also known as cancer-testis antigens, through epigenetic modifications. Theses genes are normally exclusively expressed in testes and fetal ovaries but are also present in many different tumour histologies. They pose one of the most promising therapeutic targets, as their original cells do not express MHC complexes, thus would be highly cancer specific. They have even been found in multiple different cancer types so that one therapy could be applied to many patients. [154]

There are different ways in which T cells can be genetically engineered to target a specific antigen of choice. One method is to isolate high-affinity TCRs from a patient who had a very effective antitumour response. The corresponding genes are cloned into a viral vector like a lenti- or retrovirus. These vectors can be used to infect patients autologous T cells, given that they share the same HLA restriction elements with the patient the TCRs originated from [86]. Instead of using other human patients as a source for TCRs, mice can also be used. So called humanised mice, meaning that the mouse is able to express human MHC class I and II molecules, are immunised with the tumour antigen of interest. Mouse T cells recognising said antigen are then extracted and analogous to the above mentioned method introduced into the patient via a viral vector [135]. Another approach to genetically engineer T cells, is via chimeric antigen receptors. Here, the variable regions from antibodies are engineered to encode a single chain which is fused to a TCR intracellular domain. Through viral vectors the chimeric antigen receptor is introduced into the patients T cells. Because the receptors have antibody-like specificities they are able to recognise MHC-nonrestricted peptides [161]. A visualisation of the discussed strategies for T cell genetic engineering can be seen in figure 26.

15.2.1 *Off-target effects*

Two different side-effects can cause unwanted reactions in patients treated with engineered T cells. On the one hand, on-target effects can occur, meaning that although the intended target is specifically recognised by the TCR, it occurs not only in tumour cells, but is also expressed in other tissue. On the other hand, off-target effects can lead to devastating side effects, when the transduced TCRs recognise different antigens additional to their intended target. While on-target effects can mostly be prevented by pre-screening of expression databases, off-target effects are harder to foresee, as it is not trivial how different other epitopes can be while still being detected.


a
Tumour sample from patient
TCR
T cell
T cells isolated from tumour
Avidity testing
Viral vector encoding tumour-specific TCR
Tumour-specific TCR
Human T cell engineered to express tumour-specific TCR
Expand cell populations in culture
Patient
b
Genetically engineered activating motif
Clone chimeric antigen receptor into retroviral vector
Human T cell expressing chimeric antigen receptor specific for a tumour antigen
c
Clone human MHC class I or class II genes
Mouse transgenic for human MHC class I molecules
Immunize mouse with human tumour antigens
Isolate mouse T cell specific for a human MHC class I-restricted tumour antigen
Clone mouse TCR into retroviral vector
Human T cell expressing mouse TCR

Nature Reviews | Immunology

Figure 26: Three strategies to genetically engineer T cells.
a: T cells from patients with good anti-tumour responses are cloned and brought into autologous T cells via viral vectors. **b:** Chimeric antigen receptors are generated by engrafting the variable region of an antibody onto the TCR intracellular domain, so they can recognise MHC-nonrestricted structures. **c:** TCRs are isolated from mice expressing human MHC molecules and have been immunised with tumour antigens. Specific T cells are extracted and cloned into a viral vector to infect autologous T cells. Taken from Restifo et al. [154].

Most immunotherapy studies that have gone into trial so far, have been administered to patients suffering from melanoma. It is not clear, why this type of cancer has shown the best response to natural tumour infiltrating T cells, but it has been proposed that the extremely high mutation rate of melanoma compared to other tumour histologies plays a role [194]. Some of these studies could report very good response rates und multiple patients who cleared their tumours, e.g. in one trial targeting cancer-testis antigen NY-ESO1 seven out of eleven patients showed a response, while no toxicities were reported [155]. However, multiple immunotherapy trials have resulted in minor or devastating side-effects. A study targeting the differentiation antigens MART-1 and gp100 lead to severe but treatable inflammation of skin, eyes and ears, even leading to hearing loss [86].
Two recent studies highlight that not only the expression of the direct target has to be examined across all vital tissues, but also approximate sequences have to be considered. In one study, Morgan et al.

[125] reported cross-recognition of a cancer-testis antigen MAGE-A3 TCR with a MAGE-A12 epitope that was later found to be expressed in a subset of neurons in the human brain. The MAGE-A12 epitope had one mismatch when compared to the initial target of the study, but was apparently recognised by the TCR and the treatment lead to changes in the mental status of the patients and two fell into a coma and subsequently died [125]. In another trial, Linette et al. [114] used a different MAGE-A3-specific TCR that was found to show cross-recognition of an epitope present in titin, a protein expressed in the heart. Although the titin-associated epitope had four mismatches compared with the original MAGE-A3 epitope, it was recognised by the TCR. Both patients participating in the trial suffered cardiac arrest [114].

15.3 MOTIVATION

Adoptive T cell therapies based on introduction of new T cell receptors (TCRs) into patient recipient T cells is a promising new treatment for various kinds of cancers. A major challenge, however, remains the choice of target antigens. If an engineered TCR can cross-react with antigens in healthy tissue, the side-effects can be devastating. Hence, we wanted to present the immunotherapy-community with an in silico tool for screening multiple healthy tissues for expression of their target antigens of choice in order to exclude epitope sharing before starting in vitro tests. In light of recent studies, highlighting the possibilities of TCRs recognising epitopes distinguished by multiple mismatches from the initial target, we wanted to search protein databases not only for exact matches, but also approximate ones.

Instead of just providing researchers with all string matches to their lead target, we wanted to integrate the biological processes involved in epitope processing. All results should be sorted by a score, which reflects the probability that the epitope is created by proteasomal cleavage and its affinities to the TAP transporter and the MHC class I alleles.

The single steps of these assessments have been available before, but distributed over various web servers and research groups. We wanted to integrate all searches, expression analysis and epitope scoring in a single framework in order to make it easily accessible and navigable for immunotherapy studies. With this framework we hope to provide a helpful tool to exclude potential cross-reactivity in the early stage of TCR selection for use in design of adoptive T cell immunotherapy.

16 METHODS

16.1 RNA-SEQ DATABASE

16.1.1 *Illumina Human Body Map*

As it is of utmost importance to provide the user with expression data of as many healthy human tissues as possible, we searched for a publicly available set of RNA-Seq data for multiple organ sites. Important factors in this search were the unrestricted distribution of the datasets in order to enable us to integrate the data in a web server and the fact that the origin of the samples was tissue as opposed to cell lines, as the latter can behave differently in their expression patterns due to the immortalisation they underwent.

While considering all of the above mentioned factors, we found the Illumina Human Body Map 2.0 to be the most fitting set for our purpose. It contains RNA-Seq data for 16 different normal tissues from unrelated patients. All samples have been sequenced on an Illumina HiSeq 2000 machine, which produces 50 bp long paired-end and 75 bp long single-end reads. While the raw data can be accessed easily via the Gene Expression Omnibus (GEO) [46, 9] identifier GSE30611 or the ArrayExpress [101] ID E-MTAB-513, we decided to use the already mapped data provided by Ensembl [53, 38]. They obtained the resulting fastq files directly from Illumina and used BWA 0.5.9 [108] to align the reads against human genome assembly hg19. All created bam files can be accessed via the Illumina Human Body Map section of the Ensembl ftp server.

16.1.2 *Encyclopedia of DNA Elements*

Although the intended purpose of our Expitope web server is to provide expression values of potential off-targets in healthy tissue, we decided to integrate a small number of cancer samples. This enables the user to immediately validate by eye that their original target is expressed in malignant cells. We decided to use RNA-Seq data from the Encyclopedia of DNA Elements (ENCODE) [49] Consortium for this purpose. Although these data sets originate from cell lines, we are confident that they nevertheless serve as sufficient positive controls. In fact, they might even obviate that expression values from tissue could reflect the transcript levels in the individual cancer. Additionally to cancer cell lines from liver, brain and colon tumours, we also obtained the RNA-Seq data of normal skin and lung cell lines from the ENCODE collection. The former was integrated as one of the targets with which our server was initially tested was known to be only expressed in cancer tissue and skin cells. The second serves as a control to check for expression differences that are founded in the lab of origin as opposed to the underlying diseases state. When the lung cell line from ENCODE would not be in agreement with the tissue values from the Illumina lung sample, we would know to look for more fitting sequencing data.

As many different laboratories contribute to the ENCODE project, not all of our samples originate from the same group. To keep the variation to a minimum, we took all RNA-Seq data from the Cold Spring Harbor Laboratory (CSHL) except the colon cancer cell line (HCT-116) which does belong to tier 3, meaning it is classified as a less important cell line and is only provided by one lab, the California Institute of Technology (Caltech). All bam files produced by CSHL contain 76 bp long paired-end reads produced on an Illumina GAIIx machine. Due to the used library preparation protocol the strand of origin for every read is known. For all sets Poly-A+ RNA extracted from the whole cell was chosen for analysis. All resulting reads were mapped to the hg19 reference with the mapping program Spliced Transcript Alignment and Reconstruction (STAR) [40].

The sample produced by Caltech also used poly-A+ extracted RNA from the whole cell and sequenced it on an Illumina Genome Analyzer GAI or GAIIx. The resulting 75 bp long paired-end reads were mapped against hg19 using TopHat (version 1.0.14) [188]. Contrary to the CSHL data set, Caltech's reads are not strand oriented.

All bam files integrated into our database were obtained from the ENCODE section of the University of California, Santa Cruz (UCSC) download portal [158]. In concordance with ENCODE's data quality guide lines, all RNA-Seq sets have at least two replicates.

Table 18 summarises the information about the chosen datasets.

GEO id	Tissue	Karyotype	# Reps	Origin
GSM758575	liver (HepG2)	cancer	2	CSHL ENCODE RNA-Seq
GSM981253	brain (SK-N-SH)	cancer	2	CSHL ENCODE RNA-Seq
GSM958749	colon (HCT-116)	cancer	2	Caltech ENCODE RNA-Seq
GSM981249	lung (IMR90)	normal	2	CSHL ENCODE RNA-Seq
GSM758562	skin (BJ)	normal	2	CSHL ENCODE RNA-Seq

Continued on next page

Table 18 – continued from previous page

GEO id	Tissue	Karyotype	# Reps	Origin
GSM759490, GSM759491	adipose	normal	1	Illumina Human Body Map 2.0
GSM759492, GSM759493	adrenal	normal	1	Illumina Human Body Map 2.0
GSM759520, GSM759522	blood	normal	1	Illumina Human Body Map 2.0
GSM759494, GSM759495	brain	normal	1	Illumina Human Body Map 2.0
GSM759496, GSM759497	breast	normal	1	Illumina Human Body Map 2.0
GSM759498, GSM759499	colon	normal	1	Illumina Human Body Map 2.0
GSM759500, GSM759500	heart	normal	1	Illumina Human Body Map 2.0
GSM759502, GSM759503	kidney	normal	1	Illumina Human Body Map 2.0
GSM759504, GSM759505	liver	normal	1	Illumina Human Body Map 2.0
GSM759506, GSM759507	lung	normal	1	Illumina Human Body Map 2.0
GSM759508, GSM759509	lymph	normal	1	Illumina Human Body Map 2.0
GSM759510, GSM759511	ovary	normal	1	Illumina Human Body Map 2.0
GSM759512, GSM759513	prostate	normal	1	Illumina Human Body Map 2.0
GSM759514, GSM759515	skeletal muscle	normal	1	Illumina Human Body Map 2.0
GSM759516, GSM759517	testes	normal	1	Illumina Human Body Map 2.0
GSM759518, GSM759519	thyroid	normal	1	Illumina Human Body Map 2.0

Table 18: Datasets which constitute Expitope's expression database. The first column shows were the RNA-Seq samples can be found in the Gene Expression Omnibus, the fourth shows how many replicates are available for the corresponding sample.

16.1.3 *Additional brain tissue*

The most vital tissue in which cross reaction of TCRs has to be excluded is the brain. Thus, we wanted to provide a broader spectrum of brain tissues than only the one included in the Illumina Human Body Map data. To fulfil this purpose we found a publication by Wang et al. [195] which analysed isoform regulation in the transcriptomes of nine different human tissues, five cell lines, two RNA reference compilations and, most importantly for integration in Expitope, six cerebellar cortex samples from unrelated anonymous donors. The authors provide the RPKM [126] values of all examined genes (23,115 Ensembl gene identifier) in their supplementary data. We integrated these very valuable data into our expression database.

16.1.4 *Genome Annotation*

We compared every alignment file obtained from either ENCODE or Illumina Human Body Map against the annotated transcripts of the latest human genome assembly (hg19). For this purpose we used the GenCodeV19 [70] annotation, with the coordinates for all exons of each transcript isoform. With the bamutils tool *count* of the ngsutils suite (version 0.5.5) [23] we extracted the number of reads/fragments per sample which map to every annotated exon. As raw read counts are not easily comparable between the different samples due to different library sizes, we obtained the normalisation FPKM (Fragments per kilobase of exon per million fragments mapped) which takes the read count per transcript, transcript length and overall number of mapped reads in the sample into account. We actively searched for a tool that is capable to distinguish between single- and paired-end library preparations, because programs which only calculate RPKMs would lead to an artificial doubled read count of paired-end samples. Although at this point our database only contains paired-end reads, we want to keep it extensible for additional data.

We made sure to use the parameter *-library unstranded* during read counting for the samples from the Illumina Human Body Map and Caltech, as these used an unstranded library preparation (contrary to CSHL) and the default setting of bamutils is a forward-reverse strandedness.

The created MySQL database contains one entry (row) for every GenCodeV19 transcript and a column for every analysed RNA-Seq sample. The cells contain the FPKM value for the given transcript in the corresponding experiment. For a visualisation of how Expitope's expression database was created, compare figure 27.

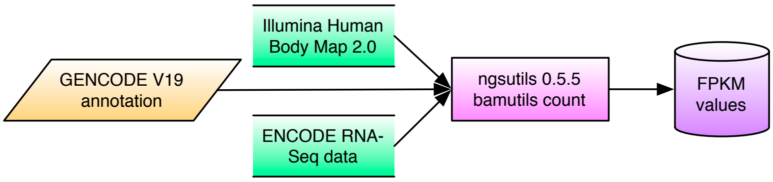


Figure 27: Schematic, showing the construction of Expitope's expression database.

16.2 EPITOPE LOOKUP

Our implementation of the epitope finding in all human proteins is designed in a way that only requires the users target epitope (String of amino acids in one letter code) and a number of allowed mismatches (integer value). We require the peptide to be at least seven amino acids long as to avoid a large amount of matches with known protein sequences by chance. The given number of mismatches has to be smaller than half the size of the provided epitope for the same reason.

We then search for occurrences of the given epitope in all available proteins contained in the NCBI protein database [147], including all annotated isoforms. All matches with zero up to the defined number of mismatches are reported and the corresponding protein ID is stored.

All obtained protein identifiers from entries of interest are mapped to Ensembl transcript identifiers via a lookup file downloaded from UCSC table browser [93] by joining three tables. In this step the same transcript ID can be reported for different protein IDs or a protein ID can result in multiple transcript IDs. The first case is caused by an ambiguous attribution of identifiers, the second is due to multiple transcript isoforms being associated with the same protein entry.

The set of transcript IDs is used to query the previously set up database for the expression values in all tissues. These results are presented to the user in a tab separated file which additionally contains the exact epitope found in a certain protein and its position relative to the complete amino acid sequence length.

The output file additionally lists proteins which contain the provided epitope but could not be matched to a transcript identifier. These are usually automatically determined proteins (recognisable through their *XP_* identifier start instead of *NP_*) which real existence is not proven yet.

16.3 SCORING SCHEME

Not all of the found peptides which are similar to the initial epitope will really be recognised by a T cell. The three important steps that will lead from full-length protein to a presented epitope are: Proteasomal cleavage, TAP transport in the endoplasmic reticulum and binding to the MHC class I proteins.

In order to sort all peptides which match the queried input with at most the given number of mismatches, we applied a scoring function proposed by Keşmir et al. [94]. It defines a combined score Q as

$$Q = \frac{P}{A_{TAP} \times A_{MHC}}$$

where P is the probability that the given peptide is cleaved from the protein and the A-terms are affinities to the complexes in subscript.

If a queried peptide could not be used for MHC affinity prediction due to reasons of length, the MHC affinity is set to one, so that the combined score is only defined by TAP affinity and proteasomal cleavage probability.

16.3.1 *Proteasomal cleavage prediction*

To calculate the first of the three terms that define the combined score Q, we used the program NetChop 3.1 [94, 129] to define the probability of the given epitope being created by proteasomal cleavage. For this purpose we ran the program on all current RefSeq protein entries and obtained a cleavage probability for every position. These values were stored in an additional database table to avoid executing NetChop for every web server query. We are using the prediction method *C-term 3.0* which is a neural network trained on a database containing 1,260 publicly available MHC class I ligands. It performs best when predicting the boundaries of cytotoxic T cell (CTL) epitopes.

When calculating the cleavage probability for the current epitope, we followed the original paper [94] and used the formula

$$P = P_c \times P_{con}$$

wherein P_c is the probability that the peptide is cleaved exactly at the C-terminus and P_{con} represents the probability of the rest of the peptide staying intact:

$$P_{con} = \prod_{O_i > t} (1 - O_i)$$

where O_i represents the output of the network for position i of the peptide. The parameter t is replaced with 0.7 in the publication, thus we use this value as default in our web server, but the user can replace it with their own threshold. When substituting t for 1, only the C-terminal probability defines the cleavage score.

Due to the overall cleavage probability being a product, it becomes very small very quickly, especially for longer input sequences. Hence, it is advisable to only rely on this score for peptides in the range of seven to eleven amino acids, as that is the epitope size for which NetChop has been most extensively tested, although the calculation itself does not limit the input to a certain length.

16.3.2 *TAP affinity prediction*

The IC_{50} *value gives the dose of the peptide which displaces 50% of a competitive ligand*

The second term that weighs into the calculation of the combined score Q is the affinity of the peptide to TAP. Peters et al. [140] have established a 9×20 matrix that contains for each amino acid at every possible epitope position (of length nine) a $\log(\mathrm{IC}_{50})$ value which can be summed up to obtain an IC_{50} (dose of peptide which displaces 50% of a competitive ligand) for the complete peptide. When testing the divergence between predicted and experimentally tested IC_{50} values, the authors concluded that the following formula returned the results with the best concordance:

$$t_{L,\alpha} = mat_{9,C} + \frac{\alpha}{L-8} \sum_{l=9}^{L} mat_{1,N1} + mat_{2,N2} + mat_{3,N3}$$

Thus only the C-terminal residue and the first three N-terminal residues are used for the affinity calculation. The parameter L represents the length of a potential precursor peptide, as it is hypothesised that the peptide which is transported by TAP is in fact not the final nonamer but a longer peptide with an elongated N terminus. Hence the formula has a fixed term for the C-terminus but sums over different precursor N-termini which can be weighted depending on their importance to the final score. Peters et al. [140] conclude that their experimentally determined best values for L is 10 and 0.2 for α.

As we do not expect the user to only subject nonamers to our webserver, we have not implemented the regard for precursors and thus do not require the parameter L. We do, however, allow for a weight α of the N-terminus which can be changed by the user. Thus, the formula used in the web server is:

$$t_{\alpha} = mat_{9,C} + \alpha \cdot (mat_{1,N1} + mat_{2,N2} + mat_{3,N3})$$

In line with the original publication the default value for α is 0.2. If only the C-terminus should be considered for the affinity prediction, α can be set to zero.

Although it is technically possible to score peptides of all length $\geqslant 4$ with this approach, it has to be kept in mind that the matrices are constructed on the basis of nonamer epitopes and have also only been extensively tested one those or with slightly longer precursors. When analysing longer peptides the returned values might not reflect the real affinity to TAP and it could be beneficial to exclude the N-terminus in those cases.

16.3.3 *MHC binding prediction*

The last term which goes into the combined score Q is the affinity of the epitopes to the major histocompatibility complex. For the prediction of this feature, we integrated the tool NetMHC 3.0 [128, 117, 118] into our web server which can make affinity predictions for a large range of HLA-alleles. The tool offers artificial neural networks trained on 55 different MHC alleles (43 human, 6 mouse, 5 rhesus macaque and 1 chimpanzee) and returns the affinity of a given peptide to the specified alleles in nM IC_{50} values. Due to the size limitations implemented in NetMHC, only peptides of a length between 8 and 14 amino acids can be used for affinity prediction. The authors want it to be noted that predictions of peptides longer than eleven positions have not been extensively validated and caution should be taken for octamer predictions, as some alleles might not bind them to any significant extend.

The user can submit a selection of multiple HLA types for affinity prediction, between one and all; the default allele is A-0201 as it is the most common MHC class I allele in Caucasian populations [24]. We report the exact IC50 values predicted by NetMHC back to the user for every MHC type that was selected in the query, but only the best (lowest) is used in the calculation of Q. The authors of NetMHC define all peptides with an affinity value below 50 nM as strong and all epitopes with a value below 500 nM as weak binders.

By definition from NetMHC, all peptides with affinity value < 50 nM are strong binder, < 500 nM weak binder

16.4 WEB SERVER IMPLEMENTATION

In order to provide the general public access to all of Expitope's functionalities, we integrated the above mentioned steps into a web site. It is hosted on the server of the institute for genome-oriented bioinformatics of the Technical University Munich and can be found under

http://webclu.bio.wzw.tum.de/expitope

The sorted list of output proteins as well as the corresponding expression values can either be viewed directly in the browser or can be downloaded as a tabular file for further downstream analysis.

All steps implemented in the web site can be seen in figure 28.

16.5 DIFFERENTIAL EXPRESSION

Our Webserver is currently used to examine certain epitopes of interest in order to ensure that they are only expressed in cancerous tissue or at least only in non vital ones. These potential antigens stem from known cancer markers and are afterwards analysed in vitro to prevent cross reactions.

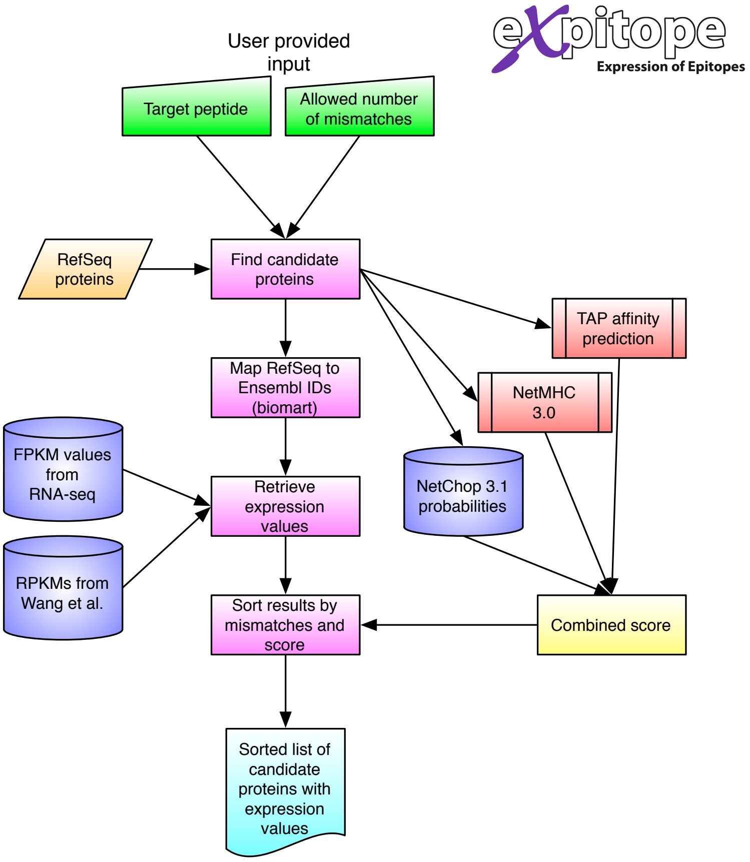

expitope
Expression of Epitopes
User provided input
Target peptide
Allowed number of mismatches
RefSeq proteins
Find candidate proteins
TAP affinity prediction
Map RefSeq to Ensembl IDs (biomart)
NetMHC 3.0
FPKM values from RNA-seq
Retrieve expression values
NetChop 3.1 probabilities
RPKMs from Wang et al.
Sort results by mismatches and score
Combined score
Sorted list of candidate proteins with expression values

Figure 28: Workflow of Expitope's functionality that is implemented in our web server.

To take the bottom-up approach, we decided to perform an unusual differential expression analysis on our database. Instead of performing pairwise comparisons between the tissues, we executed three analysis, each comparing one of the three cancerous cell types against all other healthy entries. With this approach, we will find all transcripts which are significantly differently expressed between one cancer type and healthy tissue and can further filter those results to identify transcripts which are not expressed at all in healthy cell types.

For this undertaking we utilised the R bioconductor [149, 57] package *DESeq* [3] in an analysis design as described in section 12.3.1 of this thesis.

To get an initial overview of all samples, we also normalised all datasets by size and adjusted for biological variance using the replicates. We again performed a PCA and calculated Euclidean distances to get an impression of the diversity of expression values.

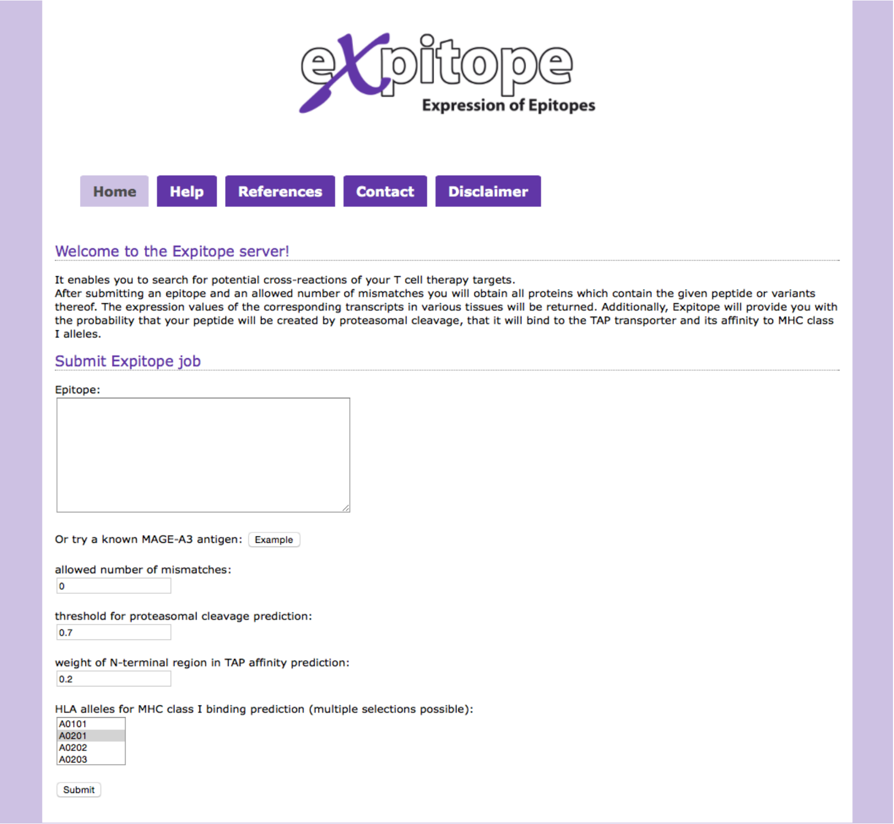
After analysing the data as a whole, we compared every three cancer cell types against all combined healthy samples. For every annotated transcript we tested if its expression was significantly different between the two compared conditions. Therefore, we compared the coverage values of the conditions normalised by the library sizes and calculated the resulting fold change. We used a negative binomial test to find significant differences in the expressions. The initial p-Values are adjusted for multiple testing using the Benjamini-Hochberg [13] procedure to control the false discovery rate (FDR). This multiple testing correction is necessary because for every single transcript the hypothesis of being differentially expressed is tested and thus we expect to see many false positives through chance when testing thousands of genes.

To filter the list of all features for the significantly expressed genes, we applied a filtering for the logarithmic fold change and the adjusted p-Value. The former had to be at least two and the latter should be smaller than 0.001 (FDR of 0.1%).

17 RESULTS

17.1 Expitope's web site is frequently visited

The web site implemented to host all functionalities of Expitope has been set up with input fields for all mandatory and optional parameters that can or must be given by the user. Where applicable, they are initialised with the published optimal values. For testing purposes, we provide a well known example input from MAGE-A3 that was used in a study where off-target recognition caused death of patients [125]. Additional to the query site that starts a database request, we provide tabs containing a help page explaining all methods given in chapter 16, references, contact details and a disclaimer informing the user that Expitope's predictions are based on computational methods and should not build the basis for treatment without further experimental validation. A screenshot of the home page of Expitope can be seen in figure 29.



Welcome to the Expitope server!

It enables you to search for potential cross-reactions of your T cell therapy targets.
After submitting an epitope and an allowed number of mismatches you will obtain all proteins which contain the given peptide or variants thereof. The expression values of the corresponding transcripts in various tissues will be returned. Additionally, Expitope will provide you with the probability that your peptide will be created by proteasomal cleavage, that it will bind to the TAP transporter and its affinity to MHC class I alleles.

Submit Expitope job

Epitope:

Or try a known MAGE-A3 antigen: Example

allowed number of mismatches:
0

threshold for proteasomal cleavage prediction:
0.7

weight of N-terminal region in TAP affinity prediction:
0.2

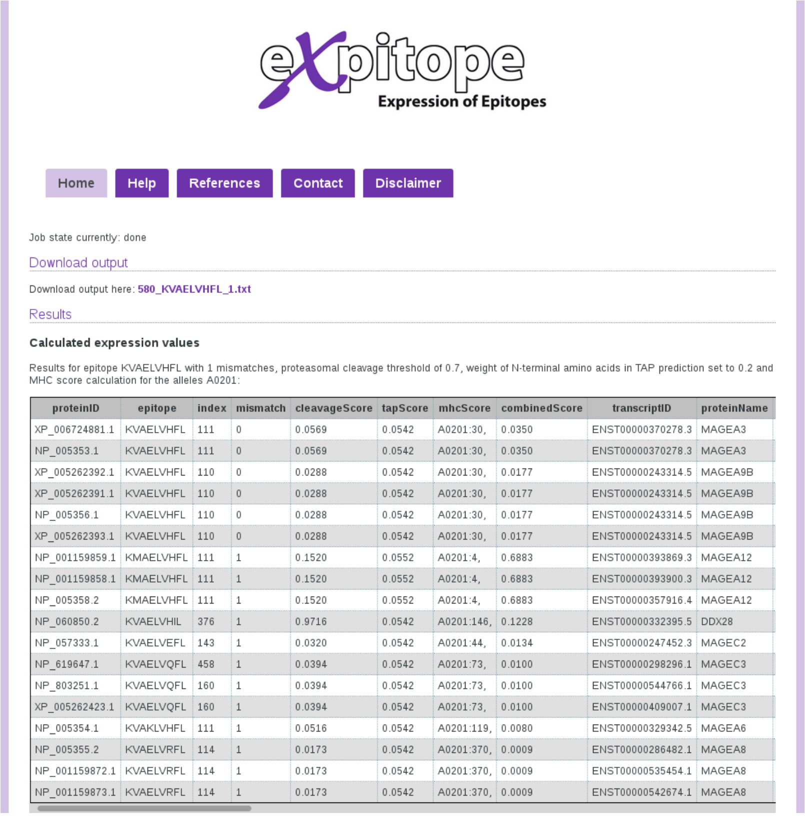
HLA alleles for MHC class I binding prediction (multiple selections possible):
A0101
A0201
A0202
A0203

Submit

Figure 29: Screenshot of Expitope's homepage.

The output is provided as a tab-separated text file for download and as a neatly formatted html table for data inspection by eye. It is sorted by ascending number of mismatches and descending combined score Q.

The columns denote the RefSeq protein identifier to which the matching peptide belongs and the exact sequence that matched the input parameters. The column titled *index* gives the starting position of the epitope in the full length RefSeq entry. The following columns denote number of mismatches which separate the current match from the input sequence and the three scores corresponding to cleavage probability, TAP affinity and MHC affinity as well as the resulting combined score. For a view of the result page of our example MAGE-A3 input compare figure 30.



Job state currently: done

Download output

Download output here: 580_KVAELVHFL_1.txt

Results

Calculated expression values

Results for epitope KVAELVHFL with 1 mismatches, proteasomal cleavage threshold of 0.7, weight of N-terminal amino acids in TAP prediction set to 0.2 and MHC score calculation for the alleles A0201:

proteinID	epitope	index	mismatch	cleavageScore	tapScore	mhcScore	combinedScore	transcriptID	proteinName
XP_006724881.1	KVAELVHFL	111	0	0.0569	0.0542	A0201:30,	0.0350	ENST00000370278.3	MAGEA3
NP_005353.1	KVAELVHFL	111	0	0.0569	0.0542	A0201:30,	0.0350	ENST00000370278.3	MAGEA3
XP_005262392.1	KVAELVHFL	110	0	0.0288	0.0542	A0201:30,	0.0177	ENST00000243314.5	MAGEA9B
XP_005262391.1	KVAELVHFL	110	0	0.0288	0.0542	A0201:30,	0.0177	ENST00000243314.5	MAGEA9B
NP_005356.1	KVAELVHFL	110	0	0.0288	0.0542	A0201:30,	0.0177	ENST00000243314.5	MAGEA9B
XP_005262393.1	KVAELVHFL	110	0	0.0288	0.0542	A0201:30,	0.0177	ENST00000243314.5	MAGEA9B
NP_001159859.1	KMAELVHFL	111	1	0.1520	0.0552	A0201:4,	0.6883	ENST00000393869.3	MAGEA12
NP_001159858.1	KMAELVHFL	111	1	0.1520	0.0552	A0201:4,	0.6883	ENST00000393900.3	MAGEA12
NP_005358.2	KMAELVHFL	111	1	0.1520	0.0552	A0201:4,	0.6883	ENST00000357916.4	MAGEA12
NP_060850.2	KVAELVHIL	376	1	0.9716	0.0542	A0201:146,	0.1228	ENST00000332395.5	DDX28
NP_057333.1	KVAELVEFL	143	1	0.0320	0.0542	A0201:44,	0.0134	ENST00000247452.3	MAGEC2
NP_619647.1	KVAELVQFL	458	1	0.0394	0.0542	A0201:73,	0.0100	ENST00000298296.1	MAGEC3
NP_803251.1	KVAELVQFL	160	1	0.0394	0.0542	A0201:73,	0.0100	ENST00000544766.1	MAGEC3
XP_005262423.1	KVAELVQFL	160	1	0.0394	0.0542	A0201:73,	0.0100	ENST00000409007.1	MAGEC3
NP_005354.1	KVAKLVHFL	111	1	0.0516	0.0542	A0201:119,	0.0080	ENST00000329342.5	MAGEA6
NP_005355.2	KVAELVRFL	114	1	0.0173	0.0542	A0201:370,	0.0009	ENST00000286482.1	MAGEA8
NP_001159872.1	KVAELVRFL	114	1	0.0173	0.0542	A0201:370,	0.0009	ENST00000535454.1	MAGEA8
NP_001159873.1	KVAELVRFL	114	1	0.0173	0.0542	A0201:370,	0.0009	ENST00000542674.1	MAGEA8

Figure 30: Screenshot of Expitope's output page defining the found targets. Shown is the output for the implemented example input KVAELVHFL from a study by Morgan et al. [125]. It can be seen that while all peptides from the MAGE family do not have a good cleavage score, they have very low IC_{50} values in the TAP and MHC prediction, i.e. high affinities to both complexes.

The second part of the results table contains information regarding the transcripts and their expression. Expitope first lists the ENSEMBL transcript identifier corresponding to the protein, its official name, followed by the RPKM values in all analysed tissues and cell lines (in alphabetical order). For this part of the output table see figure 31. The two integrated samples of healthy lung (Illumina Human Body Map lung and IMR90 from ENCODE) did not yield any conflicting results so far. While their expression values are not identical, they show the same trends for all transcripts we analysed until today. Thus it appears that our underlying expression value database is quite robust.

The results based on data from Wang et al. [195] are presented in an extra table, following the same format as described above. When comparing these data to our self-compiled database, it has to be kept in mind, that the assignment of RPKMs to proteins (as it is done by Wang et al. [195]) is slightly less exact than our own output, as only one value per gene instead of per transcript is returned. This results do not cover potential differences between multiple variants of the same gene.

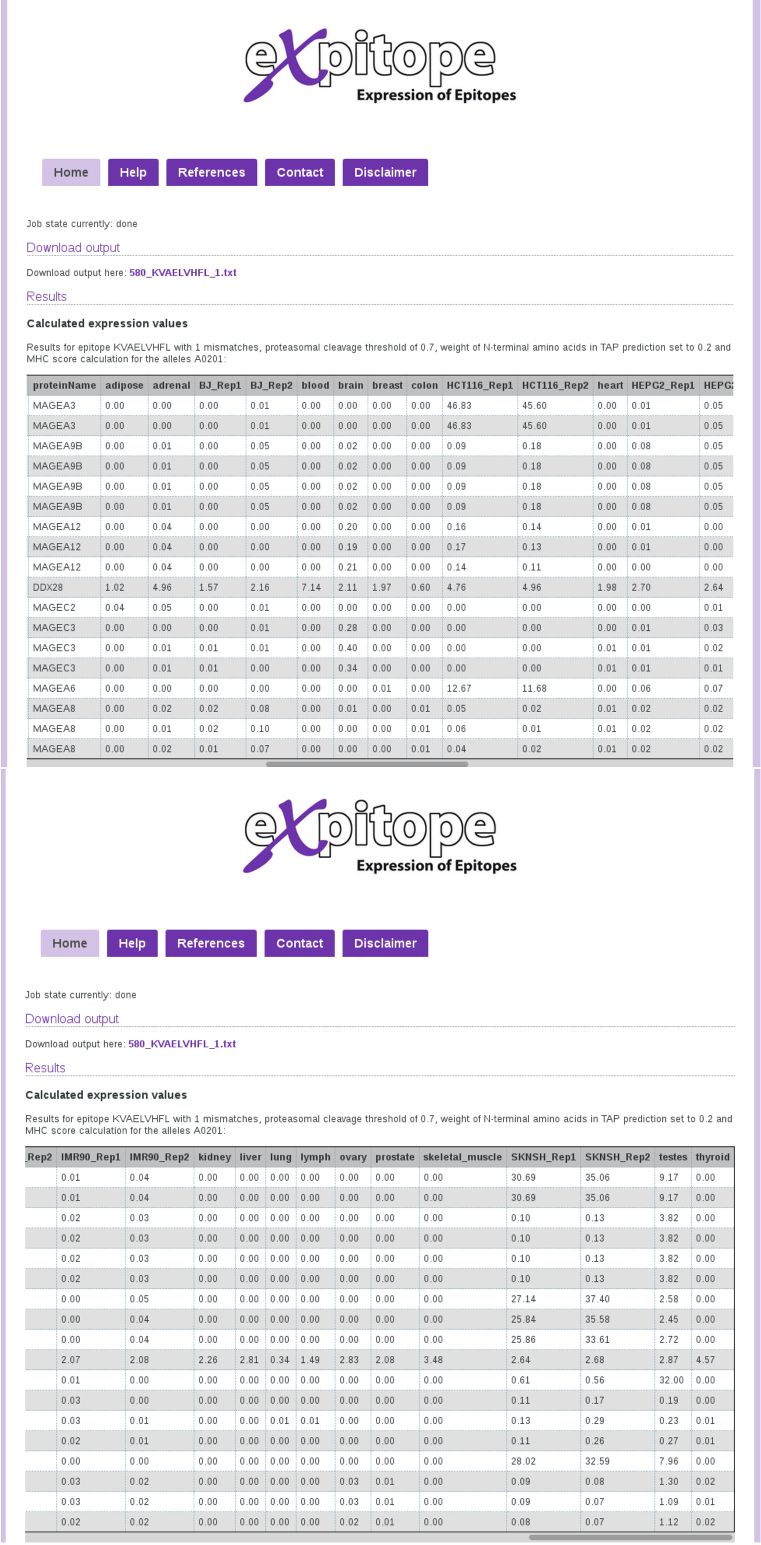
Additionally, Expitope lists proteins, which contain the provided epitope but could not be matched to a transcript identifier. We felt the need to include them in order to provide the user with as much information as possible about potential cross-reactions.

To be able to adapt and improve Expitope based on the needs and wishes of its users, we need to know the amount of visits the page has to work with and the demographic of the site's visitors. For this purpose, we imbedded google analytics [65] web tracking snippet in Expitope's source code. With this implementation we obtain access to general statistics like site visits, mean length of stay and bounce rate (fraction of visitors not reaching the output page) as well as the information about the institutes where users are located. To Expitope's frequent visitors belong multiple international high ranking universities as well as immunotherapy companies. As of 12th of March 2015, five weeks after online publication of Expitope's paper in Bioinformatics [68], the web site had already registered 1,595 page visits with a mean length of stay slightly under five minutes.

In only six weeks after publication, Expitope had already accumulated 1,595 page visits

17.2 PREVIOUS STUDIES WITH DETRIMENTAL SIDE EFFECTS CAN BE UNDERSTOOD WITH EXPITOPE

To test the capability of Expitope, we investigated two previous TCR gene therapies in which unanticipated cross-recognition of healthy tissues led to patient deaths (for details see section 15.2.1). The first study, conducted by Morgan et al. [125], is integrated into the Expitope web server as example input. It works with the nonamer peptide `KVAELVHFL` from known cancer-testis antigen MAGE-A3 and one allowed mismatch. All other parameters are set to their default values. Additionally to the intended MAGE-A3 target, MAGE-A9B is identified to contain the peptide, but these two MAGE family members were known to share the epitope. The reason for the fatal cases in the study, MAGEA-12, is also returned by Expitope and although it diverges in one position from the initial peptide (`KMAELVHFL`), it has



Job state currently: done

Download output

Download output here: 580_KVAELVHFL_1.txt

Results

Calculated expression values

Results for epitope KVAELVHFL with 1 mismatches, proteasomal cleavage threshold of 0.7, weight of N-terminal amino acids in TAP prediction set to 0.2 and MHC score calculation for the alleles A0201:

proteinName	adipose	adrenal	BJ_Rep1	BJ_Rep2	blood	brain	breast	colon	HCT116_Rep1	HCT116_Rep2	heart	HEPG2_Rep1	HEPG
MAGEA3	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	46.83	45.60	0.00	0.01	0.05
MAGEA3	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	46.83	45.60	0.00	0.01	0.05
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08	0.05
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08	0.05
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08	0.05
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08	0.05
MAGEA12	0.00	0.04	0.00	0.00	0.00	0.20	0.00	0.00	0.16	0.14	0.00	0.01	0.00
MAGEA12	0.00	0.04	0.00	0.00	0.00	0.19	0.00	0.00	0.17	0.13	0.00	0.01	0.00
MAGEA12	0.00	0.04	0.00	0.00	0.00	0.21	0.00	0.00	0.14	0.11	0.00	0.00	0.00
DDX28	1.02	4.96	1.57	2.16	7.14	2.11	1.97	0.60	4.76	4.96	1.98	2.70	2.64
MAGEC2	0.04	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
MAGEC3	0.00	0.00	0.00	0.01	0.00	0.28	0.00	0.00	0.00	0.00	0.00	0.01	0.03
MAGEC3	0.00	0.01	0.01	0.01	0.00	0.40	0.00	0.00	0.00	0.00	0.01	0.01	0.02
MAGEC3	0.00	0.01	0.01	0.00	0.00	0.34	0.00	0.00	0.00	0.00	0.01	0.01	0.01
MAGEA6	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	12.67	11.68	0.00	0.06	0.07
MAGEA8	0.00	0.02	0.02	0.08	0.00	0.01	0.00	0.01	0.05	0.02	0.01	0.02	0.02
MAGEA8	0.00	0.01	0.02	0.10	0.00	0.00	0.00	0.01	0.06	0.01	0.01	0.02	0.02
MAGEA8	0.00	0.02	0.01	0.07	0.00	0.00	0.00	0.01	0.04	0.02	0.01	0.02	0.02

Job state currently: done

Download output

Download output here: 580_KVAELVHFL_1.txt

Results

Calculated expression values

Results for epitope KVAELVHFL with 1 mismatches, proteasomal cleavage threshold of 0.7, weight of N-terminal amino acids in TAP prediction set to 0.2 and MHC score calculation for the alleles A0201:

_Rep2	IMR90_Rep1	IMR90_Rep2	kidney	liver	lung	lymph	ovary	prostate	skeletal_muscle	SKNSH_Rep1	SKNSH_Rep2	testes	thyroid
	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.69	35.06	9.17	0.00
	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.69	35.06	9.17	0.00
	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.13	3.82	0.00
	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.13	3.82	0.00
	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.13	3.82	0.00
	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.13	3.82	0.00
	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	27.14	37.40	2.58	0.00
	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.84	35.58	2.45	0.00
	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.86	33.61	2.72	0.00
	2.07	2.08	2.26	2.81	0.34	1.49	2.83	2.08	3.48	2.64	2.68	2.87	4.57
	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.56	32.00	0.00
	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.17	0.19	0.00
	0.03	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.13	0.29	0.23	0.01
	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.26	0.27	0.01
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.02	32.59	7.96	0.00
	0.03	0.02	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.09	0.08	1.30	0.02
	0.03	0.02	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.09	0.07	1.09	0.01
	0.02	0.02	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.08	0.07	1.12	0.02

Figure 31: Screenshots of Expitope's output page listing the expression values.
All members of the MAGE family are not or barely (FPKM < 1) expressed in all healthy tissues but MAGE-A3 shows strong expression in colon and brain cancer cell lines (HCT-116 and SK-N-SH, respectively) and moderate expression in testes.

a much better combined score due to its proteasomal cleavage probability being greater ($p = 0.1520$ vs. $p = 0.0569$) and the affinity to MHC allele A0201 being higher (IC_{50} of 4 vs. 30). While MAGE-A3 and MAGE-A9B are only noteworthy expressed ($FPKM > 0.1$) in the colon cancer and brain cancer cell line as well as in testes, all three variants of MAGE-A12 show additional expression in brain tissue, albeit very low, between 0.19 and 0.21 FPKM.

We also used the study of Linette et al. [114] in which they worked on another epitope from MAGE-A3 (EVDPIGHLY). We used the target they had engineered in their study as input peptide and allowed for up to four mismatches. The HLA allele was set to A0101, as the corresponding MAGE-A3 epitope is HLA-A1 restricted, all other variable input parameters were left set to their default values.

Titin, the reason for the detrimental side effects in the cited study, was found as a candidate for off-target effects by Expitope. Although the sequence has four mismatches to the initial MAGE-A3 input, the predicted affinity to MHC allele A0101 was even higher for the Titin antigen (IC_{50} of 43 vs. 38 for MAGE-A3 and Titin, respectively). Based on Expitope's reported expression values, the cardiac arrest two of the patients in Linette et al. [114]'s study suffered, can be understood. In line with the follow-up analysis in the original publication, Titin has a very high expression in the heart tissue from Illumina Human Body Map (FPKM between 97 and 122 depending on the isoform, compare figure 32).

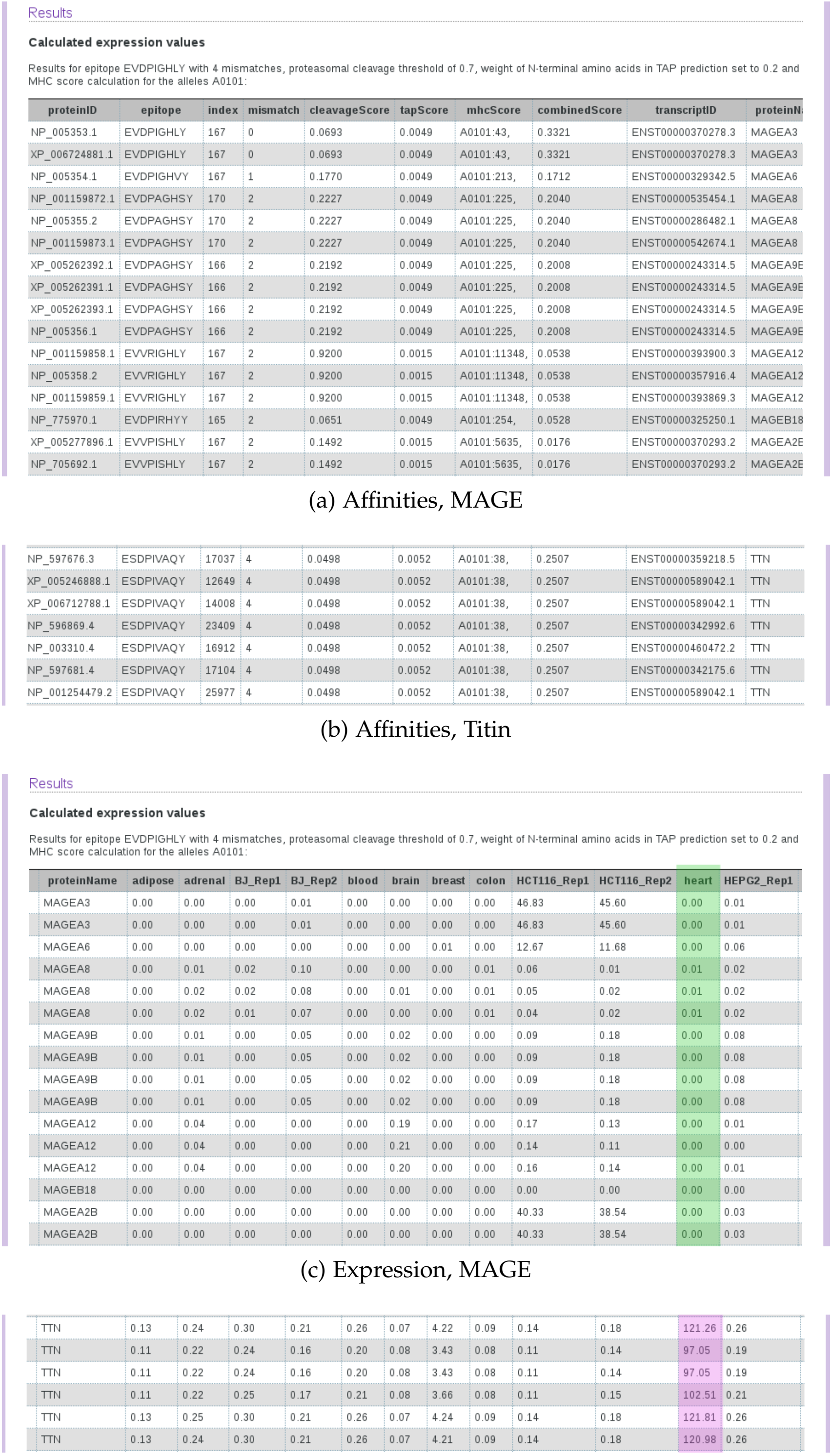
17.3 bottom-up search for potential candidates does not reveal trivial targets

In an attempt to analyse, if our database set up lends itself to do a bottom-up analysis, meaning instead of validating a candidate epitope, identifying target peptides that are not expressed in healthy tissue, we did a differential expression analysis on our RNA-Seq data from the Illumina Human Body Map and ENCODE.

To be able to compare the very diverse data sets, we normalised the read count data by library size and inherent biological variance between all samples of the same condition. We performed a principal component analysis on these transformed values to get a first overview, how the data sets relate to each other (compare figure 33). None of the first two components differentiates between cancerous and healthy tissue, the first component instead very distinctively separates the two laboratories which produced the data. All data points on the left hand side of the plot represent Illumina Human Body Map samples, the right hand side is occupied by ENCODE data sets, cancerous as well as healthy.

Principal components are the eigenvectors from the covariance matrix of the data

The calculation of the Euclidean distances between the transformed expression vectors underlined the previous finding that the two main



Results

Calculated expression values

Results for epitope EVDPIGHLY with 4 mismatches, proteasomal cleavage threshold of 0.7, weight of N-terminal amino acids in TAP prediction set to 0.2 and MHC score calculation for the alleles A0101:

proteinID	epitope	index	mismatch	cleavageScore	tapScore	mhcScore	combinedScore	transcriptID	proteinN
NP_005353.1	EVDPIGHLY	167	0	0.0693	0.0049	A0101:43,	0.3321	ENST00000370278.3	MAGEA3
XP_006724881.1	EVDPIGHLY	167	0	0.0693	0.0049	A0101:43,	0.3321	ENST00000370278.3	MAGEA3
NP_005354.1	EVDPIGHVY	167	1	0.1770	0.0049	A0101:213,	0.1712	ENST00000329342.5	MAGEA6
NP_001159872.1	EVDPAGHSY	170	2	0.2227	0.0049	A0101:225,	0.2040	ENST00000535454.1	MAGEA8
NP_005355.2	EVDPAGHSY	170	2	0.2227	0.0049	A0101:225,	0.2040	ENST00000286482.1	MAGEA8
NP_001159873.1	EVDPAGHSY	170	2	0.2227	0.0049	A0101:225,	0.2040	ENST00000542674.1	MAGEA8
XP_005262392.1	EVDPAGHSY	166	2	0.2192	0.0049	A0101:225,	0.2008	ENST00000243314.5	MAGEA9B
XP_005262391.1	EVDPAGHSY	166	2	0.2192	0.0049	A0101:225,	0.2008	ENST00000243314.5	MAGEA9B
XP_005262393.1	EVDPAGHSY	166	2	0.2192	0.0049	A0101:225,	0.2008	ENST00000243314.5	MAGEA9B
NP_005356.1	EVDPAGHSY	166	2	0.2192	0.0049	A0101:225,	0.2008	ENST00000243314.5	MAGEA9B
NP_001159858.1	EVVRIGHLY	167	2	0.9200	0.0015	A0101:11348,	0.0538	ENST00000393900.3	MAGEA12
NP_005358.2	EVVRIGHLY	167	2	0.9200	0.0015	A0101:11348,	0.0538	ENST00000357916.4	MAGEA12
NP_001159859.1	EVVRIGHLY	167	2	0.9200	0.0015	A0101:11348,	0.0538	ENST00000393869.3	MAGEA12
NP_775970.1	EVDPIRHYY	165	2	0.0651	0.0049	A0101:254,	0.0528	ENST00000325250.1	MAGEB18
XP_005277896.1	EVVPISHLY	167	2	0.1492	0.0015	A0101:5635,	0.0176	ENST00000370293.2	MAGEA2B
NP_705692.1	EVVPISHLY	167	2	0.1492	0.0015	A0101:5635,	0.0176	ENST00000370293.2	MAGEA2B

(a) Affinities, MAGE

NP_597676.3	ESDPIVAQY	17037	4	0.0498	0.0052	A0101:38,	0.2507	ENST00000359218.5	TTN
XP_005246888.1	ESDPIVAQY	12649	4	0.0498	0.0052	A0101:38,	0.2507	ENST00000589042.1	TTN
XP_006712788.1	ESDPIVAQY	14008	4	0.0498	0.0052	A0101:38,	0.2507	ENST00000589042.1	TTN
NP_596869.4	ESDPIVAQY	23409	4	0.0498	0.0052	A0101:38,	0.2507	ENST00000342992.6	TTN
NP_003310.4	ESDPIVAQY	16912	4	0.0498	0.0052	A0101:38,	0.2507	ENST00000460472.2	TTN
NP_597681.4	ESDPIVAQY	17104	4	0.0498	0.0052	A0101:38,	0.2507	ENST00000342175.6	TTN
NP_001254479.2	ESDPIVAQY	25977	4	0.0498	0.0052	A0101:38,	0.2507	ENST00000589042.1	TTN

(b) Affinities, Titin

Results

Calculated expression values

Results for epitope EVDPIGHLY with 4 mismatches, proteasomal cleavage threshold of 0.7, weight of N-terminal amino acids in TAP prediction set to 0.2 and MHC score calculation for the alleles A0101:

proteinName	adipose	adrenal	BJ_Rep1	BJ_Rep2	blood	brain	breast	colon	HCT116_Rep1	HCT116_Rep2	heart	HEPG2_Rep1
MAGEA3	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	46.83	45.60	0.00	0.01
MAGEA3	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	46.83	45.60	0.00	0.01
MAGEA6	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	12.67	11.68	0.00	0.06
MAGEA8	0.00	0.01	0.02	0.10	0.00	0.00	0.00	0.01	0.06	0.01	0.01	0.02
MAGEA8	0.00	0.02	0.02	0.08	0.00	0.01	0.00	0.01	0.05	0.02	0.01	0.02
MAGEA8	0.00	0.02	0.01	0.07	0.00	0.00	0.00	0.01	0.04	0.02	0.01	0.02
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08
MAGEA9B	0.00	0.01	0.00	0.05	0.00	0.02	0.00	0.00	0.09	0.18	0.00	0.08
MAGEA12	0.00	0.04	0.00	0.00	0.00	0.19	0.00	0.00	0.17	0.13	0.00	0.01
MAGEA12	0.00	0.04	0.00	0.00	0.00	0.21	0.00	0.00	0.14	0.11	0.00	0.00
MAGEA12	0.00	0.04	0.00	0.00	0.00	0.20	0.00	0.00	0.16	0.14	0.00	0.01
MAGEB18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MAGEA2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	40.33	38.54	0.00	0.03
MAGEA2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	40.33	38.54	0.00	0.03

(c) Expression, MAGE

TTN	0.13	0.24	0.30	0.21	0.26	0.07	4.22	0.09	0.14	0.18	121.26	0.26
TTN	0.11	0.22	0.24	0.16	0.20	0.08	3.43	0.08	0.11	0.14	97.05	0.19
TTN	0.11	0.22	0.24	0.16	0.20	0.08	3.43	0.08	0.11	0.14	97.05	0.19
TTN	0.11	0.22	0.25	0.17	0.21	0.08	3.66	0.08	0.11	0.15	102.51	0.21
TTN	0.13	0.25	0.30	0.21	0.26	0.07	4.24	0.09	0.14	0.18	121.81	0.26
TTN	0.13	0.24	0.30	0.21	0.26	0.07	4.21	0.09	0.14	0.18	120.98	0.26

(d) Expression, Titin

Figure 32: Screenshots of Expitope showing expression of Titin.
We used the peptide from the study of Linette et al. [114] (EVDPIGHLY) as input and allowed for up to four mismatches. The HLA allele was set to A0101, all other variable input parameters were left set to their default value. Notable is the higher affinity of the Titin peptide to HLA-A0101 than that of the original input from MAGE-A3 and the very high expression of Titin in heart tissue (colour coded in subfigures c and d)

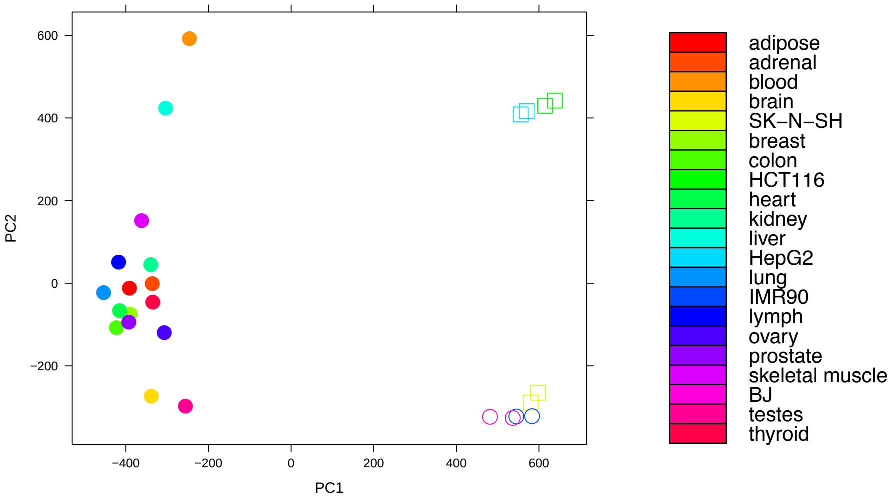


Figure 33: Principal component analysis of the transformed and normalised transcript expression values in all RNA-Seq data from our database.
The first component accounts for 30.63% of the variance, the second component for 11.78%.
Circles represent normal cell lines, squares depict cancer data sets; filled symbols show data from the Illumina Human Body Map 2.0 series, bordered ones represent ENCODE samples.

clusters visible in the database are not healthy vs. cancerous cell lines but instead depending on the laboratory of origin of the data sets. Within the Illumina Human Body Map data it is striking that the tissues testes, brain, skeletal muscles, liver and blood show greater distances to all other tissues and are thus not part of the lowest level cluster. In the ENCODE cluster, the three cell types which show the greatest similarity are skin fibroblasts, healthy lung and brain cancer (BJ, IMR90 and SK-N-SH, respectively). Thus not even within the ENCODE super cluster do tumour and healthy cell lines form their own sub cluster (compare figure 34).

All transcripts with a logarithmic fold change >2 *and an adjusted p-value* <0.001 *were classified as significantly differentially expressed*

After executing three differential expression analysis between each of the three tumour cell lines and all combined healthy tissues, we filtered the output for transcripts that showed a logarithmic fold change greater than two and an adjusted p-value smaller 0.001 which represents a false discovery rate of 0.1%. The remaining transcript were classified as significantly differentially expressed. We visualised these lists in a heatmap to see how many genes pass the filtering process and what their expression values are in the respective samples. It is striking that all transcripts which pass the filtering are over expressed in the cancer replicates and show lower expression in the healthy

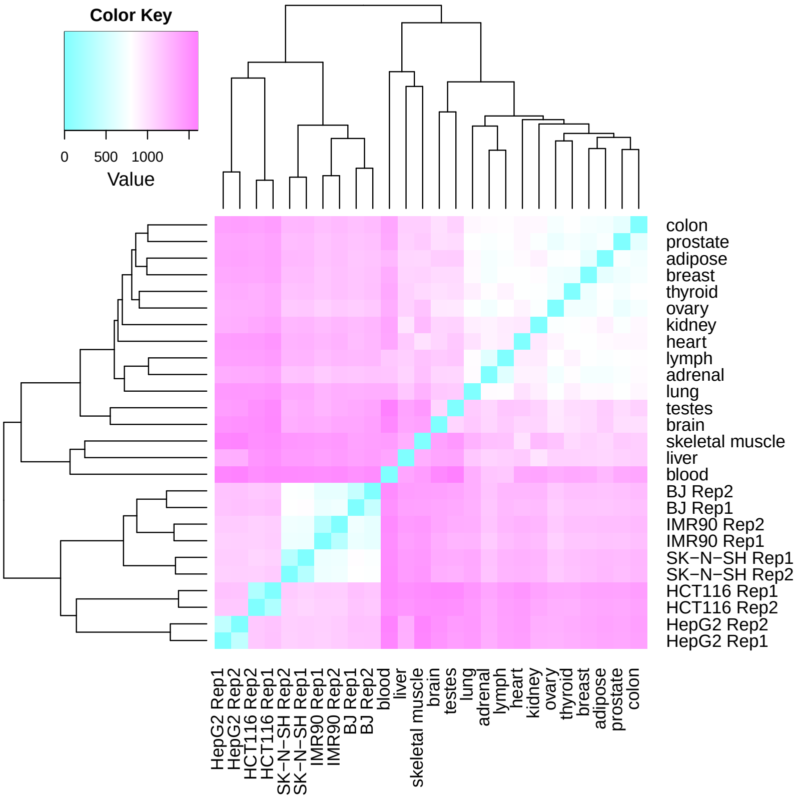
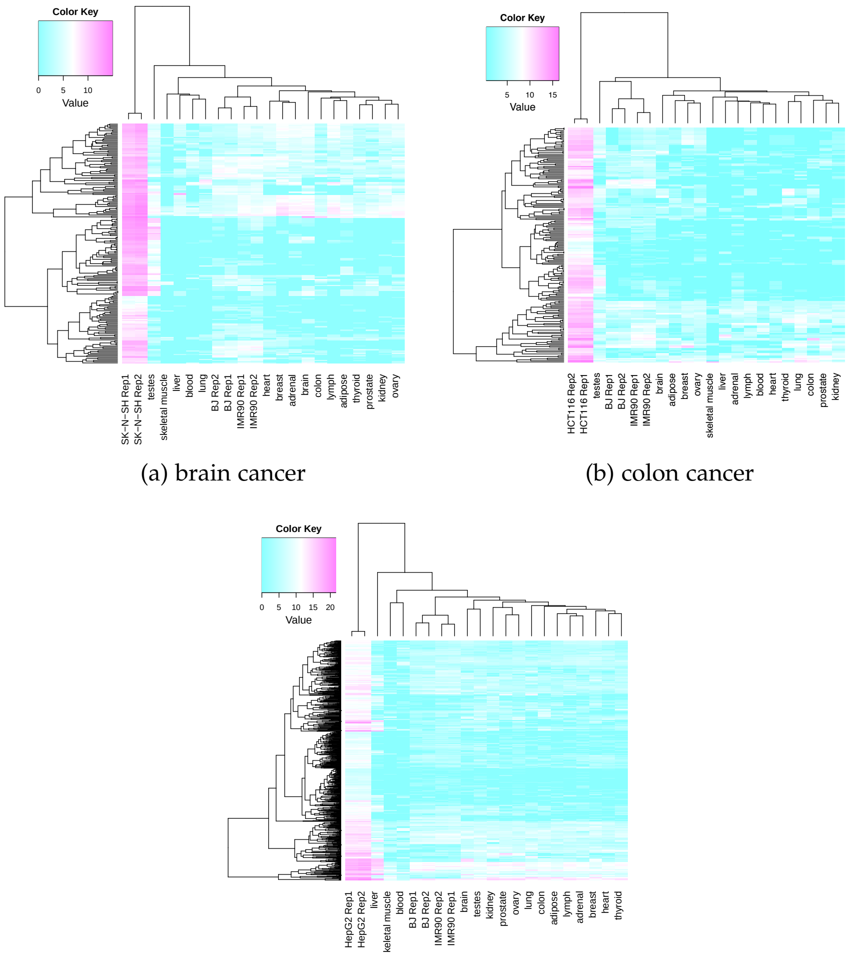


Figure 34: Heatmap showing the Euclidean distance between the transformed expression values in the database samples.
The three cancer cell lines (SK-N-SH, HepG2 and HCT-116) do not show the greatest similarity towards each other, but are clustered in an *ENCODE* cluster together with normal lung (IMR90) and skin (BJ) cell lines.

samples. It can be seen that the amount of significant differentially expressed transcripts is lower for brain and colon cancer, but much higher for the liver cancer (198, 190 and 781, respectively). Furthermore, the healthy tissue which shows the least distance to the cancerous replicates is testes in the comparison with brain and colon cancer, but the healthy liver in the liver cancer analysis (compare figure 35). This could mean that the reason for the large number of identified differentially expressed transcripts in the liver analysis is that tissue specific genes are found. The identified transcripts in the other two analyses seem to be cancer testis antigens, given that the most closely related tissue with regard to their expression are testes.
In order to identify transcripts which are not only significantly differentially expressed but also non-essential for most healthy tissues, we further filtered the list for entries that showed a mean normalised expression in all normal samples of below 0.3. We found 8, 8 and 20 transcripts in the analysis for brain, colon and liver cancer, respectively. Most of these identified transcripts turned out to be non-coding RNAs, most of them long non-coding RNAs (lncRNAs) and micro



(a) brain cancer

(b) colon cancer

(c) liver cancer

Figure 35: Heatmap of the transformed expression values of all transcripts which pass the filtering for adjusted p-Values (< 0.001) and logarithmic fold change (> 2).

RNAs (miRNAs), some small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs), one transcript belongs to a ribosomal RNA (rRNA) and one is classified as sense intronic transcript. However, one transcript in each brain and colon as well as three transcripts in liver cancer list are protein coding. The protein coding transcript from the colon sample stems from the same gene as one of the transcripts from liver, nasopharyngeal carcinoma down-regulated gene protein (NPCDR1). However, this gene does not have a corresponding protein sequence deposited in RefSeq, because it is classified by Uniprot [10] as only having *experimental evidence at transcript level*, thus the existence of the amino acid sequence has yet to be shown on protein level.

Most transcripts significantly over-expressed in cancer are non-coding RNAs

The other two protein coding transcripts with an over expression in liver cancer belong to the same gene, neuropeptide S receptor (NPSR1), a protein that is involved in the pathogenesis of asthma

[191]. The only protein coding transcript in the brain cancer analysis is pyrin domain-containing protein 2 (PYDC2). NPSR1 and PYDC2 both have associated RefSeq protein entries, NP_997055.1 and NP_001076777.1, respectively. We obtained the amino acid sequences of both proteins, split them in overlapping nonamer peptides and calculated for each peptide the probability of being created by proteasomal cleavage as described in subsection 16.3.1. We sorted the results descending based on probability and chose a cutoff at 50% to exclude the least probable peptides. This left us with 3 and 12 higher scoring potential epitopes for PYDC2 and NPSR1, respectively. We used all 15 peptides as input for Expitope to query their affinities to TAP and the MHC alleles A0101 and A0201 because these are the most common. Non of the tested nonamers reached a satisfying combined score. All reported values for MHC affinity were four-digit numbers, thus far to high to be classified as a binder.

These results show, that our Expitope framework can easily be used for a bottom-up analysis of the contained data and even identify differentially expressed proteins. Yet, in order to determine TCR targets, more background knowledge has to be integrated to find targets that are likely to be produced by the antigen presenting machinery but at the same time do not have detrimental off- or on-target effects.

18 CONCLUSION

Our herein presented Expitope web server provides the opportunity for the user to combine many functionalities that otherwise have to be combined by hand from many different tools or web sites. Our service offers an exact and approximate string search as well as the scoring of the potential target for probabilities and affinities of the epitope presentation pathway all in one place.

We are convinced that it is a helpful service to researchers in the immunotherapy community, as is shown by its rising numbers of page visitors and a high number of returning users (34% as of March 12th, 2015). The different backgrounds of the visitors represents that the functionality of Expitope is of interest to academic and industrial researchers alike.

As shown exemplary with the published cases of Linette et al. [114] and Morgan et al. [125], Expitope is a powerful tool when used as a first instance in TCR selection to identify potential detrimental off-target effects. As the number of allowed mismatches is essential in these searches and Expitope's results are based on multiple prediction steps, all of the obtained results should be verified experimentally before being used in a clinical or therapeutical setting.

Our bottom-up analysis revealed that the different steps of Expitope's workflow can also be utilised independently or in a different order. While this approach seemed promising at first, the results showed that the choice of TCR target is not trivial and requires additional knowledge other than epitope scoring and expression analysis. Yet, we are confident that Expitope's services can be of great help to the community working on cancer immunotherapy, especially since its in silico approach is automated, faster and cheaper when compared to clinical laboratory testing.

19 OUTLOOK

Due to the numerous page visits and high rate of returning users, we are convinced that Expitope is useful to the research community and we should constantly maintain and update its functionalities and web interface. In the following we will describe our visions for future updates and enhanced performance.

19.1 EXPRESSION DATABASE

19.1.1 *RNA-Seq expression*

A comprehensive database of RNA-Seq data is essential to provide maximum functionality of our web service. Thus, we invested much time and effort in finding a dataset which has a multitude of healthy human tissue as opposed to cell lines which have already undergone substantial molecular changes in order to immortalise them. We require the samples to be treated as similar as possible, to provide comparable expression values. Additionally, availability was of major importance, as most tissue samples originating from patients underlie identity protecting licenses that would prohibit us from integrating it into a freely usable web server. At the time of Expitope's implementation the Illumina Human Body data fitted all these criteria best. Yet, we are still watching out for other data sources we could use for future updates.

A future improvement of Expitope's expression database is of importance, as our currently used GENCODE annotation (V19) is the last one to be provided for human genome assembly 19. With the publication of the latest version, hg38, all new updates of GENCODE will be referencing hg38 and no new annotations for hg19 will be produced. Thus, it might be advisable to move to RNA-Seq datasets that have been mapped to the latest assembly of the human genome so we can apply the most up-to-date annotations.

In this regard we plan to utilise the next round of the ENCODE project, as the recently started phase has already begun to sequence tissue samples additionally to cell lines. Given ENCODE's free availability, this is the ideal future data source for Expitope's expression values, especially when more different tissue types have been sequenced by the same contributing lab.

Updating our expression database is fast, as all needed scripts have been prepared to make Expitope easily expandable in case of updates or specific requests by the users.

19.1.2 *Protein abundance*

In our current Expitope setup, we search for a given input peptide and report all exact and approximate hits in the whole of the human reference proteins. To define the incidence of a target protein in healthy tissue, we map it to its corresponding transcript and use expression values to define presence or absence. This approach relies heavily on the assumption, that transcript expression gives a direct measure of the abundance of protein in a tissue. Multiple publications have investigated this correlation and come to the consensus that transcript expression and protein abundance are not perfectly correlated [193, 59].

Two recent studies presenting an unprecedented comprehensive human proteome [201, 97] would give us a great data source for integrating proteomics data into Expitope. However, the question remains, if we want to use it to replace the transcript expression. On the one hand, proteomics is still not as sensitive as RNA-Seq and thus not covering the complete variety of human proteins; Wilhelm et al. [201] are estimating that they include about 84% with evidence on transcript level in their version. Due to the technicalities of mass spectrometry, the measure by which the proteome maps were constructed, different isoforms of the same proteins can only be differentiated if a very specific region of the protein is measured, hence most proteins are only defined for their standard variant.

One the other hand, if we report protein abundance additionally to transcript expression, the conflicting results, due to the low correlation, might alienate some users. This effect might even be amplified by the fact that the two measurements stem from different biological samples and thus contain inherent differences.

Ideally, we would like to integrate protein abundance data that has a matching set of RNA-Seq samples from the same background. With the current rise of multi-omics projects we hope that this data will become available in the foreseeable future.

19.2 TEXT MINING

Another enhancement we hope to add to Expitope's functionalities in the future is text mining. We imagine to search relevant publications for the mentioning of the identified potential cross-reacting peptides and/or their corresponding proteins. This way, we can provide the users with information over the immunological importance of the other targets and evaluate, if they have already been used for therapies.

One open question regarding this improvement is how we define *relevant* papers. Most openly available text mining tools, like GoPubMed [43], only index the abstracts of their publication libraries in order to save computing space and time. Before implementing it into the Expitope server we would need to determine whether abstracts are sufficient to find papers of interest or if we require full-text indexing when we want to search for peptide sequences. Because the full-text variant has intensive system requirements, we would need to find a trade-off to reduce it. Possibilities would be to either only report text mining results for the top n peptide hits from Expitope, or to use only a very limited number of journals which texts are indexed. For the latter approach, we could query Expitope's users which subset would be of interest to them and what journals belong to the most important in the field of immunotherapy.

Part V

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4th edition, 2002.

[2] D. J. Allen, J. J. Gray, C. I. Gallimore, J. Xerry, and M. Iturriza-Gómara. Analysis of amino acid variation in the P2 domain of the GII-4 norovirus VP1 protein reveals putative variant-specific epitopes. *PloS one*, 3(1):e1485, 2008.

[3] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[4] A. C. Andersson, A. C. Svensson, C. Rolny, G. Andersson, and E. Larsson. Expression of human endogenous retrovirus ERV3 (HERV-R) mRNA in normal and neoplastic tissues. *Int. J. Oncol.*, 12(2):309–313, 1998.

[5] J. M. Antony, A. M. Deslauriers, R. K. Bhat, K. K. Ellestad, and C. Power. Human endogenous retroviruses and multiple sclerosis: innocent bystanders or disease determinants? *Biochim. Biophys. Acta*, 1812(2):162–176, 2011.

[6] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe, and A. Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, 12(Suppl 6):S1, 2011.

[7] D. Baltimore. Expression of animal virus genomes. *Bacteriological reviews*, 35(3):235–41, 1971.

[8] B. Bánfai, H. Jia, J. Khatun, E. Wood, B. Risk, W. E. Gundling, A. Kundaje, H. P. Gunawardena, Y. Yu, L. Xie, K. Krajewski, B. D. Strahl, X. Chen, P. Bickel, M. C. Giddings, J. B. Brown, and L. Lipovich. Long noncoding rnas are rarely translated in two human cell lines. *Genome Research*, 22(9):1646–1657, 2012.

[9] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, 41 (Database issue):D991–995, 2013.

[10] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bing-

ley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. D. Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L. G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. MacDougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, B. E. Suzek, C. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, M. S. Yerramalla, and J. Zhang. UniProt: a hub for protein information. *Nucleic Acids Res.*, 43(Database issue):D204–212, 2015.

[11] N. Beerenwinkel and O. Zagordi. Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, 1: 1–6, 2011.

[12] F. Benachenhou, P. Jern, M. Oja, G. Sperber, V. Blikstad, P. Somervuo, S. Kaski, and J. Blomberg. Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data. *PLoS ONE*, 4 (4):e5179, 2009.

[13] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 (1):289–300, 1995.

[14] S. Best, P. Le Tissier, G. Towers, and J. P. Stoye. Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature*, 382(6594):826–829, 1996.

[15] M. Bhasin and G. P. Raghava. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.*, 13 (3):596–607, 2004.

[16] C. K. Biebricher and M. Eigen. What is a quasispecies? *Curr. Top. Microbiol. Immunol.*, 299:1–31, 2006.

[17] K. Bieda, A. Hoffmann, and K. Boller. Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. *J. Gen. Virol.*, 82(Pt 3):591–596, 2001.

[18] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega,

C. W. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. Haidar, Y. Yu, Y. Ruan, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

[19] P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*, 329(6139):506–512, 1987.

[20] J. S. Blum, P. A. Wearsch, and P. Cresswell. Pathways of antigen processing. *Annu. Rev. Immunol.*, 31:443–473, 2013.

[21] K. Bok, E. J. Abente, M. Realpe-Quintero, T. Mitra, S. V. Sosnovtsev, A. Z. Kapikian, and K. Y. Green. Evolutionary Dynamics of GII.4 Noroviruses over a 34-Year Period. *Journal of Virology*, 83(22):11890–11901, 2009.

[22] K. Boller, K. Schönfeld, S. Lischer, N. Fischer, A. Hoffmann, R. Kurth, and R. R. Tönjes. Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. *J. Gen. Virol.*, 89(Pt 2):567–572, Feb 2008.

[23] M. R. Breese and Y. Liu. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics*, 29(4):494–496, 2013.

[24] M. Browning and P. Krausa. Genetic diversity of HLA-A2: evolutionary and functional significance. *Immunol. Today*, 17(4):165–170, 1996.

[25] V. Brusic, P. van Endert, J. Zeleznikow, S. Daniel, J. Hammer, and N. Petrovsky. A neural network model approach to the study of human TAP transporter. *In Silico Biol. (Gedrukt)*, 1(2): 109–121, 1999.

[26] K. Buscher, S. Hahn, M. Hofmann, U. Trefzer, M. Ozel, W. Sterry, J. Lower, R. Lower, R. Kurth, and J. Denner. Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines. *Melanoma Res.*, 16(3):223–234, 2006.

[27] A. Buzdin, E. Kovalskaya-Alexandrova, E. Gogvadze, and E. Sverdlov. GREM, a technique for genome-wide isolation and quantitative analysis of promoter active repeats. *Nucleic Acids Res.*, 34(9):e67, 2006.

[28] Caltech RNA-Seq. Gene Expression Omnibus Entry for RNA-seq from ENCODE/Caltech (GSE33480). `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33480`, 2014. Accessed: 2014-09-06.

[29] M. O. Carneiro, C. Russ, M. G. Ross, S. B. Gabriel, C. Nusbaum, and M. A. DePristo. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, 13:375, 2012.

[30] S. E. Celniker, L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, and R. H. Waterston. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

[31] J. M. Coffin, S. H. Hughes, and H. E. Varmus. *Retroviruses*. Cold Spring Harbor Laboratory Press, New York, 1997.

[32] C. J. Cohen, W. M. Lock, and D. L. Mager. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*, 448(2):105–114, 2009.

[33] W. E. Cohn and K. Moldave. *54. Progress in Nucleic Acid Research and Molecular Biology*. Academic Press, Waltham, 1996.

[34] R. Contreras-Galindo, M. H. Kaplan, P. Leissner, T. Verjat, I. Ferlenghi, F. Bagnoli, F. Giusti, M. H. Dosik, D. F. Hayes, S. D. Gitlin, and D. M. Markovitz. Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. *J. Virol.*, 82(19):9329–9336, 2008.

[35] R. Contreras-Galindo, M. H. Kaplan, S. He, A. C. Contreras-Galindo, M. J. Gonzalez-Hernandez, F. Kappes, D. Dube, S. M. Chan, D. Robinson, F. Meng, M. Dai, S. D. Gitlin, A. M. Chinnaiyan, G. S. Omenn, and D. M. Markovitz. HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Res.*, 23(9):1505–1513, 2013.

[36] A. Cordonnier, J. F. Casella, and T. Heidmann. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *J. Virol.*, 69(9):5890–5897, 1995.

[37] S. Criscione, Y. Zhang, W. Thompson, J. Sedivy, and N. Neretti. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*, 15(1):583, 2014.

[38] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2015. *Nucleic Acids Res.*, 43(Database issue):D662–669, 2015.

[39] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, and T. R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–8, 2012.

[40] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[41] E. Domingo and J. J. Holland. RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.*, 51:151–178, 1997.

[42] E. Domingo, J. Sheldon, and C. Perales. Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.*, 76(2):159–216, 2012.

[43] A. Doms and M. Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, 33(Web Server issue):W783–786, Jul 2005.

[44] E. F. Donaldson, L. C. Lindesmith, A. D. LoBue, and R. S. Baric. Viral shape-shifting: norovirus evasion of the human immune system. *Nature reviews. Microbiology*, 8:231–41, 2010.

[45] J. Duitama, D. M. Kumar, E. Hemphill, M. Khan, I. I. Mandoiu, and C. E. Nelson. PrimerHunter: a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Res.*, 37(8):2483–2492, 2009.

[46] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, 2002.

[47] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[48] E. Eisenberg and E. Y. Levanon. Human housekeeping genes, revisited. *Trends Genet.*, 29(10):569–574, 2013.

[49] T. ENCODE Project Consortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol*, 9(4):e1001046, 2011.

[50] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, 8(3):175–185, 1998.

[51] K. Falk, O. Rötzschke, S. Stevanović, G. Jung, and H. G. Rammensee. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature*, 351(6324):290–296, 1991.

[52] J. Felsenstein. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989.

[53] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle. Ensembl 2012. *Nucleic Acids Research*, 40 (D1):D84–D90, 2012.

[54] A. Flockerzi, A. Ruggieri, O. Frank, M. Sauter, E. Maldener, B. Kopper, B. Wullich, W. Seifarth, N. Müller-Lantzsch, C. Leib-Mösch, E. Meese, and J. Mayer. Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. *BMC Genomics*, 9:354, 2008.

[55] O. Frank, C. Verbeke, N. Schwarz, J. Mayer, A. Fabarius, R. Hehlmann, C. Leib-Mösch, and W. Seifarth. Variable transcriptional activity of endogenous retroviruses in human breast cancer. *Journal of virology*, 82(4):1808–1818, 2008.

[56] Y. Gao, H. Xu, Y. Shen, and J. Wang. Transcriptomic analysis of rice (oryza sativa) endosperm using the rna-seq technique. *Plant Molecular Biology*, 81(4-5):363–378, 2013.

[57] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, 2004.

[58] M. B. Gerstein, J. Rozowsky, K. K. Yan, D. Wang, C. Cheng, J. B. Brown, C. A. Davis, L. Hillier, C. Sisu, J. J. Li, B. Pei, A. O. Harmanci, M. O. Duff, S. Djebali, R. P. Alexander, B. H. Alver, R. Auerbach, K. Bell, P. J. Bickel, M. E. Boeck, N. P. Boley, B. W. Booth, L. Cherbas, P. Cherbas, C. Di, A. Dobin, J. Drenkow, B. Ewing, G. Fang, M. Fastuca, E. A. Feingold, A. Frankish, G. Gao, P. J. Good, R. Guigo, A. Hammonds, J. Harrow, R. A. Hoskins, C. Howald, L. Hu, H. Huang, T. J. Hubbard,

C. Huynh, S. Jha, D. Kasper, M. Kato, T. C. Kaufman, R. R. Kitchen, E. Ladewig, J. Lagarde, E. Lai, J. Leng, Z. Lu, M. MacCoss, G. May, R. McWhirter, G. Merrihew, D. M. Miller, A. Mortazavi, R. Murad, B. Oliver, S. Olson, P. J. Park, M. J. Pazin, N. Perrimon, D. Pervouchine, V. Reinke, A. Reymond, G. Robinson, A. Samsonova, G. I. Saunders, F. Schlesinger, A. Sethi, F. J. Slack, W. C. Spencer, M. H. Stoiber, P. Strasbourger, A. Tanzer, O. A. Thompson, K. H. Wan, G. Wang, H. Wang, K. L. Watkins, J. Wen, K. Wen, C. Xue, L. Yang, K. Yip, C. Zaleski, Y. Zhang, H. Zheng, S. E. Brenner, B. R. Graveley, S. E. Celniker, T. R. Gingeras, and R. Waterston. Comparative analysis of the transcriptome across distant species. *Nature*, 512(7515):445–448, 2014.

[59] A. Ghazalpour, B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian, I. N. Mungrue, C. R. Farber, J. Sinsheimer, H. M. Kang, N. Furlotte, C. C. Park, P. Z. Wen, H. Brewer, K. Weitz, D. G. Camp, C. Pan, R. Yordanova, I. Neuhaus, C. Tilford, N. Siemers, P. Gargalovic, E. Eskin, T. Kirchgessner, D. J. Smith, R. D. Smith, and A. J. Lusis. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.*, 7 (6):e1001393, 2011.

[60] L. Gheorghe, S. Rugină, I. M. Dumitru, I. Franciuc, A. Martinescu, and I. Balaş. HLA class II alleles in Romanian patients with chronic hepatitis C. *Germs*, 5(2):44–49, 2015.

[61] R. Gifford and M. Tristem. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, 26(3):291–315, 2003.

[62] R. I. Glass, J. Noel, T. Ando, R. Fankhauser, G. Belliot, A. Mounts, U. D. Parashar, J. S. Bresee, and S. S. Monroe. The epidemiology of enteric caliciviruses from humans: a reassessment using new diagnostics. *J. Infect. Dis.*, 181 Suppl 2:S254–261, 2000.

[63] E. Gogvadze, E. Stukacheva, A. Buzdin, and E. Sverdlov. Human-specific modulation of transcriptional activity provided by endogenous retroviral insertions. *J. Virol.*, 83(12):6098–6105, 2009.

[64] N. L. Goodchild, D. A. Wilkinson, and D. L. Mager. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology*, 196(2):778–788, 1993.

[65] Google Developers. *Google Analytics*. Google Inc., Googleplex, Mountain View, California, U.S., 2014. URL `https://developers.google.com/analytics/`.

[66] D. Gosenca, U. Gabriel, A. Steidler, J. Mayer, O. Diem, P. Erben, A. Fabarius, C. Leib-Mösch, W.-K. Hofmann, and W. Seifarth. HERV-E-Mediated Modulation of PLA2G4A Transcription in Urothelial Carcinoma. *PLoS ONE*, 7(11):e49341, 2012.

[67] I. Göttesdorfer. *Identifizierung transkriptionell aktivierter humaner endogener Retroviren (HERVs) in Tumorerkrankungen*. PhD thesis, Technical University Munich, 2015.

[68] K. Haase, S. Raffegerst, D. Schendel, and D. Frishman. Expitope: a Web server for epitope expression. *Bioinformatics*, 31 (11):1854–1856, 2015.

[69] K. D. Hansen, S. E. Brenner, and S. Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, 38(12):e131, 2010.

[70] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard. Gencode: The reference human genome annotation for the encode project. *Genome Research*, 22(9):1760–1774, 2012.

[71] T. Hart, H. Komori, S. LaMere, K. Podshivalova, and D. Salomon. Finding the active genes in deep rna-seq gene expression studies. *BMC Genomics*, 14(1):778, 2013.

[72] C. F. Higgins. ABC transporters: from microorganisms to man. *Annu. Rev. Cell Biol.*, 8:67–113, 1992.

[73] W. Hilt and D. H. Wolf. [Proteasomes. Complex proteases lead to a new understanding of cellular regulation through proteolysis]. *Naturwissenschaften*, 82(6):257–268, 1995.

[74] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, 34(Database issue):D590–598, 2006.

[75] J. W. Ho, Y. L. Jung, T. Liu, B. H. Alver, S. Lee, K. Ikegami, K. A. Sohn, A. Minoda, M. Y. Tolstorukov, A. Appert, S. C. Parker, T. Gu, A. Kundaje, N. C. Riddle, E. Bishop, T. A. Egelhofer, S. S. Hu, A. A. Alekseyenko, A. Rechtsteiner, D. Asker, J. A. Belsky, S. K. Bowman, Q. B. Chen, R. A. Chen, D. S. Day, Y. Dong, A. C. Dose, X. Duan, C. B. Epstein, S. Ercan, E. A. Feingold, F. Ferrari, J. M. Garrigues, N. Gehlenborg, P. J. Good, P. Haseley, D. He, M. Herrmann, M. M. Hoffman, T. E. Jeffers, P. V. Kharchenko, P. Kolasinska-Zwierz, C. V. Kotwaliwale, N. Kumar, S. A. Langley, E. N. Larschan, I. Latorre, M. W. Libbrecht, X. Lin, R. Park, M. J. Pazin, H. N. Pham, A. Plachetka, B. Qin, Y. B. Schwartz, N. Shoresh, P. Stempor, A. Vielle, C. Wang, C. M. Whittle, H. Xue, R. E. Kingston, J. H. Kim, B. E. Bernstein, A. F. Dernburg, V. Pirrotta, M. I. Kuroda, W. S. Noble, T. D. Tullius, M. Kellis, D. M. MacAlpine, S. Strome, S. C. Elgin, X. S. Liu, J. D. Lieb, J. Ahringer, G. H. Karpen, and P. J. Park. Comparative analysis of metazoan chromatin organization. *Nature*, 512 (7515):449–452, 2014.

[76] H. G. Holzhütter and P. M. Kloetzel. A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. *Biophys. J.*, 79(3):1196–1205, 2000.

[77] H. G. Holzhütter, C. Frommel, 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.

[78] J. D. S. Hoser. *Analysis of Highthroughput-Sequencing Data on a large Scale*. PhD thesis, Technical University Munich, 2015.

[79] P. Huang, T. Farkas, S. Marionneau, W. Zhong, N. Ruvoen-Clouet, A. L. Morrow, M. Altaye, L. K. Pickering, D. S. Newburg, J. LePendu, and X. Jiang. Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: identification of 4 distinct strain-specific patterns. *J. Infect. Dis.*, 188(1): 19–31, 2003.

[80] A. Huda, N. Polavarapu, I. K. Jordan, and J. F. McDonald. Endogenous retroviruses of the chicken genome. *Biol. Direct*, 3:9, 2008.

[81] D. C. Hughes. Alternative splicing of the human VEGFGR-3/FLT4 gene as a consequence of an integrated human endogenous retrovirus. *J. Mol. Evol.*, 53(2):77–79, 2001.

[82] Human Genome Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[83] A. M. Hutson, R. L. Atmar, D. Y. Graham, and M. K. Estes. Norwalk virus infection and disease is associated with ABO histo-blood group type. *J. Infect. Dis.*, 185(9):1335–1337, 2002.

[84] C. A. Janeway, Jr, P. Travers, M. Walport, and M. J. Shlomchik. *Immunobiology, The Immune System in Health and Disease*. Garland Science, New York, 5th edition, 2001.

[85] P. Jern, G. O. Sperber, and J. Blomberg. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology*, 2:50, 2005.

[86] L. A. Johnson, R. A. Morgan, M. E. Dudley, L. Cassard, J. C. Yang, M. S. Hughes, U. S. Kammula, R. E. Royal, R. M. Sherry, J. R. Wunderlich, C. C. Lee, N. P. Restifo, S. L. Schwarz, A. P. Cogdill, R. J. Bishop, H. Kim, C. C. Brewer, S. F. Rudy, C. VanWaes, J. L. Davis, A. Mathur, R. T. Ripley, D. A. Nathan, C. M. Laurencot, and S. A. Rosenberg. Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood*, 114(3): 535–546, 2009.

[87] W. E. Johnson and J. M. Coffin. Constructing primate phylogenies from ancient retrovirus sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 96(18):10254–10260, 1999.

[88] C. P. Johnston, H. Qiu, J. R. Ticehurst, C. Dickson, P. Rosenbaum, P. Lawson, A. B. Stokes, C. J. Lowenstein, M. Kaminsky, S. E. Cosgrove, K. Y. Green, and T. M. Perl. Outbreak management and implications of a nosocomial norovirus outbreak. *Clinical infectious diseases*, 45:534–540, 2007.

[89] M. K. Jones, M. Watanabe, S. Zhu, C. L. Graves, L. R. Keyes, K. R. Grau, M. B. Gonzalez-Hernandez, N. M. Iovine, C. E. Wobus, J. Vinjé, S. A. Tibbetts, S. M. Wallet, and S. M. Karst. Enteric bacteria promote human and mouse norovirus infection of B cells. *Science*, 346(755):755–759, 2014.

[90] T. Junier and E. M. Zdobnov. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, 26(13):1669–1670, 2010.

[91] A. Z. Kapikian, R. G. Wyatt, R. Dolin, T. S. Thornhill, A. R. Kalica, and R. M. Chanock. Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis. *J. Virol.*, 10(5):1075–1081, 1972.

[92] H. Karlsson, J. Schroder, S. Bachmann, C. Bottmer, and R. H. Yolken. HERV-W-related RNA detected in plasma from indi-

viduals with recent-onset schizophrenia or schizoaffective disorder. *Mol. Psychiatry*, 9(1):12–13, 2004.

[93] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(suppl 1):D493–D496, 2004.

[94] C. Keşmir, A. K. Nussbaum, H. Schild, V. Detours, and S. Brunak. Prediction of proteasome cleavage motifs by neural networks. *Protein Engineering*, 15(4):287–296, 2002.

[95] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664, 2002.

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

[97] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, 2014.

[98] M. Kimura. The neutral theory of molecular evolution. *Sci. Am.*, 241(5):98–100, 1979.

[99] A. Kleiman, N. Senyuta, A. Tryakin, M. Sauter, A. Karseladze, S. Tjulandin, V. Gurtsevitch, and N. Mueller-Lantzsch. HERV-K(HML-2) GAG/ENV antibodies as indicator for therapy effect in patients with germ cell tumors. *Int. J. Cancer*, 110(3):459–461, 2004.

[100] P. M. Kloetzel and F. Ossendorp. Proteasome and peptidase function in MHC-class-I-mediated antigen presentation. *Curr. Opin. Immunol.*, 16(1):76–81, 2004.

[101] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma. ArrayExpress update-simplifying data submissions. *Nucleic Acids Res.*, 43(Database issue):D1113–1116, 2015.

[102] D. L. Kolson and F. Gonzalez-Scarano. Endogenous retroviruses and multiple sclerosis. *Ann. Neurol.*, 50(4):429–430, 2001.

[103] J. O. Koopmann, M. Post, J. J. Neefjes, G. J. Hammerling, and F. Momburg. Translocation of long peptides by transporters associated with antigen processing (TAP). *Eur. J. Immunol.*, 26 (8):1720–1728, 1996.

[104] M. Kraman, P. J. Bambrough, J. N. Arnold, E. W. Roberts, L. Magiera, J. O. Jones, A. Gopinathan, D. A. Tuveson, and D. T. Fearon. Suppression of antitumor immunity by stromal cells expressing fibroblast activation protein-alpha. *Science*, 330(6005): 827–830, 2010.

[105] C. Kuttler, A. K. Nussbaum, T. P. Dick, H. G. Rammensee, H. Schild, and K. P. Hadeler. An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol.*, 298(3):417–429, 2000.

[106] W. P. Lee, M. P. Stromberg, A. Ward, C. Stewart, E. P. Garrison, and G. T. Marth. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*, 9 (3):e90581, 2014.

[107] M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607): 682–686, 2003.

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

[109] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[110] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[111] Y. Liao, G. K. Smyth, and W. Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[112] L. Lindesmith, C. Moe, S. Marionneau, N. Ruvoen, X. Jiang, L. Lindblad, P. Stewart, J. LePendu, and R. Baric. Human susceptibility and resistance to Norwalk virus infection. *Nat. Med.*, 9(5):548–553, 2003.

[113] L. C. Lindesmith, E. F. Donaldson, A. D. LoBue, J. L. Cannon, D.-P. Zheng, J. Vinje, and R. S. Baric. Mechanisms of gii.4 norovirus persistence in human populations. *PLoS Med*, 5(2): e31, 2008.

[114] G. P. Linette, E. A. Stadtmauer, M. V. Maus, A. P. Rapoport, B. L. Levine, L. Emery, L. Litzky, A. Bagg, B. M. Carreno, P. J. Cimino, G. K. Binder-Scholl, D. P. Smethurst, A. B. Gerry, N. J. Pumphrey, A. D. Bennett, J. E. Brewer, J. Dukes, J. Harper, H. K. Tayton-Martin, B. K. Jakobsen, N. J. Hassan, M. Kalos, and C. H. June. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood*, 122(6):863–871, 2013.

[115] Y.-H. Ling, H. Xiang, Y.-S. Li, Y. Liu, Y.-H. Zhang, Z.-J. Zhang, J.-P. Ding, and X.-R. Zhang. Exploring differentially expressed genes in the ovaries of uniparous and multiparous goats using the rna-seq (quantification) method. *Gene*, 550(1):148 – 153, 2014.

[116] X. Lu, F. Sachs, L. Ramsay, P.-E. Jacques, J. Göke, G. Bourque, and H.-H. Ng. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol*, 21:423–425, 2014.

[117] C. Lundegaard, K. Lamberth, M. Harndahl, S. Buus, O. Lund, and M. Nielsen. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Research*, 36(suppl 2): W509–W512, 2008.

[118] C. Lundegaard, O. Lund, and M. Nielsen. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, 24(11):1397–1398, 2008.

[119] P. Maliniemi, M. Vincendeau, J. Mayer, O. Frank, S. Hahtola, L. Karenko, E. Carlsson, F. Mallet, W. Seifarth, C. Leib-Mösch, and A. Ranki. Expression of human endogenous retrovirus-w including syncytin-1 in cutaneous T-cell lymphoma. *PLoS ONE*, 8(10):e76281, 2013.

[120] H. Martins and P. Villesen. Improved integration time estimation of endogenous retroviruses with phylogenetic data. *PLoS ONE*, 6(3):e14745, 2011.

[121] J. Mayer and E. Meese. Human endogenous retroviruses in the primate lineage and their influence on host genomes. *Cytogenet. Genome Res.*, 110(1-4):448–456, 2005.

[122] P. G. Medd and B. M. Chain. Protein degradation in MHC class II antigen presentation: opportunities for immunomodulation. *Semin. Cell Dev. Biol.*, 11(3):203–210, 2000.

[123] S. Mi, X. Lee, X. Li, G. M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X. Y. Tang, P. Edouard, S. Howes, J. C. Keith, and J. M. McCoy. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403(6771): 785–789, 2000.

[124] F. Momburg, J. Roelse, G. J. Hammerling, and J. J. Neefjes. Peptide size selection by the major histocompatibility complex-encoded peptide transporter. *J. Exp. Med.*, 179(5):1613–1623, 1994.

[125] R. A. Morgan, N. Chinnasamy, D. Abate-Daga, A. Gros, P. F. Robbins, Z. Zheng, M. E. Dudley, S. A. Feldman, J. C. Yang, R. M. Sherry, G. Q. Phan, M. S. Hughes, U. S. Kammula, A. D. Miller, C. J. Hessman, A. A. Stewart, N. P. Restifo, M. M. Quezado, M. Alimchandani, A. Z. Rosenberg, A. Nath, T. Wang, B. Bielekova, S. C. Wuest, N. Akula, F. J. McMahon, S. Wilde, B. Mosetter, D. J. Schendel, C. M. Laurencot, and S. A. Rosenberg. Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J. Immunother.*, 36(2):133–151, 2013.

[126] 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.

[127] A. Mösch. Investigation of human endogenous retroviruses based on next-generation-sequencing data. Bachelor thesis, Technical University Munich, 2012.

[128] M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemøller, K. Lamberth, S. Buus, S. Brunak, and O. Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science*, 12(5):1007–1017, 2003.

[129] M. Nielsen, C. Lundegaard, O. Lund, and C. Keşmir. 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.

[130] A. K. Nussbaum, C. Kuttler, K. P. Hadeler, H. G. Rammensee, and H. Schild. PAProC: a prediction algorithm for proteasomal

cleavages available on the WWW. *Immunogenetics*, 53(2):87–94, 2001.

[131] M. Okada, T. Tanaka, M. Oseto, N. Takeda, and K. Shinozaki. Genetic analysis of noroviruses associated with fatalities in healthcare facilities. *Arch. Virol.*, 151(8):1635–1641, 2006.

[132] S. L. Oliver, A. M. Dastjerdi, S. Wong, L. El-Attar, C. Gallimore, D. W. Brown, J. Green, and J. C. Bridger. Molecular characterization of bovine enteric caliciviruses: a distinct third genogroup of noroviruses (Norwalk-like viruses) unlikely to be of risk to humans. *J. Virol.*, 77(4):2789–2798, 2003.

[133] E. Park, B. Williams, B. J. Wold, and A. Mortazavi. RNA editing in the human ENCODE RNA-seq data. *Genome Research*, 22(9): 1626–1633, 2012.

[134] 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.

[135] M. R. Parkhurst, J. C. Yang, R. C. Langan, M. E. Dudley, D. A. Nathan, S. A. Feldman, J. L. Davis, R. A. Morgan, M. J. Merino, R. M. Sherry, M. S. Hughes, U. S. Kammula, G. Q. Phan, R. M. Lim, S. A. Wank, N. P. Restifo, P. F. Robbins, C. M. Laurencot, and S. A. Rosenberg. T cells targeting carcinoembryonic antigen can mediate regression of metastatic colorectal cancer but induce severe transient colitis. *Mol. Ther.*, 19(3):620–626, 2011.

[136] M. M. Patel, M. A. Widdowson, R. I. Glass, K. Akazawa, J. Vinje, and U. D. Parashar. Systematic literature review of role of noroviruses in sporadic gastroenteritis. *Emerging Infect. Dis.*, 14 (8):1224–1231, 2008.

[137] J. Pačes, A. Pavlícek, and V. Pačes. Hervd: database of human endogenous retroviruses. *Nucleic Acids Research*, 30(1):205–206, 2002.

[138] J. Pačes, A. Pavlíček, R. Zika, V. V. Kapitonov, J. Jurka, and V. Pačes. Hervd: the human endogenous retroviruses database: update. *Nucleic Acids Research*, 32(suppl 1):D50, 2004.

[139] H. Perron, C. Geny, O. Genoulaz, J. Pellat, J. Perret, and J. M. Seigneurin. Antibody to reverse transcriptase of human retroviruses in multiple sclerosis. *Acta Neurol. Scand.*, 84(6):507–513, 1991.

[140] B. Peters, S. Bulik, R. Tampe, P. M. van Endert, and H.-G. Holzhütter. Identifying MHC Class I Epitopes by Predicting

the TAP Transport Efficiency of Epitope Precursors. *The Journal of Immunology*, 171(4):1741–1749, 2003.

[141] Picard. Picard: a set of tools (in java) for working with next generation sequencing data in the bam format. `http://sourceforge.net/projects/picard`, 2015. Accessed: 2015-05-13.

[142] N. Polavarapu, N. J. Bowen, and J. F. McDonald. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biol.*, 7(6):R51, 2006.

[143] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. HIV haplotype inference using a constraint-based Dirichlet process mixture model. *NIPS Workshop on Machine Learning in Computational Biology*, 2010.

[144] M. C. Prosperi, L. Yin, D. J. Nolan, A. D. Lowe, M. M. Goodenow, and M. Salemi. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Sci Rep*, 3: 2837, 2013.

[145] M. C. F. Prosperi and M. Salemi. QuRe : Software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, pages 1–3, 2011.

[146] M. C. F. Prosperi, L. Prosperi, A. Bruselles, I. Abbate, G. Rozera, D. Vincenti, M. C. Solmone, M. R. Capobianchi, and G. Ulivi. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC bioinformatics*, 12(5):1–13, 2011.

[147] K. D. Pruitt, G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. A. O'Leary, S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy, and J. M. Ostell. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, 42(Database issue):D756–763, 2014.

[148] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6): 841–842, 2010.

[149] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org`.

[150] 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.

[151] H. G. Rammensee, K. Falk, and O. Rotzschke. Peptides naturally presented by MHC class I molecules. *Annu. Rev. Immunol.*, 11:213–244, 1993.

[152] B. J. Raney, T. R. Dreszer, G. P. Barber, H. Clawson, P. A. Fujita, T. Wang, N. Nguyen, B. Paten, A. S. Zweig, D. Karolchik, and W. J. Kent. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, 30(7):1003–1005, 2014.

[153] D. Reiss and D. L. Mager. Stochastic epigenetic silencing of retrotransposons: does stability come with age? *Gene*, 390(1-2): 130–135, 2007.

[154] N. P. Restifo, M. E. Dudley, and S. A. Rosenberg. Adoptive immunotherapy for cancer: harnessing the T cell response. *Nat. Rev. Immunol.*, 12(4):269–281, 2012.

[155] P. F. Robbins, R. A. Morgan, S. A. Feldman, J. C. Yang, R. M. Sherry, M. E. Dudley, J. R. Wunderlich, A. V. Nahvi, L. J. Helman, C. L. Mackall, U. S. Kammula, M. S. Hughes, N. P. Restifo, M. Raffeld, C. C. Lee, C. L. Levy, Y. F. Li, M. El-Gamil, S. L. Schwarz, C. Laurencot, and S. A. Rosenberg. Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. *J. Clin. Oncol.*, 29(7):917–924, 2011.

[156] B. H. Rockx, H. Vennema, C. J. Hoebe, E. Duizer, and M. P. Koopmans. Association of histo-blood group antigens and susceptibility to norovirus infections. *J. Infect. Dis.*, 191(5):749–754, 2005.

[157] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, 242(1):84–89, 1996.

[158] K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, 41(Database issue):56–63, 2013.

[159] J. M. Rothberg and J. H. Leamon. The development and impact of 454 sequencing. *Nat. Biotechnol.*, 26(10):1117–1124, 2008.

[160] J. H. Russell and T. J. Ley. Lymphocyte-mediated cytotoxicity. *Annu. Rev. Immunol.*, 20:323–370, 2002.

[161] M. Sadelain, R. Brentjens, and I. Riviere. The promise and potential pitfalls of chimeric antigen receptors. *Curr. Opin. Immunol.*, 21(2):215–223, 2009.

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

[163] F. Santoni, J. Guerra, and J. Luban. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, 9(1):111, 2012.

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

[165] K. Schmitt, J. Reichrath, A. Roesch, E. Meese, and J. Mayer. Transcriptional profiling of human endogenous retrovirus group HERV-K(HML-2) loci in melanoma. *Genome Biol Evol*, 5(2):307–328, 2013.

[166] K. Schmitt. *Identifizierung und Charakterisierung transkribierter Loci der humanen endogenen Retrovirus-Gruppen HERV-K(HML-2) und HERV-W im Kontext von Tumorerkrankungen und Multipler Sklerose*. PhD thesis, Universität des Saarlandes, 2014.

[167] U. Schön, O. Diem, L. Leitner, W. H. Gunzburg, D. L. Mager, B. Salmons, and C. Leib-Mösch. Human endogenous retroviral long terminal repeat sequences as cell type-specific promoters in retroviral vectors. *J. Virol.*, 83(23):12643–12650, Dec 2009.

[168] W. Seifarth, B. Spiess, U. Zeilfelder, C. Speth, R. Hehlmann, , and C. Leib-Mösch. Assessment of retroviral activity using a universal retrovirus chip. *J. Virol. Methods*, 112(1-2):79–91, 2003.

[169] W. Seifarth, O. Frank, U. Zeilfelder, B. Spiess, A. D. Greenwood, R. Hehlmann, and C. Leib-Mösch. Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. *Journal of Virology*, 79(1):341–352, 2005.

[170] A. Shih, E. E. Coutavas, and M. G. Rush. Evolutionary implications of primate endogenous retroviruses. *Virology*, 182(2): 495–502, 1991.

[171] J. J. Siebenga, M. F. C. Beersma, H. Vennema, P. van Biezen, N. J. Hartwig, and M. Koopmans. High prevalence of prolonged norovirus shedding and illness among hospitalized patients: a

model for in vivo molecular evolution. *Journal of infectious diseases*, 198:994–1001, 2008.

[172] T. E. Spencer and M. Palmarini. Endogenous retroviruses of sheep: a model system for understanding physiological adaptation to an evolving ruminant genome. *J. Reprod. Dev.*, 58(1): 33–37, 2012.

[173] J. A. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. J. Sabo, R. Sandstrom, A. S. Stehling, R. E. Thurman, S. M. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. G. Landt, Z. Ma, B. J. Wold, J. Dekker, G. E. Crawford, C. A. Keller, W. Wu, C. Morrissey, S. A. Kumar, T. Mishra, D. Jain, M. Byrska-Bishop, D. Blankenberg, B. R. Lajoie, G. Jain, A. Sanyal, K. B. Chen, O. Denas, J. Taylor, G. A. Blobel, M. J. Weiss, M. Pimkin, W. Deng, G. K. Marinov, B. A. Williams, K. I. Fisher-Aylor, G. Desalvo, A. Kiralusha, D. Trout, H. Amrhein, A. Mortazavi, L. Edsall, D. McCleary, S. Kuan, Y. Shen, F. Yue, Z. Ye, C. A. Davis, C. Zaleski, S. Jha, C. Xue, A. Dobin, W. Lin, M. Fastuca, H. Wang, R. Guigo, S. Djebali, J. Lagarde, T. Ryba, T. Sasaki, V. S. Malladi, M. S. Cline, V. M. Kirkup, K. Learned, K. R. Rosenbloom, W. J. Kent, E. A. Feingold, P. J. Good, M. Pazin, R. F. Lowdon, and L. B. Adams. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.*, 13(8):418, 2012.

[174] C. Stocking and C. A. Kozak. Murine endogenous retroviruses. *Cell. Mol. Life Sci.*, 65(21):3383–3398, 2008.

[175] J. P. Stoye. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nat. Rev. Microbiol.*, 10(6):395–406, 2012.

[176] R. Strick, S. Ackermann, M. Langbein, J. Swiatek, S. W. Schubert, S. Hashemolhosseini, T. Koscheck, P. A. Fasching, R. L. Schild, M. W. Beckmann, and P. L. Strissel. Proliferation and cell-cell fusion of endometrial carcinoma are induced by the human endogenous retroviral Syncytin-1 and regulated by TGF-beta. *J. Mol. Med.*, 85(1):23–38, 2007.

[177] M. Tan and X. Jiang. Norovirus and its histo-blood group antigen receptors: an answer to a historical puzzle. *Trends Microbiol.*, 13(6):285–293, 2005.

[178] M. Tan and X. Jiang. Association of histo-blood group antigens with susceptibility to norovirus infection may be strain-specific rather than genogroup dependent. *J. Infect. Dis.*, 198(6):940–941, 2008.

[179] M. Tan, P. Huang, J. Meller, W. Zhong, T. Farkas, and X. Jiang. Mutations within the P2 domain of norovirus capsid affect binding to human histo-blood group antigens: evidence for a binding pocket. *Journal of virology*, 77(23):12562–71, 2003.

[180] M. Tan, M. Xia, S. Cao, P. Huang, T. Farkas, J. Meller, R. S. Hegde, X. Li, Z. Rao, and X. Jiang. Elucidation of strain-specific interaction of a GII-4 norovirus with HBGA receptors by site-directed mutagenesis study. *Virology*, 379(2):324–34, 2008.

[181] P. F. Teunis, C. L. Moe, P. Liu, S. E. Miller, L. Lindesmith, R. S. Baric, J. Le Pendu, and R. L. Calderon. Norwalk virus: how infectious is it? *J. Med. Virol.*, 80(8):1468–1476, 2008.

[182] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, 2004.

[183] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[184] C. N. Ting, M. P. Rosenberg, C. M. Snow, L. C. Samuelson, and M. H. Meisler. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev.*, 6(8):1457–1465, 1992.

[185] D. Tombácz, Z. Csabai, P. Oláh, Z. Havelda, D. Sharon, M. Snyder, and Z. Boldogkői. Characterization of novel transcripts in pseudorabies virus. *Viruses*, 7(5):2727–2744, May 2015.

[186] D. Tombácz, D. Sharon, P. Oláh, Z. Csabai, M. Snyder, and Z. Boldogkői. Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology. *Genome Announcements*, 2(4), 2014.

[187] R. R. Tönjes, R. Löwer, K. Boller, J. Denner, B. Hasenmaier, H. Kirsch, H. König, C. Korbmacher, C. Limbach, R. Lugert, R. C. Phelps, J. Scherer, K. Thelen, J. Löwer, and R. Kurth. HERV-K: the biologically most active human endogenous retrovirus family. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.*, 13 Suppl 1:S261–267, 1996.

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

[189] G. Turner, M. Barbulescu, M. Su, M. I. Jensen-Seaman, K. K. Kidd, and J. Lenz. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.*, 11(19):1531–1535, 2001.

[190] P. M. van Endert, D. Riganelli, G. Greco, K. Fleischhauer, J. Sidney, A. Sette, and J. F. Bach. The peptide-binding motif for the human transporter associated with antigen processing. *J. Exp. Med.*, 182(6):1883–1895, 1995.

[191] J. Vendelin, V. Pulkkinen, M. Rehn, A. Pirskanen, A. Raisanen-Sokolowski, A. Laitinen, L. A. Laitinen, J. Kere, and T. Laitinen. Characterization of GPRA, a novel G protein-coupled receptor related to asthma. *Am. J. Respir. Cell Mol. Biol.*, 33(3):262–270, 2005.

[192] M. Victoria, M. P. Miagostovich, M. S. Ferreira, C. B. Vieira, J. M. Fioretti, J. P. Leite, R. Colina, and J. Cristina. Bayesian coalescent inference reveals high evolutionary rates and expansion of Norovirus populations. *Infect. Genet. Evol.*, 9(5):927–932, 2009.

[193] C. Vogel and E. M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, 13(4):227–232, 2012.

[194] V. Walia, E. W. Mu, J. C. Lin, and Y. Samuels. Delving into somatic variation in sporadic melanoma. *Pigment Cell Melanoma Res*, 25(2):155–170, 2012.

[195] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[196] T. Wang, J. Zeng, C. B. Lowe, R. G. Sellers, S. R. Salama, M. Yang, S. M. Burgess, R. K. Brachmann, and D. Haussler. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U.S.A.*, 104(47):18613–18618, 2007.

[197] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.

[198] F. Wang-Johanning, A. R. Frost, B. Jian, L. Epp, D. W. Lu, and G. L. Johanning. Quantitation of HERV-K env gene expression and splicing in human breast cancer. *Oncogene*, 22(10):1528–1535, 2003.

[199] M. L. Wei and P. Cresswell. HLA-A2 molecules in an antigen-processing mutant cell contain signal sequence-derived peptides. *Nature*, 356(6368):443–446, 1992.

[200] T. Wenzel, C. Eckerskorn, F. Lottspeich, and W. Baumeister. Existence of a molecular ruler in proteasomes suggested by analysis of degradation products. *FEBS Lett.*, 349(2):205–209, 1994.

[201] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J. H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 2014.

[202] C. E. Wobus, S. M. Karst, L. B. Thackray, K. O. Chang, S. V. Sosnovtsev, G. Belliot, A. Krug, J. M. Mackenzie, K. Y. Green, and H. W. Virgin. Replication of Norovirus in cell culture reveals a tropism for dendritic cells and macrophages. *PLoS Biol.*, 2(12): e432, 2004.

[203] Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, 13(5):555–556, 1997.

[204] A. Yenigün and B. Durupinar. Decreased frequency of the HLA-DRB1*11 allele in patients with chronic hepatitis C virus infection. *J. Virol.*, 76(4):1787–1789, 2002.

[205] H. Yin, P. Medstrand, A. Kristofferson, U. Dietrich, P. Aman, and J. Blomberg. Characterization of human MMTV-like (HML) elements similar to a sequence that was highly expressed in a human breast cancer: further definition of the HML-6 group. *Virology*, 256(1):22–35, 1999.

[206] F. Yue, Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L. H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M. A. Bender, M. Zhang, R. Byron, M. T. Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y. C. Wu, M. D. Rasmussen, M. S. Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P. Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi, M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. Lacerda de Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. Ramalho Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J. Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A. P. Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. De Bruijn, L. Selleri, A. Rudensky, S. Josefowicz, R. Samstein, E. E. Eichler, S. H. Orkin, D. Levasseur, T. Papayannopoulou, K. H. Chang,

A. Skoultchi, S. Gosh, C. Disteche, P. Treuting, Y. Wang, M. J. Weiss, G. A. Blobel, X. Cao, S. Zhong, T. Wang, P. J. Good, R. F. Lowdon, L. B. Adams, X. Q. Zhou, M. J. Pazin, E. A. Feingold, B. Wold, J. Taylor, A. Mortazavi, S. M. Weissman, J. A. Stamatoyannopoulos, M. P. Snyder, R. Guigo, T. R. Gingeras, D. M. Gilbert, R. C. Hardison, M. A. Beer, and B. Ren. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364, 2014.

[207] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing dataimproper . *BMC bioinformatics*, 12(1):119, 2011.

[208] K. Zakikhany, D. J. Allen, D. Brown, and M. Iturriza Gómara. Molecular Evolution of GII-4 Norovirus Strains. *PLoS one*, 7(7): e41625, 2012.

[209] D.-P. Zheng, T. Ando, R. L. Fankhauser, R. S. Beard, R. I. Glass, and S. S. Monroe. Norovirus classification and proposed strain nomenclature. *Virology*, 346(2):312–23, 2006.

Part VI

APPENDIX

A

APPENDIX

A.1 SUPPLEMENTARY FIGURES

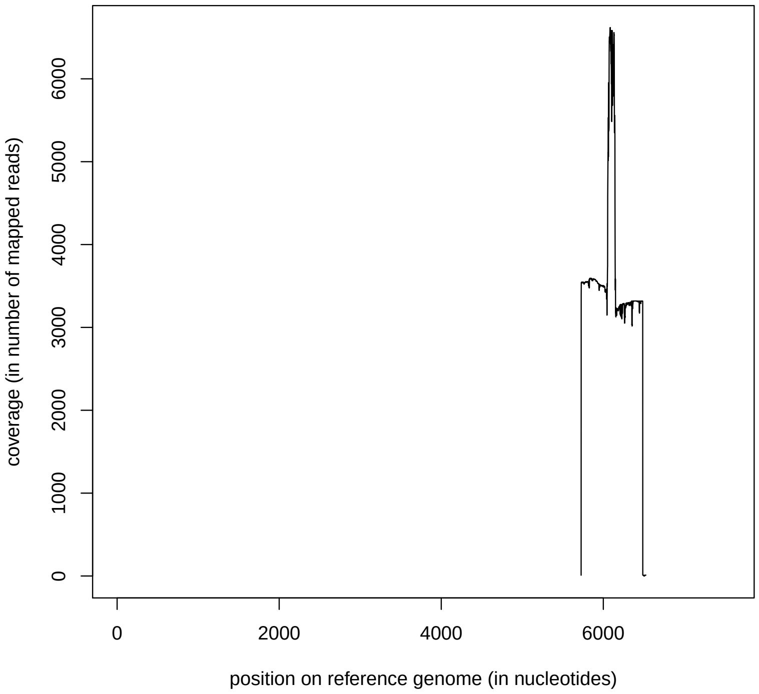
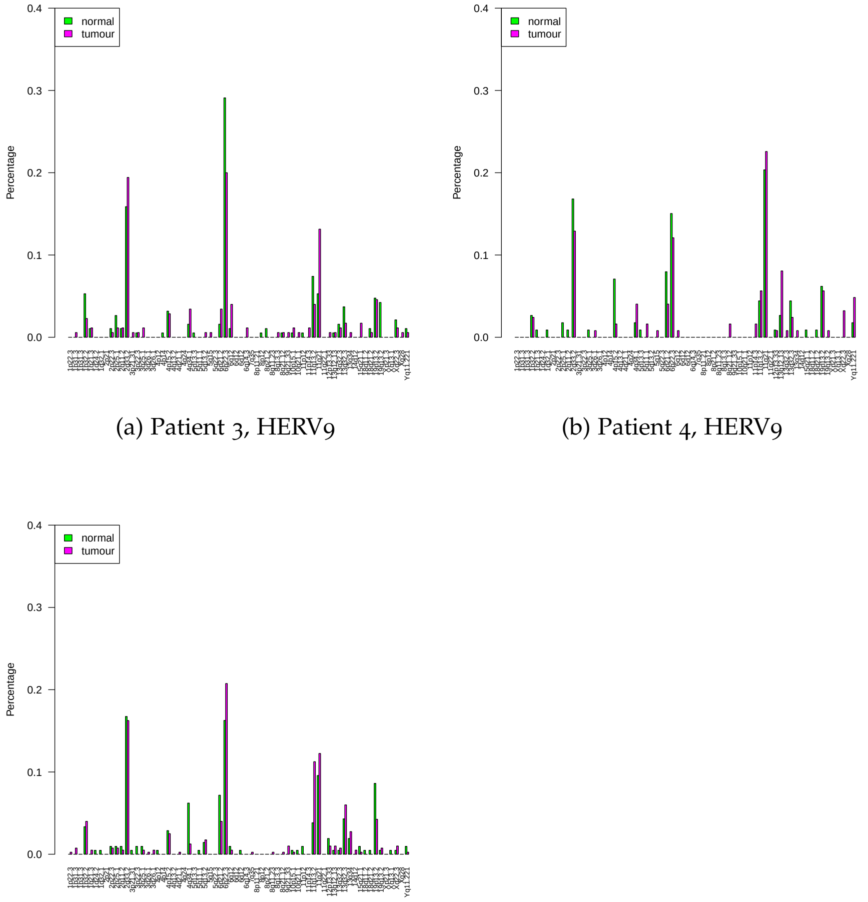


Figure A1: Example coverage of mapped reads to GenBank reference strain. The x-axis shows the length of the complete noroviral genome, reads are mapped between position 5727 and 6486. The sequencing design, two primers located at both ends of the region replicating inwards, can be recognised based on the coverage on the y-axis.
The plot shows the mapping result of all reads from the 2012 sample ATAT against GenBank reference *AY502023.1*.

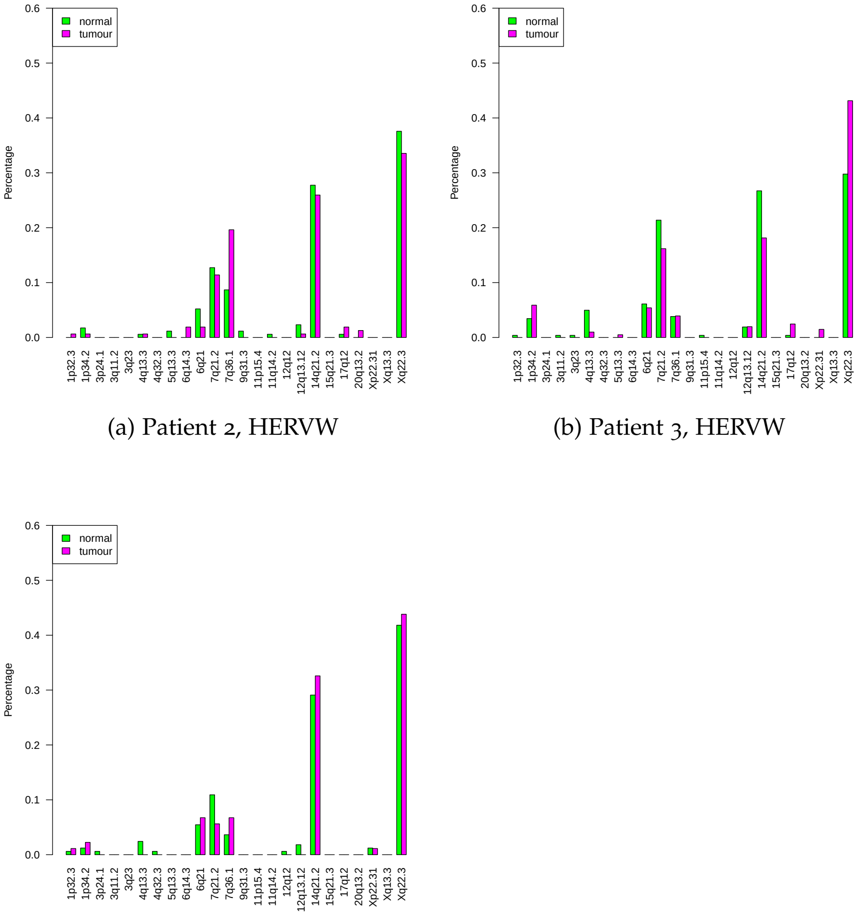


(a) Patient 3, HERV9

(b) Patient 4, HERV9

(c) Patient 5, HERV9

Figure A2: Distribution of expressed HERV9 loci in patients 3 to 5. The x-axis contains all HERV9 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.

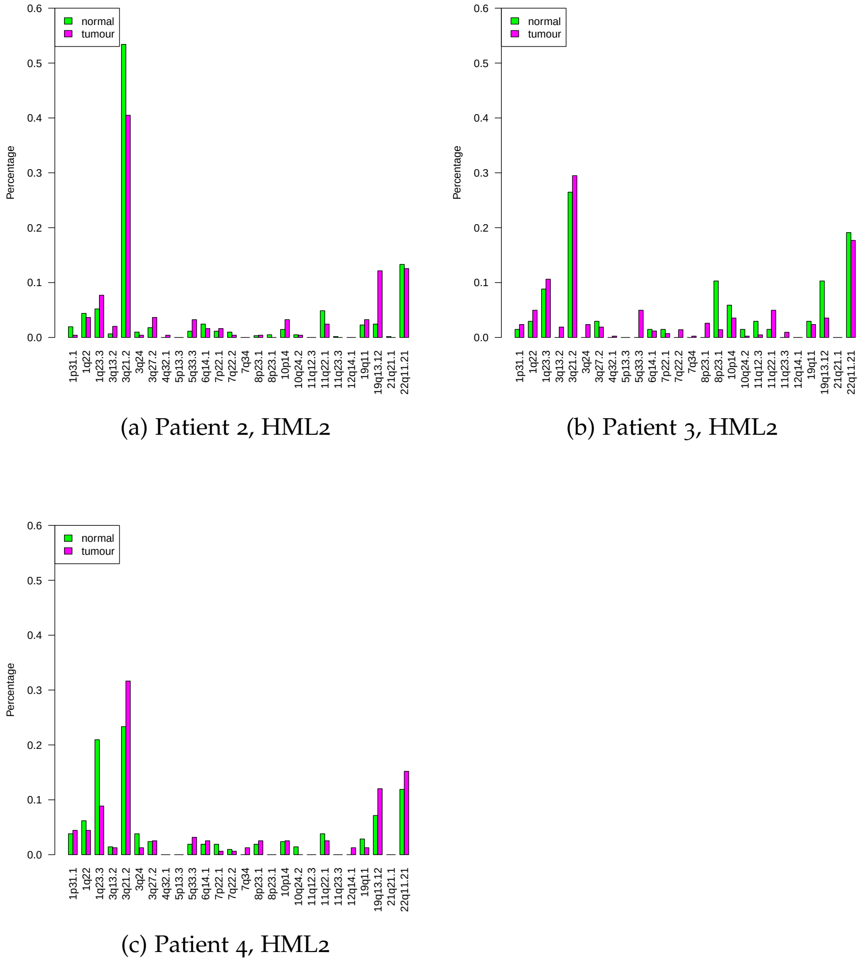


(a) Patient 2, HERVW

(b) Patient 3, HERVW

(c) Patient 4, HERVW

Figure A3: Distribution of expressed HERVW loci in patients 2 to 4. The x-axis contains all HERVW loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.

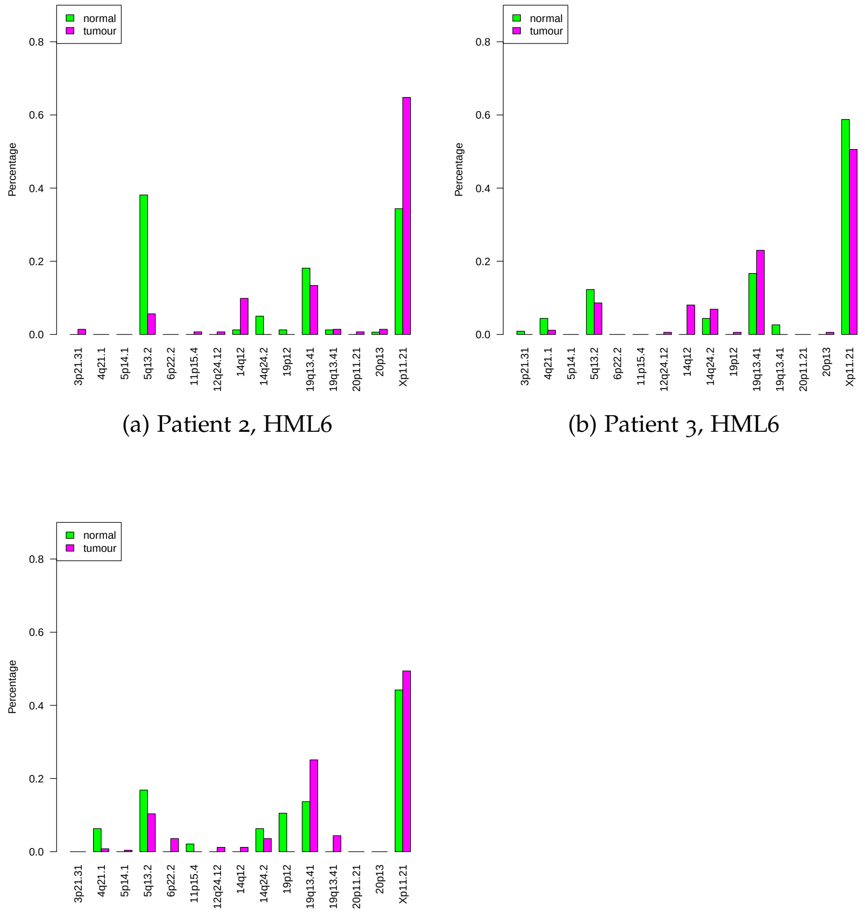


(a) Patient 2, HML2

(b) Patient 3, HML2

(c) Patient 4, HML2

Figure A4: Distribution of expressed HML2 loci in patients 2 to 4. The x-axis contains all HML2 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.

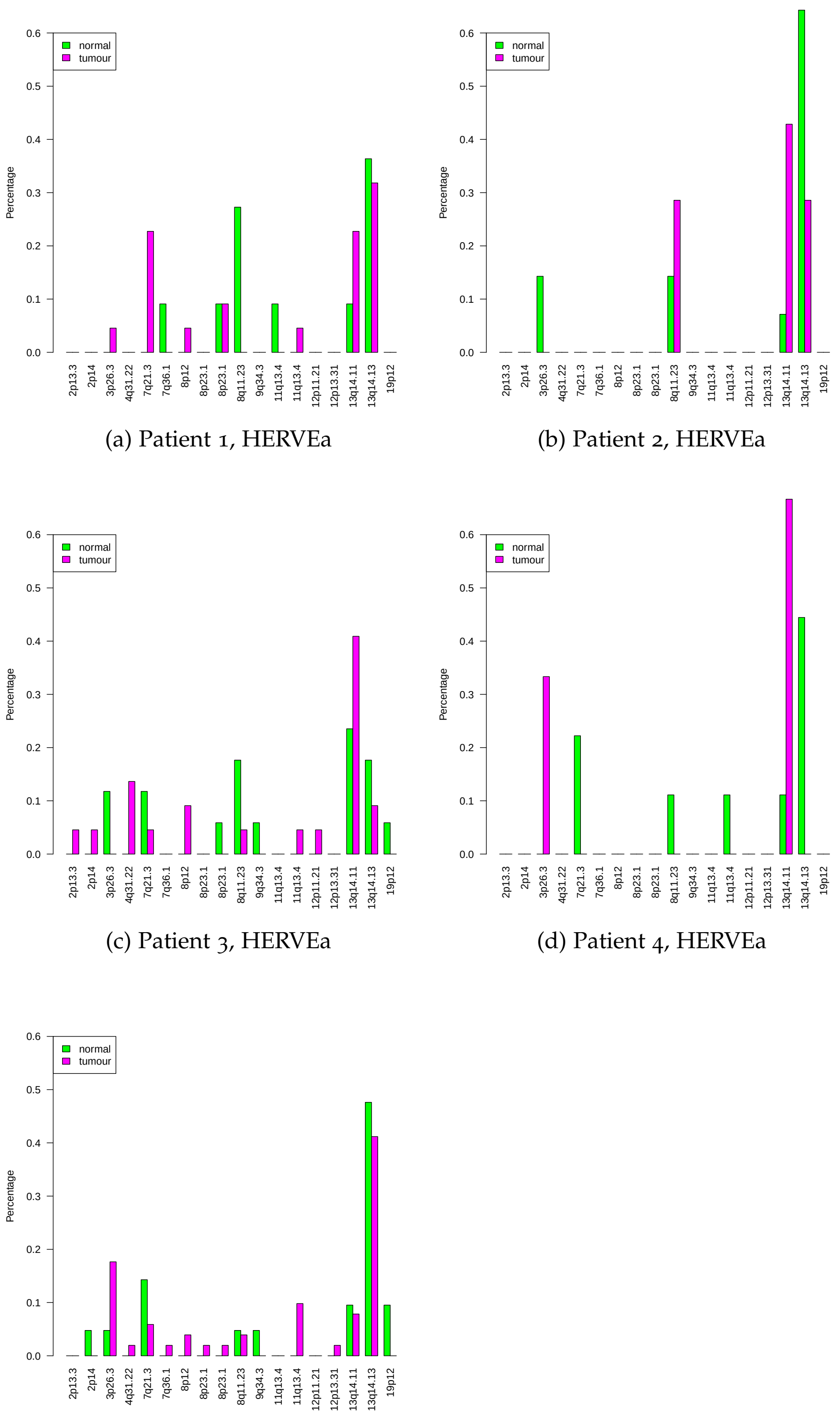


(a) Patient 2, HML6

(b) Patient 3, HML6

(c) Patient 5, HML6

Figure A5: Distribution of expressed HML6 loci in patients 2, 3 and 5. The x-axis contains all HML6 loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.



(a) Patient 1, HERVEa

(b) Patient 2, HERVEa

(c) Patient 3, HERVEa

(d) Patient 4, HERVEa

(e) Patient 5, HERVEa

Figure A6: Distribution of expressed HERVEa loci across chromosome bands.
The x-axis contains all HERVEa loci that are expressed in at least one sample, the y-axis shows which fraction of all reads mapped to members of the family arise from a specific locus.

A.2 SUPPLEMENTARY TABLES

Patient	Barcode	# Reads		Discarded Reads			Trimmed
		before	after	Ns	qual < 20	length < 50 nt	NTs
1	ACAC	4621	4051	518	3	49	432828
	AGAG	7076	6319	689	2	66	677515
	ATAT	9442	8208	1134	3	97	875091
	AGCT	10819	8580	2129	1	109	914399
	ATGA	16847	14895	1938	1	13	1592113
	TCAG	2129	1842	286	0	1	196846
2	CGCG	5712	4752	932	1	27	498922
	CTCT	12841	10893	1822	0	126	1154664
	CACA	10809	8881	1876	2	50	939822
	CTAG	4611	4047	560	0	4	427076
3	CATG	9014	7803	1124	0	87	827431
	TATA	6860	5608	1084	1	167	605529
	TCTC	6871	5834	990	3	44	617747
	TGTG	5496	4639	798	0	59	491979

(a) Run from 2012

Patient	Barcode	# Reads		Discarded Reads			Trimmed
		before	after	Ns	qual < 20	length < 50 nt	NTs
1	ACAC	7007	6377	621	0	9	652424
	AGAG	5723	4956	759	0	8	503269
	ATAT	5985	5234	739	0	12	529935
	AGCT	5352	4719	628	0	5	476723
2	ATGA	7426	6891	529	0	6	699441
	CGCG	1712	1509	202	0	1	155333
	CTCT	7328	6570	750	0	8	669802
3	CTAG	8667	7897	762	0	8	807890
	CACA	4156	3715	439	0	2	380863
4	TATA	1569	1418	149	1	1	144655
	CATG	1179	1007	169	0	3	102724
5	TCTC	18487	15561	2902	1	23	1592388
	TCAG	8676	8038	623	1	14	828165
	TGTG	4147	3637	502	0	8	373612

(b) Run from 2013

Table A1: Number of discarded reads and trimmed nucleotides per analysed sample.

Family	Orientation	Sequence	Patient 1		Patient 2		Patient 3		Patient 4		Patient 5	
			healthy	tumour	healthy	tumour	healthy	tumour	healthy	tumour	healthy	tumour
HML2	forward	CCCCCAGAAAGTCAGTATGGA	ACAC	AGAG	ATAT	CGCG	CTCT	CACA	TATA	TGTG	TCTC	ATGC
	reverse	TTTCCCAGGCTCTAAGGCAG	ACAC	AGAG	ATAT	CGCG	CTCT	CACA	TATA	TGTG	TCTC	ATGC
HERVEa	forward	TGTGGGTATAAGGTGTCCAAG	GAGA	AGAG	ATAT	CGCG	GTGT	CACA	GCGC	TGTG	TCTC	ATGC
	reverse	TGCTGCCAAGGCCCTCAAACA	ACAC	AGAG	ATAT	CGCG	GAGA	CACA	TATA	GCGC	TCTC	ATGC
HML6	forward	AAGATCTTTGGGTTTTTGAA	ACAC	AGAG	ATAT	CGCG	CTCT	GAGA	TATA	TGTG	GTGT	ATGC
	reverse	CATGTTCTTCAACATTTTAA	ACAC	AGAG	ATAT	CGCG	CTCT	GTGT	GCGC	TGTG	TCTC	ATGC
HERVW	forward	GCACAGAAATAAACACCACTTCC	ACAC	AGAG	ATAT	CGCG	GAGA	CACA	TATA	GCGC	TCTC	ATGC
	reverse	GCTGCTAGGGAGTTAAGTTG	ACAC	AGAG	ATAT	CGCG	CTCT	GTGT	GATC	TGTG	AATT	ATGC
ERV9	forward	AGACTTTGCTCTTTTCACATGC	ACAC	AGAG	ATAT	CGCG	CTCT	CACA	TATA	TACG	TCTC	ATGC
	reverse	TGCTTCCTCTGGTATTTGAGA	ACAC	AGAG	ATAT	CGCG	CTCT	CACA	TATA	TGTG	TCTC	ATGC

Table A2: Urothelium study primer and barcode design.
Shown are the primer and barcode sequences with their corresponding samples and HERV families.

Family	Orientation	Sequence	Patient 1			Patient 11			Patient 21		Patient 33		
			lymph	normal	tumour	lymph	normal	tumour	normal	tumour	lymph	normal	tumour
HML2	forward 1	CCCCCAGAAAGTCAGTATGGA	ATAT	AGAG	ACAC	CACA	CTCT	CGCG	CAAC	ATGC	TCTC	TGTG	TATA
	reverse 1	TTTCCCAGGCTCTAAGGCAG	ATAT	AGAG	ACAC	CACA	CTCT	CGCG	CAAC	ATGC	TCTC	TGTG	TATA
	forward 2	TCTCCAGAGGTTCAGTATGGA	ATAT	AGAG	ACAT	CACA	GTGT	CGCG	GAAG	ATTA	TAAT	TGTG	TATA
	reverse 2	TTCCCAGGCCCTGAGGCAA	ATAT	AGAG	ACAC	CACA	CTCT	CGCG	CAAC	ATGC	TCTC	TGTG	TATA
HML6	forward	AAGATCTTTGGGTTTTTGAA	ATAT	AGAG	ACAC	GAGA	CTCT	CGCG	CAAC	ATGC	GTGT	TGTG	TATA
	reverse	CATGTTCTTCAACATTTTAA	ATAT	AGAG	ACAC	GTGT	CTCT	CGCG	CAAC	ATGC	TCTC	TGTG	GCGC

Table A3: Mammary study primer and barcode design.
Shown are the primer and barcode sequences with their corresponding samples and HERV families.

Barcode	# Reads	Read length (nt)					Read quality (Phred score)				
		mean	median	min	max	σ	mean	median	min	max	σ
AATT	9.0	487.11	501.00	428.00	509.00	28.11	32.47	32.82	28.05	36.11	2.59
ACAC	1545.0	530.11	502.00	387.00	641.00	67.68	32.77	33.09	22.52	37.84	2.44
AGAG	2352.0	529.92	502.00	385.00	640.00	68.03	32.99	33.39	22.43	37.97	2.59
ATAT	1763.0	530.84	501.00	390.00	642.00	68.08	32.83	33.16	22.35	38.38	2.48
ATGC	2292.0	499.42	491.00	388.00	640.00	53.46	32.16	32.39	21.63	37.81	2. 49
CACA	1005.0	540.33	506.00	399.00	636.00	69.95	32.72	33.02	22.70	37.61	2.48
CGCG	1205.0	512.08	494.00	383.00	649.00	63.54	31.32	31.48	21.14	37.19	2.50
CTCT	845.0	492.26	489.00	381.00	637.00	43.66	32.26	32.60	23.25	37.11	2.28
GAGA	295.0	477.99	485.00	385.00	515.00	29.03	31.83	31.67	25.06	38.22	2.51
GATC	245.0	491.71	499.00	381.00	523.00	23.51	32.97	33.37	26.28	37.32	2.34
GCGC	92.0	479.70	481.00	406.00	517.00	22.38	32.50	33.32	24.09	36.63	2.77
GTGT	492.0	482.27	485.00	389.00	624.00	25.82	32.21	32.38	23.50	36.89	2.24
TATA	603.0	529.57	501.00	405.00	644.00	68.46	32.43	32.66	23.24	38.37	2.67
TACG	74.0	481.34	488.50	432.00	503.00	17.39	32.13	32.45	27.33	36.14	1.90
TCTC	699.0	497.41	490.00	396.00	634.00	50.15	32.07	32.27	24.33	37.62	2.41
TGTG	628.0	519.06	494.00	388.00	676.00	66.11	33.12	33.41	25.58	38.43	2.23
no Tags	843.0	510.00	494.00	379.00	644.00	60.26	32.29	32.58	20.86	37.76	2.47

Table A4: Initial statistic of urothelium sequencing run.
For every barcode and the *no tags* set the number of obtained reads are shown. Additionally, the main statistics regarding read length and quality are listed.

Family	H1-hESC		GM12878		HUVEC		K562		B-cells CD20+	SK-N-SH	MCF-7		HeLa-S3		A549	HepG2		IMR90	Mono-cytes CD14+
	†	*	†	*	†	*	†	*	*	*	†	*	†	*	*	†	*	*	*
BaEV-int	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ERVL	25	29	14	44	16	28	37	45	21	22	9	10	18	26	20	13	18	12	20
HERV15	2	1	1	1	1	2	0	0	0	1	0	0	0	1	0	1	1	0	0
HERV16	9	11	6	18	4	6	9	11	4	4	0	2	3	5	3	5	6	1	8
HERV17	5	6	2	4	2	1	5	3	0	2	1	2	1	2	4	1	1	1	1
HERV3	3	2	1	3	3	5	2	3	1	2	0	1	1	1	0	5	3	0	2
HERV30	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
HERV30I	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1
HERV9	4	6	1	3	3	5	7	4	1	1	0	1	2	2	4	10	11	0	3
HERVE	4	5	4	6	1	0	5	3	1	0	0	1	0	0	2	4	5	1	2
HERVFH19	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
HERVFH21	0	0	1	0	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0
HERVH	683	661	5	11	8	3	31	14	7	7	5	8	290	15	14	5	5	4	3
HERVH48	1	1	1	0	0	0	1	2	0	0	1	1	1	1	1	1	1	0	2
HERVI	8	6	1	3	2	2	0	0	0	0	0	0	1	0	2	1	2	0	0
HERVK	28	7	0	3	0	1	2	1	3	1	1	1	0	2	1	1	3	1	2
HERVK11	2	4	2	4	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
HERVK13	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
HERVK14	0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	0	0	0	0
HERVK14C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0

Continued on next page

Table A5 – continued from previous page

Family	H1-hESC		GM12878		HUVEC		K562		B-cells CD20+	SK-N-SH	MCF-7		HeLa-S3		A549	HepG2		IMR90	Mono-cytes CD14+
	†	*	†	*	†	*	†	*	*	*	†	*	†	*	*	†	*	*	*
HERVK22	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	1	3	0	0
HERVK22I	1	1	1	1	0	0	0	0	1	0	0	0	1	1	0	1	1	0	1
HERVK3	4	3	2	0	2	2	1	2	2	3	1	2	1	0	0	8	9	2	4
HERVK9	3	4	2	5	4	4	5	3	3	7	2	2	0	0	2	3	5	1	3
HERVKC4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0
HERVL	8	20	5	19	11	14	11	19	5	7	6	7	5	14	7	4	14	6	7
HERVL18	0	0	2	2	3	4	1	1	1	2	1	1	0	2	2	5	6	1	1
HERVL32	0	0	1	2	1	2	1	2	2	0	1	2	0	2	1	1	2	1	2
HERVL40	1	1	0	1	1	1	2	3	3	0	1	0	1	0	0	0	0	1	1
HERVL47	1	1	1	1	1	2	1	2	1	1	0	0	0	0	0	1	0	1	1
HERVL66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
HERVL74	1	1	3	8	2	3	4	5	2	3	1	1	0	0	2	1	1	3	4
HERVP71AI	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
HERVS71	9	8	1	2	0	1	4	2	2	2	1	1	1	1	1	0	1	0	1
LTR10A	0	0	1	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
LTR10B	4	4	0	0	0	1	1	2	1	1	1	1	0	0	0	1	0	0	0
LTR10C	3	3	1	1	0	1	4	2	4	1	1	2	3	1	2	1	1	1	1
LTR10D	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1
LTR10E	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0

Continued on next page

Table A5 – continued from previous page

Family	H1-hESC		GM12878		HUVEC		K562		B-cells CD20+	SK-N-SH	MCF-7		HeLa-S3		A549	HepG2		IMR90	Mono-cytes CD14+
	†	*	†	*	†	*	†	*	*	*	†	*	†	*	*	†	*	*	*
LTR11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1
LTR12	18	33	8	15	6	20	40	24	27	14	7	4	7	7	8	29	28	5	20
LTR13	3	4	1	0	1	7	1	0	6	2	1	1	2	0	2	2	1	1	8
LTR14A	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	2	2	0	0
LTR14B	0	0	0	1	0	2	1	1	1	0	0	0	0	0	0	1	1	0	3
LTR15	0	0	0	0	2	1	5	0	1	1	0	0	0	0	1	4	4	0	0
LTR16A	6	5	3	7	6	6	4	4	6	8	0	0	3	1	1	5	7	5	5
LTR16A1	2	0	2	1	6	5	4	5	4	1	0	0	2	3	1	4	2	1	7
LTR16B	4	2	2	6	5	7	4	3	1	3	1	1	2	1	1	6	6	1	4
LTR16C	11	7	7	23	4	6	7	6	17	7	0	0	2	4	2	7	9	8	17
LTR16D	1	1	0	1	1	1	1	1	0	1	0	0	1	3	1	1	1	2	2
LTR17	1	2	0	2	0	0	1	0	1	2	1	0	0	1	0	0	0	0	0
LTR18A	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LTR18B	0	0	0	1	2	2	1	0	0	1	1	1	1	0	0	2	2	0	1
LTR19-int	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LTR19A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
LTR19B	0	0	0	4	1	1	0	6	1	8	1	1	8	2	0	0	1	0	1
LTR19C	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
LTR2	5	4	4	3	2	4	2	2	1	1	1	0	0	0	0	5	5	1	3

Continued on next page

Table A5 – continued from previous page

Family	H1-hESC		GM12878		HUVEC		K562		B-cells CD20+	SK-N-SH	MCF-7		HeLa-S3		A549	HepG2		IMR90	Monocytes CD14+
	†	*	†	*	†	*	†	*	*	*	†	*	†	*	*	†	*	*	*
LTR21A	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LTR22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
LTR22A	1	1	0	3	1	2	0	6	3	6	0	0	0	0	0	0	0	0	0
LTR26	0	1	1	3	3	2	3	4	2	1	2	0	3	2	1	0	2	2	1
LTR26B	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	0	2
LTR26E	1	1	2	1	0	0	1	0	2	0	0	0	0	1	0	1	0	0	2
LTR2B	0	0	7	13	1	0	1	0	3	0	0	0	0	0	0	9	9	0	1
LTR2C	2	2	2	1	0	0	2	2	0	0	0	0	0	0	0	1	1	0	0
LTR30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
LTR32	1	1	1	2	2	3	1	1	2	1	0	0	0	0	1	2	3	1	1
LTR33	4	3	6	15	10	8	11	10	9	8	3	1	6	5	8	13	15	7	20
LTR33A	6	3	3	4	4	5	5	4	6	5	3	3	5	2	1	2	5	3	7
LTR3B	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
LTR4	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	1	1	0	0
LTR40A	0	0	2	2	1	1	0	1	0	0	0	0	2	2	0	3	4	0	0
LTR40B	1	1	3	2	0	3	1	1	1	4	1	0	1	2	2	0	0	3	1
LTR40C	0	0	1	2	1	1	2	2	1	0	2	0	1	2	4	4	6	0	2
LTR41	2	8	0	2	0	0	3	5	2	2	2	2	2	2	3	7	5	2	2
LTR42	1	0	2	1	0	0	0	0	2	0	0	0	0	1	0	1	2	1	1

Continued on next page

Table A5 – continued from previous page

Family	H1-hESC		GM12878		HUVEC		K562		B-cells CD20+	SK-N-SH	MCF-7		HeLa-S3		A549	HepG2		IMR90	Mono-cytes CD14+
	†	*	†	*	†	*	†	*	*	*	†	*	†	*	*	†	*	*	*
LTR47A	0	0	1	2	1	1	1	1	1	0	0	0	0	0	0	1	1	0	1
LTR47B	0	0	1	1	0	0	0	0	2	0	0	0	1	1	2	0	0	1	1
LTR50	3	2	0	0	4	4	2	2	1	3	0	0	2	2	1	2	2	1	0
LTR53	0	0	1	1	1	0	0	1	1	1	0	0	0	0	0	1	1	0	1
LTR57	0	0	2	3	3	1	3	2	1	1	0	0	1	0	0	2	1	1	2
LTR5B	6	8	2	1	2	3	5	6	3	8	5	4	5	2	6	8	9	2	3
LTR5_Hs	1	3	1	3	2	3	0	0	0	0	0	0	0	0	0	0	0	0	1
LTR61	1	0	1	1	1	0	1	1	0	0	1	1	1	2	0	1	0	0	2
LTR62	1	1	1	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0
LTR64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
LTR66	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
LTR67	3	5	2	8	1	2	2	2	7	5	0	0	0	1	1	3	1	1	6
LTR6A	8	1	6	4	4	1	10	3	1	2	4	1	6	1	1	11	3	0	1
LTR6B	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LTR7	44	25	1	4	2	2	2	2	3	0	0	1	3	2	1	5	2	1	2
LTR72	2	1	1	1	1	0	1	2	0	0	1	1	1	0	1	1	1	0	0
LTR75	2	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
MER11A	1	0	0	0	0	2	2	2	0	1	0	0	0	0	2	2	3	0	1
MER11B	0	0	1	3	0	2	1	0	0	0	1	1	0	0	1	0	0	1	3

Continued on next page

Table A5 – continued from previous page

Family	H1-hESC		GM12878		HUVEC		K562		B-cells CD20+	SK-N-SH	MCF-7		HeLa-S3		A549	HepG2		IMR90	Mono-cytes CD14+
	†	*	†	*	†	*	†	*	*	*	†	*	†	*	*	†	*	*	*
MER11C	3	2	1	2	3	3	2	2	1	0	0	1	0	0	1	3	3	1	4
MER11D	0	0	0	0	0	2	0	0	0	1	1	2	2	2	1	0	0	1	0
MER48	0	0	0	0	0	0	3	3	0	1	0	0	0	0	0	0	0	0	1
MER54A	1	1	2	5	6	4	6	6	2	7	1	0	0	0	7	2	1	2	1
MER54B	0	0	1	2	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1
MER70A	0	1	0	0	0	1	1	2	0	1	0	0	1	0	0	0	0	0	1
MER70B	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
MER73	0	0	1	1	0	2	0	0	1	0	0	0	0	0	0	0	0	0	3
MER74A	4	4	1	2	1	1	2	3	1	0	0	0	2	1	1	3	2	1	3
MER74B	2	0	0	1	1	2	0	0	2	0	0	1	0	0	0	1	2	0	2
MER74C	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
MER9	1	2	0	3	4	3	4	5	2	0	1	0	0	2	4	4	4	1	2
MLT2A1	2	3	0	5	4	3	1	2	4	0	0	0	2	1	0	0	0	1	3
MLT2A2	6	9	2	7	3	6	2	6	3	3	0	0	2	3	4	2	3	1	2
MLT2B1	0	0	1	8	2	1	6	7	1	3	0	0	2	1	3	3	5	0	1
MLT2B2	1	1	0	2	1	1	1	0	0	2	0	0	4	3	2	2	1	0	1
MLT2B3	4	7	3	7	2	3	7	5	3	4	3	1	1	1	3	4	2	4	4
MLT2B4	5	2	2	3	2	5	2	3	2	4	2	1	2	1	3	3	3	1	1
MLT2B5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Continued on next page

Table A5 – continued from previous page

Family	H1-hESC		GM12878		HUVEC		K562		B-cells CD20+	SK-N-SH	MCF-7		HeLa-S3		A549	HepG2		IMR90	Mono-cytes CD14+
	†	*	†	*	†	*	†	*	*	*	†	*	†	*	*	†	*	*	*
MLT2C1	2	3	1	3	3	2	2	1	3	0	2	0	5	3	2	2	5	2	1
MLT2C2	2	1	0	1	1	1	6	6	0	6	0	0	2	0	1	3	2	1	1
MLT2D	2	0	3	5	5	1	6	5	1	2	2	1	2	2	0	1	0	1	1
MLT2E	1	3	0	0	0	2	2	1	1	1	0	0	0	0	1	1	0	2	1
MLT2F	1	2	3	4	1	3	3	3	3	2	1	1	1	1	2	4	4	0	4
pTR5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Table A5: Over-expressed HERV loci per cell line.

The table shows for every HERVd family, how many loci are significantly over-expressed in the pairwise comparisons with all other cell types.

$*$ = CSHL, † = Caltech