Analysis of expression profile and gene variation via development of methods for Next Generation Sequencing data

Dissertation
for the award of the degree

Doctor rerum naturalium

of the Georg-August-Universität Göttingen

within the doctoral program *Environmental Informatics (PEI)*
of the Georg-August-University School of Science (GAUSS)

submitted by
Alexander Wolff
from
Dresden (Sachsen), Germany

Göttingen, 2018

Thesis Committee:
Prof. Dr. Tim Beißbarth
Prof. Dr. Stephan Waack

Members of the Examination Board:
1st Referee: Prof. Dr. Tim Beißbarth
2nd Referee: Prof. Dr. Stephan Waack

Further members of the Examination Board:
Prof. Dr. Edgar Wingender
Prof. Dr. Burkhard Morgenstern
Prof. Dr. Wolfgang May
Prof. Dr. Kurth

Date of oral examination: 19nd of November 2018

Acknowledgements

I want to thank the people that have supported me throughout the last years:

Thank you Tim for having always an open door for spontaneous questions, feedback and discussions.
Many thanks go to my thesis committee member, Prof. Dr. Waak for his helpful and constructive feedback and their commitment and support.

Also, I want to acknowledge the close work of all collaborators for their project contributions and their scientific work, which I enjoyed and liked to help in. In details, I would like to express my gratitude to the collaborators, who provided and created the data sets which analysis and results are part of this thesis and without them this opportunity would not have been possible.

In addition, I would like to thank all the members of the Department of Medical Statistics. Many thanks go to my colleagues for fruitful discussions, feedback and a nice working environment: Julia, Florian, Andreas, Michaela, Astrid, Frank, Klaus, Manuel, Maren and Zaynab.
Last but not least I would like to thank my wife and family for their patience and permanent support.

Contents

List of Figures

List of Tables

Abbreviations

BAFF	-	B-cell activating factor
BL	-	Burkitt lymphoma
BL2	-	Burkitt Lymphoma
bp	-	base pairs
BQSR	-	base quality score recalibration
BWT	-	Burrows-Wheeler Transformation
cDNA	-	complementary desoxyribonucleic acid
CNVs	-	copy number variations
CRC	-	Colorectal cancer
CTC	-	circulating tumor cells
DEA	-	differential expression analysis
DEGs	-	differentially expressed genes
DNA	-	desoxyribonucleic acid
DNA-Seq	-	DNA sequencing
ESTs	-	expressed sequence tags
indels	-	small DNA insertions or deletions
FDR	-	False Discovery Rate
GLM	-	generalized linear models
GO-Terms	-	Gene Ontology terms
GSEA	-	gene set enrichment analysis
GTF	-	Gene Transfer Format
NGS	-	Next-Generation Sequencing
MMP	-	matrix metallopeptidase
mRNA	-	messenger ribonucleic acid
RC	-	rectal cancer

RNA	-	ribonucleic acid
RNA-Seq	-	RNA sequencing
RPKM	-	reads per kilobase of exon model per million reads
RSEM	-	RNA-Seq by Expectation Maximization
RT	-	reverse-transcribed
RT-PCR	-	reverse transcription polymerase chain reaction
SAGE	-	serial analysis of gene expression
SNP	-	Single Nucleotide Polymorphism
SNV	-	Single Nucleotide Variation
TPM	-	transcripts per million
PCR	-	Polymerase Chain Reaction
WES	-	Whole Exome Sequencing
WESA	-	Whole Exome Sequencing Analysis
WGS	-	Whole Genome Sequencing

1 Summary

1.1 Summary in English

Since the last ten to twenty years, the cost of sequencing the human genome decreased continuously. Therefore the interest in *RNA sequencing* (RNA-Seq) rose as it can be used to discover the molecular mechanisms behind gene expression profiles of cells in different healthy or disease states (Wang et al., 2009). The intention of this dissertation is two-fold, first identify the best performing bioinformatical methods for RNA-Seq analysis at hand and based on this knowledge generate a standardised workflow, which then could be used within the MetastaSys consortium. Second, answering the question: Is it possible to detect somatic mutations in cancer based on RNA-Seq data reliably? This was of particular interest as the RNA-Seq data was already created for differential gene expression analysis. Getting further information on mutation status without the need to recreate the data for Exome-Seq would, on the one hand, save the expensive costs for Exome-Seq and would, on the other hand, save precious biological material of cancer metastases patients, which are precious to the physicians.

For the RNA-Seq workflow identification data based on the microarray and Illumina RNA-Seq platforms were created. Therefore two data sets were created: human patient data from rectal cancer metastases in the liver and human cell lines from Burkitt's Lymphoma, which was stimulated with the B-Cell activating factor BAFF. The advantages of RNA-Seq over Microarray became clear during the comparative analysis in the first publication (see 3.1). The primary focus was the performance evaluation of bioinformatical methods based on the given data sets. The workflow performance was evaluated during the *alignment*, *transcript quantification*, *differential gene expression analysis*, and *functional profiling* steps. Results showed, that despite the workflow with TopHat2 and Cufflinks, all workflows achieved nearly equally good results with a slight preference for STAR and RSEM, as STAR achieved the overall highest mapping rate and RSEM incorporated multi-mapped reads for quantification and was also capable of quantifying transcript isoforms next to genes. Afterwards, the best performing workflow pipeline was applied to mice in another study (see 3.3). The mice developed metastases in the liver from colorectal cancer. The bioinformatical approach streamlined via the workflow helped a lot in interpreting the biology behind the expression of metastasis enhancing genes. It was possible to show links of metastasis-related genes and their stimulation via the liver environment. These genes were associated with tissue

remodelling, cell proliferation, adhesion, wnt activity, transcription/regulation, and inhibition of apoptosis.
The question if a reliable identification of somatic mutation is possible in RNA-Seq is tackled by implementing Wileup, a program is written in Perl. Wileup's performance was evaluated against the state-of-the-art somatic variant caller Mutect2 from the GATK tool suite for matched RNA-Seq and Exome-Seq samples of 14 patients with either brain (seven patients) or liver (seven patients) metastases (see 3.2). Results showed that Wileup was capable of finding all somatic mutations in RNA-Seq identified by Mutect2 in Exome-Seq. In contrast, Mutect2 and Wileup identified unique germline mutation only found in either of the methods. These could be explained due to a lack of expression on the RNA-Seq data or due to too high duplication level in the Exome-Seq data. Furthermore, the somatic mutations could be independently validated by pathological annotation data. For the uniquely found germline mutations of either method, it was possible to verify all of them, as they were re-identified in the Exome-sequenced blood samples of the corresponding patients.
In conclusion, the presented studies in this thesis contribute towards establishing pipeline standards in transcriptomics, with the focus on *differential expression analysis* (DEA), and exploring the capabilities of mutation calling in RNA-Seq.

1.2 Zusammenfassung in Deutsch

Seit ca. 10 bis 20 Jahren reduzieren sich die Kosten zur Sequenzierung eines ganzen menschlichen Genoms stetig. Gleichzeitig stieg, aufgrund der Seqquenzier-Möglichkeiten das Interesse, die molekularen Mechanismen hinter den Genexpressionprofilen besser zu verstehen rapide. Die Intention dieser Dissertation war zweigeteilt,

1: die Identifizierung der am besten geeigneten bioinformatischen Methoden für die RNA-Seq Analyse. Dieses Wissen wurde dann zur Etablierung eines standardisierten Arbeitsablaufes (im Folgenden Workflow genannt) genutzt, welches im Rahmen des MetastaSys Verbundprojektes angewandt werden konnte.

2: "Ist es zuverlässig möglich somatische Mutationen auch in RNA-Seq Daten zu detektieren?". Diese Frage war von speziellem Interesse, da RNA-Seq Daten von Patienten primär zur differenziellen Genexpressionsanalyse erzeugt werden sollten. Könnte man jetzt weiterhin den Mutatationstatus der Patienten basierend auf diesen Daten ermitteln ohne die Daten neu mittels Exome-Seq sequenzieren zu müssen? Dies würde auf der einen Seite die weitaus höheren Kosten massiv reduzieren und auf der anderen Seite auch wertvolles biologisches Biopsie-Material zurückhalten, welches sonst für mögliche zukünftige Analysen aufgebraucht wäre.
Zur Identifizierung des optimalen RNA-Seq Workflows wurden Daten basierend auf Microarrays und RNA-Seq erhoben. Diese setzten sich aus zwei Datensätzen zusammen: Zum Einen metastasiertem menschlichen Lebergewebe von Patienten mit Rektumkarzinom, welches in der Leber Metastasen gebildet hatte und zum Anderen aus humanen Zelllinien von einem Burkitt's Lymphom, das mit dem B-Zellen aktivierenden Faktor BAFF stimuliert wurde.

Die Vorzüge der Analyse von RNA-Seq Daten gegenüber Microarray basierenden wurden schon früh sichtbar. Dies war der Anlass den Fokus weg von einer vergleichenden Publikation zwischen Micrarray und RNA-Seq Daten hin zu der Evaluation bioinformatischer Methoden zu verschieben (siehe 3.1). Die Methoden wurden an folgenden Arbeitsschritten des Workflows evaluiert:

- während der Zuordnung der Rohdaten in Form von Millionen von kurzen Sequenzstücken (reads) zu einer Referenz (z. Bsp. dem menschlichen Genom),
- während dem Zählen der reads, welche Gene überlappen und so die Expression des Gens abbilden,
- während der statistischen Analyse der Expressionsdifferenz zwischen Gruppen dieser Gene und
- anschliessender Analyse zugrunde liegender funktionaler genetischer Gruppen.

Resultierend kann gesagt werden, dass mit Ausnahme des Workflows bestehend aus TopHat2 und Cufflinks, allen in der Publikation beschriebenen Workflows gelungen ist, ähnlich gute Resultate zu erzielen. Dabei konnte sich der Workflow mit STAR und RSEM leicht gegenüber den anderen Workflows hervorheben. Dies erklärt sich damit, dass STAR die höchste Gesamt-Zuordnungsrate der Rohdaten erreichte und RSEM den höchsten Anteil bei der Genzuordnung verarbeiten konnte. Parallel dazu konnte RSEM die Genexpression auf die jeweilig zugehörigen Transkripte aufteilen. Nachdem STAR und RSEM als Workflow im Verbundprojekt festgelegt wurden, wurden diese unter anderem auf Maus-Daten aus dem Verbundprojekt erfolgreich angewandt und die Ergebnisse publiziert (siehe hierfür 3.3). Die Mäuse entwickelten nach einer Injektion kolorektaler Krebszelllen über die Pfortader kolorektale Metastasen in der Leber. Der etablierte bioinformatische Workflow konnte nun massiv die Interpretation der Biologie hinter der Expression von Metastasen verstärkenden Genen unterstützen und voranbringen. So konnte die Verbindung zwischen Metastasen unterstützenden Genen und ihrer Stimulation durch die Leberumgebung gezeigt werden. Eine Auswahl dieser Gene wurden mit Gewebe Umbau, Zell-Proliferation, Adhesion, Wnt Aktivität, Transkription/Regulation, sowie der Inhibition der Apoptose, dem kontrollierten Zelltod assoziiert.
Um die Frage anzugehen, ob es möglich ist somatische Mutationen zuverlässig in RNA-Seq zu detektieren, wurde Wileup, ein Programm in Perl, implementiert. Wileup's Ergebnisse wurden dann mit dem state-of-the-art Programm Mutect2 aus dem GATK Programmkatalog verglichen. Dieses wurde explizit für die Detektion von somatischen Mutationen in Krebsgewebe entworfen. Es benötigte allerdings Exome-Seq Daten, sowie zusätzlich zum sequenzierten Tumor-Gewebe eine Normal-Referenz als Abgleich. Damit von Wileup verarbeitete RNA-Seq Daten Mutect2 verarbeitete Exome-Seq Daten verglichen werden konnten, wurde ein experimentelles Design gewählt, das aus 14 Patienten bestand: Jeweils sieben Patienten mit kolorektalen Gehirnmetastasen und sieben kolorektalen Lebermetastasen. Von jedem Patienten wurden drei Sequenzierungen vorgenommen: zwei Exom-Sequenzierungen

vom Blut und dem Metastasengewebe des Patienten und eine RNA Sequenzierung des Metastasengewebes (siehe 3.2). Im Vergleich wurde Mutect2 einmal mit Normalgewebereferenz und einmal ohne benutzt, was einerseits die optimale Methode widerspiegelte und anderseits vergleichbarer mit Wileup war, welches auch ohne Normalreferenz angewandt wurde. Die detektieren somatischen Mutationen konnten in allen drei Methoden einheitlich entdeckt werden und mittels Pathologiebefunde größenteils (7/8 somatischen Mutationen) bestätigt werden. Lediglich in der Anzahl der identifierten germline Mutationen gab es Unterschiede. So wurden 36 gefundene germline Mutationen von Mutect2 im "tumor-only" Modus durch Wileup nicht identifiziert, da es an diesen Positionen an Genexpression in den RNA-Seq Daten mangelte. Dafür konnte aber mittels GATK's Haplotype caller in den Blut-Daten der Patienten unabhängig bestätigt werden, dass es sich um echte germline Mutationen und nicht um Artefakte handelte. Bei Wileup wurden fünf germline Mutationen identifiziert, die nicht im "tumor-only" Modus von Mutect2 entdeckt wurden, da in den Exome-Seq Daten an diesen Stellen zu hohe Duplikationsraten vorlagen und eine Detektion nicht möglich war. GATK's Haplotype caller konnte auch diese Mutationen im Exome-Seq des Blutes der Patienten nachweisen und bestätigen.
Abschließend lässt sich sagen, dass die hier präsentierten Publikationen zum immernoch stark aktuellem Thema der Pipeline Standardisierung im Feld der "Transciptomics", speziell der differenziellen Genexpressionsanalyse, positiv beitragen konnten. Weiterhin war es möglich die Fragen der Mutationsdetektion in RNA-Seq Daten erfolgreich zu klären. Außerdem war es möglich beide Teile meiner anfänglichen Fragestellung in dieser Arbeit erfolgreich an Echtdaten erproben zu dürfen und somit ihre Validität zu bestätigen.

2 Introduction

2.1 Motivation of this thesis

Within the last ten years of *Next-Generation Sequencing* (NGS) and specially RNA-Seq as become a lot more affordable (Wang et al., 2009). They are overtaking, step by step, the place of microarray analysis at the topic of unravelling mechanisms of gene expression. The main reasons for this are decreasing running costs, a higher dynamic range of expression and low abundance accuracy of RNA-Seq over microarray (Ozsolak and Milos, 2010). Added to the versatility of RNA-Seq a further factor for the increasing popularity is the possibility of detecting mutation not only on data derived from *DNA sequencing* (DNA-Seq), but using RNA-Seq data.

The precision of current methods for detecting mutations in RNA-Seq is not on par with state of the art DNA-Seq based methods (McKenna et al., 2010; Cibulskis et al., 2013; Xu, 2018), because of higher alignment error rates near splice junctions, RNA editing and failure of detecting mutations in gene regions of very low or no expression. Nevertheless, current methods can provide additional information, like a high expression of low-frequency variants which are hard to detect in genomic DNA. An additional benefit is the detection of possible mutation states next to standard differential expression analysis (see 2.3.1.4) and gene set enrichment analysis (see 2.3.1.5), for no additional financial and biological costs (Goya et al., 2010; Quinn et al., 2013; Tang et al., 2014). For example, these mutations can not only be associated to tumour types based on mutation patterns but also it is feasible without the need of high investments for acquiring new biological samples or redoing the experiment with DNA based mutation analysis.

This work aims to reveal the best suitable bioinformatical methods to use for standard analysis of RNA-Seq data and apply them to multiple RNA-Seq data sets. This comprises the comparison of microarray platforms to several RNA-Seq workflows, evaluation of their performance and a recommendation for the best performing workflow RNA-Seq data (see 3.1). Afterwards, the workflow is applied in another publication dealing with *Colorectal cancer* (CRC) in mice (see 3.3). This dissertation also includes methodological work of a software for the detection of mutations in RNA-Seq. It is called Wileup, and it can be applied either on a complete transcriptome or using it on a small panel of mutations with specific clinical implication on possible drug response (see 3.2). Further, this work evaluates

the performance of Wileup compared to state of the art analysis tools as a further addition to standard analysis workflows for RNA-Seq data.

2.2 High-throughput sequencing data

The basic information flow in a cell goes from *desoxyribonucleic acid* (DNA) to *ribonucleic acid* (RNA) (transcribed into *messenger ribonucleic acid* (mRNA)) and is then translated into proteins at the ribosome (Crick, 1970). These different cellular contents can also be described in the context of high-throughput sequencing data, as different subpopulations, using the term -omics (Greenbaum, 2001). Three of these -omic terms play an important role in this dissertation: the genome, the exome, and the transcriptome. The complete DNA, including genes, coding and non-coding nucleotide sequences, are described as the genome. The exome is a subset of the genome and only describes the regions in the genome, that are mainly transcribed into mRNA and therefore are protein-coding. According to Ng et al. (2009), protein coding regions are only representing 1% of the human genome. The exome defines the transcriptome, but only the transcribed parts at the time point of RNA measurement for the cellular condition. So it reflects the information of the expressed genes at that time point, which can vary through cell types and over time.
Of these three -omic data types, transcriptomic and exomic sequencing data were used in this work to first decide on the best fitting bioinformatical methods at hand and second to evaluate the performance of detecting mutations based on transcriptomic data using Wileup instead of the general approach in using exomic data for mutation calling.

2.2.1 Transcriptome profiling

The main challenge with transcriptome profiling is the identification of genes differentially expressed between two conditions. First, the tools at hand to identify a single molecule or small amounts of them were Northern blots (Alwine et al., 1977), *reverse transcription polymerase chain reaction* (RT-PCR) (Shaffer et al., 1990), *expressed sequence tags* (ESTs) (Adams et al., 1991; Nagaraj et al., 2006) and *serial analysis of gene expression* (SAGE) (Velculescu et al., 1995). Big advancements in analysing transcriptome data were made with the development of microarrays first reported in the Science journal in 1995 (Schena et al., 1995). With these chips, it became possible to analyse the expression of thousands of genes in parallel on one chip.
Agilent and Affymetrix are two of the main microarray platforms (and are analysed within the first publication, see 3.1). They differ slightly in their hybridisation techniques of the RNA and follow-up bioinformatical analysis steps, see Tan et al. (2003) for a thorough evaluation of commercial microarray platforms.

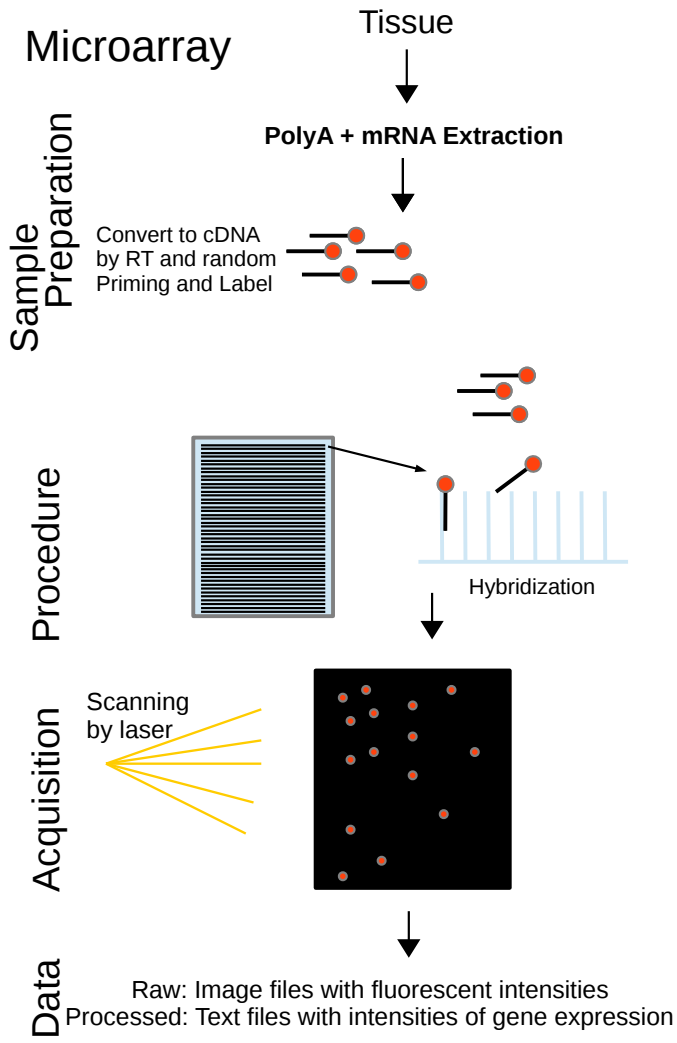

Figure 2.1. *The generalized processing work flow of microarrays, as exemplified by input tissue samples and output processed text files. Adapted from Figure 1 in Malone and Oliver (2011).*

The main idea behind microarrays is the use of short oligonucleotide probes made of genomic DNA; a general processing workflow is shown in Figure 2.1. These so-called probes are complementary to thousands of possible transcripts of interest. They are immobilised on a solid substrate. The transcripts are extracted from cells or tissues of interest and labelled with fluorescent dye(s), hybridised to the array, washed and scanned with a laser. Transcript and their complementary probe counterpart hybridise, and due to the dye(s) at the labelled transcript, light intensities are detected as a measure of gene expression. Even through microarrays have good stable analytical solutions and established quality control standards (Shi et al., 2006, 2010), a new standard for transcript analysis revolutionizedRNA analysis: RNA-Seq. It introduced several profiling opportunities that microarrays cannot provide. Because RNA-Seq is direct utilising fragments of transcriptomic sequences, investigations

of junctions between exons without prior knowledge of a gene model, RNA editing events, knowledge of polymorphisms, which can support emerging allele-specific expression, are possible. In contrast, microarray probes are designed by prior genomic sequence data, and light intensities are used as a surrogate of gene expression. Thus, microarrays will miss all these differences in gene expression without undergoing great efforts (Kapranov et al., 2007; Agarwal et al., 2010). Another big advantage of RNA-Seq is the possibility to analyse non-model organisms for quantification of individual transcript isoforms or alternative splicing events where the complete genome is not present, and therefore no complementary probes could be designed for microarrays due to missing annotation (Trapnell et al., 2010; Richard et al., 2010).
Over the years multiple companies started selling RNA-Seq platforms: Sanger, the first-generation (Metzker, 2005; Hutchison, 2007), sequencing, Roche/454, Illumina/Solexa, Life/APG (Zhou et al., 2010) for the next-generation sequencing platforms and Helicos BioSciences and nanopore (Branton et al., 2008) for the emerging third generation sequencing platforms. The following explanation of mRNA sequencing will focus on the Illumina's platform *sequence by synthesis* technique, as it is the main platform of all sequencing data presented in this dissertation. For a full review see Metzker (2010). The protocol for mRNA sequencing is described in Figure 2.2. First, the samples of interest are getting fragmented, for example by hydrolysis, and then *reverse-transcribed* (RT) to make double-stranded *complementary desoxyribonucleic acid* (cDNA) utilising random hexamer primers from Illumina. Next, the double-stranded fragments are ligated to adapters at the ends. Afterwards, they are amplified by *Polymerase Chain Reaction* (PCR) and injected into a flow cell. The flow cell is used as a solid phase and consists of a lot of oligonucleotides complementary to the adaptors ligated to the transcripts.

After the adaptors on the DNA fragments have been hybridised to the complementary oligonucleotides in the flow cell, the fragments are bridge amplified to generate significant clusters of clones to be better detected by laser, later on. Then wash and go cycles take place, where sequencing reagents are added, and in each cycle precisely one nucleotide can be added to the complementary oligonucleotide for all clusters simultaneously for millions of clusters on the flow cell. After hybridisation took place, all none added nucleotides are washed away. Finally, a laser detects hybridised nucleotides. These cycles proceed as long as the sequenced fragments should be. In this dissertation for instance, the regular nucleotide length of the fragments for RNA-Seq, so-called *reads*, were 50 bases long; therefore 50 cycles took place.

2.2.2 Exome profiling

Typical application, when profiling an exome are the detection of *Single Nucleotide Polymorphism* (SNP) genotyping and the detection of *small DNA insertions or deletions* (indels) but also *copy number variations* (CNVs), rearrangements and inversions of sequences can be of high interest. SNP are sequence alternatives, so-called alleles, at single *base pairs* (bp) positions in the genome. A single bp mutation has to be present in at least 1% of a population

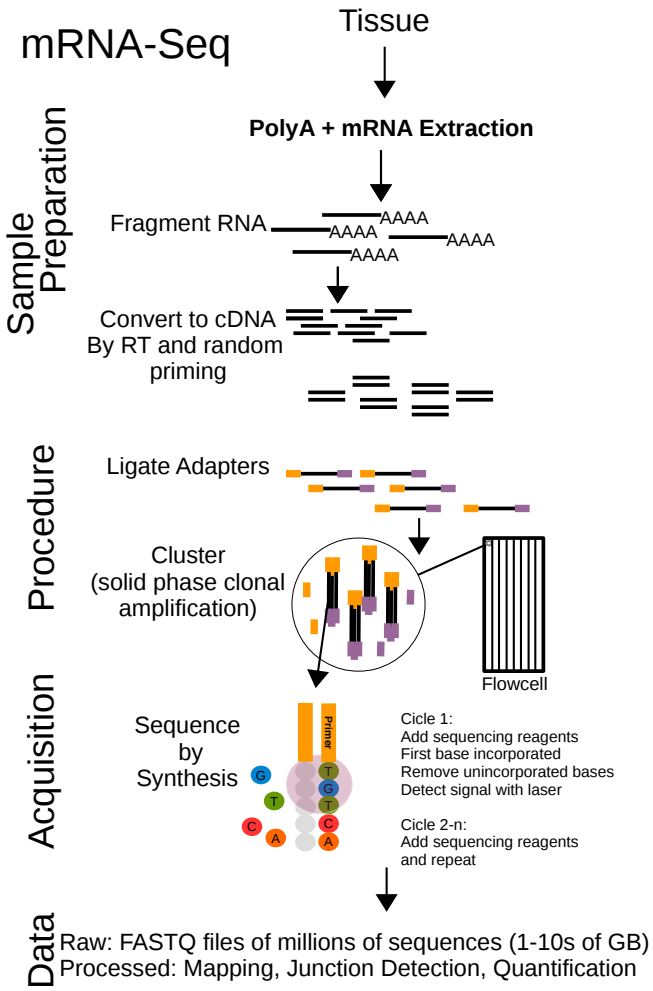

Figure 2.2. *The generalized processing workflow of RNA-Seq based on the Illumnia platform, as example inputting tissue samples and outputting processed text files. Adapted from Figure 2 in Malone and Oliver (2011).*

to be called a SNP (Brookes, 1999). Therefore the correct naming of mutations, where it is not known, if at least 1% of the population has it, is *Single Nucleotide Variation* (SNV), but are often intermixed. SNVs and SNPs are of special interest, as they are associated with heritable phenotypes, multifactorial diseases, cancer, drug responses to cancer (Gray et al., 2000) or most prominent in Mendelian diseases (Ng et al., 2010). The most common used method to identify SNPs is *Whole Genome Sequencing* (WGS) and *Whole Exome Sequencing* (WES) (Belkadi et al., 2015) but is also well described for microarrays (Kwok and Chen, 2003). RNA-Seq as a method to identify SNPs (Quinn et al., 2013; Cirulli et al., 2010) has become a valuable addition, which should not be overseen and is the focus in this dissertation at publication 3.2.

SNP genotyping with microarrays and DNA-Seq on DNA samples is similar to the described instruction from Figure 2.1 and 2.2 with the difference, that the starting material is DNA

instead of RNA and used chemicals have to be adapted for it. There are microarray-based methods for SNP genotyping from Affymetrix as well as from Illumina, which are successful and still used. Today SNP arrays are capable of detecting more than one million different human SNPs. The fraction of SNPs on the array that can be reliably called exceeds 99.5% according to Bumgarner (2013). Nevertheless, the analysis focus remains on exomic data based on Exome-Seq from Illumina machines.

2.3 Analysis of Sequencing data

With the new possibilities of data creation, there is a rising need for new methods and software to manage all the "-omics" data. This part will consist of an overview of current bioinformatical tools for dealing with RNA-Seq and Exome-Seq data in a workflow-like structure of subsections. It will only touch the microarray analysis for transcriptomic data shortly, as it is mentioned in the publication in section 3.1 and of minor importance nowadays for the analysis of transcriptomic data.
A crucial prerequisite for a successfully RNA-Seq study is a good experimental design involving library type, sequencing depth and numbers of replicates dependent on the biological system under study. Knowledgeable execution of the sequencing experiment itself ensures that data acquisition does not become undermined by unnecessary biases (Auer and Doerge, 2010). This statement holds true for Exome-Seq based experiments as well.

2.3.1 Transcriptomic Analysis

The possibilities of analysing RNA-Seq data are as versatile as there are applications of this technology (Williams et al., 2017). This section addresses the major analysis steps for *differential gene expression* followed by interpretations of the differentially expressed genes via *functional profiling*. The main steps for a typical transcriptomic analysis are quality control of the raw input data (see 2.3.1.1), read alignment against a reference genome (see 2.3.1.2), obtaining metrics for gene and transcript expression, so-called *counts* (see 2.3.1.3), and detection of differential gene expression (see 2.3.1.4) followed by functional enrichment of genes into groups (like over-representation of genes in pathways or functional gene-ontology terms (**?**)) (see 2.3.1.5). For microarrays, the steps are preprocessing of the raw intensity values from the chip, normalisation of raw intensities followed by differential gene expression analysis with optional functional enrichment of significant genes (using Limma (Ritchie et al., 2015)). There are also different analysis options for applications on RNA-Seq data involving alternative splicing, identification of fusion transcripts, and small RNA expression, which are not covered here. For a survey of possible analysis strategies on RNA-Seq data and accompanying tools see Conesa et al. (2016). An illustration of the most relevant tools for this dissertation are shown in Figure 2.3.

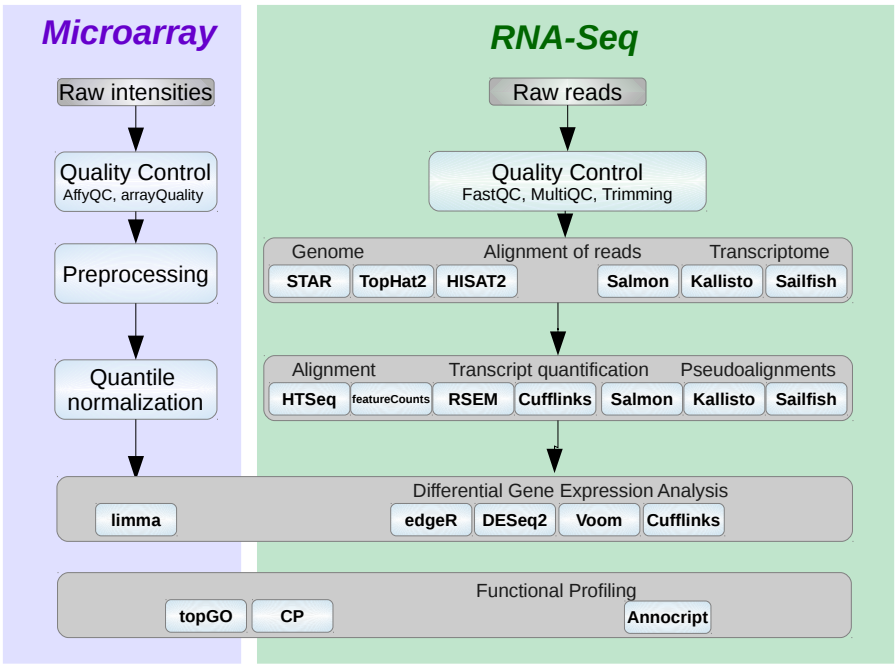

Figure 2.3. *The workflow for microarray and RNA-Seq analysis structured with the different processing steps with a selection of tools commonly used in the bioinformatical community surrounded by light blue bubbles.*

2.3.1.1 Quality Control of raw reads

Raw reads are stored in text files. They are called FASTQ files and are a widely accepted and standardised file format. Each entry of a read in a FASTQ file consists of four lines: sequence identifier, sequence, quality score identifier line and Phred-scaled base quality scores. Quality control of raw reads involves the analysis of sequence quality, GC content, the presence of adaptors, overrepresented k-mers, duplicated reads, and PCR artefacts or contaminations to evaluate the technical quality of the samples. The range of acceptable k-mer or GC content levels are dependent on experiment and organisms and should always be checked individually on each experimental setup. FastQC (Andrews, 2010) is a prominent tool to quality check Illumina reads, whereas MultiQC (Ewels et al., 2016) is perfect to combine sample based results into one big overview file from multiple tools across many samples for different steps during the analysis, not only when doing quality control of raw reads. Software tools like the FASTX-Toolkit (Hannon, 2009) and Trimmomatic (Bolger et al., 2014) can be used to trim (discard) low-quality reads, trim adaptor sequences, and eliminate poor-quality bases. Discarding parts of reads with lousy quality prevents aligning reads to wrong positions against the reference and therefore erroneously altering the read count in regions where the reads misalign due to detection errors in the nucleotide sequence during sequencing.

2.3.1.2 Alignment of reads

Reads are typically mapped to either a genome or transcriptome directly. Each of them has advantages and disadvantages. Prerequisite for mapping against the transcriptome is a highly annotated genome to extract the regions with coding sequences, therefore alignments against a transcriptome are computational quicker.
A good quality check for alignment accuracy is the percentage of mapped reads. Diverging percentage from 80 % and higher up to nearly 100% total matched reads are observable for alignments against the genome of well-studied species like mouse and human (Engström et al., 2013). Around 55% or more for (see results in publication 3.1) mapping against the transcriptome. As the transcriptome is only a small part of the genome and alignment rates go down as a consequence. This is due to the fact that a percentage of the reads cannot find a matching counterpart to the referenced transcriptome and stay unmapped. On the contrary, this means that reads also map to noncoding regions when mapped against the genome, which are not part of the annotated transcriptome. In turn, this could lead to de novo findings of transcripts or simple stay unannotated for the time being. Reads can map to only one position at the reference (*uniquely mapped*) or to multiple positions (*multi-mapped* reads). Genomic multi-mapped reads mostly come from repetitive sequences or shared domains of paralogous genes. They contain valuable information for further downstream analysis and should not be discarded. When the reference is a transcriptome, multi-mapping occurs a lot more because a read that would have been uniquely mapped on the genome would map equally well to all transcripts of a gene in the transcriptome that share the same exon. Nowadays the decision to map against genome is quite clear, as it allows to identify novel splice sites and isoforms and does not force reads into the predefined borders of the transcriptome when they arise from somewhere else in the unannotated genome. At the beginning of RNA-Seq analysis mapping against the transcriptome was the first thing to do, as it circumvented the problem to deal with splice junction and reads spanning them over multiple exons. It was easy to directly apply genomic aligner (bwa (Li and Durbin, 2009), Bowtie (Langmead et al., 2009),) to the transcriptome. Today the problem is solved and good and fast splice aware aligners are available (STAR (Dobin et al., 2013), TopHat(2) (Kim et al., 2013) and his successor HISAT2 (Kim et al., 2015), GSNAP (Wu and Nacu, 2010), MapSplice (Wang et al., 2010) and PALMapper (Jean et al., 2010)).
In general, parameters of alignment tools can be tweaked according to the mapping against the foward or reversed strand, the number of mismatches, the length and type of reads (single-end or paired-end) and the length of sequenced fragments the reads are derived from. Single-end reads are reads derive from only one side of a fragment during the library preparation step before sequencing. Whereas, paired-end reads are derived from both ends of a fragment and therefore contain distance information between that pair that should not change when aligned, which can be used to gain knowledge about sequence alterations from the reference. Besides, existing gene models annotation files can be provided to some aligners to map exon coordinates accurately and help in identifying splice sites. The addition of gene model annotation can have a crucial effect (Zhao and Zhang, 2015) on the quantification (see 2.3.1.3)

of reads towards the gene model (reflected as counts for a gene) and differential expression analysis (see 2.3.1.4). For a well rounded and comprehensive comparison of RNA-Seq mapper see Engström et al. (2013). Worth noting are pseudo aligners like Sailfish (Patro et al., 2014) its successor Salmon (Patro et al., 2017) and especially Kallisto (Bray et al., 2016), which are incredibly fast in performing the (pseudo-)alignment and read quantification (see 2.3.1.3) on the fly in a couple of minutes. These methods usually achieve this by building a coloured de Bruijn graph from all indexed k-mer possibilities of the read input data, which form nodes in the graph and the coloured paths are possible transcripts found in the transcriptome. The coloured paths generate k-compatibility classes for each k-mer, which can be interpreted as sets of potential transcripts. To achieve this, it is a prerequisite to have a high-quality transcriptome.

2.3.1.3 Transcript quantification

The next step in analysing transcriptomic data is transcript quantification. As noted at the end of 2.3.1.2, there are algorithms which rely on counting k-mers without the need of mapping them to a reference; still, the majority of tools for read-counting relies on aligned reads. The easiest way to count reads is to aggregate only the uniquely mapped reads, done in HTSeq-count (Anders et al., 2015), featureCounts (Liao et al., 2014) or the –quant option in STAR. For the quantification on gene-level a *Gene Transfer Format* (GTF) file, containing the genome coordinates of exons and genes, is needed.
More advanced algorithms are needed and have been developed for the estimation of transcript-level expression by tackling the problem of related transcripts sharing most of their reads. Cufflinks (Trapnell et al., 2012) estimates transcript expression from the mapping information to the genome obtained from mappers such as TopHat using an expectation-maximisation approach, which estimates transcript abundances within their so-called tuxedo workflow. It can estimate transcript *de novo* from the mapping data alone. Further algorithms that quantify expression from transcriptome mappings are *RNA-Seq by Expectation Maximization* (RSEM) (Li and Dewey, 2011), eXpress (Roberts and Pachter, 2013), Sailfish (Patro et al., 2014), Salmon (Patro et al., 2017) and Kallisto (Bray et al., 2016) among others. These methods use multi-mapped reads among transcripts as well and output within-sample normalised values corrected for sequencing. Additionally, the RSEM algorithm uses an expectation maximisation approach that also returns *transcripts per million* (TPM) values which are a more sophisticated way of normalising counts within samples than by Cufflinks *reads per kilobase of exon model per million reads* (RPKM) values used for removing gene-length and library-size effects as discussed by Wagner et al. (2012).

2.3.1.4 Differential gene expression analysis

Within sample normalisations like RPKM or TPM fail when used between samples due to different transcripts/ count distributions (Bullard et al., 2010) or biases, like Illumina's random hexamer priming used in their transcriptome sequencing protocols (Hansen et al.,

2010). Differential gene expression tools have to take between samples effects into account, for example by modelling the read count distribution as negative binomial distributed and using *generalized linear models* (GLM) followed by, for example, likelihood ratio test, test statistics. EdgeR (Robinson et al., 2010) or DESeq (Anders and Huber, 2010) and its successor DESeq2 (Love et al., 2014) are prominent examples for this. Limma followed by voom, as the successor of limma, the standard microarray analysis tool, is also applicable for count data and even allows to account for the correction of inter-gene correlation structures (Law et al., 2014) before testing for differential gene expression. Cufflinks is another option when positional biases in the coverage of transcripts are a problem.
Another big issue, when doing differential expression analysis, is the occurrence of batch effects. Sequencing of all samples from one experiment in the same machine run is often not possible. Usually, samples are collected over time and distributed in sequencing runs. These effects should be accounted for by minimising them via appropriate experimental designs at the beginning of project planning, being included in experimental designs as cofactors afterwards (Auer and Doerge, 2010) or removed by batch correction methods such as COMBAT (Johnson et al., 2007) or ARSyN (Nueda et al., 2012). These batch correction methods, although initially developed for microarray data, are still applicable for RNA-Seq. At the end of DEA a list of *differentially expressed genes* (DEGs) between conditions emitted with statistical annotations, like log_2 foldchange, p-value and adjusted p-values.

2.3.1.5 Functionial profiling

Functional annotation of differentially expressed genes is the last step for a standard transcriptomic analysis. It can be divided into two main approaches 1) check for DEGs if they are enriched in functional annotations, and 2) *gene set enrichment analysis* (GSEA), which is based on ranking the expressed genes from the differential expression analysis for enrichment of genes belonging to a common group or pathway. GSEA was initially developed for microarrays. RNA-Seq data biases such as gene length have to be accounted for and RNA-Seq specific tools have been developed. ClusterProfiler (Yu et al., 2012) can be applied on *Gene Ontology terms* (GO-Terms) and KEGG pathways and topGO (Alexa and Rahnenfuhrer, 2010) provides multiple algorithms to deal with the inherent correlation structure of ontologies, GOstat (Beissbarth and Speed, 2004; Beissbarth, 2006) is a program which automatically obtains the GO annotations from a database and generates statistics of overrepresented annotation from the analysed list of genes. ClusterProfiler and topGO are available via Bioconductor (https://www.bioconductor.org/) for the programming language R (R Core Team, 2013) and GOstat at http://gostat.wehi.edu.au/.
Functional annotation of not annotated species is possible as well by performing orthology searches via BLAST (Altschul et al., 1990) analysis based on similarity on sequence or protein databases like UniProt (Bateman et al., 2017), SwissProt (Bairoch and Apweiler, 1998) or Pfam (Finn et al., 2016). A handy workflow tool, written in Perl for performing these tasks is Annocript available at GitHub (https://github.com/frankMusacchia/Annocript).

2.3.2 Whole Exome Sequencing Analysis

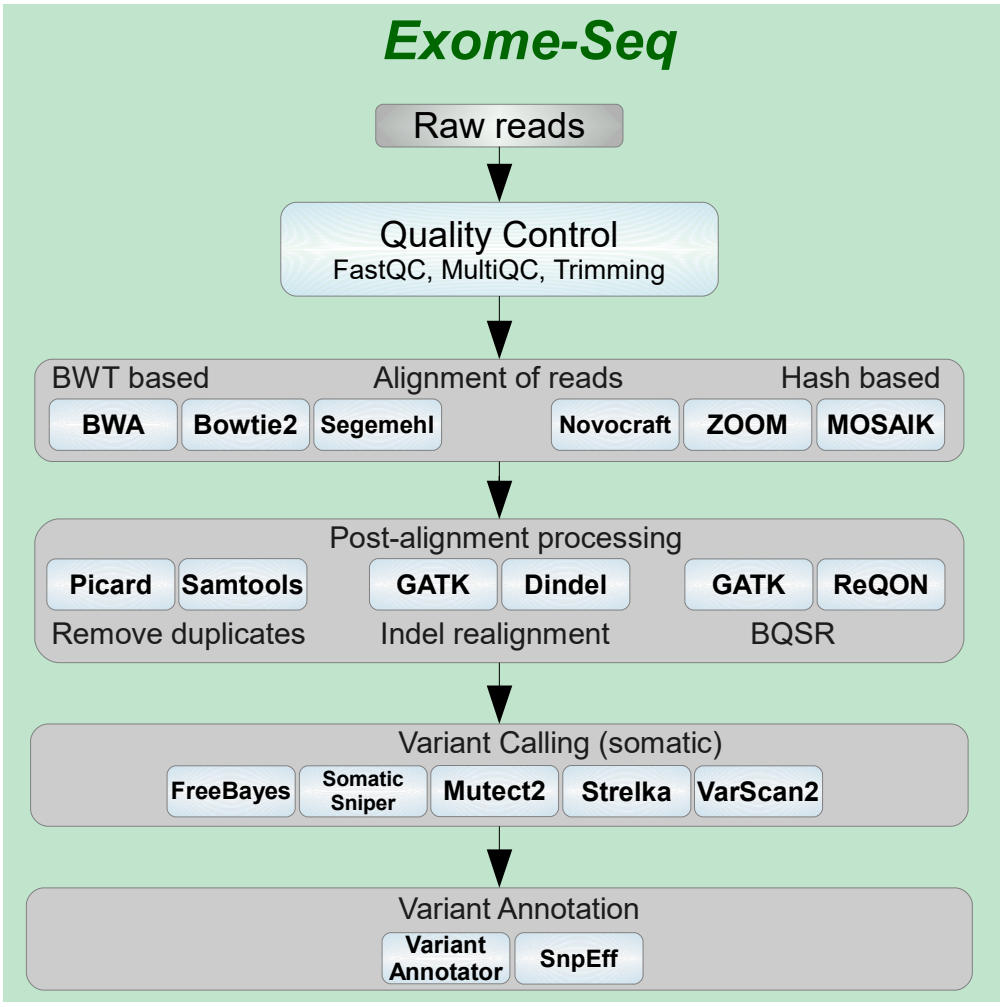

Figure 2.4. *The workflow for somatic Variation Calling structured with the different processing steps (grey boxes) with a selection of tools commonly used in the bioinformatical community (blue boxes).*

Performing a *Whole Exome Sequencing Analysis* (WESA) has been proven to be a successful alternative to WGS in detecting SNV genotypes and indels for a fraction of the sequence needed and therefore correlated with a fraction of the costs (Ng et al., 2010). Variants have been described to have a more likely neutral or weak effects on phenotypes in non-coding regions, even in well conserved non-coding sequences (Chen et al., 2007; Kryukov et al., 2005). Therefore the WESA is suitable to find high impacting variants enriched in a well-selected subset of the genome.

Another typical use case of WESA is the detection of somatic, here referring to the cell specific variations, which are likely to have a higher frequency of occurrence in cancer tissue samples. To differentiate mutations which are germline from mutations which a somatic, samples are sequenced in pairs. The healthy control sample can be extracted from blood

(or other non-cancer tissues) to detect germline mutations. The second pair is the sample coming from the tumour tissue. With these matched samples for one patient, it is possible to identify a tumour-specific somatic mutation, for the simple reason that the variation does not occur in the blood sample it is likely to be unique from the tumour. The main steps for a general WESA are quality control of the raw reads, alignment of reads, post-alignment processing, variant calling and variant annotation. An exhaustive review of various somatic mutation callers with descriptions of their underlying algorithms is given by Xu (2018). An illustration of the workflow for WESA with the most relevant tools for this dissertation is accompanied again in Figure 2.4.

2.3.2.1 Quality Control of raw reads

Raws reads for Exome-Seq analysis are stored in the same FASTQ file format as stated in section 2.3.1.1 and equal quality control measures can be applied for them. Minor differences to transcriptomic reads quality control exist, nevertheless. As the focus in Exome-Seq is more on single base resolution and detection of indels or complete CNVs it is a lot more urgent to be strict on selecting qualitative bases of reads and trim more aggressively lousy quality parts of reads. This has to be counteracted by increased the overall length of reads during sequencing via increasing the number of sequencing cycles and by sequencing both ends of a fragment to generate so-called paired-end reads. They shared the same fragment length distribution and based on this information, conclusions on events like indels as well as CNVs can be drawn (Bao et al., 2014). In other words, instead of using information on 50 base-pair long reads for mapping as it is often done for transcriptomic analysis, paired-end reads with at least 100 base pairs are suggested for WESA (as it has been done in the second publication of this dissertation at 3.2).

2.3.2.2 Alignment of reads

As the reads come from DNA and not from mRNA (with already spliced transcripts), the reads do not have to span splice junction and alignments become easier than for RNA-Seq splice aware aligners. Typical genome aligners are BWA (Li and Durbin, 2009), bowtie 2 (Langmead and Salzberg, 2012), segemehl (Hoffmann et al., 2009), which are based on the *Burrows-Wheeler Transformation* (BWT) (Burrows and Wheeler, 1994) or on hashing methods, like Eland (AJ Cox, Illumina, San Diego), ZOOM (Lin et al., 2008), SOAP2 (Li et al., 2009b), MOSAIK (Lee et al., 2014), Novocraft (http://www.novocraft.com) for DNA indexing and searching. Hash-based methods are stated to outperform BWT based methods in speed at the cost of memory usage, whereas BWT based methods offer sensitivity at the cost of flexibility (Lee et al., 2014).

2.3.2.3 Post-alignment processing

After alignment of reads, the post-alignment processing takes place to increase the quality of downstream variant calling. It consists of read duplicate removal, indel realignment, and *base*

quality score recalibration (BQSR). Also, there is not a general best practice in particular for each combination of methods out there as aligner and variant caller as well as variation and sequencing depth of the data play a role. Nevertheless, it is advised to be used and should be checked in each particular case.
Reads are considered duplicates when they have the same mapping coordinates against the reference. It is not possible to distinguish them from true DNA material or derived from the PCR amplification step during sequencing (see 2.2.1). For WES analysis, it is recommended to remove duplicates before variant calling, to reduce bias due to uneven amplification of DNA fragments (Xu, 2018). Programs such as Picard MarkDuplicates (http://picard.sourceforge.net) and SAMtools (Li et al., 2009a) can be used to remove duplicates.
After duplicate reads are removed areas in the genome, that contain indels has to be found and realigned to improve the overall alignment quality. The main issue here is that each read is getting aligned independently resulting in alignments with different mismatch positions with equal alignment score in gapped regions like indels, which lead to possible artificial mutation calls. Therefore the quality of alignment for this regions can be improved by doing local realignments considering all reads at once, so-called multiple sequence alignments. Programs implementing these are for example Dindel (Albers et al., 2011) and GATK's (DePristo et al., 2011) Unified Genotyper and the original Mutect (Cibulskis et al., 2013).
Each read has a Phred-scaled quality score attached, generated by the sequencing machine as the confidence of the called base at each position of the read. However, the scores generated during sequencing can be biased (Minoche et al., 2011) and need to be corrected, if possible. Therefore BQSR is recommended to increase the accuracy of confidence scores before calling variants. For each base of a read alignment, a corrected Phred-scaled quality score is calculated assuming that all observed differences between the aligned reads and the reference genome are sequencing errors. Also, it is necessary to exclude known variants before score recalibration, as they are true genomic variations and should not be considered as sequencing errors. GATK BaseRecalibrator (McKenna et al., 2010) and the Bioconductor package ReQON (Cabanski et al., 2012), which uses logistic regression for recalibration of the base quality scores, are available for this.

2.3.2.4 Variant Calling

When talking about variant detection, it is important to differentiate between germline and somatic variants. The first is most likely an inherited mutation related to family history and the second one is only present in tissue-specific cells. Therefore, both play wide-ranging roles in tumour development and this should be reflected in the selection of tools used. Popular somatic SNV caller are MuTect2 (Cibulskis et al., 2013), Strelka (Saunders et al., 2012), SomaticSniper (Larson et al., 2012) and VarScan2 (Koboldt et al., 2012) utilizing paired tumour-normal samples, whereas popular germline caller include GATK, SAMtools and FreeBayes (Garrison and Marth, 2012) utilizing variant detection either on multiple samples or sample-wise.

2.3.2.5 Variant Annotation

Identifying biological relevant mutations, like disease-causing mutations, from random errors or polymorphisms is the aim of variant annotation. Attributes, that can be annotated are a genomic feature, gene symbol, exonic function and amino acid changes. It has been described that most disease-causing mutations in Mendelian disorders and many disease-predisposing SNPs throughout the genome are non-synonymous SNVs and indels in the protein-coding regions (Rabbani et al., 2014). Therefore many public databases for additional information of variants have been set up. Public databases such as CIVIC (Griffith et al., 2017), ClinVar (Landrum et al., 2014), COSMIC (Forbes et al., 2008), RegulomeDB (Boyle et al., 2012), dbSNP (Sherry et al., 1999), PolyPhen (Adzhubei et al., 2010) and HaploReg (Ward and Kellis, 2012) are a source of additional knowledge when deciding for the pathology of found mutations. Programs, that perform these annotations are VariantAnnotator from GATK and SnpEff (Cingolani et al., 2012) equipped with read filters, Variant filtration, and prioritisation.

2.4 Biology of cancer

2.4.1 What is Cancer?

In healthy mammalian cells, embryogenesis, growth, and tumourigenesis require cell proliferation. During the acquirement of a broader and more profound knowledge of proliferation-state regulations, growth-factor signal transduction and transcriptional networks necessary to initiate and maintain cell cycling, it became known that proliferation relies highly on the metabolic activities of the cell. The cell requires a high amount of nutrition to divide into daughter cells, which is a metabolic challenge and vulnerable to disruptive influences. The cellular uptake and metabolism of nutrients, therefore, depends on extracellular signals, like hormones, cytokines, and growth factors (Hedeskov, 1968; Whetton et al.; Bauer et al., 2005; Berridge and Tan, 1995). These extracellular signals also drive cell growth, proliferation, and survival.

When these extracellular stimuli start to malfunction, for example, introduced by mutations in genes in pathways responsible for signal transduction, they can initiate permanent cell growth leading to cancer. Normally, uncontrolled growth should activate controlled cell death, called apoptosis regulated by the tumour suppressor p53 to prevent the cell from replicating. On the other hand, there are oncogenes and proto-oncogenes, if they are altered by mutations or overexpressed, they can counteract cell death and promote cell proliferation ultimately leading to rapid growth, cancer (Todd and Wong, 1999; DeBerardinis et al., 2008).

Cancer can be divided into different sub-types based on their origin. Carcinomas, like in colorectal, lung breast or prostate cancer usually starts from the skin or the tissue surface of inner organs and form solid tumours. They are the most common type of cancer. Sarcomas, they start from tissue connecting the body, like fat, muscles, nerves, tendons, joints, blood vessels, lymph vessels, cartilage, or bone. Leukaemias are cancer originating from the blood. Healthy blood cells change and grow uncontrollably. They can be divided into four types:

acute lymphocytic leukaemia, chronic lymphocytic leukaemia, acute myeloid leukaemia, and chronic myeloid leukaemia. Lymphomas, are cancer originating in the lymphatic system, which is a system of connected vessels and glands to help fight infections. There are two main lymphomas: Hodgkin lymphoma and non-Hodgkin lymphoma (ASCO, 2012).

2.4.2 Cancer Progression and Metastases

A clonal evolutionary model of cancer development was first proposed by Nowell (1976) and elaborates upon Darwinian models of natural selection and connects cancer as an asexually reproducing, unicellular, quasi-species (Greaves and Maley, 2012). Its a repeating process of clonal expansion, genetic diversification due to occurring gene modifications and mutations, and selection of subclones in heterogeneous cancer, which may influence the growth of an entire tumour and thereby actively maintain tumour heterogeneity (Heppner and Miller, 1983). The next step in cancer progression is called metastasation.

At an early stage of cancer, the cancerous cells are bound to a primary site within tissue boundaries. If the cancer cells manage to dislocate from the primary tumour tissue and penetrate the walls of lymphatic or blood vessel they can circulate through the body. Of these *circulating tumor cells* (CTC) only a minor fraction of only 0.01% manages to establish distant tumours in organs or tissues. These distant tumours are the fittest subpopulation of the primary tumour as they managed to survive the circulation and the establishment in the new microenvironment and therefore are a turning point to the worse as the patient cannot be cured by local therapy anymore (Liotta and Kohn, 2000). Nowadays genes are known, which initiate tumour progression and metastasation. They can promote cell motility, epithelial-mesenchymal transition (EMT), extracellular matrix degradation, angiogenesis or evasion of the immune system. Prominent examples here are Twist1, Snai1 and Snai2 as aberrantly regulated transcription factors, or modulators of invasion associated pathways like hepatocyte growth factor (HGF), VEGF and ERK pathways. It is also known that the suppression of non-coding RNAs (like miR-126 and miR-335 in breast and gastric carcinomas) promotes metastatic growth (Yang et al., 2011; Feng et al., 2010).

Metastasis suppressor genes, on the other hand, can inhibit metastasis at any step of the metastatic cascade. To date, some metastasis suppressor genes are known, such as nonmetastatic gene 23 (NM23), Kangai 1 (KAI1), KISS1, mitogen-activated protein kinase 4 (MKK4), breast cancer metastasis suppressor 1 (BRMS1), Rho GDP dissociation inhibitor 2 (RhoGDI2), cofactor required for Sp1 transcriptional activation subunit 3 (CRSP3) and Vitamin D3 up-regulated protein 1 (VDUP1) (Martin et al., 2013).

In this dissertation, multiple datasets of different metastatic tissues were used. In all three publications, metastatic tissue from different patients and animal having CRC was used. CRC is the third most common type of cancer and the five years survival rate for patients is roughly 50% (Machii and Saika, 2014), given that half of them develop distant metastasis in the liver the survival rate goes down further towards approximately 30% (Leporrier et al., 2006).

2.4.3 Treatment of Cancer

Treatment options for cancer typically consist of surgery, systemic and radiation therapy or a combination of them. Treatment for metastasis is different, as the therapeutic window of opportunity in which to treat a driver mutation, before clonal expansion and divergence occur (Gerlinger and Swanton, 2010; Galvão and Newton, 2005; Shah et al., 2002), can have closed. Resulting in narrowed therapeutic options as possible drug resistance in the fittest metastatic subclones could emerge. Therefore, physicians try different combinations of chemotherapy, radiation and surgery to remove the metastases. Segal et al. (2003) reported that resistance exercises reduces fatigue and improves quality of life and muscular fitness. Therefore his form of exercise can be an essential component of supportive care for these patients. Another therapy option are targeted therapies. They attempt to exploits a tumour's dependence on proliferation and survival pathways. However, it has been shown that this only showed high success rates in a range of solid tumour types and fails in advanced disease cases, explainable by the tumour heterogeneity via subclone progression of the cancer (Gore and Larkin, 2011; Diaz et al., 2012).

Therefore the identification of biomarkers that inform about prognosis, or identify low frequent clonal subgroups of heterogeneous cancer and metastasis which drive the final disease outcome is the primary challenge sequencing analysis, and bioinformatics should overcome.

2.5 Aim and structure of this work

The global objective of this work is to understand the capabilities of RNA-Seq analysis methods and their limits. Two specific aims are defined:

- Evaluating different bioinformatical tools on microarray and RNA-Seq (see 3.1), to establish a quality standard when performing differential gene expression analysis followed by gene set enrichment with RNA-Seq samples structured as a workflow.
- Develop a method to detect mutations in RNA-Seq data, as the detection of mutations is an integrated part of gaining knowledge about the cause of cancer and metastasis next to differential expression analysis. Therefore the performance and limits of this application in RNA-Seq had to be checked (see 3.2).

The last paper deals with the successful application of the knowledge for the best-performing tools from the first paper and utilises the complete RNA-Seq analysis workflow in the background of colorectal cancer metastasising into the liver in mice (see 3.3).

3 Cumulative part of the dissertation

3.1 A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal-cancer patients and Burkitt Lymphoma cells

Reference:
Wolff A, Bayerlova M, Gaedcke J, Kube D, Beißbarth T (2018) A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal-cancer patients and Burkitt Lymphoma cells. PLoS ONE 13(5): e0197162.https://doi.org/10.1371/journal.pone.0197162

3.1.1 Summary and discussion

This work aimed to establish a quality standard when performing differential gene expression analysis followed by gene set enrichment with *RNA sequencing* (RNA-Seq) samples. Different methods were compared according to their performance during different steps of analysis based on real datasets for RNA-Seq and microarray. The best performing pipeline for RNA-Seq analysis was then used as standard for the analysis of all RNA-Seq samples within the MetastaSys Consortium (for an application on mice see Paper 3.3).
Matched datasets were prepared to reflect the spectrum of possible platforms to derive datasets for RNA-Seq (see Table 3.1) and microarray analysis. The *rectal cancer* (RC) patient dataset consists of patients having a good and a bad prognosis of recurring metastasis. Initially planned for an equal samples size of five samples in each group, the prognosis of one patient changed for the worse resulting into a comparison of four versus six patients. The data is available at GEO under the accession number GSE99897 for RNA-Seq and GSE100109 for microarray. The second data set is from Burkitt Lymphoma cell lines comprising three replicates of control samples and three replicates of cell line stimulated with *B-cell activating factor* (BAFF).
RNA-Seq data was quality checked, aligned, qualities of mapped reads were manually investigated, reads were counted and tested for differential gene expression. The microarray data was preprocessed, quantile normalised and differentially expressed genes were detected.

Finally, GO-term enrichment analysis was performed on the results of all used pipelines (see Figure 1 from the Manuscript). In total, five pipelines were evaluated P1 to P4 for RNA-Seq pipelines and on the left side of Figure 1 the microarray analysis pipeline using limma.

Dataset	Microarray	RNA-Seq	samples
BL2	Affymetrix	mRNA	3 CTL and 3 BAFF samples
RC	Agilent	total RNA	4 good prog. and 6 bad prog. patient samples

Table 3.1. *Overview of the datasets and platforms that were spotted or sequenced in the publication.*

Three aligners were evaluated based on their total mapping rates as the alignments should recover as much information from the reads as possible by finding the right locations in the reference genome. The STAR aligner could map the most reads with a total of 98.98 (± 0.05) % of reads for BL2 and 98.49 (± 0.35) % for RC. TopHat2 could map 97.02 (±0.1) % and 96.73 (±0.4) % respectively with a slightly higher proportion of unique mapped reads (see 3.2 for full details). Sailfish cannot deal with any multi-mapping at all. The RC dataset was made of total-RNA and had a higher percentage of non-coding RNA fragments, which are not annotated in databases yet and therefore cannot be matched by Sailfish.
Next, the normalised count values derived from HTSeq, RSEM, Sailfish, Cufflinks and the normalised intensities from the microarray were compared via correlation analysis using $1 - PearsonCorrelationCoefficient$ as a distance measure for complete linkage clustering. The average correlation coefficients between the RNA-Seq methods and microarray were also observed in other studies (Marioni et al. (2008); Bradford et al. (2010)), but a surprise was the low correlation results of Cufflinks which were worse than the technical differences explainable by the different techniques from microarray and RNA-Seq. It was expected, that the difference between the two different platforms used is higher than the methodological within RNA-Seq methods (see Figure 3 from the Manuscript).

Mapping-rates	STAR		TopHat2		Sailfish	
	BL2	RC	BL2	RC	BL2	RC
unique	78.61 ± 0.55	83.13 ± 1.85	84.25 ± 0.54	85.40 ± 1.62	84.81 ± 0.85	54.44 ± 3.71
multi	20.37 ± 0.52	15.36 ± 1.69	12.78± 0.48	11.32 ± 1.46	0	0
total	98.98 ± 0.05	98.49 ± 0.35	97.02 ±0.1	96.73 ±0.4	84.81 ±0.85	54.44 ± 3.71

Table 3.2. *Overview of unique, multi and total mapping rates in percentages for the three aligners STAR, TopHat2 and Sailfish.*

For *differential expression analysis* (DEA) the overall number of *differentially expressed genes* (DEGs) for the RNA-Seq pipelines were much higher than observed for the microarray analysis. Moreover, the overlaps of DEGs between the pipelines P1-P4 were compared. DEGs found by at least two other pipelines were named as 'consensus DEGs' and genes found by only one pipeline are called 'DEGs unique'. In the absence of the ground truth, this setup

uses these two measures as an indicator of pipeline performance. Having a high overlap with other existing pipelines should generally hint towards a more robust method for pipeline usage. Also, these proportion are only indicators and are by no means actual true-positives and false-negatives, but more reflecting the general agreement with other compared methods and the inconsistency or uniqueness of each method.

For DEG evaluations, the microarray results could not be addressed as the experiment had no DEGs left after multiple test correction in both datasets. Strikingly, P4(Cuff) had the lowest Consensus DEGs reflected by both datasets (BL2: 45%, RC: 16.88%) and at the same time, the highest rate of uniquely found DEGs (BL2: 55%, RC: 79.87%). If we take into account the low correlation results from before, it seems that this pipeline performance is far worse than the rest of the tested pipelines. On the other hand P1(HTSeq) closely followed by P2(RSEM) and P3(Sailfish) had a high number of agreeing DEGs resulting in overall consistent results (Consensus DEGs: BL2: (169-211 genes), RC: (48-51 genes)) between all three pipelines. Taking the RC dataset as an example, the pipelines differ by their number of unique DEGs. P1(HTSeq) had only 9 out of 71 DEGs unique. This little number could be explained by the conservative approach of HTSeq only utilising unique reads for counting. P2(RSEM) had 28 out of 96 genes unique, most likely due to integrating multi-mapped reads as well, resulting in more read counts, which support the easier identification of DEGs by edgeR. Sailfish had 78 out of 146 genes unique, resulting in a rather high percentage of 53% unique DEGs, which might be due to the kmer exact match approach from sailfish is not robust enough against sequencing errors and base mutations in genes, resulting in higher number of misplaced gene counts, ultimately resulting in higher numbers of identified DEGs. For a complete overview of the number and percentages as well as the individual overlaps, have a look at Table 3 (https://doi.org/10.1371/journal.pone.0197162.t003) and figure 4 (https://doi.org/10.1371/journal.pone.0197162.g004) from the publication.

The last evaluation of pipeline performance was done with functional enrichment on *Gene Ontology terms* (GO-Terms). As the Microarray results did not show significant genes after multiple test correction, the threshold was lowered here to less than five % p-value. This was done to check if at least in the gene ranks amongst the top, have interpretable biological results, despite not being significant. For the BL2 dataset it is known, that analysis involving BL2 cell lines stimulated with BAFF, GO-Terms related to immune response should show up (Mackay and Browning, 2002). In the top 20 significantly enriched GO-Terms four (GO:0060333 interferon-gamma-mediated signalling pathway, GO:0019886 antigen processing and presentation of exogenous peptide antigen via MHC class II, GO:0050852 T cell receptor signalling pathway, GO:0031295 T cell costimulation) were identified with this background in all pipelines P1-P4. In total 13 out of the 20 GO-terms were linked to the immune system but only found by P1-P3 consistently. Also, four GO-terms are related to metabolism, two to cell signalling and the last one to biological regulation. For the microarray data, only one of these GO-term was significant, and seven were not detected at all.

When comparing GO-term analysis results of the rectal cancer patient group, it was expected to see GO-terms related to metastases formation, like increased proliferation, cell rearrangements, changes in cell organisation and to a specific extent immune response as well. Therefore

it was checked again if several of these assumptions within the top 20 significantly enriched GO-terms could be observed. From these, GO-terms (GO:0090263, GO:0002158, GO:0090090) linked to cellular proliferation, GO-terms (GO:0030199, GO:0022617, GO:0071711) linked to cellular rearrangements and GO-terms (GO:0060337, GO:2000551, GO:0030853) related to immune system response could be identified. However, a lot of significant GO-terms could not be related to any of the expected cellular response classes. The full setup of each of the 20 GO-Terms for both datasets is present in the publication at figure 5 (https://doi.org/10.1371/journal.pone.0197162.g005)

In conclusion, the combination of STAR aligner with HTSeq-Count followed by STAR aligner with RSEM and Sailfish generated differentially expressed genes best suited for the dataset at hand and in agreement with most of the other transcriptomic pipelines. As RSEM is utilising multi-mapped in addition to unique mapped counts for isoform and gene expression, the final pipeline for MetastaSys consisted of STAR for alignments, RSEM for read counting followed by edgeR for DEA and topGO for functional level annotation of gene ontologies. If your interest is not in high variable cancer data and no high-performance cluster for analysis is available, Sailfish is a good option to do the CPU and memory intense alignment and counting part, resulting in a reliable and profound analysis suitable for a typical desktop setup.

3.1.2 Declaration of my contribution

Conceptualization: Alexander Wolff, Michaela Bayerlova, Tim Beißbarth.
Data curation: Alexander Wolff, Jochen Gaedcke, Dieter Kube, Tim Beißbarth.
Formal analysis: Alexander Wolff, Michaela Bayerlova.
Investigation: Alexander Wolff.
Methodology: Alexander Wolff.
Project administration: Alexander Wolff, Tim Beißbarth.
Resources: Jochen Gaedcke, Dieter Kube, Tim Beißbarth.
Software: Alexander Wolff, Michaela Bayerlova.
Visualization: Alexander Wolff.
Writing– original draft: Alexander Wolff, Michaela Bayerlova.
Writing – review & editing: Alexander Wolff, Michaela Bayerlova, Jochen Gaedcke, Dieter Kube, Tim Beißbarth

PLOS ONE

RESEARCH ARTICLE

A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal-cancer patients and Burkitt Lymphoma cells

Alexander Wolff[1], Michaela Bayerlová[1], Jochen Gaedcke[2], Dieter Kube[3], Tim Beißbarth[1]*

1 Dept. of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany, **2** Dept. of General-, Visceral- and Pediatric Surgery, University Medical Center Göttingen, Göttingen, Germany, **3** Dept. of Hematology and Oncology, University Medical Center Göttingen, Göttingen, Germany

* Tim.Beissbarth@ams.med.uni-goettingen.de

Citation: Wolff A, Bayerlová M, Gaedcke J, Kube D, Beißbarth T (2018) A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal-cancer patients and Burkitt Lymphoma cells. PLoS ONE 13(5): e0197162. https://doi.org/10.1371/journal.pone.0197162

Editor: Petr V Nazarov, Luxembourg Institute of Health, LUXEMBOURG

Received: June 26, 2017

Accepted: April 27, 2018

Published: May 16, 2018

Data Availability Statement: All expression data are available from the GEO (https://www.ncbi.nlm.nih.gov/geo) database (accession numbers GSE99768, GSE100111, GSE99897, GSE100109).

Funding: Funding for this study provide by: Bundesministerium für Bildung und Forschung, MetastaSys (0316173A), Prof. Tim Beißbarth; Bundesministerium für Bildung und Forschung (DE), MyPathSem (031L0024), Prof. Tim Beißbarth; Bundesministerium für Bildung und Forschung, HER2LOW (031A429), Prof. Tim

Abstract

Background

Pipeline comparisons for gene expression data are highly valuable for applied real data analyses, as they enable the selection of suitable analysis strategies for the dataset at hand. Such pipelines for RNA-Seq data should include mapping of reads, counting and differential gene expression analysis or preprocessing, normalization and differential gene expression in case of microarray analysis, in order to give a global insight into pipeline performances.

Methods

Four commonly used RNA-Seq pipelines (STAR/HTSeq-Count/edgeR, STAR/RSEM/edgeR, Sailfish/edgeR, TopHat2/Cufflinks/CuffDiff)) were investigated on multiple levels (alignment and counting) and cross-compared with the microarray counterpart on the level of gene expression and gene ontology enrichment. For these comparisons we generated two matched microarray and RNA-Seq datasets: Burkitt Lymphoma cell line data and rectal cancer patient data.

Results

The overall mapping rate of STAR was 98.98% for the cell line dataset and 98.49% for the patient dataset. Tophat's overall mapping rate was 97.02% and 96.73%, respectively, while Sailfish had only an overall mapping rate of 84.81% and 54.44%. The correlation of gene expression in microarray and RNA-Seq data was moderately worse for the patient dataset ($\rho = 0.67–0.69$) than for the cell line dataset ($\rho = 0.87–0.88$). An exception were the correlation results of Cufflinks, which were substantially lower ($\rho = 0.21–0.29$ and 0.34–0.53). For both datasets we identified very low numbers of differentially expressed genes using the microarray platform. For RNA-Seq we checked the agreement of differentially expressed genes identified in the different pipelines and of GO-term enrichment results.

Beißbarth; Bundesministerium für Bildung und Forschung, MMML-Demonstrators (031A428), Dieter Kube; Deutsche Forschungsgemeinschaft, Open Access Publication Funds of the Göttingen University, Not applicable.

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

Conclusion

In conclusion the combination of STAR aligner with HTSeq-Count followed by STAR aligner with RSEM and Sailfish generated differentially expressed genes best suited for the dataset at hand and in agreement with most of the other transcriptomics pipelines.

Introduction

Transcriptomics as an area in the research field of functional genomics has always been a key player for identifying interactions and regulations of gene expression. Over the last two decades it was common practice to use microarrays for any investigation in transcriptomics. Within the last ten years the next generation sequencing (NGS) and especially RNA sequencing (RNA-Seq), became widely available [1]. These technologies are gradually replacing microarrays, when analyzing and identifying complex mechanism in gene expression. Decreasing running costs, higher dynamic range of expression and higher accuracy in low abundance measurements [2] are the main factors for this fast development of NGS and increasing use of RNA-Seq over microarray.

The versatility in using RNA-Seq, like discovering novel small RNAs (smRNA), microRNA (miRNA), long-non-coding RNAs (lncRNA) or alternative splicing events [3], is a further factor for an increasing popularity of this profiling approach. Another advantage is the currently highly discussed variant calling [4] [5] [6] based on RNA-Seq data, which makes this technology even more attractive. The developments of new technologies, like Pacific Bioscience or Nanopore [7], can further contribute in the field of RNA-Seq and transcriptomics in form of more detailed annotation databases in the future.

A typical application for RNA-Seq is the differential gene expression analysis. First, millions of short reads are produced, which are mapped to a reference genome. Subsequently, the amount of reads mapping to a genomic feature of interest (for example a gene, transcript or exon) is measured as the abundance of these features [8]. The abundance per feature is used as an input for differential expression analysis.

Still, microarrays are widely used because of their lower costs compared to the RNA-Seq technology. Moreover, there are large and well maintained repositories, such as ArrayExpress [9] and Gene Expression Omnibus (GEO) [10], that have collected the microarray data over long time periods. RNA-Seq data collections are increasing in GEO and the The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/).

While the preprocessing and analysis steps of microarray data are mostly standardized, the establishment of RNA-Seq data analysis methodology and standards is still ongoing in the field of transcriptomics. A lot of efforts have been performed into method comparison studies to change this [11] [12] [13] [14]. The quality evaluation of different RNA-Seq (pre-)processing methods are one important step to establish a quality standard. Great effort in this field have been accomplished, for instance by Sequencing Quality Control (SEQC) consortium [11] and has already been done for microarrays years ago in the MAQC-I and MAQC-II projects [15] [16].

We aim to investigate commonly used RNA-Seq pipelines on multiple levels (alignment, counting) and cross-compare the results with the microarray counterpart on the level of gene expression and gene ontology enrichment. For these evaluations we generated two matched microarray and RNA-Seq datasets: rectal cancer (RC) patient data (good versus bad prognosis patients) and Burkitt Lymphoma (BL2) cell line data (control versus stimulated cells).

Materials and methods

Burkitt Lymphoma cell-line data (BL2)

BL2 cells were cultivated as described previously at cell densities between 2×10^5 and 1×10^6 cells/ml [17]. For stimulation studies, cells were cultured in cell culture medium supplemented with 10 mM HEPES at 1×10^6 cells/ml and incubated with B-cell activating factor (BAFF) for up to 24 hrs instead of 9hrs [18]. RNA was isolated with RNAeasy Plus Mini Kit (Qiagen) according to the manufacturer's instructions and labeled using Affymetrix GeneChip IVT Labelling Kit (Affymetrix). Fragmentation and hybridization on Human ST1.0 Arrays were processed according to manufacturer's recommendations by the TAL (UMG, Germany). Microarray based profiling was performed using Affymetrix GeneChip Human Gene 1.0 ST array in three independent replicates of the experiment with the stimulated versus unstimulated cell line. For RNA-Seq, single-end sequencing on an Illumina HiSeq 2000 machine with the poly-A capturing protocol with 43 base pairs read length was used. The RNA was isolated using Trizol reagent including a DNase I (Roche, Mannheim, Germany) digestion step and Library preparation was performed using the TruSeq Stranded Sample Preparation Kit (Illumina, RS-122-2201) starting from 1000 ng of total RNA. Accurate quantitation of cDNA libraries was performed using the QuantiFluor TM dsDNA System (Promega). The size range of nal cDNA libraries was determined applying the SS-NGS-Fragment 1–6000 bp Kit on the Fragment Analyzer from Advanced Analytical (320 bp). cDNA libraries were amplified and sequenced by using the cBot and the HiSeq2000 from Illumina. The BL2 dataset is accessible through GEO Series accession number GSE99768 for the RNA-Seq dataset and GSE100112 for the microarray data.

Rectal cancer patient data (RC)

The rectal cancer patient dataset consists of 10 patients from a clinical study at the Surgery department of the University Medical Center Göttingen collected over a longer time. Patients were chosen based on the follow-up time and development of a distant metastasis. First a balanced sample size of five versus five patients with and without a metastatic event was intended. A later development of metastasis of one of the good prognosis patients changed the sample size to 6 versus 4 patients. The study is approved from the Ethic commission of the University medical centre Göttingen, ethic number: 9/8/08. Biopsies were immediately stored in RNAlater (Qiagen, Hilden, Germany). Subsequently, for microarray RNA was isolated using TRIzol (Invitrogen, Carlsbad, CA) according to the manufacturer´s instructions. Nucleic acid quantity, quality and purity were determined using a spectrophotometer (Nanodrop, Rockland, DE) and a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). 1 μg of total RNA was labeled with Cy3 using the Low RNA Input Fluorescent Linear Amplification Kit according to the manufacturer's recommendations (Agilent Technologies, Santa Clara, CA). Quantity and efficiency of the labeled amplified cRNA were determined using the NanoDrop ND-1000 UV-VIS Spectrophotometer version 3.2.1. 1.5 mg of Cy3-labeled cDNA was hybridized to an oligonucleotide-based Whole Human Genome Microarray (4x44K, Agilent Technologies) and incubated at 65˚C for 17 hours. Slides were washed and scanned using an Agilent G2565BA scanner.

For RNA-Seq single-end sequencing for 50 base pair reads the RNA was isolated using Trizol reagent including a DNase I (Roche, Mannheim, Germany) digestion step. Library preparation for RNA-Seq was performed using the TruSeq Stranded Sample Preparation Kit (Illumina, RS-122-2201) starting from 1000 ng of total RNA. Accurate quantitation of cDNA libraries was performed using the QuantiFluor TM dsDNA System (Promega). The size range

of nal cDNA libraries was determined applying the SS-NGS-Fragment 1–6000 bp Kit on the Fragment Analyzer from Advanced Analytical (320 bp). cDNA libraries were amplified and sequenced by using the cBot and the HiSeq2000 from Illumina for single end reads with a base pair length of 50.

The RC dataset is accessible through GEO Series accession number GSE99897 for the RNA-Seq dataset and GSE100110 for the microarray data.

Microarray data preprocessing and analysis

All preprocessing and statistical analyses of microarray data were performed using R statistical computing environment [19]. Affymetrix BL2 data was processed using the custom CDF file (*hugene10st_Hs_ENTREZG), getting the most complete gene meta data annotation for the affymetrix probe ids. Afterwards* the Robust Multi-array Average (RMA) algorithm was applied [20]. Adittional quality control metrics for BL2 can be found in the supplements (S1 File). Both datasets were log2 transformed and quantile normalized. In case of several probes corresponding to the same Ensembl gene identifier, the probe with median expression intensities was chosen to represent the gene level expression. Differential expression analysis was performed by fitting linear models using empirical Bayes method as implemented in the limma r-package [21] and p-values were adjusted for multiple testing using Benjamini-Hochberg (BH) method [22].

NGS data preprocessing and analysis

NGS quality control. The raw reads from both datasets were quality assessed using fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Beside an agglomeration of nucleotides with slightly lower quality at the starting positions than in the middle of reads, no major quality issues were observed (S1 Table, S1a and S1b Fig). For each samples the distribution of unique, multi- and unmapped reads were checked for high proportion of unmapped or multi mapped reads, which were not explainable by the underlying alignment methods (S1 Appendix).

Generation of alignments. Different state-of-the-art RNA-Seq aligners were compared: STAR, TopHat2 and Sailfish.

STAR (v2.4.0h) is a splice-aware ultrafast universal RNA-Seq aligner, which utilizes a sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and a stitching procedure [23]. TopHat2 (v2.0.13) is as well a splice-aware RNA-Seq aligner which uses a two-step approach: 1. detecting potential splice sites for introns, 2. using these candidate splice sites in a subsequent step to correctly align multi exon-spanning reads [24]. Sailfish (v0.6.3) works differently and is not directly an aligner, since it avoids mapping of reads entirely and utilizes the observation of k-mers occurring in reads instead of alignments of reads [25].

Reads obtained from RNA sequencing were mapped against the reference genome of Homo sapiens Ensembl Version GRCh38.76 utilizing further information from the gene transfer format (.gtf) annotation from Ensembl version GRCh38.76. In case of Sailfish, it required a precomputed set of transcripts in fasta format. This was done with RSEM's rsem-prepare-reference function providing the reference and the .gtf annotation.

Generation of counts. Multiple tools for counting of reads overlapping gene features were utilized: HTSeq, RSEM, Sailfish, and Cufflinks.

HTSeq-Count is a tool from the Python Toolbox HTSeq (v0.5.4p1) for counting reads overlapping into a specific feature (gene) [26]. RSEM (v1.2.19) is a software package for quantification of gene and isoform abundance estimation, utilizing an expectation maximization

algorithm [27]. Sailfish an alignment-free tool to estimate isoform abundances via an expectation maximization algorithm, directly from a set of reference sequences, using k-mers as main transcript coverage unit. Cufflinks (v2.0.13) performs estimation of abundance with a likelihood based approach for simultaneous estimation of bias parameters and expression levels [28].

Later comparisons are based on tpm (transcript per million) values. Therefore, after statistical testing, the fragments per kilobase per million (fpkm) values and normal read count data were transformed (Supplement S2 Appendix) to tpm values for comparability in figures, using the R programming language.

Correlation analysis. The correlation analysis were done by taking the mean of each sample wise correlation test between pipeline methods. As distance measure we took 1-Pearson correlation, followed by complete linkage hierarchical clustering of the samples. All calculation were performed in R.

Analysis of differential gene expression. After counting reads, all abundance values were compared with edgeR, performing a likelihood ratio test (glmLRT). This R-package implements a range of statistical methods based on the negative binomial distributions, like empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests [29]. Cufflinks does not deliver read count data and therefore had to be tested by Cufflinks cuffDiff [30].

As cutoff for significantly differentially expressed genes after multiple testing correction (BH), a false discovery rate (FDR) of five percent was used. All results after differential gene expression were transformed into tpm (S2 Appendix) and the significant genes of each Pipeline result were used as input for gene ontology enrichment analysis.

Pipelines. Based on the described steps and tools used, 4 different pipelines were set up (Fig 1) and named as follows: P1(HTSeq) including STAR, HTSeq-Count and edgeR; P2 (RSEM) consisting of STAR, RSEM and edgeR; P3(Sail) with Sailfish and edgeR; P4(Cuff) consisting of TopHat2, Cufflinks and CuffDiff.

Gene ontology enrichment analysis (GO analysis). Genes with an FDR smaller than 5% where selected and sorted as a gene-list in ascending order. These gene-lists derived from edgeR, CuffDiff and limma were used as input for the weighted fisher-exact test implemented in the package TopGO version 1.0 [31] to calculate the enrichment for each GO-category. The GO-Enrichment analysis tests if a selected feature set (gene-list of DEGs) falls into a Gene Ontology category more often than expected by chance. GO-terms with a p-value smaller than 5% were considered significant and used subsequently for visualizations.

Results and discussion

The aim of this study was to evaluate common analysis methods for RNA-Seq differential gene expression and cross-compare them with well established analysis methods for microarray. The comparisons were evaluated based on matched microarray and RNA-Seq profiles of two datasets: 1.) rectal cancer patient dataset comprising four patients with a good prognosis and six patients with a bad prognosis (referred to as RC dataset). 2.) Burkitt Lymphoma cell line dataset comprising three replicates of control cell line and three replicates of cell line stimulated with BAFF (referred to as BL2 dataset). RNA-Seq data was quality checked, aligned, qualities of mapped reads were manually investigated, reads were counted and analyzed for differential gene expression. The microarray data was preprocessed, quantile normalized and differentially expressed genes were detected. Finally a GO-term enrichment analysis was performed on the results of all used pipelines (Fig 1).

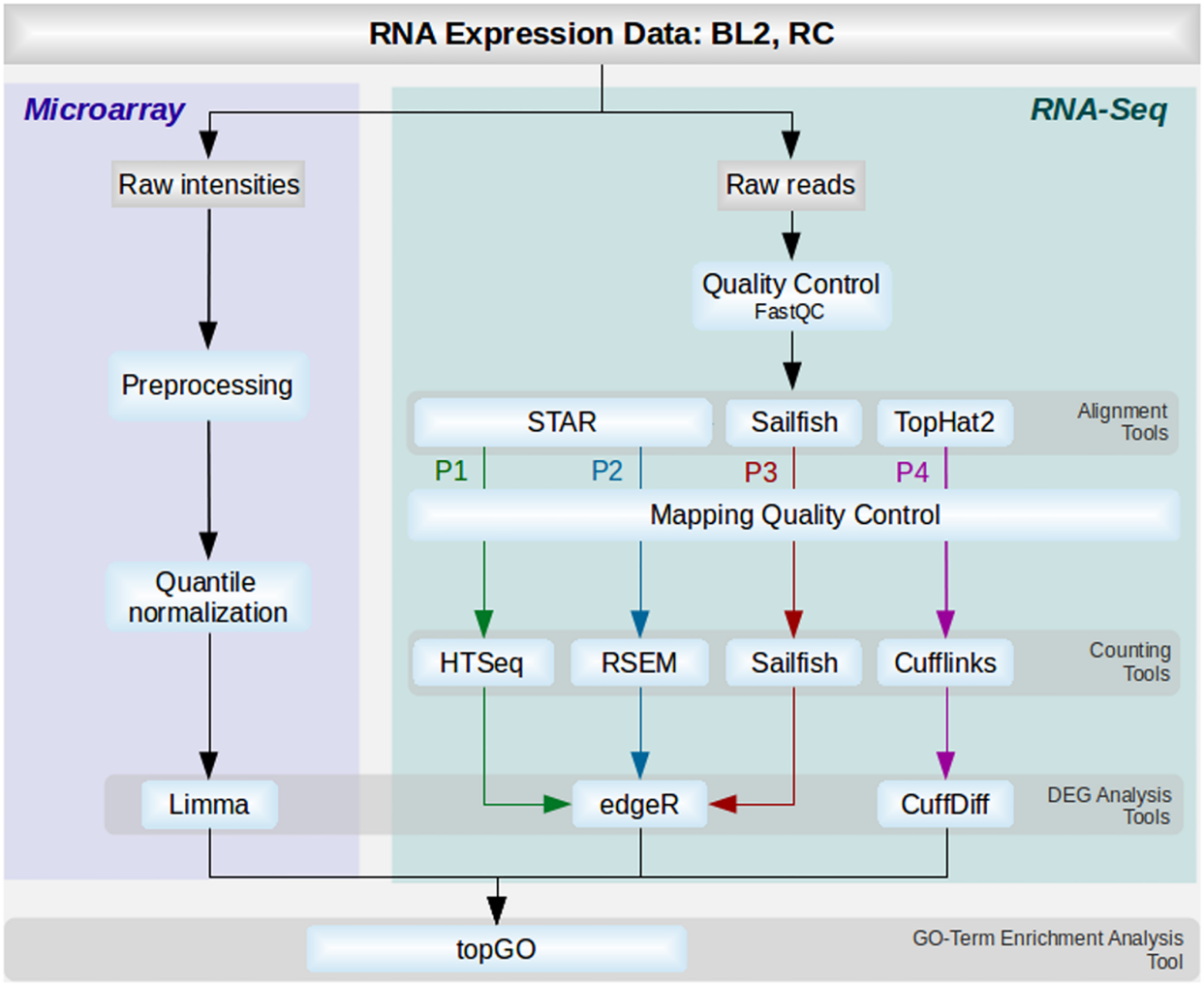

Fig 1. The different analysis pipelines. The flowchart describes the different tools and steps used for microarray (blue) and RNA-Seq analysis (green). Tasks and tools used at different steps are colored in light blue. Tools corresponding to the same steps are grouped and colored as follows: green (P1 with STAR, HTSeq and EdgeR), blue (P2 with STAR, RSEM and edgeR), red (Sailfish and edgeR) and purple (TopHat2, Cufflinks and CuffDiff).

https://doi.org/10.1371/journal.pone.0197162.g001

We evaluated four RNA-Seq pipelines (P1 –P4) based on different analysis steps: aligning (section 'Performance of Alignment tools'), then we cross-compared these pipeline and microarray results (MA) based on correlation of expression levels (section 'Gene-wise correlation of RNA-Seq and microarray data'), differential gene detection (section 'Results of differential gene expression') and pathway enrichment detection (section 'GO-Enrichment analysis').

Performance of alignment tools

The BL2 data was profiled by poly-A-mRNA sequencing whereas RC data by total-RNA sequencing. For investigating the mapping performance on BL2 and RC, three different aligners for read mapping, STAR, TopHat2 and Sailfish, were investigated and the results compared.

The three aligners were evaluated based on their total mapping rate (see Table 1), where the aim should be to always map as much data correctly as possible. In the BL2 datasets the

Table 1. Overview of total mapping rates over all samples in % for the different RNA-Seq aligner. Displayed are the mean mapping rates over the complete dataset with the variance in brackets.

Dataset\Tools	STAR	TopHat2	Sailfish
BL2	98.98 (±0.05)	97.02 (±0.1)	84.81 (±0.85)
RC	98.49 (±0.35)	96.73 (±0.4)	54.44 (±3.71)

https://doi.org/10.1371/journal.pone.0197162.t001

proportion of mapping rate results were close together. For the RC data, aligners that map to the genome (TopHat2, STAR) performed much better than an aligner mapping to a transcriptome, like Sailfish.

Therefore we took a closer look into the proportions of unique and multi mapping rates, which together result in the overall mapping rate (Fig 2). STAR showed an overall lower unique mapping rate (BL2 78.61%, RC 83.13%) than TopHat2 (BL2 84.81%, RC 85.4%), but got a higher total mapping rate of reads, due to a higher multimapping rate (STAR (BL2 20.37%, RC 15.36%), TopHat2 (BL2 12.78%, RC 11.32%)). For a full list of the complete mapping-performance see S1 Appendix. Depending whether or not to use multi mapped reads for later counting of features, in case of using them STAR performs slightly better than TopHat2. If only unique mapped reads are utilized, Sailfish performed the best for the BL2 dataset, which is based on poly-A mRNA sequencing. The performance of Sailfish on the RC dataset is a lot worse than for STAR or TopHat2. This is due to multiple reasons: as Sailfish is mapping against known transcripts only, its performance is based on the quality of the species reference

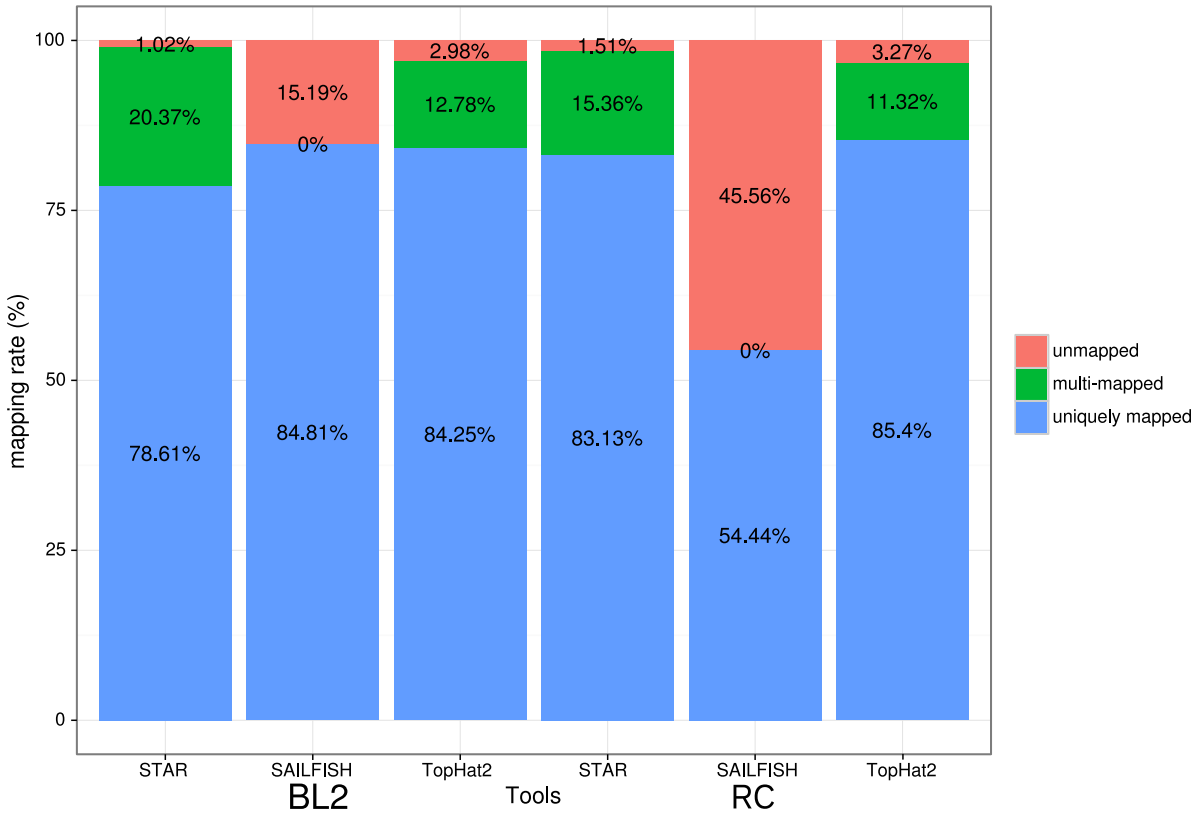

Fig 2. Mapping distribution. Mapping distribution in % for all three Aligners for both datasets. On the left the BL2 dataset and on the right the RC dataset is shown.

https://doi.org/10.1371/journal.pone.0197162.g002

transcriptome. For the BL2 dataset the unique mapping rate is 84.81%, which is better than STAR's (78.61%) and TopHat's unique mapping rate (84.25%). Sailfish allows only for perfect (unique) matched reads (kmers) and all multi mapped reads are inherently discarded during the processing by Sailfish. Variant rich data, contradicts the unique matching dogma of Sailfish, which could lead to less mapped reads overall. This phenomenon described is illustrated in the RC data by only 54.44% of mapped reads. The data is based on total-RNA sequencing, where only around 54% of the dataset is annotated when building the transcriptome, as such a large proportion of it seems still unknown and therefore cannot be mapped with Sailfish, yet.

Nevertheless, in terms of time required for aligning the data, the performance of Sailfish was the fastest. As an example: the read mapping for RC took Sailfish 6–7 minutes per dataset including the counting step, whereas STAR took 6–10 minutes per aligning sample and TopHat2 up to three hours.

Evaluation of RNA-Seq pipelines and cross-comparison with microarray

Gene-wise correlation of RNA-Seq and microarray data. We performed a gene-wise correlation analysis based on expression levels after counting. We correlated the different quantification levels after they were transformed into log2 tpm (RNA-Seq) and log2 quantile normalized expression values (microarray).

The correlation heatmaps shown in Fig 3 were done by taking the mean of each sample wise correlation test between pipeline methods. Last, they were clustered based on complete linkage with the distance 1-Pearson correlation. Overall we observed a high correlation on all performed RNA-Seq Pipeline runs together with the corresponding microarray values, which was observed similarly in other studies as well [32] [33] [34] [35] for different datasets. The correlation of microarray and RNA-Seq data is moderately worse for the RC data (ranges of 0.67 to 0.68) than for the BL2 data (0.87 to 0.88), which can be expected, since the overall biological variability of patient data is higher than in cell lines. The overall difference in correlation between microarray and RNA-Seq can be explained by their technological difference in the quantification of the gene expression. For RNA-Seq analysis the Pipeline P4 utilizing Cufflinks and CuffDiff was a big surprise since the mean correlation coefficients where quite low, even when correlating the replicate of the same method with each other. Microarray methods measure the intensities of fluorescence, which mirrors the associated gene expression, whereas RNA-Seq methods measure read counts as associated relative abundance measure for gene expression levels. Interestingly, the correlation between the different RNA-Seq tools is high (BL2: 0.97 to 0.99, RC: 0.94 to 0.98). Only a minor impact of mapping and counting approaches is observed in correlation coefficients. RSEM shows the highest correlation with the microarray data on both sets closely followed by Sailfish, HTSeq-Count and Cufflinks.

Results of differential gene expression. On the differential gene expression level all pipelines were compared based on the number of significantly differentially expressed genes (DEGs). P1, P2 and P3 were tested for differential expression with edgeR, P4 with the CuffDiff script of Cufflinks and limma was used for microarray analysis.

Overall the number of DEGs for the RNA-Seq pipelines were much higher than observed for the microarray analysis. Moreover, we evaluated overlaps of DEGs between the P1-P4 pipelines. In particular, we focused on the subsets of genes that were detected by at least two out of four pipelines ('consensus DEGs') and the subset of genes that were detected solely by only one pipeline ('DEGs unique'). In the absence of the 'ground truth' we use these measures ('consensus DEGs' and 'unique DEGs') as indicator of pipeline performance in terms of identifying potential true-positive results and false-positive results. Also these proportion are only indicators and are by no means actual true-positives and false-negatives, but more reflecting

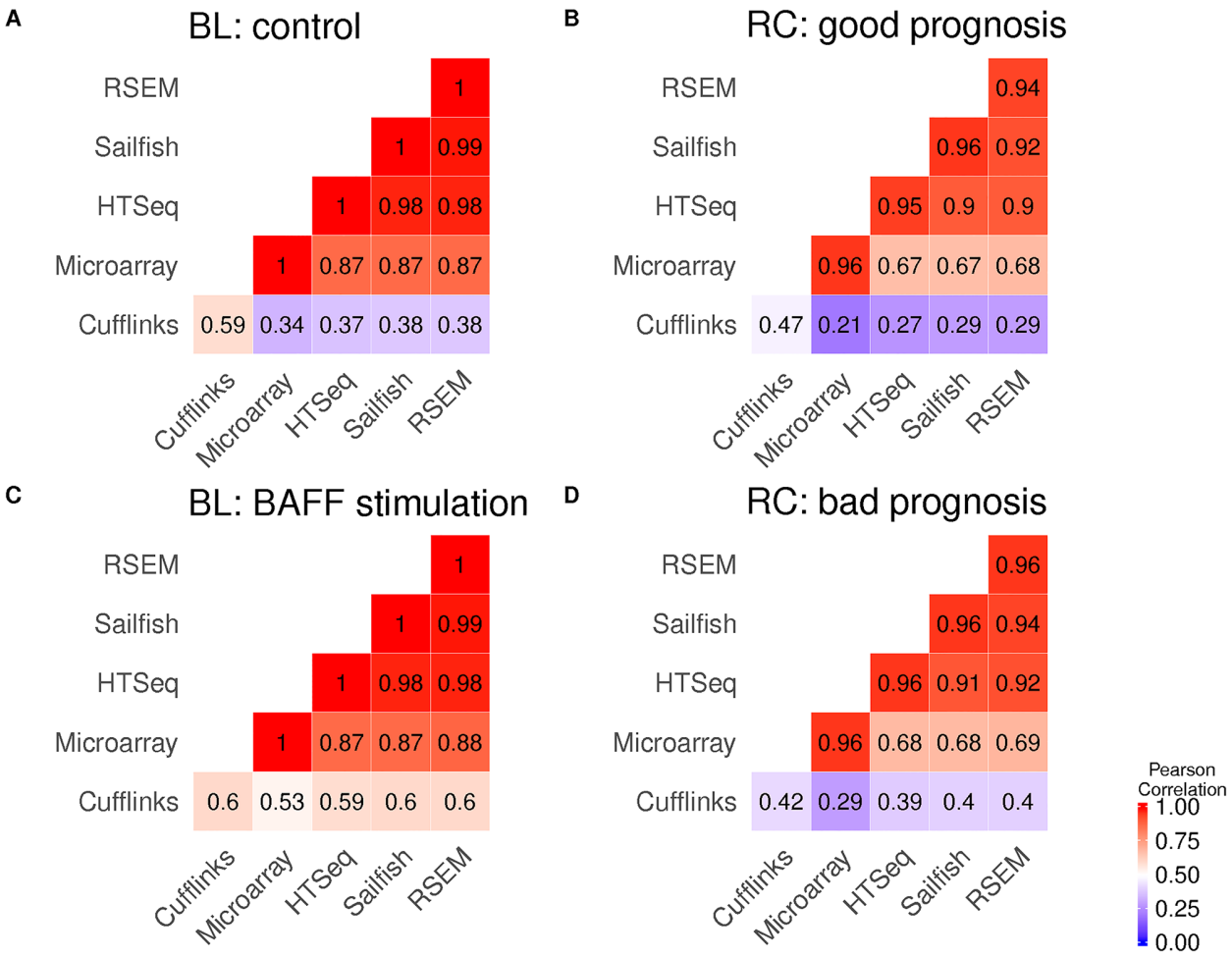

Fig 3. Correlation of all samples after analysis. The heatmap describes the combined Pearson correlation coefficient over all pairwise correlation tests of normalized gene expression against all replicates between groups. For RNA-Seq all expression values are normalized to tpm (transcript per million), to be able to compare them. Fig. 3A and 3C show the BL2 dataset and the correlation of all samples before (A) and after (C) BAFF stimulation for each analysis tool used. Fig 2B and 2D show the correlation of patient samples for the two groups with good prognosis of distant free metastases (good) and a bad prognosis (bad) together with the different analysis pipelines used.

https://doi.org/10.1371/journal.pone.0197162.g003

the general agreement with the other compared methods and the inconsistency or uniqueness of each individual method.

Microarray results. For the two investigated datasets we identified very low numbers of DEGs using the microarray platform. For BL2, out of 1196 genes significant on p-value level only 1 gene (*GPER*) remained significant after FDR correction. For the RC dataset, out of 1285 genes significant on p-value level no gene remained significant after FDR correction (S2b Fig). We checked these genes with the results of differential expression analysis for microarray reported in Schrader et al. [18], which did a similar study with nearly the same condition. We could reidentify 26 genes based on the gene symbols in common (see supplement S2 Table). A reason for this small number of overlapping genes could be attributed to the difference in power of the analysis and experimental conditions. Our BL2 dataset was newly resequenced, incubated for 24h instead of 9h and the microarray chip used was the HG-U133_Plus_2 chip instead of U133 plus 2.0.

Table 2. Overview of the number genes and GO-terms significant (p-value <5%) and after FDR correction for P1-4 and microarray. The GO-terms for microarray are in bold, because the p-value was used as a cutoff instead of the FDR.

Pipelines	Number of DEGs (p-value)		Number of DEGs (FDR)		Number of sign. GO Terms	
	BL2	RC	BL2	RC	BL2	RC
P1(HTSeq)	2299	3377	287	71	138	127
P2(RSEM)	2329	3646	340	96	131	111
P3(Sail)	2410	1285	375	146	158	127
P4(Cuff)	316	1398	20	154	89	96
Microarray	1196	1289	**1**	**0**	**116**	**148**

https://doi.org/10.1371/journal.pone.0197162.t002

RNA-Seq Pipeline comparison on BL2. P1 to P3 found the highest amount of DEGs (287, 340, 375) after FDR correction. For P4, only 20 genes were left and for the microarray results one gene was left significant after p-value adjustment (Table 2). 29 of these FDR corrected genes could also be reidentified from a former study from Schrader et al. [18] and are commonly shared by the pipelines of P1, P2 and P3 (see supplement S2 Table). When adding P4, the number reduced to 4.

Next, we identified the different overlaps of DEGs for the individual pipelines (P1-4 in Fig 4 BL2). Since we don't know the true calls for the dataset we utilized 'consensus DEGs' and 'unique DEGs' as surrogate measures supporting the interpretation of the overlaps.

Considering all four RNA-Seq pipelines a total of 9 genes can be found in common by all pipelines for BL2. P4(Cuff) pipeline detected 20 DEGs from which 9 were found also by the other pipelines. These genes are: *TSPAN11*, *PFKFB4*, *SGK1*, *CCR7*, *NFKBIE*, *CCDC28B*, *HLA-DQA1*, *HLA-DRB5*, *HLA-DQA2*.

For evaluation purposes, we consider the genes found by the majority of tools as promising candidates for true findings. Therefore, we are looking at the overall agreement between the tools in form of overlaps of genes in common by at least 2 other pipelines (Table 3) as a measurement of potentially true findings for the tools. It can be seen that P1(HTSeq) has the largest number of significant genes also found by others (211/287), which is 73.52% of their complete findings. P1 got the lowest percentage of genes found unique (19.16%), which translates to 55 out of 287 genes. Genes not found by other pipelines can either way be interpreted as false calls or as simply missed by the other pipelines or most likely a mixture of both. Since finding false

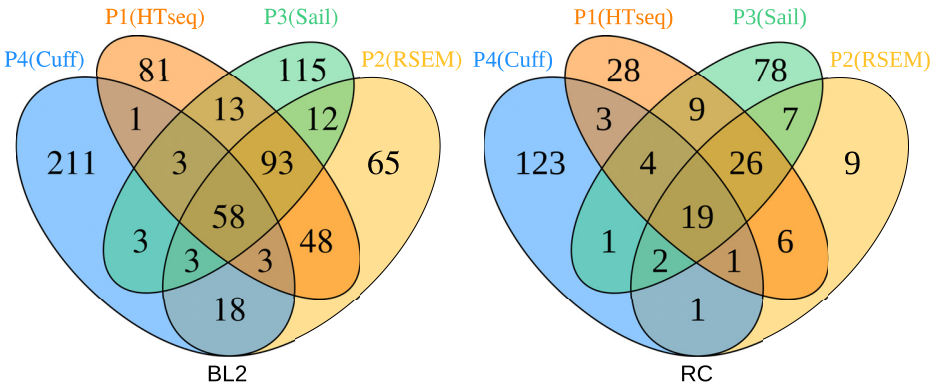

Fig 4. Significant overlapping genes for the different strategies after multiple test adjustment. Shown are two venn diagrams, one for each dataset (BL2 Fig. 4A and RC Fig. 4B). The different pipelines used here are: TopHat2 and Cufflinks (T&C), STAR and HTSeq-Count (S&HT), Sailfish (Sa), STAR and RSEM (S&R). The microarray data is not included, because there were close to no significant genes after FDR adjustment.

https://doi.org/10.1371/journal.pone.0197162.g004

Table 3. Overview of the proportion of genes and corresponding percentage of differential expressed genes for each pipeline after multiple testing adjustment. 'consensus' stands for the amount of genes shared with at least two other pipelines and 'unique' for genes not found by any other Pipeline from the total amount of genes found by each Pipeline.

Pipelines	Consensus DEGs				DEGs unique			
	BL2		RC		BL2		RC	
P1(HTSeq)	73.52%	(211/287)	67.60%	(48/71)	19.16%	(55/287)	12.68%	(9/71)
P2(RSEM)	49.70%	(169/340)	52.08%	(50/96)	29.41%	(100/340)	29.17%	(28/96)
P3(Sail)	52.80%	(198/375)	34.93%	(51/146)	41.60%	(156/375)	53.42%	(78/146)
P4(Cuff)	45.00%	(9/20)	16.88%	(26/154)	55.00%	(11/20)	79.87%	(123/154)

https://doi.org/10.1371/journal.pone.0197162.t003

calls in general is not desired, we tend to consider a low amount of unique genes found as positive. Following this interpretation, the quality of the pipelines in their outcome for the BL2 dataset can be ordered as follows: P1(HTSeq), P2(RSEM), P3(Sail), P4(Cuff).

RNA-Seq pipeline comparison on RC. The ordering of the highest amount of significant genes after FDR correction flipped in this dataset for P1-4. This time P4 found the largest number of significant genes (154), followed by P3(146), P2(96) and P1(71). P1 has the highest percentage of consensus DEGs 67.60% (48/71), followed by P2 52.08% (50/96), P3 34.93% (51/146) and P4 16.88% (26/154) (Table 3). A total of 19 genes (Fig 4 RC) can be found in common, namely: *EYA1*, *NPR3*, *MUC5B*, *RSAD2*, *IGF2BP3*, *ITGA11*, *IFI44L*, *IFI44*, *ASZ1*, *MX1*, *CTHRC1*, *FAM3B*, *POU5F1B*, *COL11A1*, *C1QC*, *SLC35D3*, *ZFHX4*, *MMP11*, *ANO1*. P4 has the highest number of DEGs, but only 26 of them found by others, whereas a total of 79.87% (123/154) of the genes cannot be found by any of the other pipelines.

This is highest rate of unique genes found, as well as the lowest rate of Consensus DEGs consistent in both datasets, despite here having the most genes found. Notably, Cufflinks is coupled with CuffDiff for the differential expression analysis, so the results of Cufflinks are as well influenced by differences in the statistical analysis. Overall P4 is the most divergent from all others, whereas P1 provides the results most concordant with the other pipelines, followed by P2.

GO-Enrichment analysis. To evaluate the results from BL2 and RC datasets after differential gene expression analysis enriched GO-terms were investigated. Microarray results showed close to no significant genes after FDR correction. To nevertheless generate GO-term enrichment results for microarray based datasets, their significance threshold for the enrichment test was altered to <5% of the p-value instead of <5% FDR value (Table 2). Fig 5 shows the top 20 significant GO-terms from the different pipelines and microarray datasets. Hierarchical clustering of all GO-terms was applied to investigate the similarity of the different pipelines based on the enrichment scores. The complete set of significant pathways is depicted in S3a Fig for dataset BL2 and in S3b Fig for dataset RC.

BL2 dataset. In the context of comparing control Burkitt Lymphoma cell-line with the BAFF stimulated BL2 cells, it is to be expected to detect GO-terms related to the immune response as significantly enriched [36].

We checked whether we find this biological context in the top 20 significantly enriched GO -terms. In Fig 5A we can observe four highly enriched GO-terms related to immune response (GO:0060333, GO:0019886, GO:0050852, GO:0031295), primarily for pipelines P1-P4. In total we see 13 out of the 20 GO-terms linked to the immune system. In addition, four of the depicted GO-terms are related to metabolism, two to cell signaling and the last one to biological regulation. The enriched terms based on the microarray data leads to only one GO-term being highly significant, while 7 were not detected to be significantly enriched at all. Fig 5a shows that P1 to P3 perform similarly in terms of additional enrichment analysis.

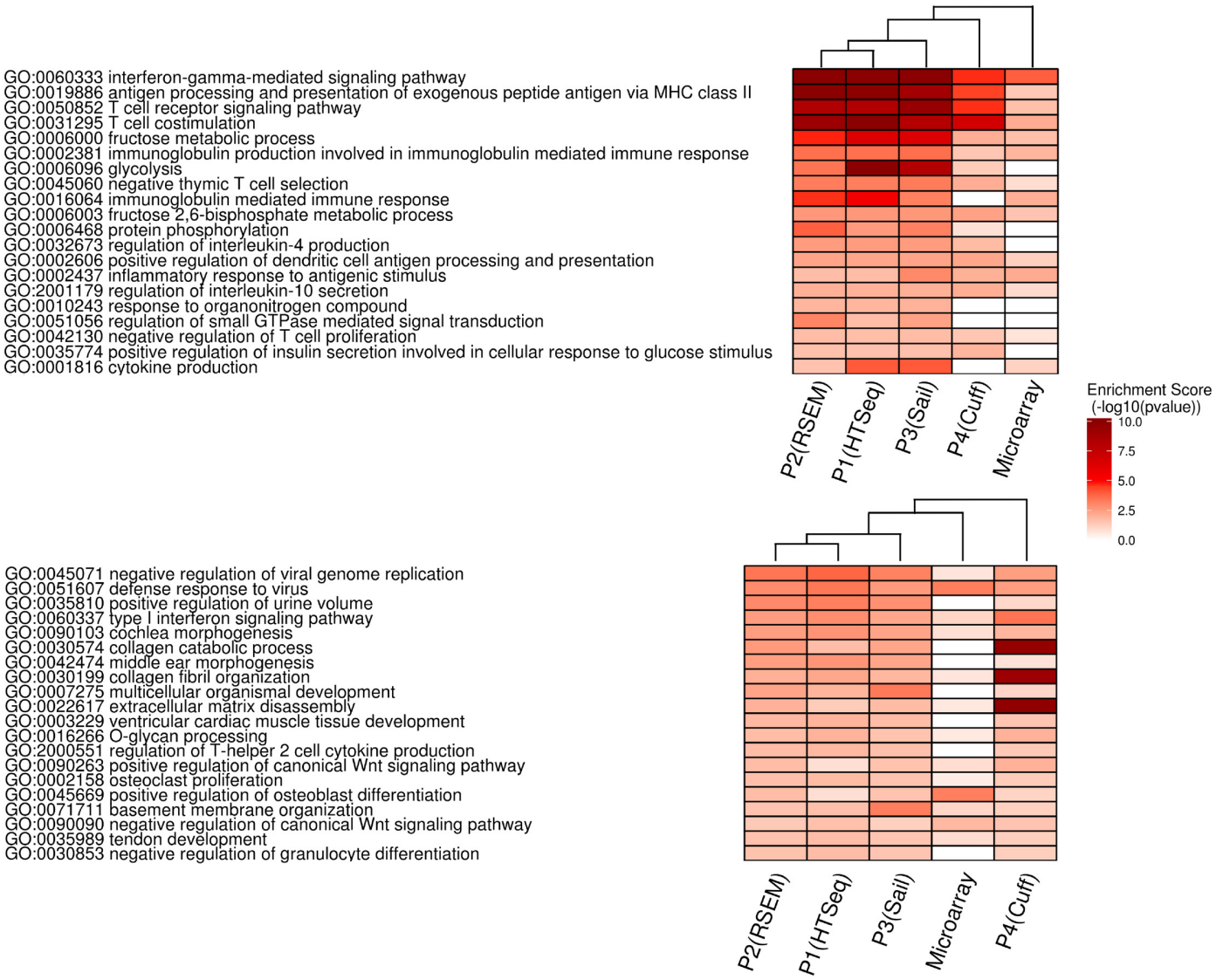

Fig 5. Top20 significant enriched GO-categories for BL2 and RC. In all shown RNA-Seq and microarray strategies the visualized GO-categories were enriched (p-value smaller than five percent and the pathway was bigger than four genes). The enrichment of GO-terms is shown in red: the higher the intensity of red, the lower the p-value. For better scalability of colors the negative log 10 was chosen. The pathways agreeing the most amongst all pipelines are shown at the top.

https://doi.org/10.1371/journal.pone.0197162.g005

However Pipeline P4 and the microarray datasets show highly different enrichment scores. In summary, the deregulated GO-terms associated to immune response are identified by the majority of the pipelines.

RC dataset. When comparing GO-term analysis results of the rectal cancer patient group, we expect to see GO-terms related to metastases formation, like increased proliferation, cell rearrangements, changes in cell organization and to a certain extent immune response as well. Therefore we checked again if we observe several of these assumptions within the top 20 significantly enriched GO-terms. In these we observed GO-terms (GO:0090263, GO:0002158, GO:0090090) linked to cellular proliferation, GO-terms (GO:0030199, GO:0022617, GO:0071711) linked to cellular rearrangements and GO-terms (GO:0060337, GO:2000551, GO:0030853) related to immune system response. However a lot of significant GO-terms

could not be related to any of the expected cellular response classes. As previously discussed, these results might be due to sequencing of total-RNA in case of the RC dataset in comparison to polyA-mRNA sequencing. In addition, the biological variability of human patient data is much higher in contrast to cell line data. Based on the data processed by P4 three GO-terms were identified to be highly enriched (GO:0030574, GO:0030199, GO:0022617). These are linked to (extra-)cellular rearrangements and fit well into our expectations, however could only be detected with such high enrichment score based on P4. In comparison, P1-P3 look very similar as the BL2 dataset.

Conclusion

This study presents a comparison of RNA-Seq specific pipelines as well as a cross comparison with matched microarray data. For the investigated realistic datasets microarray analysis was inferior to the used RNA-Seq analysis strategies and only a minor proportion of DEGs already reported by Schrader et al. could reproduced. Pipelines P1 to P3 performed rather similar when looking at the correlation results, with a small lead in regard to utilization of raw data for P3. In contrast, P1 outperformed the rest in terms of the highest agreement with the other pipelines in the detection of differentially expressed genes. Results from P4 varied a lot, presumably due to the use of the internal Cufflinks statistics in the tool suite.

Supporting information

S1 File. Additional quality control metric for the BL2 and RC data sets.
(GZ)

S1 Appendix. Mean readmapping-rates and their standard-deviation.
(DOC)

S2 Appendix. R-functions for converting count and fpkm values into TPM.
(DOC)

S1 Table. Summary of RNA-Seq data.
(DOC)

S2 Table. Overview of genes shared in common between our results and DEG results reported by Schrader et al.
(DOC)

S1 Fig. Overview of qualities metrics for the BL2 (A) and RC (B) RNA-Seq data set.
(PDF)

S2 Fig. Overview of the amount of FDR corrected significant genes overlapping each pipeline and microarray for the BL2 (A) and RC (B) dataset.
(PDF)

S3 Fig. The complete list of GO-Terms for all 5 pipelines for the BL2 (A) and RC (B) dataset.
(PDF)

Acknowledgments

The Transcriptome and Genome Analysis Laboratory (TAL) for RNA sequencing and microarray data creation. Astrid Wachter and Júlia Perera-Bel for plenty fruitful discussions.

Author Contributions

Conceptualization: Alexander Wolff, Michaela Bayerlová.

Data curation: Alexander Wolff.

Formal analysis: Alexander Wolff, Michaela Bayerlová.

Funding acquisition: Tim Beißbarth.

Investigation: Alexander Wolff.

Methodology: Alexander Wolff.

Project administration: Alexander Wolff, Tim Beißbarth.

Resources: Jochen Gaedcke, Dieter Kube, Tim Beißbarth.

Software: Alexander Wolff, Michaela Bayerlová.

Supervision: Tim Beißbarth.

Visualization: Alexander Wolff.

Writing – original draft: Alexander Wolff, Michaela Bayerlová, Tim Beißbarth.

Writing – review & editing: Alexander Wolff, Michaela Bayerlová, Jochen Gaedcke, Dieter Kube, Tim Beißbarth.

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57–63. https://doi.org/10.1038/nrg2484 PMID: 19015660
2. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011; 12: 87–98. https://doi.org/10.1038/nrg2934 PMID: 21191423
3. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet. 2014; 15: 423–437. https://doi.org/10.1038/nrg3722 PMID: 24776770
4. Goya R, Sun MGF, Morin RD, Leung G, Ha G, Wiegand KC, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 2010; 26: 730–736. https://doi.org/10.1093/bioinformatics/btq040 PMID: 20130035
5. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, et al. Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. Futscher BW, editor. PLoS ONE. 2013; 8: e58815. https://doi.org/10.1371/journal.pone.0058815 PMID: 23555596
6. Tang X, Baheti S, Shameer K, Thompson KJ, Wills Q, Niu N, et al. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. Nucleic Acids Res. 2014; 42: e172–e172. https://doi.org/10.1093/nar/gku1005 PMID: 25352556
7. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010; 11: 31–46. https://doi.org/10.1038/nrg2626 PMID: 19997069
8. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010; 11: 1.
9. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2015; 43: D1113–D1116. https://doi.org/10.1093/nar/gku1057 PMID: 25361974
10. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013; 41: D991–D995. https://doi.org/10.1093/nar/gks1193 PMID: 23193258
11. Su Z, Łabaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014; 32: 903–914. https://doi.org/10.1038/nbt.2957 PMID: 25150838

12. Xu W, Seok J, Mindrinos MN, Schweitzer AC, Jiang H, Wilhelmy J, et al. Human transcriptome array for high-throughput clinical studies. Proc Natl Acad Sci. 2011; 108: 3707–3712. https://doi.org/10.1073/pnas.1019753108 PMID: 21317363

13. Rehrauer H, Opitz L, Tan G, Sieverling L, Schlapbach R. Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching. BMC Bioinformatics. 2013; 14: 1.

14. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. Zhang S-D, editor. PLoS ONE. 2014; 9: e78644. https://doi.org/10.1371/journal.pone.0078644 PMID: 24454679

15. Shi L, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006; 24: 1151–1161. https://doi.org/10.1038/nbt1239 PMID: 16964229

16. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010; 28: 827–838. https://doi.org/10.1038/nbt.1665 PMID: 20676074

17. Vockerodt M, Pinkert D, Smola-Hess S, Michels A, Ransohoff RM, Tesch H, et al. The Epstein-Barr virus oncoprotein latent membrane protein 1 induces expression of the chemokine IP-10: Importance of mRNA half-life regulation. Int J Cancer. 2005; 114: 598–605. https://doi.org/10.1002/ijc.20759 PMID: 15578697

18. Schrader A, Meyer K, von Bonin F, Vockerodt M, Walther N, Hand E, et al. Global gene expression changes of in vitro stimulated human transformed germinal centre B cells as surrogate for oncogenic pathway activation in individual aggressive B cell lymphomas. Cell Commun Signal. 2012; 10: 43. https://doi.org/10.1186/1478-811X-10-43 PMID: 23253402

19. R Development Core Team. R: A Language and Environment for Statistical Computing. 2008; http://www.R-project.org

20. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4: 249–264. https://doi.org/10.1093/biostatistics/4.2.249 PMID: 12925520

21. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Stat Appl Genet Mol Biol. 2004; 3: 1–25. https://doi.org/10.2202/1544-6115.1027 PMID: 16646809

22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995; 289–300.

23. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

24. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14: R36. https://doi.org/10.1186/gb-2013-14-4-r36 PMID: 23618408

25. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014; 32: 462–464. https://doi.org/10.1038/nbt.2862 PMID: 24752080

26. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31: 166–169. https://doi.org/10.1093/bioinformatics/btu638 PMID: 25260700

27. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12: 323. https://doi.org/10.1186/1471-2105-12-323 PMID: 21816040

28. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28: 511–515. https://doi.org/10.1038/nbt.1621 PMID: 20436464

29. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

30. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7: 562–578. https://doi.org/10.1038/nprot.2012.016 PMID: 22383036

31. Alexa A, Rahnenfuhrer J. topGO: topGO: Enrichment analysis for Gene Ontology. 2010.

32. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18: 1509–1517. https://doi.org/10.1101/gr.079558.108 PMID: 18550803

33. Bradford JR, Hey Y, Yates T, Li Y, Pepper SD, Miller CJ. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC Genomics. 2010; 11: 282. https://doi.org/10.1186/1471-2164-11-282 PMID: 20444259

34. 't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. 2008; 36: e141–e141. https://doi.org/10.1093/nar/gkn705 PMID: 18927111

35. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y. Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. Provero P, editor. PLoS ONE. 2013; 8: e71462. https://doi.org/10.1371/journal.pone.0071462 PMID: 23977046

36. Mackay F, Browning JL. BAFF: A fundamental survival factor for B cells. Nat Rev Immunol. 2002; 2: 465–475. https://doi.org/10.1038/nri844 PMID: 12094221

3.2 Using RNA-Seq data for the detection of a panel of clinically relevant mutations

Reference

A Wolff, J Perera-Bel, HU Schildhaus, K Homayounfar, B Schatlo, A Bleckmann, T Beißbarth: Using RNA-Seq data for the detection of a panel of clinically relevant mutations. Studies in Health Technology and Informatics 2018, DOI: 10.3233/978-1-61499-896-9-217

3.2.1 Summary and discussion

RNA-Seq became the most popular technology to detect the expression of genes and isoforms, and members of the MetastaSys consortium raised the question if it is possible to detect mutations reliably on RNA-Seq data, on top of DEA, resulting in a sub-project and this publication. After the establishment of a standard pipeline (see 3.1) the next step was to answer the question if RNA-Seq has the potential for a cost-effective identification of single nucleotide variants (SNVs). For that reason, Wileup a tool written in Perl to call SNVs in RNA-Seq for a tumour only sample set is presented. In this publication, the focus was on a panel of 442 SNVs with high clinical interest. Furthermore, this was evaluated on a matched dataset from RNA-Seq and DNA-Seq of 14 patients. SNVs detected in RNA-Seq data by Wileup were compared to the clinical standard of calling mutations from tumour samples and subtracting mutations from a blacklist found in normal samples of the same patient. By doing that it is possible to make a distinction between somatic mutations (tumour tissue-specific) and germline mutations (occur in every cell, therefore the normal samples) without using prior knowledge. This experimental setup resulted in 42 samples, three samples for each of the 14 patients (one DNA tumour tissue sample, one DNA blood sample as a reference called normal and one RNA tumour sample). The last method to be compared was detecting mutations in DNA-Seq without the support from samples of normals. The complete experimental setup is illustrated in figure 3.1.

WES and RNA-Seq reads were quality assessed using fastqc. Amongst the 14 patients where seven patients with brain metastases and seven patients with liver metastases. All WES brain samples showed high duplication levels and a drop in quality at the end of the reads due to high levels of contamination with Nextera adapters. Hence, TrimGalore-0.4.3 (`https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`) was applied to all brain samples. Between 11% and 38.4% of base pairs were trimmed.

For the analysis of the DNA-Seq data the WES paired-end reads were aligned against the reference genome of Homo sapiens version GRCh38.91 with bowtie2 (version 2.2.3). Samtools was used to create bam files, and Picard (version 2.0.1) to mark duplicates. Then, GATK (v3.8.0) best practices were followed to perform read realignment (IndelRealigner) and base recalibration (BaseRecalibrator). RNA-Seq single-end reads were aligned against the reference genome of Homo sapiens Ensembl Version GRCh38.91 with the splice-aware aligner

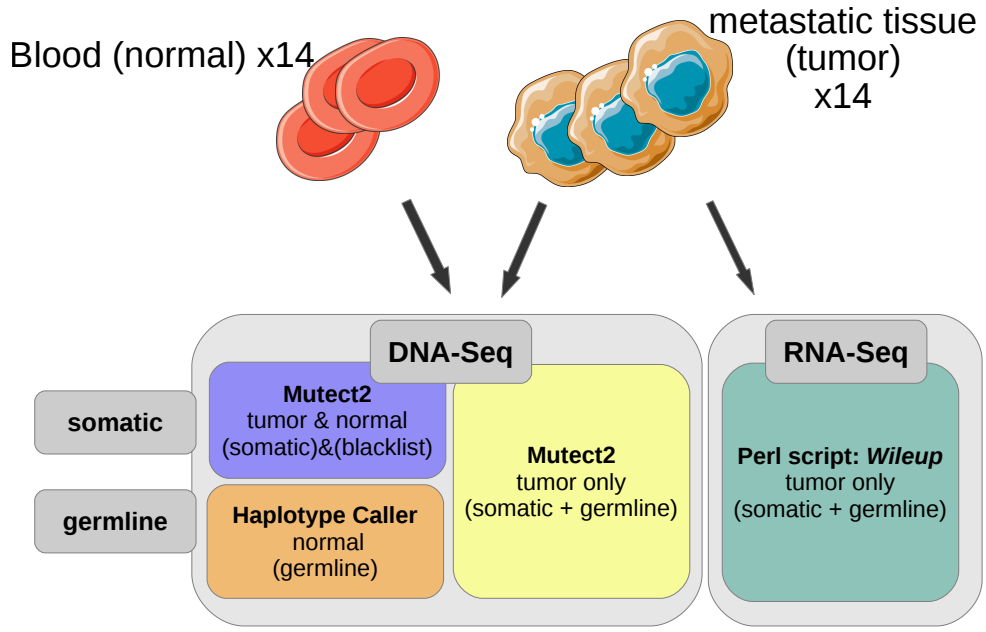

Figure 3.1. *Illustration of the collected samples of 14 patients as input for the DNA-Seq and RNA-Seq based analysis. The blood samples and metastatic tissue samples were exome sequenced and used for somatic and germline mutation calling, using either Mutect2 (somatic caller) or the Haplotype Caller of GATK (germline caller). Furthermore, the metastatic tissue was also RNA sequenced and used for the mutation calling via Wileup, resulting in a total of 42 samples sequenced.*

STAR (v2.5.2b). Picard (version 2.0.1) was used to remove duplicates. Mutect2 (GATK v3.8.0, beta version) to detect somatic variants in WES data using matched tumour-normal samples (referred to as "clinical standard") as well as only tumour samples ("tumour-only" mode). Cosmic (version 83, Coding and Non Coding vcf files) and dbSNP (version 138) were provided as input to Mutect2 to adjust the threshold for evidence of a variant in the normal sample. To confirm germline mutations detected by tumour-only samples and variants found only in the RNA-Seq samples, GATK Haplotypecaller (v3.8.0) with dbSNP (version 138) was used.

For the analysis of RNA-Seq data Wileup (see figure 3.2) was used. After Quality Control, alignment with STAR (v2.5.2b) and Duplicate removal Two modes are available for Wileup: 1) a complete mode, for this the mpileup function of Samtools is used to derive all nucleotide bases at each position from the bam files in the RNA-Seq data, saved as mpileup format. Then the distribution of bases in each position from the mpileup format is parsed, saved and annotated with also parsed database information of CiVIC, ClinVar and Cosmic. Resulting in a filtered output file of all annotated mutations found in at least one of the databases. 2) Panel mode, in which a list of 442 nucleotide positions is selected. They are predictive (only drug responses) SNVs with implications for cancer therapy from both CGI and CIViC, comprising somatic, germline and germline polymorphisms. Here the mpileup is done only for these positions resulting in a speedup from 2 hours to roughly 15 minutes per bam file. A minimum of three reads or 10% of the reads supporting the alternative variant is used

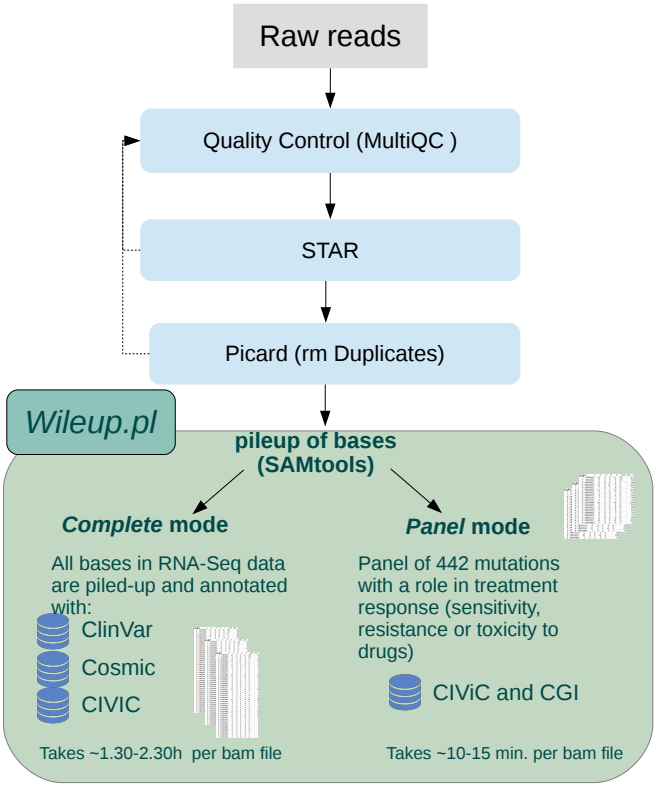

Figure 3.2. *Workflow for detecting mutations in RNA-Seq using Wileup.*

as default threshold. The output comprises the distribution of bases at each position, the decision whether the position contains an SNV, and the clinical annotations from CIViC and CGI. As shown in figure 3.3A 109 SNVs were found in all samples by the three methods: 10 in the gold standard, 104 in tumor-only and 73 in the RNA-Seq. The 10 SNVs detected by the clinical standard were also detected in WES tumor-only and RNA-Seq. As the primary purpose of the clinical standard is to find somatic mutations only, it differs a little bit from Wileup, as Wileup is also annotating germline and germline polymorphism with clinical relevance on top. The highest overlap was found between WES tumor-only and RNA-Seq (68 SNVs). The 36 variants unique to the WES tumor-only analysis were undetected in RNA-Seq due to low expression. The 5 variants unique to RNA-Seq were undetected in WES due to high duplication levels but could be validated by using the Haplotype Caller from GATK on the normal samples of the individual patients. Seven out of the eight validated mutations from pathology (bottom color bar of the heatmap in figure 3.3B) could be reliably identified by all three methods (shown as validated in dark green and dark violet in the heatmap). The missing mutation E545K of PIK3CA could not be identified by any of the methods, due to high duplication levels for WES and missing expression in RNA-Seq. Nonetheless, all somatic SNVs detected by the clinical standard were also detected by tumour-only and RNA-Seq (figure 3.3B).

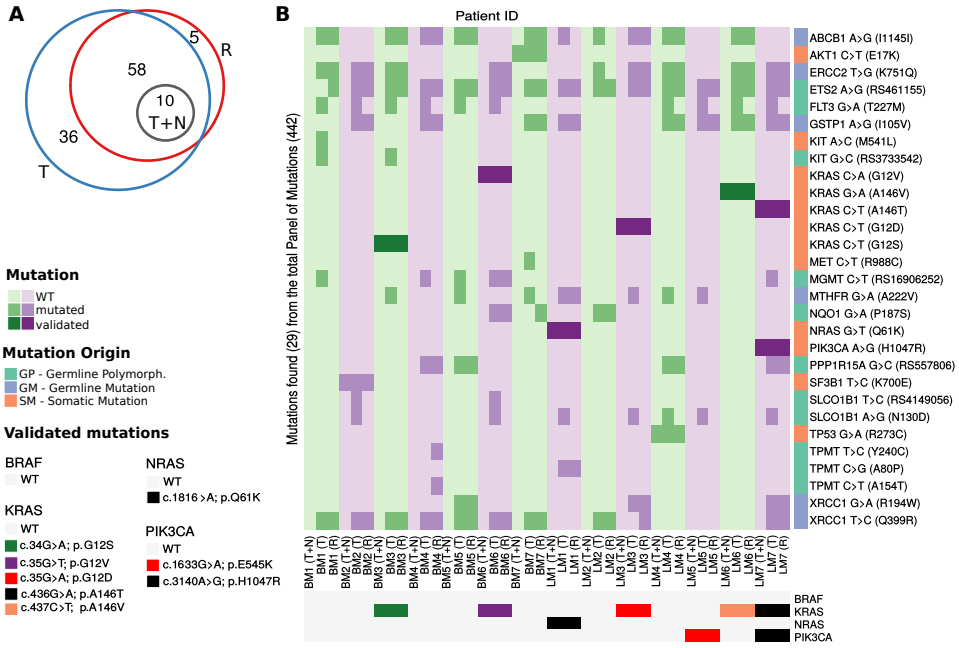

Figure 3.3. *A) Venn diagram depicting the number of SNVs identified by each method across all samples (T+N: Mutect2, Tumor+Normal samples, T: Mutect2, Tumor samples, R: Wileup RNA-Seq). In total 109 SNVs were found (29 unique mutations from the panel of 442 clinical relevant mutations). B) Heatmap visualisation of 29 unique SNVs which were found by at least one of the methods in any of the 14 patients. Wild Type (WT) mutations are shown light green and purple, mutations found by the methods are in green and purple, mutation agreeing with the pathological annotation (validated) are marked in dark green and purple. The details of the pathological mutations are described in the annotation bars at the bottom of the figure. The origin of the mutation is annotated in the bar at the right sight of the heatmap.*

In conclusion, RNA-Seq appears to be a reliable approach for detecting the selected panel of SNVs with clinical relevance, as confirmed by the pathologically validated data (for somatic variants) and by the analysis of normal WES samples (for germline variants). A high overlap between RNA-Seq and tumour-only WES was shown. Previous studies reported high numbers of false positives in RNA-Seq data, however, by using a whitelist of well-defined SNVs this problem could be avoided. In this setting, detecting SNVs in RNA-Seq data is a comparable approach to WES tumor-only; yet, in RNA-Seq, it is regarded as an additional analysis next to DEA, which can be quickly performed (average of 11-15 min/sample) for no extra biological sample cost. Of course, the user has to accept false negatives in non expressed genes, but that is inherent to RNA-Seq data. For future implementations, it would be essential to consider RNA editing processes as well as including indels in the analysis.

3.2.2 Declaration of my contribution

Conceptualization: Alexander Wolff, Júlia Perera-Bel, Tim Beißbarth.
Data curation: Alexander Wolff, Júlia Perera-Bel, Tim Beißbarth
Formal analysis: Alexander Wolff, Júlia Perera-Bel
Investigation: Alexander Wolff, Júlia Perera-Bel

Methodology: Alexander Wolff.
Project administration: Alexander Wolff, Júlia Perera-Bel, Tim Beißbarth.
Resources: Hans-Ulrich Schildhaus, Kia Homayounfar, Bawarjan Schatlo, Annalen Bleckmann, Tim Beißbarth
Software: Alexander Wolff.
Visualization: Alexander Wolff, Julia Perera
Writing– original draft: Alexander Wolff, Julia Perera.
Writing – review & editing: Alexander Wolff, Júlia Perera-Bel, Hans-Ulrich Schildhaus, Kia Homayounfar, Bawarjan Schatlo, Annalen Bleckmann, Tim Beißbarth

Using RNA-Seq Data for the Detection of a Panel of Clinically Relevant Mutations

Alexander WOLFF[a], Júlia PERERA-BEL[a], Hans-Ulrich SCHILDHAUS[c], Kia HOMAYOUNFAR[d], Bawarjan SCHATLO[e], Annalen BLECKMANN[a,b] and Tim BEISSBARTH[a,1#]

[a]*Department of Medical Statistics, University Medical Center Göttingen*
[b]*Department of Hematology and Oncology, University Medical Center Göttingen*
[c]*Department of Pathology, University Medical Center Göttingen*
[d] *Department of General, Visceral and Pediatric Surgery, University Medical Center Göttingen*
[e] *Department of Neurosurgery, University Medical Center Göttingen*

Abstract. Somatic single nucleotide variants (SNVs) are genomic events with increasing implications in cancer treatment. The clinical standard for SNVs detection is whole genome/exome sequencing (WGS/WES) in matched tumor-normal samples. Yet, this is a very costly approach both economically and biologically and very often only tumor samples are sequenced. On the other hand, RNA sequencing (RNA-Seq) is the most popular technology to study gene expression, and has also the potential for a cost-effective identification of SNVs as an alternative to tumor-only WES. Here we present a method for the identification of SNVs in tumor-only RNA-Seq data putting a special focus on a small panel of clinically relevant SNVs. For evaluation purposes, we analyzed matched tumor-normal WES and tumor-only RNA-Seq data from 14 cancer patients. We compared SNVs detected in i) RNA-Seq by our method, ii) WES tumor-only by Mutect2 and iii) WES matched tumor-normal by Mutect2. We did a detailed evaluation for a reduced panel of clinically relevant SNVs and reliably identified in RNA-Seq data a subset of mutations for which we had pathological annotation. Hence, RNA-Seq rises as a cost-effective option to detect in parallel gene expression as well as a small panel of clinically relevant SNVs in research.

Keywords: RNA-Seq, SNVs, Mutect2, variant calling, GATK

1. Introduction

Somatic single nucleotide variants (SNVs) are genomic events known to drive cancer. Whole genome and exome sequencing (WGS, WES) in matched tumor-normal samples are the clinical standard for detecting somatic SNVs. There are many tools for identifying SNVs on WGS or WES data, thoroughly compared in different contexts [1-3]. According to these studies, two tools outperform the rest: Mutect [4] and Varscan2 [5]. The first performs better at identifying SNVs with low allele frequencies, whereas the latter detects the highest number of SNVs and outperforms any tool at positions with high coverage. On the other hand, RNA sequencing (RNA-Seq) has become the

1 Corresponding Author, Tim Beissbarth, University Medical Center Göttingen, Humboldtallee 32 D-37073 Göttingen, Tim.Beissbarth@ams.med.uni-goettingen.de

most popular technology -after replacing microarrays- to study gene expression. Unlike microarrays, RNA-Seq can easily be used to detect alternative splicing, RNA editing, fusion genes, other RNA species, and, potentially, SNVs. Calling somatic SNVs in RNA-Seq data has been done in some studies by applying tools specific for WES/WGS data [6-8]. Besides obvious false negatives produced in regions with low or no expression, these studies reported false positive SNV calls in RNA-Seq data mainly due to: PCR cycle bias, strand bias, RNA editing and difficulty to align the transcriptome to the reference genome due to splicing. Sheng and colleagues tried to address some of these issues both on DNA and RNA-Seq [9]. An added problem is the fact that clinical samples are usually limited to tumor-only profiling. Detection of somatic SNVs in WES tumor-only samples is challenging and has been addressed with machine learning approaches [10] or the use of whitelists and blacklists as in Mutect [4]. However, the same has not yet been attempted for RNA-Seq data. All in all, its cheaper cost compared to WES/WGS together with all its possible applications makes RNA-Seq a technology with high interest for clinical use (e.g. parallel detection of SNVs and functional activation of genes). It seems worthwhile developing a method to call SNVs in RNA-Seq data optimized for a panel of well-known SNVs.In this study we present a method to call SNVs in RNA-Seq tumor-only samples. We assess its performance putting special focus on optimizing the method for a panel of known SNVs with high clinical interest. We compare our method's performance on a matched dataset comprising RNA-Seq and WES data. We chose Mutect2 to detect SNVs in WES data.We compare the SNVs detected in RNA-Seq data by our method to tumor-only and tumor-normal results by Mutect2.

2. Materials and Methods

2.1. *Databases*

The panel of known SNVs with high clinical interest was based on the Clinical Interpretation of Variants in Cancer (CIViC, version from 01/06/2017) [11] and Cancer Genome Interpreter (CGI, downloaded on 11/09/2017, last updated 02/08/2017) [12]. In both cases, we filtered for SNVs *predictive* of drug response. Genomic coordinates were transformed from hg19 to hg38 built using the *rtracklayer* R package. Both databases were merged by aggregating duplicate entries. The panel of actionable variants contains information on 442 variants in 92 genes.

2.2. *Collection of Patient Samples*

Tissue samples were collected by the surgery departments of the University Medical Center Göttingen. The collected tissues are from seven metastatic brain and seven metastatic liver tumours with origin from either colorectal or breast cancer. Fresh frozen tissue samples for WES and RNA-Seq were separated. From these 14 patients we collected EDTA-blood for WES as well. EDTA-blood samples served as control

samples to differentiate between germline and somatic mutations. In total 42 samples were sequenced, 3 samples per patient. The study is approved by the Ethics Committee of the University Medical Centre Göttingen, application number 21/3/11 and 14/10/05.

2.3. *Data Preprocessing and Analysis*

WES and RNA-Seq reads were quality assessed using fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All WES brain samples showed high duplication levels and a drop in quality at the end of the reads due to high levels of contamination with nextera adapters. Hence, TrimGalore-0.4.3 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was applied to all brain samples. Between 11% and 38.4% of base pairs were trimmed. WES paired-end reads were aligned against the reference genome of Homo sapiens version GRCh38 with bowtie2 (version 2.2.3). Samtools was used to create bam files, and Picard (version 2.0.1) to mark duplicates. Then, GATK (v3.8.0) best practices were followed to perform read realignment (IndelRealigner) and base recalibration (BaseRecalibrator). RNA-Seq single-end reads were aligned against the reference genome of Homo sapiens Ensembl Version GRCh38.91 with the splice-aware aligner STAR (v2.5.2b). Picard (version 2.0.1) was used to remove duplicates.

We used Mutect2 (GATK v3.8.0, beta version) to detect somatic variants in WES data using matched tumor-normal samples (referred to as “clinical standard” in the text) as well as only tumor samples (“tumor-only” mode). Cosmic (version 83, Coding and Non Coding vcf files) and dbSNP (version 138) were provided as input to Mutect2 to adjust the threshold for evidence of a variant in the normal sample. To confirm germline mutations detected by tumor-only samples, GATK Haplotypecaller (v3.8.0) with dbSNP (version 138) was used.

3. Results

3.1. *Detection of SNVs in RNA-Seq Data*

For calling RNA-Seq specific mutations we implemented a variation of pileuping nucleotide bases at each position in the transcriptome, utilizing mpileup from samtools. In case all possible mutations in the RNA-Seq data should be checked, a mpileup of all base positions in the RNA-Seq data is performed. Afterwards the distribution of bases in each position is saved and evaluated to annotate them at positions where supplemental database information (CIVIC, ClinVAR, Cosmic), parsed accordingly in the script, is available. In case the -panelmode flag is selected, a list of 442 nucleotide positions is selected and the mpileup is called only for these positions. A minimum of 3 reads or 10% of the reads supporting the alternative variant are used as default thresholds. The output comprises the distribution of bases at each position, the decision whether the position contains an SNV, and clinical annotations from CIViC and CGI.

The possibility to detect well-known SNVs with very little extra effort in RNA-Seq data is of high interest. For that, it is crucial to define the set of SNVs one wants to

detect. In this study we designed a panel that covers a curated list of variants with a role in treatment response (i.e. biomarkers of drug response). The panel is used as a whitelist of mutations known to have implications in cancer therapy. It comprises 442 SNVs and on 92 genes. The variants can have different origins: somatic, germline mutations or germline polymorphisms.

3.2. *Comparison of WES and RNA-Seq Data in Detecting SNVs*

We generated WES matched tumor-normal and RNA-Seq tumor-only data from 14 cancer patients. We applied a standard pipeline to detect somatic SNVs in WES matched tumor-normal samples, referred to as clinical standard. We also applied a tumor-only mode to WES tumor samples, referred to as tumor-only. Finally, we applied our method to detect SNVs in RNA-Seq tumor samples. SNVs detection was focused on 442 cancer-specific variants with clinical interest.

As shown in Figure 1A we found 109 SNVs in all samples by the three methods: 10 in the gold standard, 104 in tumor-only and 73 in the RNA-Seq (average of 0.7, 7.4 and 5.2 SNVs/sample, respectively). The 10 SNVs detected by the clinical standard were also detected in WES tumor-only and RNA-Seq. We found a higher overlap between WES tumor-only and RNA-Seq (68 SNVs) than between the two methods on WES (10 SNVs). This finding is explained by the fact that our panel includes germline mutations and polymorphisms; the clinical standard is optimized to reliably detect mutations only present in the tumor sample by filtering out any mutation present in the normal sample. Accordingly, the clinical standard only detected SNVs known to be somatic. Nonetheless, all somatic SNVs were also detected by tumor-only and RNA-Seq (Figure 1B).

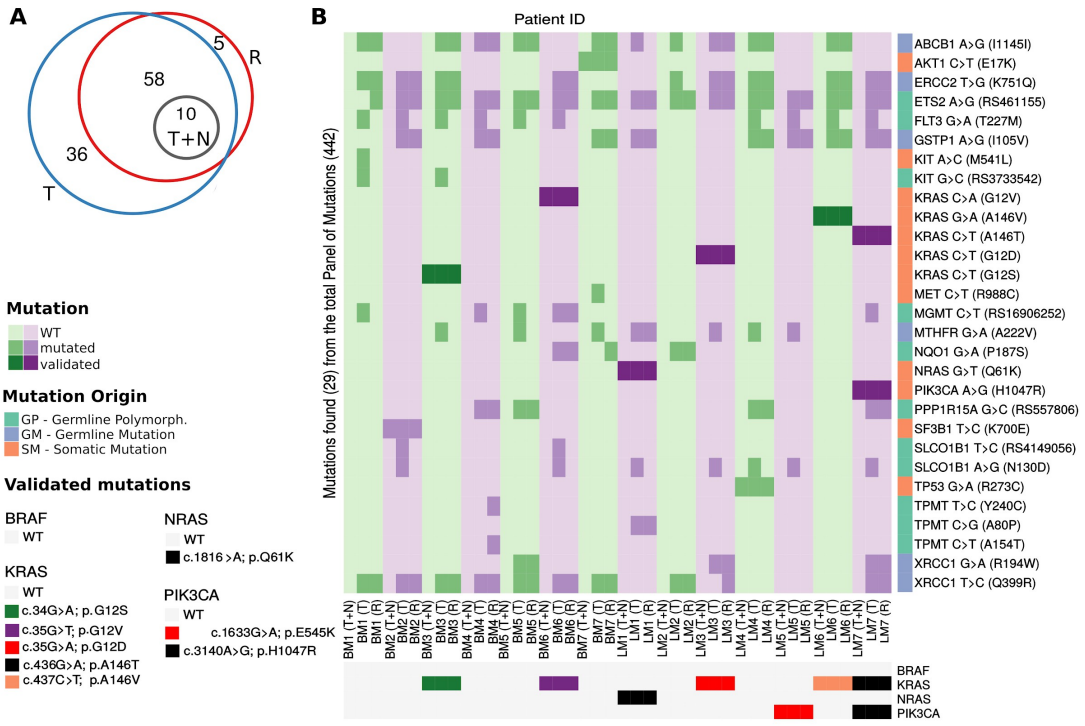

Figure 1: A) Venn diagram depicting the number of SNVs identified by each method across all samples (T+N: Mutect2, Tumor+Normal samples, T: Mutect2, Tumor samples, R: Wileup RNA-Seq). In total 109 SNVs were found (29 unique mutations from the panel of 442 clinical relevant mutation). B) Heatmap visualization of 29 unique SNVs which were found by at least one of the methods in any of the 14 patients.

Wild Type (WT) mutations are shown light green and purple, mutations found by the methods are in green and purple, mutation agreeing with the pathological annotation (validated) are marked in dark green and purple. The details of the pathological mutations are described in the annotation bars at the bottom of the figure. The origin of the mutation is annotated in the bar at the right sight of the heatmap.

The variants uniquely detected by WES tumor-only (36 SNVs) could be explained in the majority of the cases due to low expression in the RNA-Seq data. The only exception presenting high expression was the *MGMT* promoter SNP rs16906252. Yet, this SNP is known to be associated with low MGMT expression, leading to allele specific expression [13]. On the other hand, only 5 SNVs were exclusively detected by RNA-Seq. Two of them - *TPMT* Y240C and *TPMT* A154T - are known to be an haplotype of the TPMT enzyme (TPMT*3A) [14] and were indeed found in the same patient (BM4). These haplotype was not confirmed by WES tumor-only due to high duplication levels, which did not pass Mutect2 filters. The other three mutations (*ETS2* mutation in patient BM1, *NQO1* in patient BM7 and *XRCC1* mutation in patient LM3) were not found in WES tumor-only also due to the same reason. As a matter of fact, these 5 germline polymorphisms detected exclusively in RNA-Seq data could be confirmed by a germline SNV caller (Haplotypecaller) in normal samples.

We had pathological data on 4 routinely tested biomarkers (*BRAF*, *KRAS*, *NRAS* and *PIK3CA*) as part of the dataset. 7 out of the 8 pathologically validated mutations were consistently detected by the three methods (Figure 1). *PIK3CA* E545K mutation in patient LM5 was not detected by any method; WES data presented high duplicated regions in that position, whereas in RNA-Seq data *PIK3CA* was not expressed.

4. Discussion

We showed a high overlap between RNA-Seq and tumor-only WES. Previous studies reported high numbers of false positives in RNA-Seq data, however, by using a whitelist of well-defined SNVs we avoid this problem. In this setting, detecting SNVs in RNA-Seq data is a comparable approach to WES tumor-only; yet, in RNA-Seq it is regarded as an extra analysis which can be quickly performed (average of 11-15 min/sample) for no extra cost. More important, RNA-Seq appears to be a reliable approach for detecting the selected panel of clinically relevant SNVs, as confirmed by the pathologically validated data (for somatic variants) and by the analysis of normal WES samples (for germline variants). Of course, the user has to accept false negatives in non expressed genes, but that is inherent to RNA-Seq data. For future implementations, it would be important to consider RNA editing processes as well as including indels in the analysis.

5. Conflict of Interest

The authors declare no conflict of interest

3.3 The adaptation of colorectal cancer cells when forming metastases in the liver: expression of associated genes and pathways in a mouse model

Reference

Derya Bocuk, Alexander Wolff, Petra Krause, Gabriela Salinas, Annalen Bleckmann, Christina Hackl, Tim Beissbarth and Sarah Koenig: The adaptation of colorectal cancer cells when forming metastases in the liver: expression of associated genes and pathways in a mouse model. BMC Cancer 2017 17:342, https://doi.org/10.1186/s12885-017-3342-1D

3.3.1 Summary and discussion

In this publication, the resulting standard pipeline consisting of STAR, RSEM, edgeR and topGO of the first publication 3.1 were applied on a dataset of mice having one million *Colorectal cancer* (CRC) cells of cell line CMT-93 injected into the liver, as a novel syngeneic and orthotopic mouse model of CRC liver metastasis. By using RNA-Seq data it was possible to evaluate the difference in the expression profile of the cell line when migrating into the liver to become metastases. The tumour-free liver samples were used as a blacklist for liver tissue-specific expression, which does not occur in the cell line. Next, to standard quality checks on the sequencing data itself, quality checks on the count data showed a higher separation of the tumour samples (figure 3.4). After further investigations by checking either liver or colorectal tissue-specific gene expression, it could be shown, that three of the samples represented false biopsies and were excluded due to massive infiltration with liver (4 to 20 times the expression of liver enzymes) and low content of the colorectal tissue. This resulted in an experimental setup of 3 sample replicates of the cell line CMT93, seven biological replicates of normal liver and four biological replicates of the metastatic tissues in the liver. After mapping with STAR and read counting with RSEM, edgeR was used to detect DEGs, comparing CMT-93 cells with the liver metastases. Afterwards, a geneset for gene ontologies was defined consisting of all the DEGs identified by the comparison mentioned above. The R package topGo was used to identify over- or under-represented GO terms using the weighted Fisher-exact test algorithm from the package.

This analysis resulted in a total of 3329 (1174 down-regulated (35%), 2155 up-regulated (65%))) DEGs. The top five dysregulated genes were *matrix metallopeptidase* (MMP) 7, keratin 20 an epithelial colorectal cancer marker, Wnt inhibitory factor 1, MMP 9, and chemokine receptor 4. Furthermore, the Gene ontology analysis gave a more precise picture of the significant genes related to functional groups. The top changes of the cell line, when emitted into the liver tissue, where differences in gene expression related to inflammatory response, angiogenesis and signal transduction, positive regulation of transcription from RNA polymerase II promoter, transmembrane receptor protein tyrosine kinase signalling

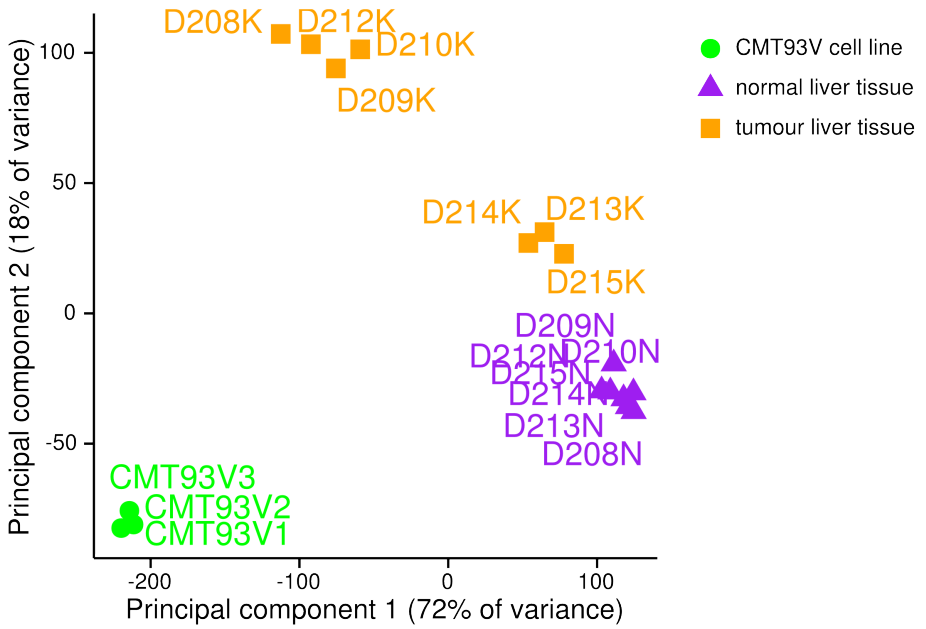

Figure 3.4. *PCA plot with the original data, showing parts of the tumour samples shifted towards normal liver samples.*

pathway, and positive regulation of ERK1 and ERK2 cascade. The complete and in-depth interpretation of the results supporting an invasion-metastasis cascade with notable changes in the expression profile can be found in the discussion part of the paper.
In conclusion, the bioinformatical approaches applied here where a crucial step for linking the most-relevant metastasis-related genes to reveal, that the liver environment stimulates the CMT-93 cells into the expression of metastasis enhancing genes. These genes were associated with tissue remodelling, cell proliferation, adhesion, wnt activity, transcription/regulation, and inhibition of apoptosis.

3.3.2 Declaration of my contribution

Conceptualization: Sarah Koenig, Derya Bocuk, Alexander Wolff, Christina Hackl, Tim Beissbarth, Annalen Bleckmann.
Data curation: Alexander Wolff, Gabriela Salinas, Sarah Koenig, Christina Hackl
Formal analysis: Alexander Wolff, Derya Bocuk
Investigation: Sarah Koenig, Derya Bocuk, Annalen Bleckmann
Methodology: Derya Bocuk, Alexander Wolff.
Project administration: Sarah Koenig, Petra Krause, Annalen Bleckmann, Tim Beißbarth.
Resources: Christina Hackl, Gabriela Salinas
Software: Alexander Wolff.
Visualization: Alexander Wolff, Derya Bocuk
Writing– original draft: Derya Bocuk, Sarah Koenig, Alexander Wolff.

Writing – review & editing: Derya Bocuk, Alexander Wolff, Petra Krause, Gabriela Salinas, Annalen Bleckmann, Christina Hackl, Tim Beißbarth and Sarah Koenig

Bocuk et al. BMC Cancer (2017) 17:342
DOI 10.1186/s12885-017-3342-1

BMC Cancer

RESEARCH ARTICLE Open Access

The adaptation of colorectal cancer cells when forming metastases in the liver: expression of associated genes and pathways in a mouse model

Derya Bocuk[1], Alexander Wolff[2], Petra Krause[1], Gabriela Salinas[3], Annalen Bleckmann[2,4], Christina Hackl[5], Tim Beissbarth[2] and Sarah Koenig[1,6*]

Abstract

Background: Colorectal cancer (CRC) is the second leading cause of cancer-related death in men and women. Systemic disease with metastatic spread to distant sites such as the liver reduces the survival rate considerably. The aim of this study was to investigate the changes in gene expression that occur on invasion and expansion of CRC cells when forming metastases in the liver.

Methods: The livers of syngeneic C57BL/6NCrl mice were inoculated with 1 million CRC cells (CMT-93) via the portal vein, leading to the stable formation of metastases within 4 weeks. RNA sequencing performed on the Illumina platform was employed to evaluate the expression profiles of more than 14,000 genes, utilizing the RNA of the cell line cells and liver metastases as well as from corresponding tumour-free liver.

Results: A total of 3329 differentially expressed genes (DEGs) were identified when cultured CMT-93 cells propagated as metastases in the liver. Hierarchical clustering on heat maps demonstrated the clear changes in gene expression of CMT-93 cells on propagation in the liver. Gene ontology analysis determined inflammation, angiogenesis, and signal transduction as the top three relevant biological processes involved. Using a selection list, matrix metallopeptidases 2, 7, and 9, wnt inhibitory factor, and chemokine receptor 4 were the top five significantly dysregulated genes.

Conclusion: Bioinformatics assists in elucidating the factors and processes involved in CRC liver metastasis. Our results support the notion of an invasion-metastasis cascade involving CRC cells forming metastases on successful invasion and expansion within the liver. Furthermore, we identified a gene expression signature correlating strongly with invasiveness and migration. Our findings may guide future research on novel therapeutic targets in the treatment of CRC liver metastasis.

Keywords: Colorectal cancer (CRC), RNA-sequencing, Gene expression, Liver metastasis

* Correspondence: koenig_sarah@ukw.de
[1]Department of General, Visceral and Paediatric Surgery, University Medical Centre, Georg – August – University Goettingen, Göttingen, Germany
[6]Medical Teaching and Medical Education Research, University Hospital Wuerzburg, Julius-Maximilians-University Wuerzburg, Josef-Schneider-Str. 2/D6, 97080 Wuerzburg, Germany
Full list of author information is available at the end of the article

Background

Colorectal cancer (CRC) is the third most common type of cancer in the Western world and the second most common cause of cancer-related death in both genders. The overall relative 5-year survival of CRC patients is approximately 50% [1]. Almost half of all patients suffering from CRC are confronted with liver metastasis either at the time of diagnosis (15 to 20%), or later during the course of the disease (25%) [2]. Given the rather poor 5-year survival rate of patients who develop liver metastasis (approx. 30%), it is vital that we develop and evaluate new therapeutic strategies. In particular, the knowledge of molecular changes to CRC cells that end up in the liver may enable us to search for new target options far more selectively.

Metastasis is frequently a final and fatal step in the progression of solid malignancies. The nature and time of onset of the changes that provide tumour cells with metastatic functions are still largely unknown. Furthermore, there has been an ongoing debate to this end for more than a 100 years. In 1889, Stephen Paget noticed that the pattern of metastases produced by different neoplasms was not random. In his 'seed and soil' hypothesis, Paget claimed that certain tumour cells ('seeds') have an affinity for the microenvironment of specific organs ('soil'), and only when the 'seed' and the 'soil' are compatible can metastasis occur [3].

With respect to the "seed", it is widely accepted these days that cancer is attributed to the accumulation of genetic alterations in cells. Thus, to understand the molecular mechanisms of cancer metastasis, it is indispensable to identify not only the genes whose alterations accumulate during cancer progression but also those genes whose expression is responsible for the acquisition of metastatic potential in cancer cells [4]. Indeed, comparative analyses of the gene expression profiles of metastatic and non-metastatic cells have revealed that various genes are differentially expressed in association with the metastatic potential of cancer cells [4]. Conversely, the existence of genes expressed by rare cellular variants that specifically mediate metastasis has been disputed [5]. Transcriptomic profiling of primary human carcinomas has identified gene expression patterns that, when present in the primary tumour, predict a poor prognosis for patients [6, 7]. The existence of such signatures can be interpreted in the sense that genetic lesions acquired early on in tumorigenesis may prove sufficient for the metastatic process, and that consequently no metastasis-specific genes exist.

There is growing evidence that the development of or progression to metastases is also dependent on the "soil". Tumour cell circulation, extravasation into a distant organ, angiogenesis, and uninhibited growth also provide essential hints as to the metastatic process [8]. The molecular requirements for some of the steps involved may be highly tissue specific. For example, the proclivity that tumours have for specific organs, such as breast carcinomas for bone and lung, was noted more than a century ago [9]. Moreover, the potential of tumour cells to metastasize depends on their interaction with homeostatic factors in the target organ that promote tumour-cell growth: survival, angiogenesis, invasion, and progression. It seems that the intrinsic cellular heterogeneity within tumour populations evolves through an extrinsic selection process, which is based on more or less infrequent cellular variants with augmented metastatic abilities and which finally mediates the outgrowth in distant sites [9]. Of note, the mechanism that enables the liver microenvironment to influence the behaviour of CRC cells is still only poorly understood.

The most common site for CRC metastasis is the liver [10]. Many patients still suffer from recurrence of the primary and/or distant metastasis, even after undergoing liver resection combined with adjuvant approaches such as chemo- and radiotherapy. Nonetheless, only a minority of patients actually survive for years [11]. Therefore, the a priori or early inhibition of metastasis could prove to be a key step towards the curative treatment of patients. We have to assume that each organ places different demands on circulating cancer cells for the homing and subsequent outgrowth of metastases.

To clarify this issue, we established a novel syngeneic and orthotopic mouse model of CRC liver metastasis. This model comprises the injection of cells from a known CRC cell line to mimic the spread of the primary tumour and thus to investigate the invasion and expansion of CRC cells in the liver on the gene expression level. The goal here was to identify genes that contribute to this process of adapting to the new "soil" and thereby the metastatic progression of the disease. The fundamental aim of the study was to identify new candidate markers or molecular mechanisms in the diagnosis of liver metastasis resulting from CRC, as well as therapeutic targets effectively inhibiting CRC metastasis in the liver.

Methods

Reagents and antibodies

Unless specified otherwise, all chemicals and reagents were supplied by Life Technologies (Darmstadt, Germany). Foetal Bovine Serum Superior (FBS) was purchased from Biochrom (Berlin, Germany) and trypsin 10-fold was supplied by PAA (Pasching, Austria). Antibodies for immunolabelling purposes were purchased and used as illustrated in Table 1.

Cell lines and culture

The cell line CMT-93 (isolated from a mouse colorectal adenocarcinoma) was kindly donated by Christina Hackl

Table 1 Antibodies used in immunolabelling analysis

Antigen	Species	Dilution	Catalogue	Manufacturer
β-catenin	Rabbit	1:50	14–6765	eBioscience, Frankfurt a.M., Germany
CD44	Rat	1:1000	550,538	BD Pharmingen, Heidelberg, Germany
Ki-67	Rabbit	1:200	275R-14	Cell Marque, California, United States
E-cadherin	Rabbit	1:50	sc-7870	Santa Cruz Biotechnology, Heidelberg, Germany
Vimentin	Rabbit	1:1000	ab92547	Abcam, Cambridge, UK
Anti-rat biotinylated	Donkey	1:200	RPN1004	GE Healthcare, Freiburg, Germany
Avidin HRP		1:400	18–4100-94	eBioscience, Frankfurt a.M., Germany
HRP Labelled anti-rabbit	Goat	Ready to use	K4002	Dako, Hamburg, Germany

and her workgroup in Regensburg, Germany. On testing, the cells were found to be negative for mycoplasma by RT-PCR.

CRC cells were expanded and stored in frozen aliquots (– 70 °C). After thawing, the cells were routinely cultured in 75 cm^2 culture flasks in DMEM high glucose, supplemented with 10% FBS, 1% L-glutamine, 1% sodium pyruvate and 1% penicillin/streptomycin at 37 °C and 5% CO_2 in a humidified incubator. Tumour cells were passaged once (following 3 days in culture), cultured for a further 4 days, and then trypsinized for subsequent implantation studies. Tumour cells from the same passage were used for all the implantation experiments. Additionally, aliquots of the cell line were snap frozen and processed for transcriptome sequencing analysis (RNA-seq).

Animals and procedures

Ten-week-old female C57BL/6NCrl mice (mass 18–22 g) were purchased from Charles River (Sulzfeld, Germany). Animals were kept on a 12-h day/night rhythm and fed with a phytoestrogen-reduced mouse diet (ssniff, Soest, Germany).

Prior to (surgical intervention) surgery, animals received a subcutaneous application of carprofen (Rimadyl®, Pfizer, Berlin, Germany) (5 mg/kg body mass). Animals were anaesthetized under constant sevoflurane inhalation (Sevorane®, Abbott, Wiesbaden, Germany). After median laparotomy, the hilum of the liver was exposed to access the portal vein. One million tumour cells in a volume of 100 μl PBS buffer were injected slowly into the portal vein using a 30 G needle.

In the study group, seven animals were implanted with tumour cells. The control group encompassed five animals which underwent the same procedures (sham-OP), but were only injected with buffer solution. All animals were sacrificed after 4 weeks. Explanted livers were sliced for macroscopic assessment, photographic documentation of the section planes, and further processing. Tissue samples from the tumour core of the liver metastases derived from CMT-93 as well as matched unharmed liver tissue (macroscopically tumour-free liver) were excised, snap frozen for whole transcriptome sequencing analysis (RNA-seq), or frozen in 2-methylbutane at −70 °C for immunolabelling.

Immunolabelling

Cryostat sections (5 μm) were fixed in ice-cold acetone for 10 min and were stored at −80 °C. After rehydration in Tris/HCl buffer (pH 7.6), sections were incubated with the primary antibodies (see Table 1) overnight at 4 °C. Endogenous peroxidase was inactivated by incubation with 0.3% H_2O_2 in 70% methanol and 30% Tris/HCl buffer for 20 min at RT. The HRP-labelled goat anti-rabbit IgG secondary antibody (DakoCytomation K4002, Carpinteria, USA, ready-to-use reagent) was used to identify β-catenin, Ki-67, E-cadherin, and vimentin. To immunolabel CD44, sections were exposed to an avidin/biotin blocking step (Life Technologies, Darmstadt, Germany) followed by incubation with the primary antibody (overnight at 4 °C). This antigen was identified by the secondary antibodies donkey anti-rat biotinylated (1:200, 1 h at RT) and avidin-horseradish peroxidase (HRP) (1:400, 1 h at RT). 3-amino-9-ethyl-carbazole (AEC) solution (BD Pharmingen, Heidelberg, Germany) and haematoxylin counterstaining were used for visualization by light microscopy. Negative controls were carried out for each antibody by omitting the primary antibody from the protocol. Samples were covered with 50 μl of the aqueous mounting agent Aquatex (Merck, Darmstadt, Germany) and evaluated under a light microscope (LEICA DM IRE2, Bensheim, Germany).

RNA isolation

For RNA sequencing purposes (RNA-seq), three aliquots of the cell line and specimens (tumour core and liver) from seven animals were collected. The RNA purification system PeqGold TriFast (Peqlab, Erlangen, Germany) was used to isolate RNA from metastatic liver tissue. Briefly, specimens were defrosted in peqGold TriFast (1 ml/100 mg tissue) and then homogenized using TissueLyser LT (Qiagen, Hilden, Germany) at

50 Hz. Total RNA was isolated according to the manufacturer's instructions and stored at −80 °C. In addition, the High Pure RNA Isolation Kit (Roche, Grenzach-Wyhlen, Germany) was used to isolate RNA from CMT-93 cells according to the manufacturer's recommendations. The quantity and integrity of the isolated RNA was assessed in a NanoDrop ND − 1000 spectrophotometer, version 3.5.2 (Peqlab, Erlangen, Germany), using the 260 nm/280 nm absorbance ratio and was further analysed with an Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, California, USA) as a quality check. RNA-seq was performed at the Transcriptome and Genome Analysis Laboratory in Goettingen, Germany, using an Illumina HiSeq2000 sequencer (Illumina, Inc., San Diego, California, USA).

Deep sequencing analysis

As starting material for the library preparation, 0.5 μg of total RNA was used. The libraries were generated according to the TruSeq mRNA Sample Preparation Kits v2 Kit from Illumina (Cat. N°RS − 122-2002). The fluorometric based QuantiFluor™ dsDNA System from Promega (Mannheim, Germany) was used for accurate quantitation of cDNA libraries. The size of final cDNA libraries was determined by using the Fragment Analyzer from Advanced Bioanalytical. cDNA libraries were amplified and sequenced by using the cBot and HiSeq2000 from Illumina (SR; 1 × 50 bp; ca. 30 Mio reads per sample). Sequence images were transformed to bcl files using Illumina software BaseCaller, which were demultiplexed to fastq files with CASAVA v1.8.2 and quality checks were done via fastqc.

Statistics

Preparation of data/statistical model

An in-house RNA-seq analysis pipeline employing the STAR-aligner (version 2.4.0 h) [12] for the mapping and counting of reads with the expectation-maximization algorithm implemented in the software package RSEM (version 1.2.19) [13] was used for counting reads. Ensembl *Mus musculus* GRCh38 Version 78 was considered as the reference for mapping and further annotations.

Following RNA-seq, all seven tumour probes derived from CMT-93 underwent quality control measures. Employing the corresponding RNA-seq data, they were checked for the expression of CK 20 as surrogate parameter for colorectal tissue or liver-specific gene expression to identify liver-specific genes, such as phosphoenolpyruvate-carboxykinase 1 (PCK1), cytochrome p450 (CYP), and carbamoyl phosphate synthase1 (CPS1). Three tumour probes (D213K, D214K, D215K) representing false biopsies were excluded from further analysis owing to a strong infiltration of liver (approx. 4 to 20 times the elevated expression levels of liver enzymes) and low content of colorectal tissue.

Principal component analysis (PCA) was performed in R, the programming language and environment (version 3.2), to visualize the underlying structure of the dataset by calculating the eigenvectors and plotting those two components with the highest variance in the data.

Focussing on the comparison between the cell line and metastases, we filtered out differential genes specific to liver tissue, which we considered as 'liver tissue effect'. Thus, differentially expressed genes (DEGs) were identified as either up- or down-regulated when comparing the CMT-93 cell line with the unharmed liver tissue. Subsequently, these differences relating to the normal liver background were excluded from the gene expression results between the cell line and liver metastases. This filtering step was done in order to identify genes representing differences in cell line versus tumour, instead of general differences in cell lines versus normal liver.

Significant differential gene analysis

Basing on the read counts attained from RSEM, the R package EdgeR [14] was used to calculate the mean intensities as well as the p-value and the log fold change (logFC) for each DEG, comparing the CMT-93 cells with the liver metastases formed. Thus, gene differences between the two groups were identified by fitting a negative binomial generalized linear model implemented in EdgeR. Expression results were reported as mean transcripts per million (TPM) values for each group.

A list was created comprising 119 genes associated with metastasis, based on the genes described in the Tumor Metastasis RT2 Profiler PCR Array by Qiagen Hilden, Germany (Additional file 1). This list was applied as a filter following completion of the analysis of the DEGs to profile the expression of these genes in our dataset.

Gene ontology (GO) analysis

A gene set was defined, comprising all the DEGs identified in the comparison of CMT-93 cells and the liver metastases that formed, corrected for the liver background and with a false discovery rate of less than 5% (FDR < 0.05). This gene set was employed in the gene ontology and pathway analysis. This method, implemented in the R package topGO [15], allows us to identify GO terms that are over-represented (or under-represented) using the annotations for that gene set taken from the Gene Ontology Database (http://www.geneontology.org/). The significant level of GO terms for the DEGs was analysed with the weighted Fisher's exact test in the package. We computed p-values for all the

DEGs in the GO category "biological processes"; the threshold of significance was defined as p-value <0.05.

Results

Syngeneic mouse model of CRC metastasis in the liver

Following the injection of CMT-93 cells via the portal vein and their subsequent expansion, 70–80% of animals developed liver metastases in any number of liver lobules during the course of the study, as seen on macroscopic assessment (Fig. 1a). Within the life span of the animals, CMT-93 cells had colonized about 30–50% of the mouse liver with the tumour spots increasing to approximately 5 to 10 mm in diameter. However, the spread of the tumour burden resulting from CMT-93 proved to be inhomogeneous when comparing the right and left liver lobules of injected mice. On the microscopic level, immunohistochemical staining 4 weeks after tumour cell injection into the portal vein was used to assess phenotypic expression of colorectal carcinoma markers. Liver metastases displayed features of a moderately differentiated colorectal adenocarcinoma revealing complex glandular structures (40–75% gland formation) in a desmoplastic stroma (Fig. 1 b, c). An important feature of invasion is the presence of this desmoplasia or desmoplastic reaction, a type of fibrous proliferation surrounding tumour cells and secondary to the invasive tumour growth. The glandular structures expressed epithelial markers such as membrane-bound β-catenin (Fig. 1d) and E-cadherin (Fig. 1e). The latter was expressed mostly in the cytoplasm, which indicates that this wnt marker was inactive. The hyaluronic acid receptor CD44 (a putative marker of 'stemness' in CRC) was also present and staining was detected in nearly all of the tumour cells we investigated (Fig. 1f). The unsystematic arrangements of gland formations also expressed the mesenchymal marker vimentin (Fig. 1g). The proliferation marker Ki-67 was expressed abundantly in more than 75% of all liver tumours in a random pattern (Fig. 1h).

Adaptation of CMT-93 cells when forming metastases in the liver

To assess the factors involved in the formation of liver metastases, we performed RNA-seq analysis on both the liver metastases as well as unaffected liver tissue. Figure 2 summarizes the structure of the gene expression data. The first principal component (PC) is plotted on the x-axis and captures 74% of the variance. The second PC is plotted on the y-axis and captures 19% of the variance. The PCA plot clearly portrays the separation of the CMT-93 cell line samples from those of the corresponding macroscopically tumour-free liver, as well as from the derived liver metastases. As assumed, the cluster associated with CMT-93 was found to be located in close proximity to the cluster relating to the derived

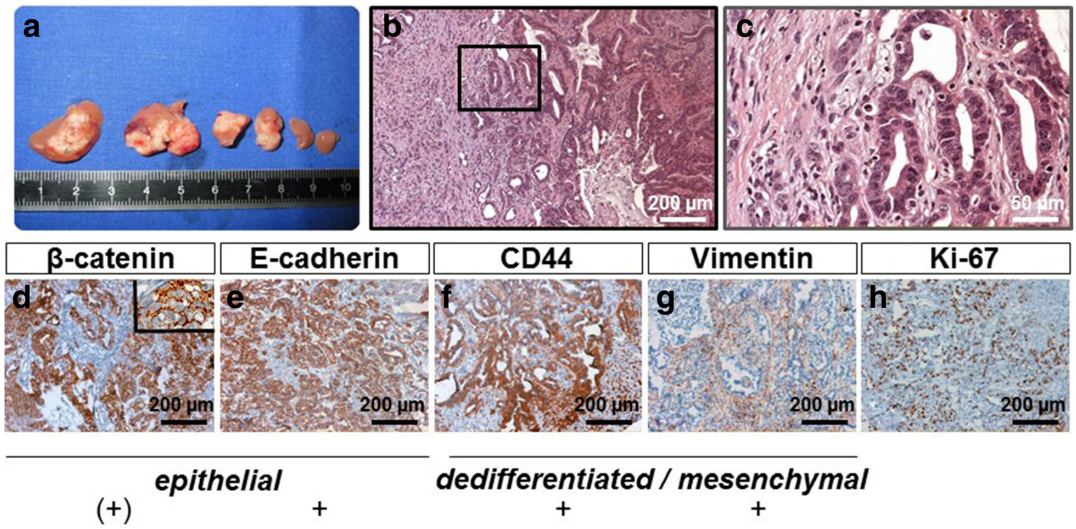

Fig. 1 Macroscopic and microscopic aspects of liver metastases derived from CMT-93. Macroscopic overview of metastatic growth 4 weeks following implantation of CMT-93 cells via the portal vein in syngeneic C57BL/6NCrl mice (treatment group). Explanted livers were sliced for macroscopic assessment (**a**) and subjected to H&E staining (**b, c**). Liver metastases represented a moderately differentiated colorectal adenocarcinoma revealing complex glandular structures (50–75% gland formation) in a desmoplastic stroma. Subpanel **d** depicts the results of immunolabelling to detect ß-catenin (the inlay illustrates the mostly cytoplasmic location of expression) and **e** confirms E-cadherin, both markers indicative of the glandular structures. The dedifferentiated/mesenchymal differentiation was identified by staining with anti-vimentin (**g**) and anti-CD44 (**f**). Ki-67 was used as a proliferation marker as illustrated in **h** Scales are as indicated

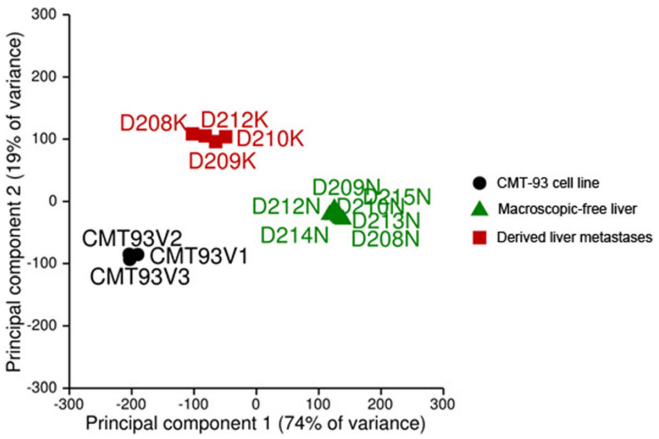

Fig. 2 PCA was used to visualize the underlying structure of the dataset. CMT-93 cells as well as tumour probes (*tumour liver tissue*) formed clusters clearly separate from the corresponding macroscopically tumour-free liver. Each group was tightly clustered

metastases. We also plotted liver specimens originating from sham-operated animals (injection of buffer alone). When overlaid on the PCA plot in Fig. 2, these samples lay in exactly the same position as the cluster of the macroscopically tumour-free liver samples (data not illustrated).

Table 2 lists the DEGs during the propagation of CMT-93 cells in the liver. A total of 5297 genes were down-regulated and 6597 were up-regulated when CMT-93 cells propagated in the liver. The elimination of DEGs relating to the liver background (see materials and methods) reduced the total number of DEGs to 1174 down-regulated genes (35%) and 2155 up-regulated genes (65%).

The results of hierarchical cluster analysis to assess the relatedness of the 120 DEGs displaying the greatest differences in expression are presented in Fig. 3. The heat map reveals systematic and fairly clearly distinguished variations in the expression of genes between the original CMT-93 cells, the derived liver metastases, and tumour-free liver. Of note, the changes in expression of the CMT-93 cells outgrowing as liver metastases is clearly apparent. However, the hierarchical clusters relating to the metastases and the tumour-free liver are somewhat closer to each other owing to the fact that the implanted CMT-93 cells forming metastases infiltrate the liver tissue. The gene expression profile is therefore bound to reflect the colonization of the hepatic tissue.

We then applied a filter set of 119 selected genes associated with metastasis to our RNA-seq data set following the elimination of DEGs determining the liver background. Thus, 32 relevant genes were identified with a threshold of FDR < 5%, of which 23 were up-regulated and 9 down-regulated. Additional file 2 contains the count data of the samples. Figure 4 illustrates the heat map of the filtered DEGs within the CMT-93 cell line and the liver metastasis samples. Hierarchical clustering confirmed the clear changes between DEGs of the CMT-93 cells and the derived metastases. Table 3 presents an overview of these genes which were also ranked by *p*-value. The top five dysregulated genes were found to be matrix metallopeptidase (MMP) 7, keratin 20 as an epithelial marker of colorectal carcinoma, wnt inhibitory factor 1, MMP 9, and chemokine receptor 4.

With respect to functional gene groups, the 23 up-regulated genes were attributed to extracellular matrix proteins (6 genes), cell growth and proliferation (5 genes), cell adhesion (4 genes), wnt signalling (3 genes), transcription factors/regulators and epithelial-mesenchymal transition (EMT) (2 genes each), and CRC-related genes (1 gene). The 9 down-regulated genes were assigned to cell growth and proliferation, wnt signalling and

Table 2 Differences in gene expression among the sample groups (DEGs)

	CMT-93 vs. liver metastases	Macroscopic-free liver vs. liver metastases	CMT-93 vs. macroscopic-free liver	CMT-93 vs. liver metastases corrected for the liver background
−1 = down-regulated	5297	5163	5764	1174
0 = unregulated	3667	4157	2684	0
1 = up-regulated	6597	5749	6187	2155

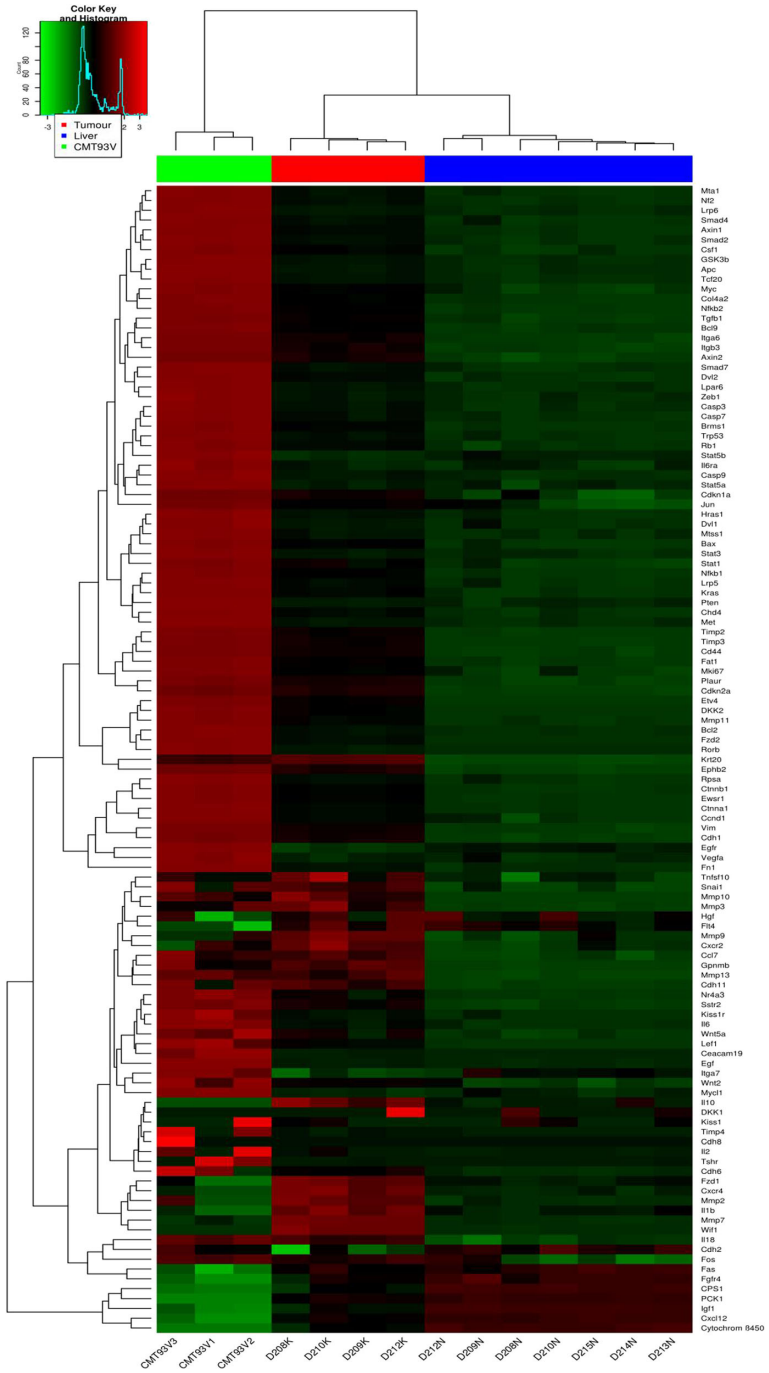

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Top 120 DEGs between CMT-93 cells, liver metastases derived from CMT-93 and macroscopically tumour-free liver. Expression data are depicted as a data matrix in which each row represents a gene and each column represents a sample. The colour coding *bar above the heat map marks* the samples from the CMT-93 cell line (samples CMT93V1–3) as *green*, those from the liver metastases derived from CMT-93 (samples D208-210 K, D212K) in *red* and those of macroscopically tumour-free liver (samples D208-210 N, D212-215 N) in *blue*. Expression levels are depicted according to the *colour scale* presented in the *top left* corner. *Red* indicates expression levels *above* and *green below* the median, respectively. The magnitude of deviation from the median is represented by the colour saturation. The hierarchical clustering is visualized by the dendrogram at the *top*, which illustrates the degree of relatedness in gene expression

transcription factors/regulator (2 genes each), as well as apoptosis genes, extracellular matrix proteins, and EMT (1 gene each).

Biological processes involved

To address the pathways and processes involved, significant DEGs with FDR < 5% were selected and tested against the background set of all genes with GO annotation (Fig. 5). The most relevant GO terms for biological processes enriched were "inflammatory response", "angiogenesis", "signal transduction", "positive regulation of transcription from RNA polymerase II promoter", "transmembrane receptor protein tyrosine kinase signalling pathway", and "positive regulation of ERK1 and ERK2 cascade".

Discussion

CRC is a common disease whose considerable metastatic potential highlights the urgency and necessity to develop novel therapeutic approaches to prevent or treat tumour progression and metastasis. In this study, we set out to analyse the changes in gene expression and pathways that play a role in the colonisation of mouse liver by the cell line CMT-93, mimicking the processes that lead to the ultimate formation of liver metastases secondary to CRC. To this end, we first had to establish the in vivo metastatic mouse model. CMT-93 cells demonstrate a strikingly efficient tumorigenic capacity following their implantation via the portal vein, the common route CRC cells take when colonising the liver. Furthermore, the CMT-93 cell line originated from the mouse strain we employ, C57BL/6NCrl; the two are thus syngeneic and this enables us to circumvent rejection responses as complications. Moreover, the mice are immunocompetent, which allows us to investigate the normal inflammatory response to tumour growth. Of note, the outgrowing metastases in this model are reproducible and display a number of prototypic features (structure and markers) common to human CRC liver metastases. The model itself acts as a paradigmatic proof of principle for genetic alteration when forming liver metastases. A xenograft model involving human CRC cell lines would not have been able to fulfil our criterion of immunocompetence to mimic the seed and soil theory in humans. While searching for suitable models we tried three combinations in total, based on literature research and commercial availabilities. Although CT-26 (ATCC® CRL-2639™) and the mouse strain Balb/c resulted in tumour formation in the liver, these were found to be mesenchymally de-differentiated in nature. The cell line APC1638-NT (kindly donated by Prof. R. Smits, Rotterdam, Netherlands) with mouse strains C57BL/6 N or C57BL/6 J resulted in no detectable tumour growth whatsoever (data not shown). The CMT-93/C57BL/6NCrl model proved to be the only one demonstrating reproducible liver metastatic growth with the histological features of colorectal cancer.

Disseminating tumour cells need to adapt to surrounding tissues in a continuous fashion [16]. For example, CRC cells of the primary tumour have to avoid succumbing to any immune response, detach from the primary, migrate into the portal system, arrive in the liver, traverse the endothelial barrier of the portal vessels (extravasation), overcome hypoxia on integration into the liver parenchyma, adapt to the new environment, initiate angiogenesis, and finally expand as metastases (metastatic colonisation) [17–19]. Taken together, these numerous influences during the process of metastatic spread into clinically detectable macroscopic disease lead to marked changes in gene expression. This new gene signature is the result of a multi-step process in which carcinoma cells progress along the "invasion-metastasis cascade" [18]. The results of our study visibly support the notion that CRC cells certainly undergo a number of clear changes in the liver environment.

It goes without saying that every single finding from the dataset following RNA-seq analysis cannot be commented on in this highly specific context during propagation in the liver. However, there are some definite hints as to which genes and pathways have to be addressed when aiming to treat CRC liver metastasis.

Our RNA-seq expression analysis and subsequent filtering with the selection list of the most-relevant metastasis-related genes reveals that the liver environment stimulates CMT-93 cells into expressing a number of genes enhancing metastasis. These genes are associated with functions such as tissue remodelling, cell proliferation, adhesion, wnt activity, transcription/regulation, and inhibition of apoptosis, which all contribute to metastatic activity and tumour cell invasion.

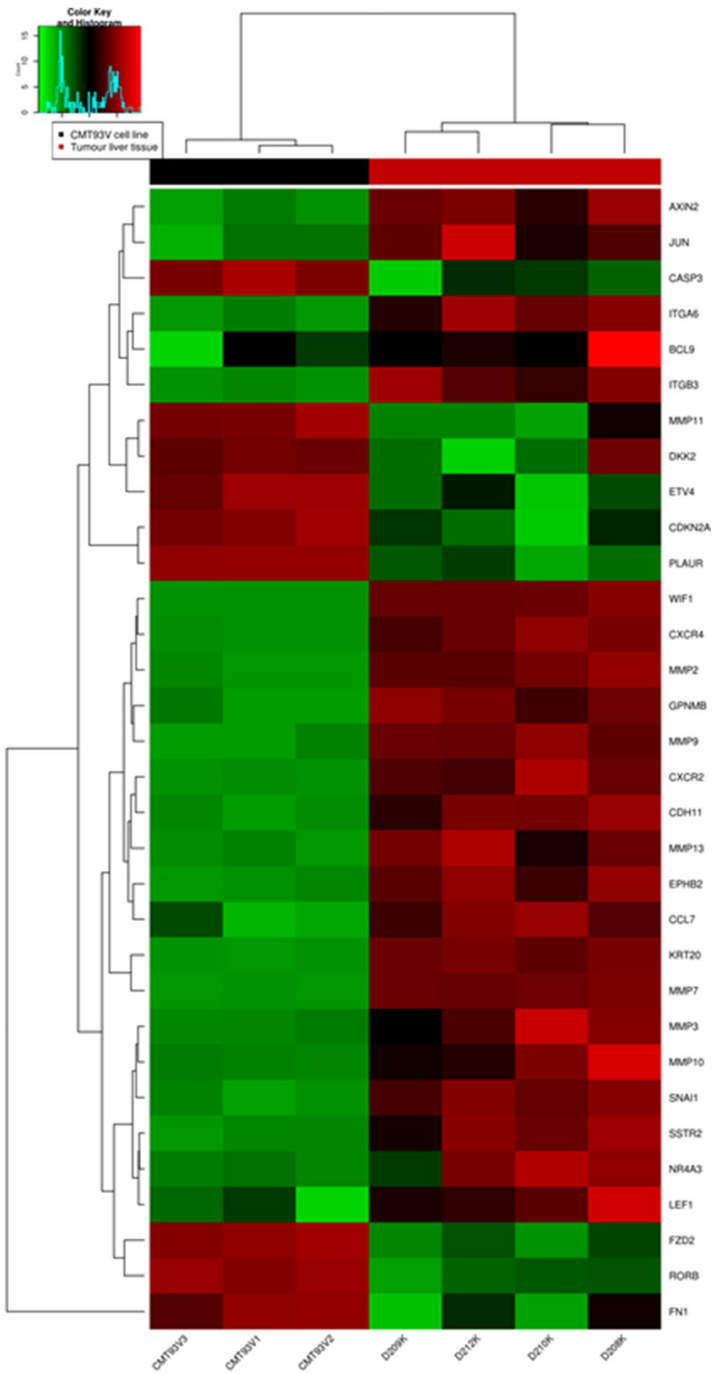

Fig. 4 Heat map illustrating the expression signature of 32 genes related to liver metastasis development. Unsupervised analysis was performed on the data set using our filtered gene list (119 genes associated with metastasis). Depicted are 32 representative genes which were expressed differentially during the propagation of CMT-93 cells in the liver

Table 3 Overview of the most relevant genes associated with metastasis during liver colonisation (FDR < 5%)

Gene	Ensembl ID	Gene Symbol	Functional Gene Group	*p*-value	Mean intensity CMT-93	Mean intensity liver metastases	logFC*	logFC^	FDR
Matrix metallopepti-dase 7	ENSMUSG00000018623	MMP7	Extracellular Matrix Proteins	$1{,}7 \times 10^{-180}$	0,02	72,56	8,97	11,69	$4{,}9 \times 10^{-178}$
Keratin 20	ENSMUSG00000035775	KRT20	CRC-Related Genes	$2{,}9 \times 10^{-174}$	0,81	51,99	5,8	6,12	$7{,}7 \times 10^{-172}$
Wnt inhibitory factor 1	ENSMUSG00000020218	WIF1	Wnt Signalling (canonical)	$3{,}7 \times 10^{-126}$	0,00	20,68	7,37	12,85	$3{,}3 \times 10^{-124}$
Matrix metallopepti-dase 9	ENSMUSG00000017737	MMP9	Extracellular Matrix Proteins	$3{,}8 \times 10^{-72}$	0,10	14,16	5,95	7,92	9×10^{-71}
Chemokine (C-X-C motif) receptor 4	ENSMUSG00000045382	CXCR4	Cell Growth and Proliferation Genes	$1{,}5 \times 10^{-71}$	0,02	19,29	7,06	10,50	$3{,}4 \times 10^{-70}$
Matrix metallopepti-dase 2	ENSMUSG00000031740	MMP2	Extracellular Matrix Proteins	$1{,}8 \times 10^{-70}$	0,09	27,83	7,04	8,33	$4{,}2 \times 10^{-69}$
Eph receptor B2	ENSMUSG00000028664	EPHB2	Cell Growth and Proliferation Genes	$4{,}8 \times 10^{-59}$	1,06	10,86	3,2	3,65	8×10^{-58}
Chemokine (C-X-C motif) receptor 2	ENSMUSG00000026180	CXCR2	Cell Growth and Proliferation Genes	$1{,}1\ 10^{-40}$	0,04	7,53	5,51	7,75	$9{,}7 \times 10^{-40}$
Cadherin 11	ENSMUSG00000031673	CDH11	Cell Adhesion Genes	$8{,}8 \times 10^{-40}$	0,12	6,32	4,67	5,75	$7{,}9 \times 10^{-39}$
Axin2	ENSMUSG00000000142	AXIN2	Wnt Signalling (canonical)	$2{,}2 \times 10^{-39}$	20,31	72,10	1,82	1,87	2×10^{-38}
Integrin beta 3	ENSMUSG00000020689	ITGB3	Cell Adhesion Genes	$1{,}6 \times 10^{-33}$	5,54	25,48	2,17	2,35	$1{,}2 \times 10^{-32}$
Matrix metallopeptidase 13	ENSMUSG00000050578	MMP13	Extracellular Matrix Proteins	$8{,}1 \times 10^{-29}$	0,22	7,45	4,43	5,20	$4{,}8 \times 10^{-28}$
Glycoprotein (transmembrane) nmb	ENSMUSG00000029816	GPNMB	Cell Adhesion Genes	$3{,}9 \times 10^{-27}$	0,21	14,50	5,45	6,25	$2{,}1 \times 10^{-26}$
Snail family zinc finger 1	ENSMUSG00000042821	SNAI1	EMT Transition	$2{,}9 \times 10^{-26}$	0,12	3,05	3,62	4,66	$1{,}5 \times 10^{-25}$
Plasminogen activator, urokinase receptor	ENSMUSG00000046223	PLAUR	Cell Growth and Proliferation Genes	$8{,}9 \times 10^{-20}$	191,38	71,50	−1,42	−1,28	$3{,}6 \times 10^{-19}$
Matrix metallopeptidase 3	ENSMUSG00000043613	MMP3	Extracellular Matrix Proteins	$3{,}7 \times 10^{-17}$	0,05	2,70	3,95	5,74	$1{,}3 \times 10^{-16}$
Integrin alpha 6	ENSMUSG00000027111	ITGA6	Cell Adhesion Genes	5×10^{-17}	12,03	31,23	1,36	1,32	$1{,}8 \times 10^{-16}$
Matrix metallopeptidase 10	ENSMUSG00000047562	MMP10	Extracellular Matrix Proteins	$7{,}6 \times 10^{-17}$	0,06	2,68	3,83	5,37	$2{,}7 \times 10^{-16}$
Somatostatin receptor 2	ENSMUSG00000047904	SSTR2	Cell Growth and Proliferation Genes	$5{,}7 \times 10^{-15}$	0,22	1,57	2,19	3,17	$1{,}8 \times 10^{-14}$
Jun proto-oncogene	ENSMUSG00000052684	JUN	Transcription Factors and Regulators	$7{,}8 \times 10^{-13}$	22,13	44,89	1,01	1,16	$2{,}3 \times 10^{-12}$
Frizzled homolog 2 (Drosophila)	ENSMUSG00000050288	FZD2	Wnt Signalling (canonical)	$1{,}1 \times 10^{-12}$	6,88	2,16	-1,7	-1,54	3×10^{-12}
Chemokine (C-C motif) ligand 7	ENSMUSG00000035373	CCL7	Cell Growth and Proliferation Genes	$2{,}4 \times 10^{-12}$	0,98	9,98	3,18	3,47	$6{,}9 \times 10^{-12}$
Nuclear receptor subfamily 4, group A, member 3	ENSMUSG00000028341	NR4A3	Transcription Factors and Regulators	$3{,}5 \times 10^{-12}$	0,16	1,22	2,1	3,31	1×10^{-11}
RAR-related orphan receptor beta	ENSMUSG00000036192	RORB		$2{,}6 \times 10^{-11}$	2,57	0,70	-1,94	-1,68	7×10^{-11}

Table 3 Overview of the most relevant genes associated with metastasis during liver colonisation (FDR < 5%) *(Continued)*

			Cell Growth and Proliferation Genes						
Cyclin-dependent kinase inhibitor 2A	ENSMUSG00000044303	CDKN2A	Transcription Factors and Regulators	$1{,}9 \times 10^{-8}$	179,73	91,40	−0,98	−0,84	$4{,}4 \times 10^{-8}$
Matrix metallopeptidase 11	ENSMUSG00000000901	MMP11	Extracellular Matrix Proteins	$6{,}6 \times 10^{-7}$	12,30	5,36	−1,21	−1,19	$1{,}4 \times 10^{-6}$
Ets variant 4	ENSMUSG00000017724	ETV4	Transcription Factors and Regulators	$9{,}5 \times 10^{-7}$	21,60	10,13	−1,1	−0,93	$1{,}9 \times 10^{-6}$
Lymphoid enhancer binding factor 1	ENSMUSG00000027985	LEF1	EMT Transition	$5{,}3 \times 10^{-5}$	0,37	1,19	1,26	1,69	$9{,}3 \times 10^{-5}$
Caspase 3	ENSMUSG00000031628	CASP3	Apoptosis Genes	0,00027	48,58	29,55	−0,72	−0,57	0,00045
Fibronectin 1	ENSMUSG00000026193	FN1	EMT Transition	0,00085	1051,50	805,60	−0,38	−0,74	0,00134
B cell CLL/lymphoma 9	ENSMUSG00000038256	BCL9	Wnt Signalling (canonical)	0,00222	15,50	19,20	0,3	0,42	0,00336
Dickkopf homolog 2 (*Xenopus laevis*)	ENSMUSG00000028031	DKK2	Wnt Signalling (canonical)	0,02671	9,31	6,03	−0,65	−0,50	0,03566

LogFC* describes changes between/comparing the mean intensities calculated as the log fold changes of base 2 referring to the TPM-values, logFC^ represents the log fold change of base 2 from the R package EdgeR

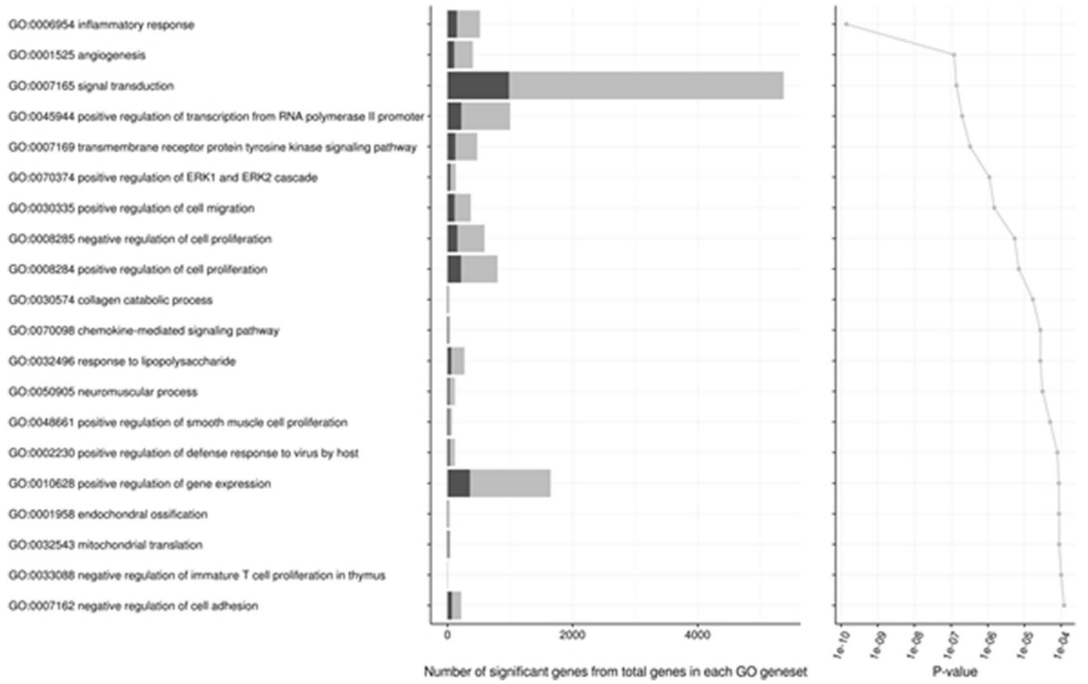

Fig. 5 Bar plot of GO pathways. Top 20 Gene Ontology categories from biological processes comparing the CMT-93 *cell line* with the liver metastases derived thereof. GO categories are depicted and sorted by *p*-value. The *dark grey bar* indicates the number of significant genes from all genes in that GO category (*dark and light grey*). Together with the *light grey* region of the bar, the plot summarises the total number of genes on the pathway

Specifically, MMPs, chemokine receptors, and integrins are predominantly up-regulated in liver metastases derived from CMT-93.

Interactions between carcinoma cells and stromal cells play a vital role during the invasion of the new anatomic metastatic site. Tumour cells must first traverse the basement membrane and then create space for further expansion. Components of the extracellular matrix (ECM) contain a repository of growth factor molecules that can be liberated by proteases secreted by carcinoma tissue. Moreover, the basement membrane also plays crucial roles in signal transduction events within carcinoma cells via pathways initiated by integrin-mediated cell-matrix adhesions, leading to alterations in cell polarity, proliferation, invasiveness, and survival [20]. Additionally, the entry of CRC cells into the hepatic microvasculature can also initiate the pro-inflammatory cascade that results in Kupffer cells being triggered to secrete chemokines [21]. Those are known to up- and also down-regulate various vascular adhesion receptors, thereby enabling adhesion of CRC cells in the microvasculature of the fibroblasts and myofibroblasts, endothelial cells, adipocytes, and various bone-marrow-derived cells – including macrophages and other immune cells [22].

GO analysis is widely recognized as the premier tool in the organization and functional annotation of molecular aspects of cellular systems [23]. We determined the significant GO categories based on a threshold of significance of $p < 0.05$. Our results reveal that the GO terms inflammatory response, angiogenesis, and signal transduction were the most relevant biological processes involved in the propagation of CRC in the liver. Most recently, Becht et al. reported that CRC molecular subgroups and micro-environmental signatures were highly correlated [24]. More precisely, he stated that the mesenchymal subtype of CRC was characterized by a high density of fibroblasts that most likely produces the chemokines and cytokines which favour tumour-associated inflammation and support angiogenesis, resulting in a poor prognosis. Looking into features of assessable inflammatory state in patients, Hamilton et al. were able to link elevated serum levels of C - reactive protein (CRP) with increased circulating pro-inflammatory cytokines. Those patients with colorectal liver metastases were attributed with shorter disease-free and overall survival following surgical resection [25]. Most recently, the inflammatory milieu of CRC liver metastases was used to investigate a new treatment option based on TIE2-

expressing monocytes/macrophages (TEMs), a myeloid cell subset. Adopting the concept of gene transfer, TEMs located in peritumoral sites and exerted an anti-tumour effect through the release of interferon-alpha (IFNα) [26]. Utilizing this strategy in mouse models of CRC liver metastasis, TEMs accumulate in the proximity of hepatic metastatic areas and the TEM-mediated delivery of IFNα inhibits tumour growth. In our study, we could not detect IFNα as being deregulated to a significant level in the liver metastases. However, there were a number of DEGs associated with interferon (e.g. interferon-activated gene 205 (IFI205), interferon-induced transmembrane protein 1 (IFITM1) and interferon gamma inducible protein 30 (IFI30)), which were not members of the top 100 list (data not shown).

Different angiogenic factors have been related to metastasis formation because they promote primary tumour growth and increase the likelihood that tumour cells come into contact with blood and thus disseminate [27]. In particular, the liver is known to be a permissive soil with respect to angiogenesis. The liver parenchyma adjacent to the synchronous liver metastases provides an angiogenically favourable environment for metastatic tumour growth [28]. On the individual level, there was significant correlation between primary CRCs and matched liver metastases with respect to vascular endothelial growth factor (VEGF) mRNA expression. VEGF mRNA levels in patients with two or more liver metastatic tumours were significantly higher than those in patients with only solitary liver metastases [29]. To date, oxaliplatin- and irinotecan-based chemotherapy regimens combined with monoclonal antibody treatment in the form of bevacizumab (anti-VEGF) have proved to be efficient as first-line therapy of metastatic colorectal cancer [30, 31].

Invasion processes are crucial to the formation of liver metastases in CRC and regularly involve a variety of MMPs leading to the degradation and remodelling of the extracellular matrix (ECM) [18, 27]. CRC liver metastases express MMP7 more intensely than normal liver [32]. Our results support the notion that MMP7 is one of the significant players enhancing invasiveness in CRC [33–35]. It is worth noting here that targeted therapy in this matter is difficult. There is evidence in some pre-clinical models that MMP inhibitors (MMPIs) are effective at multiple stages of CRC tumour progression, inhibiting both establishment and growth of primary CRC tumours, as well as reducing metastasis in the lungs and liver [36]. However, clinical trials with MMPIs have been largely unsuccessful as therapeutic agents in CRC so far. A recent study in pre-clinical mouse models of metastatic CRC suggests that ulinastatin (an intrinsic trypsin inhibitor) and natural polyphenol curcumin are capable of inhibiting CRC liver metastases via modulation of MMP9 and E-cadherin expression [37].

The model developed certainly proved to be suitable, as the invasion and expansion of CMT-93 following their injection via the portal vein leading to liver metastasis was reproducible. Our results clearly support the notion of an invasion-metastasis cascade with notable changes to the expression profile of CMT-93 cells on entering and expanding in the liver. Although we were perhaps able to shed some light on the bigger picture, the relative importance of distinct events, interactions, and the molecular drive that all serve to facilitate organ-specific colonisation will require further investigation.

Conclusions

Our work demonstrates that the gene expression in tumour cells is clearly altered during and following the process of metastasis. Here, bioinformatics greatly assists in the analysis of large amounts of data derived from RNA-seq. Through rigorous experimental planning and sophisticated statistical analysis, we are a step closer to elucidating the factors and processes involved during the liver metastasis of CRC. One or more of these dysregulated genes may prove to be a worthy target and enable us effectively to switch off a CRC cell's capacity to act as seed in the formation of metastases. Such a development, at best during the early stages of disease progression, for example prior to the outgrowth of tumour cells within the target soil, could well have a markedly positive effect on the prognosis as well as overall survival of CRC patients.

Additional files

Additional file 1: Supplement 1. List of 119 genes associated with metastasis. The dataset was filtered using this list to identify DEGs between the CMT-93 cell line and liver metastases derived from CMT-93. Gene names and Ensembl IDs are shown in Additional file 1: Supplement 1. (XLSX 11 kb)

Additional file 2: Supplement 2. Count data of relevant genes for the propagation of CMT-93 cells in the liver. Thirty-two DEGs were identified with a threshold of FDR < 5%, of which 23 were up-regulated and 9 down-regulated. Gene names, count data and Ensembl IDs are shown in Additional file 2: Supplement 2. (XLSX 191 kb)

Abbreviations

AEC: 3-amino-9-ethyl-carbazole; BMBF: German Ministry of Education and Research; CPS1: Carbamoyl phosphate synthase1; CRC: Colorectal Cancer; CRP: C-reactive protein; CYP: Cytochrome p450; DEG: Differentially expressed genes; ECM: Extracellular matrix; EdgeR: Empirical Analysis of Digital Gene Expression Data in R; EMT: Epithelial-mesenchymal transition; FBS: Foetal bovine serum; FDR: False discovery rate; GO: Gene ontology; HRP: Horseradish peroxidase; IFI205: Interferon-activated gene 205; IFI30: Interferon gamma inducible protein 30; IFITM1: Interferon-induced transmembrane protein 1; IFNα: Interferon-alpha; LogFC: Log fold change; MMP: Matrix metallopeptidase; MMPIs: MMP inhibitors; PBS: Phosphate-buffered saline; PC: Principal component; PCA: Principal component analysis; PCK1: Phosphoenolpyruvate-carboxykinase 1; R: Programming language; RSEM: RNA-seq by Expectation-Maximization; RT: Room temperature;

TEMs: TIE2-expressing monocytes/macrophages; TPM: Transcripts per million; VEGF: Vascular endothelial growth factor

Acknowledgements
The authors would like to thank Tobias Pukrop and Florian Klemm for their assistance in developing the study design and throughout the process of data retrieval and discussion. They would also like to thank Jetcy Arackal for preparing the CMT-93 cell cultures and Sabine Wolfgramm for her excellent work in preparing the immunolabelling figures for this article. They furthermore express their gratitude to Andrew Entwistle for his critical review of the manuscript.

Ethics approval
All animal breeding, care, and experimentation procedures were in accordance with the German national and regional legislation on animal protection approved by the Lower Saxony State Office for Consumer Protection and Food Safety (Approval/Reference number: 33.12 42,502–04 – 13/1047).

Funding
This work was supported by the ebio initiative of the German Ministry of Education and Research (BMBF), DB and AW were funded by the MetastaSys project (0316173A) within the ebio initiative. This funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Authors' contributions
SK directed the study, reviewed data, and wrote the manuscript; DB designed and performed experiments, as well as contributed to the writing of the manuscript; PK coordinated the study and performed experiments, analysed the data and reviewed the manuscript; CH donated the CMT-93 cells and advised on establishing the syngeneic mouse model of liver metastasis; AB contributed to the development of the study design, GS performed the RNA-seq analysis, AW and TB analysed the data and added to the Methods section of the manuscript. All the authors read and approved the final manuscript.

Competing interests
The authors disclose that there are no conflicts of interest.

Consent for publication
Not applicable.

Publisher's Note

Author details
[1]Department of General, Visceral and Paediatric Surgery, University Medical Centre, Georg – August – University Goettingen, Göttingen, Germany. [2]Statistical Bioinformatics, Department of Medical Statistics, University Medical Centre, Georg – August – University Goettingen, Göttingen, Germany. [3]Microarray and Deep-Sequencing Core Facility, Institute for Developmental Biochemistry, University Medical Centre, Georg – August – University Goettingen, Göttingen, Germany. [4]Department of Haematology and Medical Oncology, University Medical Centre, Georg – August – University Goettingen, Göttingen, Germany. [5]Department of Surgery, University Hospital Regensburg, Regensburg, Germany. [6]Medical Teaching and Medical Education Research, University Hospital Wuerzburg, Julius-Maximilians-University Wuerzburg, Josef-Schneider-Str. 2/D6, 97080 Wuerzburg, Germany.

Received: 8 November 2016 Accepted: 11 May 2017
Published online: 19 May 2017

References

1. Machii R, Saika K. Five-year relative survival rate of colon cancer in the USA, Europe and Japan. Jpn J Clin Oncol. 2014;44(1):105–6.
2. Leporrier J, Maurel J, Chiche L, Bara S, Segol P, Launoy G. A population-based study of the incidence, management and prognosis of hepatic metastases from colorectal cancer. Br J Surg. 2006;93(4):465–74.
3. Paget S. The distribution of secondary growths in cancer of the breast. Cancer Metastasis Rev. 1989;8(2):98–101.
4. Yokota J. Tumor progression and metastasis. Carcinogenesis. 2000;21(3):497–503.
5. Bernards R, Weinberg RA. A progression puzzle. Nature. 2002;418(6900):823.
6. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet. 2003;33(1):49–54.
7. Lim B, Mun J, Kim JH, Kim CW, Roh SA, Cho DH, Kim YS, Kim SY, Kim JC. Genome-wide mutation profiles of colorectal tumors and associated liver metastases at the exome and transcriptome levels. Oncotarget. 2015;6(26):22179–90.
8. Chambers AF, Groom AC, MacDonald IC. Dissemination and growth of cancer cells in metastatic sites. Nat Rev Cancer. 2002;2(8):563–72.
9. Talmadge JE, Fidler IJ. AACR centennial series: the biology of cancer metastasis: historical perspective. Cancer Res. 2010;70(14):5649–69.
10. Clark ME, Smith RR. Liver-directed therapies in metastatic colorectal cancer. J Gastrointest Oncol. 2014;5(5):374–87.
11. Mi K, Kalady MF, Quintini C, Khorana AA. Integrating systemic and surgical approaches to treating metastatic colorectal cancer. Surg Oncol Clin N Am. 2015;24(1):199–214.
12. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
13. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
14. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
15. Alexa A, Rahnenfuhrer J. topGO: topGO: enrichment analysis for Gene ontology. R Packag Version. 2010;2:20,0.
16. Al-Taee KK, Ansari S, Hielscher T, Berger MR, Adwan H. Metastasis-related processes show various degrees of activation in different stages of pancreatic cancer rat liver metastasis. Oncol Res Treat. 2014;37(9):464–70.
17. Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nat Rev Cancer. 2003;3(6):453–8.
18. Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell. 2011;147(2):275–92.
19. Gupta GP, Massague J. Cancer Metastasis: building a framework. Cell. 2006;127(4):679–95.
20. Bissell MJ, Hines WC. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. Nat Med. 2011;17(3):320–9.
21. Wen SW, Ager EI, Christophi C. Bimodal role of Kupffer cells during colorectal cancer liver metastasis. Cancer Biol Ther. 2013;14(7):606–13.
22. Joyce JA, Pollard JW. Microenvironmental regulation of metastasis. Nat Rev Cancer. 2009;9(4):239–52.
23. Lovering RC, Camon EB, Blake JA, Diehl AD. Access to immunology through the Gene ontology. Immunology. 2008;125(2):154–60.
24. Becht E, de Reynies A, Giraldo NA, Pilati C, Buttard B, Lacroix L, Selves J, Sautes-Fridman C, Laurent-Puig P, Fridman WH. Immune and stromal classification of colorectal cancer is associated with molecular subtypes and relevant for precision immunotherapy. Clin Cancer Res. 2016;22(16):4057–66.
25. Hamilton TD, Leugner D, Kopciuk K, Dixon E, Sutherland FR, Bathe OF. Identification of prognostic inflammatory factors in colorectal liver metastases. BMC Cancer. 2014;14:542.
26. Catarinella M, Monestiroli A, Escobar G, Fiocchi A, Tran NL, Aiolfi R, Marra P, Esposito A, Cipriani F, Aldrighetti L, et al. IFNalpha gene/cell therapy curbs colorectal cancer colonization of the liver by acting on the hepatic microenvironment. EMBO Mol Med. 2016;8(2):155–70.
27. Nadal C, Maurel J, Gascon P. Is there a genetic signature for liver metastasis in colorectal cancer? World J Gastroenterol. 2007;13(44):5832–44.
28. van der Wal GE, Gouw AS, Kamps JA, Moorlag HE, Bulthuis ML, Molema G, de Jong KP. Angiogenesis in synchronous and metachronous colorectal liver metastases: the liver as a permissive soil. Ann Surg. 2012;255(1):86–94.
29. Kuramochi H, Hayashi K, Uchida K, Miyakura S, Shimizu D, Vallbohmer D, Park S, Danenberg KD, Takasaki K, Danenberg PV. Vascular endothelial

growth factor messenger RNA expression level is preserved in liver metastases compared with corresponding primary colorectal cancer. Clin Cancer Res. 2006;12(1):29–33.
30. Kocakova I, Melichar B, Kocak I, Bortlicek Z, Buchler T, Dusek L, Petruzelka L, Kohoutek M, Prausova J, Finek J, et al. Bevacizumab with FOLFIRI or XELIRI in the first-line therapy of metastatic colorectal carcinoma: results from Czech observational registry. Anticancer Res. 2015;35(6):3455–61.
31. Yamazaki K, Nagase M, Tamagawa H, Ueda S, Tamura T, Murata K, Eguchi Nakajima T, Baba E, Tsuda M, Moriwaki T, et al. Randomized phase III study of bevacizumab plus FOLFIRI and bevacizumab plus mFOLFOX6 as first-line treatment for patients with metastatic colorectal cancer (WJOG4407G). Ann Oncol. 2016;27(8):1539–46.
32. Zeng ZS, Shu WP, Cohen AM, Guillem JG. Matrix metalloproteinase-7 expression in colorectal cancer liver metastases: evidence for involvement of MMP-7 activation in human cancer metastases. Clin Cancer Res. 2002;8(1):144–8.
33. Lee SK, Han YM, Yun J, Lee CW, Shin DS, Ha YR, Kim J, Koh JS, Hong SH, Han DC, et al. Phosphatase of regenerating liver-3 promotes migration and invasion by upregulating matrix metalloproteinases-7 in human colorectal cancer cells. Int J Cancer. 2012;131(3):E190–203.
34. Ochiai H, Nakanishi Y, Fukasawa Y, Sato Y, Yoshimura K, Moriya Y, Kanai Y, Watanabe M, Hasegawa H, Kitagawa Y, et al. A new formula for predicting liver metastasis in patients with colorectal cancer: immunohistochemical analysis of a large series of 439 surgically resected cases. Oncology. 2008; 75(1–2):32–41.
35. Fang YJ, Lu ZH, Wang GQ, Pan ZZ, Zhou ZW, Yun JP, Zhang MF, Wan DS. Elevated expressions of MMP7, TROP2, and survivin are associated with survival, disease recurrence, and liver metastasis of colon cancer. Int J Color Dis. 2009;24(8):875–84.
36. Wagenaar-Miller RA, Gorden L, Matrisian LM. Matrix metalloproteinases in colorectal cancer: is it worth talking about? Cancer Metastasis Rev. 2004; 23(1–2):119–35.
37. Shen F, Cai WS, Li JL, Feng Z, Liu QC, Xiao HQ, Cao J, Xu B. Synergism from the combination of ulinastatin and curcumin offers greater inhibition against colorectal cancer liver metastases via modulating matrix metalloproteinase-9 and E-cadherin expression. Onco Targets Ther. 2014; 7:305–14.

4 Outlook

The field of transcriptomics is another step to further understanding the mechanisms in cells. The possibilities of performing a variety of analysis strategies are rapidly increasing over time and the newest advancements for mutation calling further support this, thus becoming more efficient and cheaper. The three publications included in this dissertation aim at contributing towards a better understanding of cell mechanisms, especially for cancer, by providing an optimised standard of analysis which is up to date with current research, add the benefits of mutation calling, typically done in Exome-Seq, and apply both on real datasets, to find biomarkers linking to possible drug responses.

The first study assessed the use of RNA-Seq and microarray analysis. Two datasets were used: human RC patient data metastasised into liver and cell lines from Burkitt's lymphoma for the microarray and RNA-Seq platform respectively. Moreover, multiple methods for RNA-Seq analysis were compared, which resulted in the recommendation of STAR together with RSEM, followed by edgeR and topGO.

The second study consisted of the development of Wileup, a tool for a cost-effective identification of single nucleotide variants (SNVs) in tumour-only RNA-Seq samples. The publication assessed the use case of checking a panel of mutations and had equal results than state of the art tools. Furthermore, it is possible to detect unannotated variations as well, despite increasing the chance of finding a false positive mutation over the panel mode in Wileup, as they are not backed up with annotation data. Improvements for the future would be the detection of RNA-editing events (Gott and Emeson, 2000), resulting in false positive SNVs, and indels. The third publication (3.3) was an application of the resulting pipeline from paper one on real data from mice developing metastasis in liver from colorectal cancer. The bioinformatical approaches condensed via the pipeline revealed essential mechanisms behind the expression of metastasis enhancing genes supporting a better biological interpretation.

In conclusion, the presented studies in this thesis contribute to the highly ongoing topic of transcriptomics with the focus on DEA and mutation calling in RNA-Seq. All related findings in the RNA-Seq and Exome-Seq data would benefit from further studies and validations. Finally, the provided software tool should be further improved and upgraded according to future changing user requirements and technology development. Together with the establishment of a standard analysis of RNA-Seq data as well as the newly emerging field of mutation for RNA-Seq as an addition next to Exome-Seq, this dissertation is a valuable

addition to strategies tackling the problem of finding new biomarkers and corresponding drug responses in the research field of cancer.

Herewith I declare, that I prepared this PhD thesis on my own and with no other sources and aids than quoted.

Göttingen, November, 2018 Alexander Wolff

Lebenslauf

Alexander Wolff

Persönliche Daten

Ahlestr. 41
99974 Müehlhausen
Deutschland
E-Mail: alexander.wolff@med.uni-goettingen.de

Geburtsdatum: 6. März 1987
Geburtsort: Dresden
Familienstand: verheiratet, 2 Kinder (3 Jahre, 2 Monate)

Bildung und Forschung

02/2013 – jetzt **Doktorand der Bioinformatik an der Universitätsmedizin in Göttingen, angestellt im Projekt MetastaSys, Göttingen, Deutschland**
PhD thesis 'Analysis of expression profile and gene variation via development of methods for Next Generation Sequencing data'

10/2007 – 06/2012 **Studium der Bioinformatik an der Friedrich-Schiller-Universität Jena (Abschluss: 2,1)**
Diplomarbeit: 'Entwicklung einer Prozessierungspipeline zur Analyse und Klassifikation humaner genomischer Varianten'
Zentrum für Bioinformatik, Hamburg

Berufserfahrung

02/2013 – jetzt **Wissenschaftlicher Mitarbeiter, Statistische Bioinformatik Gruppe**
Institut für Medizinische Statistik, Universitätsmedizin Göttingen
Göttingen, Deutschland

03/2017 – jetzt **Wissenschaftlicher Mitarbeiter, Zentrale Serviceeinheit Medizinische Biometrie und Statistische Bioinformatik**
Institut für Medizinische Statistik, Universitätsmedizin Göttingen
Göttingen, Deutschland

Veröffentlichungen

Wolff A, Bayerlová M, Gaedcke J, Kube D, Beißbarth T. A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal- cancer patients and Burkitt Lymphoma cells, PLOS ONE 13, e0197162 (2018)., doi: 10.1371/journal.pone.0197162

Wolff A, Perera-Bel J, Schildhaus HU, Homayounfar K, Schatlo B, Bleckmann A, Beißbarth T. Using RNA-Seq Data for the Detection of a Panel of Clinically Relevant Mutations. Stud Health Technol Inform 253, 217?221 (2018). doi: 10.3233/978-1-61499-896-9-217

Bocuk D, **Wolff A**, Krause P, Salinas G, Bleckmann A, Hackl C, et al. (2017). The adaptation of colorectal cancer cells when forming metastases in the liver: expression of associated genes and pathways in a mouse model. BMC Cancer. 2017;17. doi:10.1186/s12885-017-3342-1

Raquel Blazquez, Darius Wlochowitz, **Alexander Wolff**, Stefanie Seitz, Astrid Wachter, Julia Perera, Annalen Bleckmann, Tim BeiSSbarth, Gabriela Salinas, Markus Riemenschneider, Martin Proescholdt, Matthias Evert, Laila Siam, Bawarjan Schatlo, Christine Stadelmann, Hans-Ulrich Schildhaus, Ulrike Korf, Eileen Reinz, Stefan Wiemann, Elena Vollmer, Mathias Schulz, Uwe Ritter, Uwe K. Hanisch, Tobias Pukrop (2018) PI3K is a master regulator of metastasis-promoting macrophages/microglia during CNS colonization of breast cancer cells GLIA. 2018. doi: 10.1002/glia.23485

Wlochowitz D, Haubrock M, Arackal J, Bleckmann A, **Wolff A**, Beißbarth T, et al. (2017) Computational Identification of Key Regulators in Two Different Colorectal Cancer Cell Lines. Front Genet. 2016;7. doi:10.3389/fgene.2016.00042

Bayerlová M, Menck K, Klemm F, **Wolff A**, Pukrop T, Binder C, et al. (2017) Ror2 Signaling and Its Relevance in Breast Cancer Progression. Front Oncol. 2017;7. doi:10.3389/fonc.2017.00135

von der Heyde S, Wagner S, **Czerny A**, Nietert M, Ludewig F, Salinas-Riester G, Arlt D, BeiSSbarth T. (2015) mRNA profiling reveals determinants of trastuzumab efficiency in HER2- positive breast cancer. Plos One 2015, doi: 10(2):e0117818.

References

MD Adams, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merril, A Wu, B Olde, RF Moreno, and al. et. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651, June 1991. doi: 10.1126/science.2047873. URL `http://science.sciencemag.org/content/252/5013/1651.abstract`.

Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–249, April 2010. ISSN 1548-7091. doi: 10.1038/nmeth0410-248. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2855889/`.

Ashish Agarwal, David Koppstein, Joel Rozowsky, Andrea Sboner, Lukas Habegger, LaDeana W Hillier, Rajkumar Sasidharan, Valerie Reinke, Robert H Waterston, and Mark Gerstein. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, 11(1):383, 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-383. URL `http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-383`.

Cornelis A. Albers, Gerton Lunter, Daniel G. MacArthur, Gilean McVean, Willem H. Ouwehand, and Richard Durbin. Dindel: Accurate indel calls from short-read data. *Genome Res*, 21(6):961–973, June 2011. ISSN 1088-9051. doi: 10.1101/gr.112326.110. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3106329/`.

Adrian Alexa and Jorg Rahnenfuhrer. *topGO: topGO: Enrichment analysis for Gene Ontology*. 2010. 00002 R package version 2.22.0.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, October 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

J. C. Alwine, D. J. Kemp, and G. R. Stark. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences*, 74(12):5350–5354, December 1977. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.74.12.5350. URL http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5350.

S. Anders, P. T. Pyl, and W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu638. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu638. 01608.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, October 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-10-r106. URL https://doi.org/10.1186/gb-2010-11-10-r106.

S. Andrews. FastQC A quality control application for high throughput sequence data, 2010. URL http://www.bioinformatics.babraham.ac.uk/projects/download.html. 00000.

ASCO. What is Cancer?, August 2012. URL https://www.cancer.net/navigating-cancer-care/cancer-basics/what-cancer.

Paul L. Auer and R. W. Doerge. Statistical Design and Analysis of RNA Sequencing Data. *Genetics*, 185(2):405–416, June 2010. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.110.114983. URL http://www.genetics.org/content/185/2/405.

A Bairoch and R Apweiler. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res*, 26(1):38–42, January 1998. ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC147215/.

Riyue Bao, Lei Huang, Jorge Andrade, Wei Tan, Warren A. Kibbe, Hongmei Jiang, and Gang Feng. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Cancer Informatics*, 13s2:CIN.S13779, January 2014. ISSN 1176-9351, 1176-

9351. doi: 10.4137/CIN.S13779. URL http://journals.sagepub.com/doi/10.4137/CIN.S13779.

Alex Bateman, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Andrew Cowley, Alan Da Silva, Maurizio De Giorgi, Tunca Dogan, Francesco Fazzini, Leyla Garcia Castro, Luis Figueira, Penelope Garmiri, George Georghiou, Daniel Gonzalez, Emma Hatton-Ellis, Weizhong Li, Wudong Liu, Rodrigo Lopez, Jie Luo, Yvonne Lussi, Alistair MacDougall, Andrew Nightingale, Barbara Palka, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Alexandre Renaux, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Elena Speretta, Edward Turner, Nidhi Tyagi, Vladimir Volynkin, Tony Wardell, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard de Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh, and Jian Zhang. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, 45 (D1):D158–D169, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1099. URL https://academic.oup.com/nar/article/45/D1/D158/2605721.

Daniel E. Bauer, Georgia Hatzivassiliou, Fangping Zhao, Charalambos Andreadis, and Craig B. Thompson. ATP citrate lyase is an important component of cell growth and transformation. *Oncogene*, 24(41):6314–6322, September 2005. ISSN 1476-5594. doi: 10.1038/sj.onc.1208773. URL https://www.nature.com/articles/1208773.

T. Beissbarth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, June 2004. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bth088. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth088.

Tim Beissbarth. Interpreting experimental results using gene ontologies. *Meth. Enzymol.*, 411:340–352, 2006. ISSN 0076-6879. doi: 10.1016/S0076-6879(06)11018-6.

Aziz Belkadi, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B. Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean-Laurent Casanova, and Laurent Abel. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*, 112(17):5473–5478, April 2015. ISSN 0027-8424. doi: 10.1073/pnas.1418631112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4418901/.

M V Berridge and A S Tan. Interleukin-3 facilitates glucose transport in a myeloid cell line by regulating the affinity of the glucose transporter for glucose: involvement of protein phosphorylation in transporter activation. *Biochem J*, 305(Pt 3):843–851, February 1995. ISSN 0264-6021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1136336/.

Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu170. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/.

Alan P. Boyle, Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, Julie Park, Benjamin C. Hitz, Shuai Weng, J. Michael Cherry, and Michael Snyder. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*, 22(9):1790–1797, September 2012. ISSN 1088-9051. doi: 10.1101/gr.137323.112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431494/.

James R. Bradford, Yvonne Hey, Tim Yates, Yaoyong Li, Stuart D. Pepper, and Crispin J. Miller. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics*, 11(1):282, 2010. URL http://www.biomedcentral.com/1471-2164/11/282. 00110.

Daniel Branton, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, Stevan B Jovanovich, Predrag S Krstic, Stuart Lindsay, Xinsheng Sean Ling, Carlos H Mastrangelo, Amit Meller, John S Oliver, Yuriy V Pershin, J Michael Ramsey, Robert Riehn, Gautam V Soni, Vincent Tabard-Cossa, Meni Wanunu, Matthew Wiggin, and Jeffery A Schloss. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26:1146, 2008. URL http://dx.doi.org/10.1038/nbt.1495.

Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34 (5):525–527, May 2016. ISSN 1546-1696. doi: 10.1038/nbt.3519. URL https://www.nature.com/articles/nbt.3519.

Anthony J. Brookes. The essence of SNPs. *Gene*, 234(2):177–186, July 1999. ISSN 0378-1119. doi: 10.1016/S0378-1119(99)00219-X. URL http://www.sciencedirect.com/science/article/pii/S037811199900219X.

James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, February 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-94. URL https://doi.org/10.1186/1471-2105-11-94.

Roger Bumgarner. DNA microarrays: Types, Applications and their future. *Curr Protoc Mol Biol*, 0 22:Unit–22.1., January 2013. ISSN 1934-3639. doi: 10.1002/0471142727.mb2201s101. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4011503/.

M Burrows and David J Wheeler. A block-sorting lossless data compression algorithm. *CA: Digital Equipment Corporation*, page 24, 1994.

Christopher R. Cabanski, Keary Cavin, Chris Bizon, Matthew D. Wilkerson, Joel S. Parker, Kirk C. Wilhelmsen, Charles M. Perou, JS Marron, and

D. Neil Hayes. ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data. *BMC Bioinformatics*, 13(1): 221, September 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-221. URL https://doi.org/10.1186/1471-2105-13-221.

Christina T. L. Chen, Jen C. Wang, and Barak A. Cohen. The Strength of Selection on Ultraconserved Elements in the Human Genome. *Am J Hum Genet*, 80(4):692–704, April 2007. ISSN 0002-9297. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1852725/.

Kristian Cibulskis, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31 (3):213–219, March 2013. ISSN 1546-1696. doi: 10.1038/nbt.2514. URL https://www.nature.com/articles/nbt.2514.

Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*, 6(2):80–92, April 2012. ISSN 1933-6934. doi: 10.4161/fly.19695. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679285/.

Elizabeth T Cirulli, Abanish Singh, Kevin V Shianna, Dongliang Ge, Jason P Smith, Jessica M Maia, Erin L Heinzen, James J Goedert, and David B Goldstein. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. page 8, 2010.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biol*, 17, 2016. ISSN 1474-7596. doi: 10.1186/s13059-016-0881-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/.

Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227:561, August 1970. URL http://dx.doi.org/10.1038/227561a0.

Ralph J. DeBerardinis, Julian J. Lum, Georgia Hatzivassiliou, and Craig B. Thompson. The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation. *Cell Metabolism*, 7(1):11–20, January 2008. ISSN 1550-4131. doi: 10.1016/j.cmet.2007.10.002. URL `http://www.sciencedirect.com/science/article/pii/S1550413107002951`.

M.A. DePristo, E. Banks, R.E. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A Rivas, M. Hanna, A. McKenna, T.J. Fennell, A.M. Kernytsky, A.Y. Sivachenko, K. Cibulskis, S.B. Gabriel, D. Altshuler, and M.J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43 (5):491–498, May 2011. ISSN 1061-4036. doi: 10.1038/ng.806. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/`.

Luis A. Diaz, Richard T. Williams, Jian Wu, Isaac Kinde, J. Randolph Hecht, Jordan Berlin, Benjamin Allen, Ivana Bozic, Johannes G. Reiter, Martin A. Nowak, Kenneth W. Kinzler, Kelly S. Oliner, and Bert Vogelstein. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, 486(7404):537–540, June 2012. ISSN 1476-4687. doi: 10.1038/nature11219.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts635. URL `http://bioinformatics.oxfordjournals.org/content/29/1/15`. 00326.

Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, and Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods*, 10(12):1185–1191, December 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2722. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4018468/`.

Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, October 2016. ISSN 1367-4803. doi: 10.

1093/bioinformatics/btw354. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5039924/.

Runhua Feng, Xuehua Chen, Yingyan Yu, Liping Su, Beiqin Yu, Jianfang Li, Qu Cai, Min Yan, Bingya Liu, and Zhenggang Zhu. miR-126 functions as a tumour suppressor in human gastric cancer. *Cancer Lett.*, 298(1):50–63, December 2010. ISSN 1872-7980. doi: 10.1016/j.canlet.2010.06.004.

Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*, 44(D1):D279–D285, January 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1344. URL https://academic.oup.com/nar/article/44/D1/D279/2503120.

S.A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J.W. Teague, P.A. Futreal, and M.R. Stratton. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet*, CHAPTER:Unit–10.11, April 2008. ISSN 1934-8266. doi: 10.1002/0471142905.hg1011s57. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705836/.

Daniel A. Galvão and Robert U. Newton. Review of Exercise Intervention Studies in Cancer Patients. *JCO*, 23(4):899–909, February 2005. ISSN 0732-183X. doi: 10.1200/JCO.2005.06.085. URL http://ascopubs.org/doi/full/10.1200/JCO.2005.06.085.

Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*, July 2012. URL http://arxiv.org/abs/1207.3907. arXiv: 1207.3907.

M. Gerlinger and C. Swanton. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *British Journal of Cancer*, 103(8):1139–1143, October 2010. ISSN 1532-1827. doi: 10.1038/sj.bjc.6605912. URL https://www.nature.com/articles/6605912.

M. E. Gore and J. M. G. Larkin. Challenges and opportunities for converting renal cell carcinoma into a chronic disease with targeted therapies. *Br. J. Cancer*, 104(3):399–406, February 2011. ISSN 1532-1827. doi: 10.1038/sj.bjc.6606084.

J. M. Gott and R. B. Emeson. Functions and mechanisms of RNA editing. *Annu. Rev. Genet.*, 34:499–531, 2000. ISSN 0066-4197. doi: 10.1146/annurev.genet.34.1.499.

Rodrigo Goya, Mark G.F. Sun, Ryan D. Morin, Gillian Leung, Gavin Ha, Kimberley C. Wiegand, Janine Senz, Anamaria Crisan, Marco A. Marra, Martin Hirst, David Huntsman, Kevin P. Murphy, Sam Aparicio, and Sohrab P. Shah. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–736, March 2010. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btq040. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq040`.

I. C. Gray, D. A. Campbell, and N. K. Spurr. Single nucleotide polymorphisms as tools in human genetics. *Hum. Mol. Genet.*, 9(16):2403–2408, October 2000. ISSN 0964-6906.

Mel Greaves and Carlo C. Maley. CLONAL EVOLUTION IN CANCER. *Nature*, 481(7381):306–313, January 2012. ISSN 0028-0836. doi: 10.1038/nature10762. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367003/`.

D. Greenbaum. Interrelating Different Types of Genomic Data, from Proteome to Secretome: 'Oming in on Function. *Genome Research*, 11(9):1463–1468, September 2001. ISSN 10889051. doi: 10.1101/gr.207401. URL `http://www.genome.org/cgi/doi/10.1101/gr.207401`.

Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, Erica K Barnell, Alex H Wagner, Zachary L Skidmore, Amber Wollam, Connor J Liu, Martin R Jones, Rachel L Bilski, Robert Lesurf, Yan-Yang Feng, Nakul M Shah, Melika Bonakdar, Lee Trani, Matthew Matlock, Avinash Ramu, Katie M Campbell, Gregory C Spies, Aaron P Graubert, Karthik Gangavarapu, James M Eldred, David E Larson, Jason R Walker, Benjamin M Good, Chunlei Wu, Andrew I Su, Rodrigo Dienstmann, Adam A Margolin, David Tamborero, Nuria Lopez-Bigas, Steven J M Jones, Ron Bose, David H Spencer, Lukas D Wartman, Richard K Wilson, Elaine R Mardis, and Obi L Griffith. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat*

Genet, 49(2):170–174, January 2017. ISSN 1061-4036. doi: 10.1038/ng.3774. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5367263/.

Lab Hannon. FASTX-Toolkit, 2009. URL http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):e131–e131, July 2010. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gkq224. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq224.

C. J. Hedeskov. Early effects of phytohaemagglutinin on glucose metabolism of normal human lymphocytes. *Biochem J*, 110(2):373–380, November 1968. ISSN 0264-6021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1187214/.

Gloria H. Heppner and Bonnie E. Miller. Tumor heterogeneity: biological implications and therapeutic consequences. *Cancer Metast Rev*, 2(1):5–23, March 1983. ISSN 1573-7233. doi: 10.1007/BF00046903. URL https://doi.org/10.1007/BF00046903.

Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M. Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. *PLoS Comput Biol*, 5(9), September 2009. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1000502. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2730575/.

C. A. Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35(18):6227–6237, August 2007. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkm688. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm688.

Géraldine Jean, André Kahles, Vipin T. Sreedharan, Fabio De Bona, and Gunnar Rätsch. RNA-Seq Read Alignments with PALMapper. *Current Protocols in Bioinformatics*, 32(1):11.6.1–11.6.37, December 2010. ISSN 1934-340X. doi: 10.1002/0471250953.bi1106s32. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1106s32.

W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, January 2007. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj037. URL https://academic.oup.com/biostatistics/article/8/1/118/252073.

Philipp Kapranov, Aarron T. Willingham, and Thomas R. Gingeras. Genome-wide transcription and the implications for genomic organization. *Nature Reviews Genetics*, 8:413, 2007. URL http://dx.doi.org/10.1038/nrg2083.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14: R36, April 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-4-r36. URL https://doi.org/10.1186/gb-2013-14-4-r36.

Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12:357, 2015. URL http://dx.doi.org/10.1038/nmeth.3317.

Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22(3):568–576, March 2012. ISSN 1088-9051. doi: 10.1101/gr.129684.111. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3290792/.

Gregory V. Kryukov, Steffen Schmidt, and Shamil Sunyaev. Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet*, 14 (15):2221–2229, August 2005. ISSN 0964-6906. doi: 10.1093/hmg/ddi226. URL https://academic.oup.com/hmg/article/14/15/2221/551730.

Pui-Yan Kwok and Xiangning Chen. Detection of Single Nucleotide Polymorphisms. page 19, 2003.

Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, 42(Database issue):D980–D985, January 2014. ISSN

0305-1048. doi: 10.1093/nar/gkt1113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965032/.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, March 2012. ISSN 1548-7091. doi: 10.1038/nmeth.1923. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, March 2009. ISSN 1474-760X. doi: 10.1186/gb-2009-10-3-r25. URL https://doi.org/10.1186/gb-2009-10-3-r25.

David E. Larson, Christopher C. Harris, Ken Chen, Daniel C. Koboldt, Travis E. Abbott, David J. Dooling, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, February 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr665. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3268238/.

Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, 15(2):R29, 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-2-r29. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053721/.

Wan-Ping Lee, Michael P. Stromberg, Alistair Ward, Chip Stewart, Erik P. Garrison, and Gabor T. Marth. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *PLoS One*, 9 (3), March 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0090581. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3944147/.

J. Leporrier, J. Maurel, L. Chiche, S. Bara, P. Segol, and G. Launoy. A population-based study of the incidence, management and prognosis of hepatic metastases from colorectal cancer. *BJS*, 93(4):465–474, April 2006. ISSN 1365-2168. doi: 10.1002/bjs.5278. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bjs.5278.

Bo Li and Colin N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12

(1):323, 2011. URL http://www.biomedcentral.com/1471-2105/12/323/. 00614.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705234/.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/.

Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, August 2009b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp336. URL https://academic.oup.com/bioinformatics/article/25/15/1966/212427.

Yang Liao, Gordon K. Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt656. URL https://academic.oup.com/bioinformatics/article/30/7/923/232889.

Hao Lin, Zefeng Zhang, Michael Q. Zhang, Bin Ma, and Ming Li. ZOOM! Zillions of oligos mapped. *Bioinformatics*, 24(21):2431–2437, November 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn416. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732274/.

Lance A. Liotta and Elise C. Kohn. *Invasion and Metastases*. BC Decker, 2000. URL https://www.ncbi.nlm.nih.gov/books/NBK20786/.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL https://doi.org/10.1186/s13059-014-0550-8.

Ryoko Machii and Kumiko Saika. Five-Year Relative Survival Rate of Colon Cancer in the USA, Europe and Japan. *Jpn J Clin Oncol*, 44(1):105–106,

January 2014. ISSN 0368-2811. doi: 10.1093/jjco/hyt227. URL https://academic.oup.com/jjco/article/44/1/105/872378.

Fabienne Mackay and Jeffrey L. Browning. BAFF: A fundamental survival factor for B cells. *Nature Reviews Immunology*, 2(7):465–475, July 2002. ISSN 14741741. doi: 10.1038/nri844. URL http://www.nature.com/doifinder/10.1038/nri844. 00535.

John H Malone and Brian Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1):34, 2011. ISSN 1741-7007. doi: 10.1186/1741-7007-9-34. URL http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-9-34.

J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, July 2008. ISSN 1088-9051. doi: 10.1101/gr.079558.108. URL http://www.genome.org/cgi/doi/10.1101/gr.079558.108. 01481.

Tracey A. Martin, Lin Ye, Andrew J. Sanders, Jane Lane, and Wen G. Jiang. *Cancer Invasion and Metastasis: Molecular and Cellular Perspective*. Landes Bioscience, 2013. URL https://www.ncbi.nlm.nih.gov/books/NBK164700/.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, September 2010. ISSN 1088-9051. doi: 10.1101/gr.107524.110. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.107524.110.

Michael L. Metzker. Emerging technologies in DNA sequencing. *Genome Res.*, 15(12):1767–1776, December 2005. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.3770505. URL http://genome.cshlp.org/content/15/12/1767.

Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, January 2010. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2626. URL http://www.nature.com/doifinder/10.1038/nrg2626. 03400.

André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol*, 12(11):R112, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-11-r112. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334598/`.

S. H. Nagaraj, R. B. Gasser, and S. Ranganathan. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*, 8(1):6–21, May 2006. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbl015. URL `https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbl015`.

Sarah B. Ng, Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E. Eichler, Michael Bamshad, Deborah A. Nickerson, and Jay Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, September 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature08250. URL `http://www.nature.com/doifinder/10.1038/nature08250`.

Sarah B. Ng, Kati J. Buckingham, Choli Lee, Abigail W. Bigham, Holly K. Tabor, Karin M. Dent, Chad D. Huff, Paul T. Shannon, Ethylin Wang Jabs, Deborah A. Nickerson, Jay Shendure, and Michael J. Bamshad. Exome sequencing identifies the cause of a Mendelian disorder. *Nat Genet*, 42 (1):30–35, January 2010. ISSN 1061-4036. doi: 10.1038/ng.499. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847889/`.

P. C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194 (4260):23–28, October 1976. ISSN 0036-8075.

Maria j Nueda, Alberto Ferrer, and Ana Conesa. ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics*, 13(3):553–566, July 2012. ISSN 1465-4644. doi: 10.1093/biostatistics/kxr042. URL `https://academic.oup.com/biostatistics/article/13/3/553/248541`.

Fatih Ozsolak and Patrice M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12:87, 2010. URL `http://dx.doi.org/10.1038/nrg2934`.

Rob Patro, Stephen M. Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, May 2014. ISSN 1546-1696. doi: 10.1038/nbt.2862. URL https://www.nature.com/articles/nbt.2862.

Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods*, 14(4):417–419, April 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4197. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5600148/.

Emma M. Quinn, Paul Cormican, Elaine M. Kenny, Matthew Hill, Richard Anney, Michael Gill, Aiden P. Corvin, and Derek W. Morris. Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLOS ONE*, 8(3):e58815, March 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0058815. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0058815.

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/.

Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, 59(1):5–15, January 2014. ISSN 1435-232X. doi: 10.1038/jhg.2013.114. URL https://www.nature.com/articles/jhg2013114.

Hugues Richard, Marcel H. Schulz, Marc Sultan, Asja Nürnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A. Haas, and Marie-Laure Yaspo. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, 38(10):e112–e112, June 2010. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gkq041. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq041.

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, April

2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/.

Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, January 2013. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2251. URL http://www.nature.com/articles/nmeth.2251.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp616. URL http://bioinformatics.oxfordjournals.org/content/26/1/139. 01889.

Christopher T. Saunders, Wendy S. W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, July 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts271. URL https://academic.oup.com/bioinformatics/article/28/14/1811/218573.

Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, October 1995. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.270.5235.467. URL http://science.sciencemag.org/content/270/5235/467.

Roanne J. Segal, Robert D. Reid, Kerry S. Courneya, Shawn C. Malone, Matthew B. Parliament, Chris G. Scott, Peter M. Venner, H. Arthur Quinney, Lee W. Jones, Monika E. Slovinec D'Angelo, and George A. Wells. Resistance Exercise in Men Receiving Androgen Deprivation Therapy for Prostate Cancer. *JCO*, 21(9):1653–1659, May 2003. ISSN 0732-183X. doi: 10.1200/JCO.2003.09.534. URL http://ascopubs.org/doi/abs/10.1200/JCO.2003.09.534.

A. L. Shaffer, W. Wojnar, and W. Nelson. Amplification, detection, and automated sequencing of gibbon interleukin-2 mRNA by Thermus aquaticus DNA polymerase reverse transcription and polymerase chain reaction. *Analytical Biochemistry*, 190(2):292–296, November 1990. ISSN 0003-2697. doi: 10.1016/0003-2697(90)90196-G. URL http://www.sciencedirect.com/science/article/pii/000326979090196G.

Neil P. Shah, John M. Nicoll, Bhushan Nagar, Mercedes E. Gorre, Ronald L. Paquette, John Kuriyan, and Charles L. Sawyers. Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell*, 2(2):117–125, August 2002. ISSN 1535-6108. doi: 10.1016/S1535-6108(02)00096-X. URL `http://www.sciencedirect.com/science/article/pii/S153561080200096X`.

S. T. Sherry, M. Ward, and K. Sirotkin. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, 9 (8):677–679, August 1999. ISSN 1088-9051.

Leming Shi, Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise de Longueville, Ernest S Kawasaki, Kathleen Y Lee, Yuling Luo, Yongming Andrew Sun, James C Willey, Robert A Setterquist, Gavin M Fischer, Weida Tong, Yvonne P Dragan, David J Dix, Felix W Frueh, Federico M Goodsaid, Damir Herman, Roderick V Jensen, Charles D Johnson, Edward K Lobenhofer, Raj K Puri, Uwe Scherf, Jean Thierry-Mieg, Charles Wang, Mike Wilson, Paul K Wolber, Lu Zhang, Shashi Amur, Wenjun Bao, Catalin C Barbacioru, Anne Bergstrom Lucas, Vincent Bertholet, Cecilie Boysen, Bud Bromley, Donna Brown, Alan Brunner, Roger Canales, Xiaoxi Megan Cao, Thomas A Cebula, James J Chen, Jing Cheng, Tzu-Ming Chu, Eugene Chudin, John Corson, J Christopher Corton, Lisa J Croner, Christopher Davies, Timothy S Davison, Glenda Delenstarr, Xutao Deng, David Dorris, Aron C Eklund, Xiao-hui Fan, Hong Fang, Stephanie Fulmer-Smentek, James C Fuscoe, Kathryn Gallagher, Weigong Ge, Lei Guo, Xu Guo, Janet Hager, Paul K Haje, Jing Han, Tao Han, Heather C Harbottle, Stephen C Harris, Eli Hatchwell, Craig A Hauser, Susan Hester, Huixiao Hong, Patrick Hurban, Scott A Jackson, Hanlee Ji, Charles R Knight, Winston P Kuo, J Eugene LeClerc, Shawn Levy, Quan-Zhen Li, Chunmei Liu, Ying Liu, Michael J Lombardi, Yunqing Ma, Scott R Magnuson, Botoul Maqsodi, Tim McDaniel, Nan Mei, Ola Myklebost, Baitang Ning, Natalia Novoradovskaya, Michael S Orr, Terry W Osborn, Adam Papallo, Tucker A Patterson, Roger G Perkins, Elizabeth H Peters, Ron Peterson, Kenneth L Philips, P Scott Pine, Lajos Pusztai, Feng Qian, Hongzu Ren, Mitch Rosen, Barry A Rosenzweig, Raymond R Samaha, Mark Schena, Gary P Schroth, Svetlana Shchegrova, Dave D Smith, Frank Staedtler,

Zhenqiang Su, Hongmei Sun, Zoltan Szallasi, Zivana Tezak, Danielle Thierry-Mieg, Karol L Thompson, Irina Tikhonova, Yaron Turpaz, Beena Vallanat, Christophe Van, Stephen J Walker, Sue Jane Wang, Yonghong Wang, Russ Wolfinger, Alex Wong, Jie Wu, Chunlin Xiao, Qian Xie, Jun Xu, Wen Yang, Liang Zhang, Sheng Zhong, Yaping Zong, and William Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24 (9):1151–1161, September 2006. ISSN 1087-0156. doi: 10.1038/nbt1239. URL `http://www.nature.com/doifinder/10.1038/nbt1239`. 01611.

Leming Shi, Gregory Campbell, Wendell D Jones, Fabien Campagne, Zhining Wen, Stephen J Walker, Zhenqiang Su, Tzu-Ming Chu, Federico M Goodsaid, Lajos Pusztai, John D Shaughnessy, André Oberthuer, Russell S Thomas, Richard S Paules, Mark Fielden, Bart Barlogie, Weijie Chen, Pan Du, Matthias Fischer, Cesare Furlanello, Brandon D Gallas, Xijin Ge, Dalila B Megherbi, W Fraser Symmans, May D Wang, John Zhang, Hans Bitter, Benedikt Brors, Pierre R Bushel, Max Bylesjo, Minjun Chen, Jie Cheng, Jing Cheng, Jeff Chou, Timothy S Davison, Mauro Delorenzi, Youping Deng, Viswanath Devanarayan, David J Dix, Joaquin Dopazo, Kevin C Dorff, Fathi Elloumi, Jianqing Fan, Shicai Fan, Xiaohui Fan, Hong Fang, Nina Gonzaludo, Kenneth R Hess, Huixiao Hong, Jun Huan, Rafael A Irizarry, Richard Judson, Dilafruz Juraeva, Samir Lababidi, Christophe G Lambert, Li Li, Yanen Li, Zhen Li, Simon M Lin, Guozhen Liu, Edward K Lobenhofer, Jun Luo, Wen Luo, Matthew N McCall, Yuri Nikolsky, Gene A Pennello, Roger G Perkins, Reena Philip, Vlad Popovici, Nathan D Price, Feng Qian, Andreas Scherer, Tieliu Shi, Weiwei Shi, Jaeyun Sung, Danielle Thierry-Mieg, Jean Thierry-Mieg, Venkata Thodima, Johan Trygg, Lakshmi Vishnuvajjala, Sue Jane Wang, Jianping Wu, Yichao Wu, Qian Xie, Waleed A Yousef, Liang Zhang, Xuegong Zhang, Sheng Zhong, Yiming Zhou, Sheng Zhu, Dhivya Arasappan, Wenjun Bao, Anne Bergstrom Lucas, Frank Berthold, Richard J Brennan, Andreas Buness, Jennifer G Catalano, Chang Chang, Rong Chen, Yiyu Cheng, Jian Cui, Wendy Czika, Francesca Demichelis, Xutao Deng, Damir Dosymbekov, Roland Eils, Yang Feng, Jennifer Fostel, Stephanie Fulmer-Smentek, James C Fuscoe, Laurent Gatto, Weigong Ge, Darlene R Goldstein, Li Guo, Donald N Halbert, Jing Han, Stephen C Harris, Christos Hatzis, Damir Herman, Jianping Huang, Roderick V Jensen, Rui Jiang, Charles D Johnson,

Giuseppe Jurman, Yvonne Kahlert, Sadik A Khuder, Matthias Kohl, Jianying Li, Li Li, Menglong Li, Quan-Zhen Li, Shao Li, Zhiguang Li, Jie Liu, Ying Liu, Zhichao Liu, Lu Meng, Manuel Madera, Francisco Martinez-Murillo, Ignacio Medina, Joseph Meehan, Kelci Miclaus, Richard A Moffitt, David Montaner, Piali Mukherjee, George J Mulligan, Padraic Neville, Tatiana Nikolskaya, Baitang Ning, Grier P Page, Joel Parker, R Mitchell Parry, Xuejun Peng, Ron L Peterson, John H Phan, Brian Quanz, Yi Ren, Samantha Riccadonna, Alan H Roter, Frank W Samuelson, Martin M Schumacher, Joseph D Shambaugh, Qiang Shi, Richard Shippy, Shengzhu Si, Aaron Smalter, Christos Sotiriou, Mat Soukup, Frank Staedtler, Guido Steiner, Todd H Stokes, Qinglan Sun, Pei-Yi Tan, Rong Tang, Zivana Tezak, Brett Thorn, Marina Tsyganova, Yaron Turpaz, Silvia C Vega, Roberto Visintainer, Juergen von Frese, Charles Wang, Eric Wang, Junwei Wang, Wei Wang, Frank Westermann, James C Willey, Matthew Woods, Shujian Wu, Nianqing Xiao, Joshua Xu, Lei Xu, Lun Yang, Xiao Zeng, Jialu Zhang, Li Zhang, Min Zhang, Chen Zhao, Raj K Puri, Uwe Scherf, Weida Tong, and Russell D Wolfinger. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28 (8):827–838, August 2010. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.1665. URL `http://www.nature.com/doifinder/10.1038/nbt.1665`. 00002.

Paul K. Tan, Thomas J. Downey, Edward L. Spitznagel, Pin Xu, Dadin Fu, Dimiter S. Dimitrov, Richard A. Lempicki, Bruce M. Raaka, and Margaret C. Cam. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, 31(19):5676–5684, October 2003. ISSN 0305-1048. doi: 10.1093/nar/gkg763. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC206463/`.

Xiaojia Tang, Saurabh Baheti, Khader Shameer, Kevin J. Thompson, Quin Wills, Nifang Niu, Ilona N. Holcomb, Stephane C. Boutet, Ramesh Ramakrishnan, Jennifer M. Kachergus, Jean-Pierre A. Kocher, Richard M. Weinshilboum, Liewei Wang, E. Aubrey Thompson, and Krishna R. Kalari. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic Acids Research*, 42(22):e172–e172, December 2014. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gku1005. URL `http://academic.oup.com/nar/article/42/22/e172/2410988/The-eSNVdetect-a-computational-system-to-identify`.

R. Todd and D. T. Wong. Oncogenes. *Anticancer Res.*, 19(6A):4729–4746, December 1999. ISSN 0250-7005.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.1621. URL `http://www.nature.com/doifinder/10.1038/nbt.1621`. 04678.

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2012.016. URL `http://www.nature.com/doifinder/10.1038/nprot.2012.016`. 01931.

Victor E. Velculescu, Lin Zhang, Bert Vogelstein, and Kenneth W. Kinzler. Serial Analysis of Gene Expression. *Science*, 270(5235):484, 1995. doi: 10.1126/science.270.5235.484. URL `http://science.sciencemag.org/content/270/5235/484.abstract`.

Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, 131(4):281–285, December 2012. ISSN 1431-7613, 1611-7530. doi: 10.1007/s12064-012-0162-3. URL `https://link.springer.com/article/10.1007/s12064-012-0162-3`.

Kai Wang, Darshan Singh, Zheng Zeng, Stephen J. Coleman, Yan Huang, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178–e178, October 2010. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq622. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq622`.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009. ISSN 1471-0056. doi: 10.1038/nrg2484. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/`.

Lucas D. Ward and Manolis Kellis. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*, 40(Database issue):D930–D934, January 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr917. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245002/.

A. D. Whetton, G. W. Bazill, and T. M. Dexter. Stimulation of hexose uptake by haemopoietic cell growth factor occurs in WEHI-3b myelomonocytic leukaemia cells: A possible mechanism for loss of growth control. *Journal of Cellular Physiology*, 123(1):73–78. ISSN 1097-4652. doi: 10.1002/jcp.1041230112. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcp.1041230112.

Claire R. Williams, Alyssa Baccarella, Jay Z. Parrish, and Charles C. Kim. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, 18(1):38, January 2017. ISSN 1471-2105. doi: 10.1186/s12859-016-1457-z. URL https://doi.org/10.1186/s12859-016-1457-z.

Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, April 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq057. URL https://academic.oup.com/bioinformatics/article/26/7/873/212606.

Chang Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24, January 2018. ISSN 2001-0370. doi: 10.1016/j.csbj.2018.01.003. URL http://www.sciencedirect.com/science/article/pii/S2001037017300946.

Rongxi Yang, Michelle Dick, Frederik Marme, Andreas Schneeweiss, Anne Langheinz, Kari Hemminki, Christian Sutter, Peter Bugert, Barbara Wappenschmidt, Raymonda Varon, Sarah Schott, Bernhard H. F. Weber, Dieter Niederacher, Norbert Arnold, Alfons Meindl, Claus R. Bartram, Rita K. Schmutzler, Heiko Müller, Volker Arndt, Hermann Brenner, Christof Sohn, and Barbara Burwinkel. Genetic variants within miR-126 and miR-335 are not associated with breast cancer risk. *Breast Cancer Res. Treat.*, 127(2):549–554, June 2011. ISSN 1573-7217. doi: 10.1007/s10549-010-1244-x.

Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.

OMICS, 16(5):284–287, May 2012. ISSN 1536-2310. doi: 10.1089/omi.2011.0118. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339379/.

Shanrong Zhao and Baohong Zhang. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1):97, 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1308-8. URL http://www.biomedcentral.com/1471-2164/16/97.

Xiaoguang Zhou, Lufeng Ren, Qingshu Meng, Yuntao Li, Yude Yu, and Jun Yu. The next-generation sequencing technology and application. *Protein & Cell*, 1(6):520–536, June 2010. ISSN 1674-800X, 1674-8018. doi: 10.1007/s13238-010-0065-3. URL http://link.springer.com/10.1007/s13238-010-0065-3.