Computational methods for studying posttranscriptional gene regulation in cancer from PAR-CLIP and RNA-Seq data

---

Kumulative Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Andreas Klötgen

aus Essen

Düsseldorf, September 2016

aus der Arbeitsgruppe für Bioinformatik der Infektionsforschung
am Helmholtz-Zentrum für Infektionsforschung, Braunschweig
und der Klinik für Kinder-Onkologie, -Hämatologie und Klinische Immunologie
am Universitätsklinikum Düsseldorf der Heinrich-Heine Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referentin: Prof. Dr. Alice McHardy
Koreferent: Prof. Dr. Arndt Borkhardt
Koreferent: Prof. Dr. Holger Schwender

Tag der mündlichen Prüfung:

Eigenständigkeitserklärung

Hiermit erkläre ich, Andreas Klötgen, dass ich die vorliegende Dissertation selbstständig verfasst und bei keiner anderen Universität oder Fakultät in der vorgelegten oder ähnlichen Form eingereicht habe. Für die Anfertigung der Dissertation habe ich keine anderen als die angegebenen Hilfsmittel verwendet. Die Stellen, die anderen Arbeiten dem Wortlaut oder dem Sinn nach entnommen sind, wurden unter Angabe der dazugehörigen Quelle kenntlich gemacht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 28.09.2016

(Andreas Klötgen)

Statement of authorship

I, Andreas Klötgen, hereby certify that this thesis is the result of my own work and was not submitted for consideration to any other University or Faculty in the same or similar form. No other person's work has been used without acknowledgement. All parts which are taken from other publications in meaning or wording are cited accordingly. I have not previously failed a doctoral examination procedure.

Table of Contents

Deutsche Zusammenfassung

Krebs ist eine komplexe Krankheit, die durch eine maligne Transformation von Zellen verschiedensten Ursprungs entsteht. Krebs verursacht etwa 13% aller Todesfälle weltweit pro Jahr. Die entscheidenden molekularen Veränderungen, die einen Tumor auslösen können, sind zwischen Tumoren aus verschiedenen Geweben sehr unterschiedlich. Die molekularen Veränderungen sind unter anderem chromosomale Verkürzungen, Punktmutationen, abnormale DNA-Methylierungsmuster, Gen-Verlust/Duplizierung oder abnormale Expressions-Profile von RNAs oder Proteinen. Die betroffenen Gene gehören überwiegend zu den sogenannten Onkogenen und Tumorsuppressoren, deren Funktionen alle Ebenen der Genregulation betreffen, von transkriptioneller Kontrolle und posttranskriptioneller Regulation bis zu translationaler Kontrolle. Die vorliegende Doktorarbeit beschäftigt sich dabei ausschließlich mit der Analyse von posttranskriptioneller Genregulation in Tumoren, welche überwiegend durch RNA-bindende Protein (RBPs) und nicht-kodierende RNAs beeinflusst wird. Allerdings erschwert das komplexe Zusammenspiel dieser verschiedenen Regulationsebenen eine genomweite Vorhersage für entscheidende genomische Abnormalitäten. Daher ist eine Datenintegration von Datensätzen, die verschiedene Ebenen der Genregulation analysieren, von entscheidender Bedeutung um tiefere Einblicke in die Tumor-Biologie zu erhalten.

In den letzten Jahren haben Verbesserungen in Sequenziertechnologien weitreichendere Analysen ermöglicht. Die Sequenziertechnologien der sogenannten nächsten Generation ermöglichen so heute genomweite Vorhersagen für Mutationen, transkriptomweite Charakterisierung von Expressionsprofilen, genomweite Identifikation von RNA-Bindestellen eines RBPs und vieles mehr. Diese Doktorarbeit beschäftigt sich mit Verbesserungen für die computergestützte Analyse von „photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation“ (PAR-CLIP) Daten, welche das gebundene „RNA-Netzwerk“ eines RBPs erfassen können. Diese Arbeit beinhaltet zudem die Analyse von RNA-Sequenzdaten (RNA-Seq) für

kodierende und nicht-kodierende Gene und die Integration von PAR-CLIP und RNA-Seq Daten zur Verknüpfung der verschienenen Ebenen der Genregulation. Die Modifikation des Sequenzier-Read Aligners BWA (Burrows-Wheeler Aligner) für die hier entwickelte PAR-CLIP analyzer suite (PARA-suite) ist speziell für die Datenauswertung von PAR-CLIP konzipiert. Der modifizierte Algorithmus der PARA-suite bezieht ein empirisch ermitteltes Fehlerprofil in das Alignment mit ein, um PAR-CLIP spezifische Nukleotid-Konversionen im Alignmentprozess besser zu berücksichtigen. Die PARA-suite konnte damit die Detektion von RNA-Bindestellen für RBPs aus PAR-CLIP Daten entscheidend verbessern.

Die PARA-suite wurde zur Analyse eines AGO-PAR-CLIP Experimentes im Rahmen des International Cancer Genome Consortium Project "Determining Molecular Mechanisms in Malignant Lymphoma by Sequencing" (ICGC MMML-Seq) verwendet. Dadurch konnten gebundene mRNAs von *AGO2*, einem Protein des sogenannten „RNA-induced silencing complex" (RISC), identifiziert werden. Das RISC ist entscheidend für miRNA-induzierte posttranskriptionelle Genregulation. Durch dieses Projekt konnten posttranskriptionelle Regulationen durch miRNAs erkannt werden, dessen Ziel-mRNAs zuvor bereits in verschiedenen Lymphom-Typen mit der Tumorgenese assoziiert wurden. Die weitere RNA-Sequenzierung einer großen Kohorte an Lymphom-Patienten ermöglichte die Erkennung differentiell exprimierter miRNAs und mRNAs. Durch die Kombination all dieser Informationsebenen konnten negative Expressions-Korrelationen zwischen den verifizierten miRNA-mRNA Regulationspaaren erkannt werden, welche relevant für die Tumorentwicklung sind.

Diese Arbeit beschäftige sich zudem mit der differenzierenden Analyse von T Zell akuter lymphoblastischer Leukämie (T-ALL) in Neugeborenen im Vergleich zu jugendlichen Patienten. Basierend auf RNA-Sequenz Daten für kodierende und nicht-kodierende Gene konnten negative Expressions-Korrelationen zwischen miRNA-mRNA Paaren erkannt werden. Diese zeigten ebenso die Bedeutung von epigenetischen Regulationen von Onkogenen und Tumorsuppressoren in T-ALL auf.

Summary

Cancer is a complex disease that arises from malignant transformations of cells from various origins. It is one of the leading causes for deaths and accounts for approximately 13% of all deaths per year world-wide. The driving molecular dysfunctions leading to tumors are diverse and vary among tumors originating from different tissues. These driving dysfunctions include amongst others chromosomal shortenings, point mutations, aberrant DNA methylation patterns, gene gains and losses or aberrant expression profiles of RNAs and proteins. The affected genes mainly belong to the classes of oncogenes and tumor suppressors. Their functions include all levels of gene regulation, from transcriptional control and posttranscriptional regulation to translational control. This thesis focuses on posttranscriptional gene regulation in cancer, which is mainly mediated by RNA-binding proteins (RBPs) and non-coding RNAs. However, the complex interplay between these regulators makes it difficult to predict the consequences of driving aberrations (i.e. affected oncogenes or tumor suppressors). Thus data integration of datasets analyzing different levels of regulation is important for gaining deeper insights into tumor biology.

During recent years, the rapid advances in sequencing technologies and its diverse applications enabled a wide range of analyses. Next-generation sequencing is nowadays available for the genome-wide identification of mutations, for the transcriptome-wide characterization of expression profiles, the genome-wide identification of RNA binding sites for RPBs and much more. This thesis deals with the improvement of computational methods for photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) data analysis, which reveals the bound "RNA network" of a given RBP. The second part focuses on the analysis of RNA sequencing expression data (RNA-Seq) from coding and non-coding genes and the integration of PAR-CLIP and RNA-Seq data, which identified connections between the different regulatory levels. A modification of the sequencing read aligner BWA (Burrows-Wheeler Aligner) specific for PAR-CLIP data, was implemented in the here described

PAR-CLIP analyzer suite (PARA-suite). The modified algorithm takes an empirical error profile into account to accommodate for PAR-CLIP specific nucleotide conversions. The PARA-suite obtained a higher accuracy than other read aligners in the detection of RNA-binding sites on the basis of simulated PAR-CLIP data.

The PARA-suite was applied to an AGO-PAR-CLIP dataset obtained within a project of the International Cancer Genome Consortium Project "Determining Molecular Mechanisms in Malignant Lymphoma by Sequencing" (ICGC MMML-Seq). This analysis revealed mRNAs targeted by *AGO2*, which is a core member of the RNA-induced silencing complex (RISC). The RISC is important for miRNA-mediated posttranscriptional gene regulation. Thereby, mRNA targets were identified that were specifically regulated by miRNAs and were recently associated with lymphomagenesis of different lymphoma subtypes. Further RNA-Seq analysis of a large cohort of lymphoma patients revealed differentially expressed miRNAs and mRNAs between heterogeneous lymphoma subtypes. The RNA-Seq results were combined with the miRNA-mRNA interaction pairs from the AGO-PAR-CLIP results. This enabled the calculation of negative expression correlations between the differentially expressed miRNAs and mRNAs on the basis of the RNA-Seq expression data. This approach identified lymphoma relevant miRNA-mRNA correlation pairs.

This thesis also includes a differential analysis of infant T-cell acute lymphoblastic leukemia (T-ALL) in comparison to incidences during childhood. On the basis of RNA-Seq data for coding and non-coding genes, negative correlations of miRNA-mRNA expressions were measured between the two cohorts. This analysis revealed important epigenetic regulations of oncogenes and tumor suppressors in T-ALL.

List of abbreviations

Molecular biology

CLIP	Crosslinking and immunoprecipitation
DNA	Deoxyribonucleic acid
HITS-CLIP	High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
miRNA	microRNA
mRNA	Messenger RNA
ncRNA	Non-coding RNA
PAR-CLIP	Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation
qRT-PCR	Quantitative real-time polymerase chain reaction
RBP	RNA-binding protein
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
T–C conversion	Thymidine to cytidine conversions
UV light	Ultra-violet light

Bioinformatics tools

BWA	Burrows-Wheeler Aligner
FM index	Ferragina and Manzini index
PARalyzer	PAR-CLIP data analyzer
PARA-suite	PAR-CLIP analyzer suite
PSSM	Positions specific scoring matrix

Cancer subtypes

ALL	Acute lymphoblastic leukemia
BL	Burkitt's lymphoma
DLBCL	Diffuse large B-cell lymphoma
FL	Follicular lymphoma
NHL	Non-Hodgkin lymphoma
T-ALL	T-cell acute lymphoblastic leukemia
iT-ALL	Infant T-cell acute lymphoblastic leukemia

Genes

AGO2	Argonaute-2
BRCA1/BRCA2	Breast cancer 1/2
HER2	Human epidermal growth factor receptor 2
MYC	Myelocytomatosis
RAS family	Retrovirus-associated DNA sequences family
RB	Retinoblastoma

Miscellaneous

ICGC MMML-Seq	International Cancer Genome Consortium Project "Determining Molecular Mechanisms in Malignant Lymphoma by Sequencing"
UCSC	University of California, Santa Cruz

1 Introduction

1.1 Cancer

Cancer is a complex disease that is influenced by many factors, including the individual's genetic background, endogenous and environmental factors. In addition to the inherited genomic alterations that are advantageous for tumor formation, further mutations are acquired and accumulate during lifetime. On the one hand, these mutations are caused by endogenous factors such as reactive oxygen species. On the other hand, environmental factors that cause genetic mutations include UV irradiation (e.g. sun exposure), radiation (e.g. X-rays), genotoxic agents (e.g. smoking) and many more. In conjunction, such accumulated genetic alterations create an instable genome that is predisposed to certain diseases including cancer (Friedberg, McDaniel et al. 2004).

DNA damage, genetic mutations and genome instability

The term genome stability describes the cell's ability to perform its normal function through a healthy genome. Genetic mutations and DNA damage, however, can disrupt the regular functionality in diverse ways, leading to cells that may be defective in controlled DNA replication and transcription. The accumulation of several crucial genomic alterations is termed genome instability (Hoeijmakers 2001). This particular state describes the cell's inability to react to external growth factors and other signals and thus its ability to proliferate uncontrolled. Genomic alterations may be introduced in different ways and are approximately divided into two types:

a) DNA damage describes the structural or physical damage of a particular DNA strand, which can be identified and repaired by enzymes.
b) Genetic mutations, which are encoded on both strands, cannot be recognized or repaired by enzymes.

DNA damage includes the chemical change of a particular base (e.g. by UV-B light causing thymidine dimers) and single- or double-strand breaks. The latter can lead to shortened chromosomes, the so-called chromosome abbreviations, or even to genomic translocations, which is the joining of two different chromosomes (De Bont and Van Larebeke 2004, Jackson and Bartek 2009). Single-strand and double-strand breaks resulting in chromosome abbreviations and genomic translocations in general have a more profound impact on the cellular integrity and are often linked to certain diseases. This is due to the fact that such aberrations affect multiple genes at once. But chemical modifications of certain bases might not be at all or be falsely repaired and can thus result in genetic mutations on both strands. A genetic mutation is irreversible and is passed on to the next generation by cell division subsequently affecting an entire population of cells. Genetic mutations include nucleotide exchanges and short insertions or deletions (together called indels) of usually a few nucleotides. They can also affect genes and may change the encoded protein, either by amino-acid exchanges or even by a truncation or elongation due to affected stop codons. Mutations can either be classified as somatic (not passed on to offspring) or germline. If a mutation further becomes present in an entire subpopulation, it is called a single nucleotide polymorphism (SNP). SNPs are curated in publicly available databases (e.g. 1000 genomes (GenomesProjectConsortium 2015) or HapMap (GenomesProjectConsortium 2012)).

Due to accumulations of DNA errors, the downstream effects in cells might be even more disruptive. By defects of the DNA repair or DNA replication machinery, further errors can accumulate much faster. A defective DNA replication might result in aneuploidy, which is an aberrant copy-number of an entire chromosome. This has drastic effects on the transcription of all genes encoded on the affected chromosomes (Rowley 1998, Bergsagel and Kuehl 2001, Greaves and Wiemels 2003). The aberrant transcription caused by genome instability in general has a profound impact on tumor formation and was recently classified as one of the hallmarks of cancer (Hanahan and Weinberg 2011). But DNA errors can also alter the encoded protein function, for

instance making proteins unable to bind their specific ligands or rendering them constitutively active. The hallmarks of cancer include six important features, as first postulated by Hannahan and Weinberg (Figure 1):

- Sustaining proliferation signals
- Evading growth suppressor signals
- Resisting cell death signals
- Enabling replicative immortality
- Inducing angiogenesis
- Activating invasion and metastasis

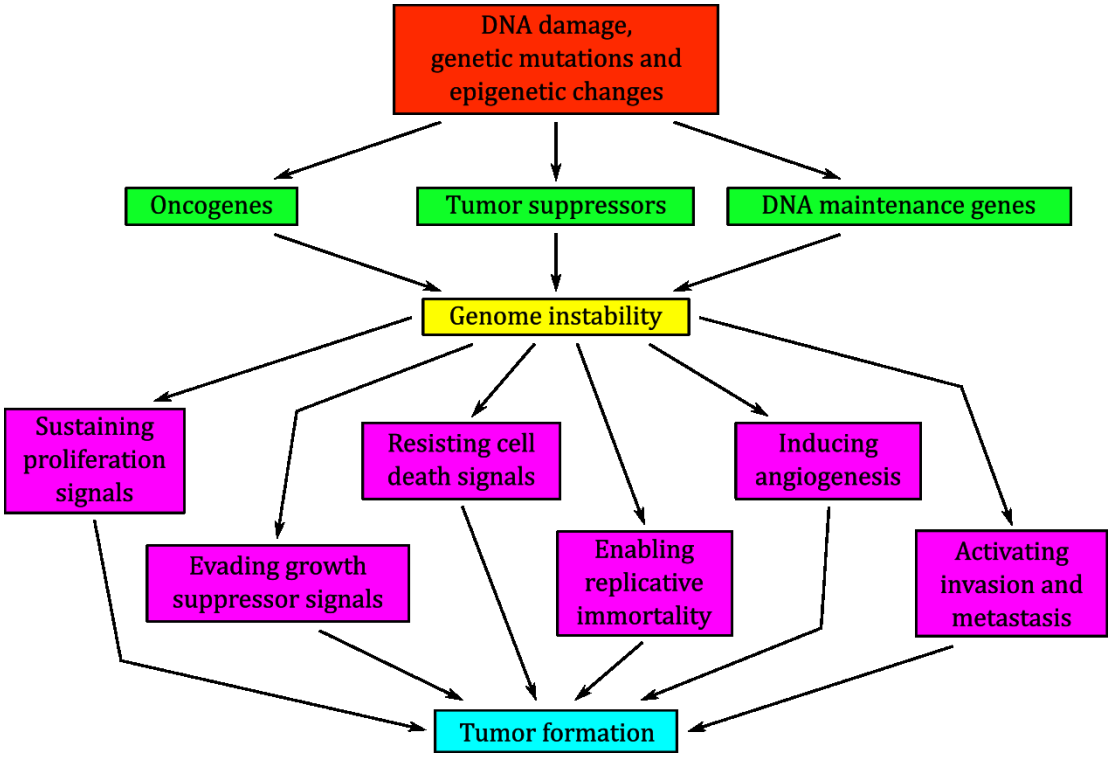


Figure 1: Steps from DNA errors to genomic instability to tumor formation, including the hallmarks of cancer as proposed by Hannahan and Weinberg (Hanahan and Weinberg 2000, Hanahan and Weinberg 2011).

In healthy cells, the aforementioned cellular functions contributing to genome stability are controlled by a few crucial genes, the so called proto-oncogenes and tumor suppressors. If these are affected by DNA errors, leading to aberrant expressions or altered protein functions, the cells become predisposed to cancer.

Oncogenes and tumor suppressors

The two types of genes generally related to cancer, the oncogenes and tumor suppressors, are extensively reviewed in (Weinberg 2013). On the one hand, oncogenes are genes, which support tumor formation upon „activation“, due to an aberrant high expression or mutation. In most (if not all) cases, a single aberrant allele of a proto-oncogene (the natural precursor of an oncogene) is sufficient for its enhancing impact on tumor formation. Upon activation, they contribute to a more instable genome. Most proto-oncogenes are related to cellular growth or differentiation, for example *HRAS*/*KRAS* (Bos 1989) or *MYC* (Nesbit, Tersak et al. 1999). The proto-oncogene *MYC* is tightly controlled by extracellular signals and is frequently turned on and off, either being expressed or repressed in healthy cells. It is an important regulator of cell growth and proliferation. However, the aberrant and constitutive high expression of its oncogenic form escaping any mitogenic growth signals is associated with tumor formation. Thus this oncogene specifically drives uncontrolled cell proliferation and the predisposition can be identified by RNA-Seq expression analysis (Weinberg 2013). The *RAS* family members are turned into oncogenes by a single point mutation. Interestingly, many investigated cancers (up to 20%) showed the same point mutation in *KRAS*, which led to an amino-acid exchange in the encoded protein (Weinberg 2013).

Tumor suppressors on the other hand are genes that inhibit malignant cell development and are usually inactivated in both alleles in tumors. These mutations are often considered loss of function mutations or result in a profound reduction of the gene’s expression. Tumor suppressors are often classified as cell cycle regulators or apoptosis regulators, making them important gatekeepers within cells. In addition, so-called caretaker genes related to DNA maintenance also have a suppressive role in

tumor formation (not discussed here). A prominent example for a tumor suppressor is the gene *RB*, which was first described in the 1980s (Weinberg 1995). The recessive phenotype of the oncogenic form requires both alleles to be knocked out to result in the formation of a retinoblastoma. Thus, an inherited genetic mutation in *RB* represents a predisposition for retinoblastomas.

Due to DNA lesions in these two classes of genes, tumor cells often gain an advantageous proliferation status and can divide uncontrolled, displacing healthy cells in the respective tissue (Hanahan and Weinberg 2000). Many therapeutic drugs target oncogenes to repress their activity or tumor suppressors to restore their natural function (Dietlein, Thelen et al. 2014). However, both oncogenes and tumor suppressors are oftentimes mutated simultaneously in a single tumor entity, making targeted therapy more complicated.

Cancer diagnostics and therapy

The early diagnosis of a tumor is crucial for prognosis and patient outcome. The exact diagnosis of the specific subtype has a profound impact on the treatment decision (Anderson, Schwab et al. 2014). Thus so called biomarkers that can be assayed at best from non-invasively collected biofluids (e.g. blood) are biologically measurable indicators defining a particular state of cancer in a patient (Henry and Hayes 2012). This includes DNA alterations (e.g. mutations or translocations) and specifically expressed non-coding RNAs, proteins or hormones. Prominent examples are certain *BRCA1* and *BRCA2* mutations in breast and ovarian cancers, as well as a *BCR-ABL* translocation that is often found in chronic myeloid leukemia or *HER2* overexpressing breast cancers. In medicine, biomarkers for tumors are roughly used in the following ways according to Henry and Hayes (Henry and Hayes 2012):

- Predicting risk of developing cancer (i.e. predisposition)
- Improving diagnostics/prognostics
- Predicting aggressiveness of the tumor and patient outcome
- Predicting treatment response
- Monitoring relapse and treatment response in metastases

For example, the aberrant high expression of *HER2* in breast cancer is a sign for a very aggressive form of the tumor. It is expressed in about 15-30% of all breast cancers and patients suffering from *HER2* overexpressing breast cancers usually have a very poor prognosis and an increased rate of relapse (Slamon, Leyland-Jones et al. 2001). Many biomarkers are already used to adjust treatment options for individual patients. In the case of *HER2* overexpressing breast cancers, specific *HER2* antibodies are available to treat *HER2* overexpressing breast cancers (Slamon, Leyland-Jones et al. 2001).

The most basic treatment options for cancers are surgery, radiotherapy and chemotherapy. Oftentimes a combined treatment is applied, e.g. a surgical removal of the tumor mass followed by radiotherapy or chemotherapy. This is called an adjuvant therapy, in which ensuing treatment helps killing the remainder of the tumor cells. Nowadays, more selective treatment options have become available, due to the identification of precise biomarkers. These for example include the aforementioned anti-*HER2* therapy specifically for *HER2* overexpressing breast cancers. However, distinction between cancer subtypes remains difficult due to a lack of precise biomarkers for diagnostics and treatment options. This is especially a problem for tumors, which were recently (or still are today) classified only on the basis of their morphological appearance, because this does not reveal genetic and epigenetic differences that might have an impact on prognostics and treatment (Quackenbush 2006). Thus the genome-wide identification of mutations and aberrant expression patterns of certain genes by next-generation sequencing is important for precise biomarker determination. Also, although many cancer biomarkers are known, reagents specifically targeting those biomolecular lesions are still not available in many cases (Hamburg and Collins 2010). The complex interplay between the genetic lesions makes it difficult to identify such reagents or drugs. Clinical trials furthermore suffer from long periods until the drugs are approved for clinical use.

1.1.1 Non-Hodgkin lymphomas

Lymphomas in general are heterogenic lymphoid neoplasms, and, depending on the affected cell type, their molecular and phenotypic appearance differs. Classifications of different lymphomas are made on morphologic as well as immunophenotypic features (based on the World Health Organization (WHO) classifications). However, due to the heterogeneity of this disease, certain diagnostic features are shared between distinct lymphoma subgroups, making a classification sometimes difficult. Hence, further biomarkers separating lymphoma subgroups are required to improve diagnostics, therapeutics and thereby patient outcome.

Non-Hodgkin lymphomas (NHLs) are a specific subgroup of lymphomas and may affect different cell types. B-cell NHLs occur during different stages of B-cell development and account for approximately 85% of all NHLs in the USA (WHO 2008). According to the WHO classification, NHLs are further classified into Burkitt's lymphoma (BL), diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL) and many more.

Approximately one out of three NHL cases is classified as a DLBCL, which makes it the most common subtype. DLBCLs are classified according to their morphological appearance, which does not account for the two distinct subtypes, namely germinal center B-cell like (GCB) and activated B-cell like (ABC). Newer approaches now identified these subtypes to be molecularly heterogenic, because of different cells of origin (Alizadeh, Eisen et al. 2000). BL in contrast is a rare NHL subtype accounting for 1-2% of all lymphomas in the USA in adults. In developed countries, BL is usually not caused by an Epstein-Barr viral infection but occurs spontaneously. As mentioned before, an aberrant high expression of the tumor suppressor *MYC* is frequently observed among different cancers and was shown to be important in many BLs. Also, the miRNAs *hsa-let-7a* (Metzler, Wilda et al. 2004) and *hsa-miR-155* (Sampson, Rong et al. 2007) are associated with BL, but transcriptome-wide predictions for important miRNAs are still missing. Besides all these facts, differences in aggressiveness and responses to treatment call for additional biomarkers molecularly separating NHL subtypes (Campo, Swerdlow et al. 2011, Sinha, Nastoupil et al. 2012). Especially, a

clinical significance for oncogenic or tumor suppressive microRNAs (miRNAs) in BLs, DLBCLs and FLs has not been shown yet. A miRNA expression classifier might represent another accurate tool for discriminating the NHL subtypes on an epigenetic level rather than on morphology.

1.1.2 T-cell acute lymphoblastic leukemia

Acute lymphoblastic leukemia (ALL) is a malignant disease of the hematopoietic system and accounts for approximately one out of three cancers in children, the most frequent cancer during childhood. Patient outcome is high for ALL patients older than 1 year showing a 5-year event-free survival of about 80% (Pui, Carroll et al. 2011). However, ALL in infants (age 0-1 year) is associated with a high rate of treatment failure of around 60% (Hilden, Dinndorf et al. 2006). ALL can arise from B- and T-cells, with T-cell ALL (T-ALL) accounting for only 10-15% of all ALL cases (Goldberg, Silverman et al. 2003). T-ALL in general is a more aggressive disease compared to B-ALL and classified as high risk (Mansur, Delft et al. 2015). Recently, the worse outcome of incidences during infancy than in older patients was particularly shown for T-ALL (Mansur, Delft et al. 2015). Thus it remains unclear whether infant T-ALL (iT-ALL) is a molecularly different disease than T-ALL in older patients.

1.2 Posttranscriptional gene regulation

Gene expression in its widest sense is the flow of genetic information within cells. It is regulated on different levels, from transcriptional regulation, posttranscriptional regulation to posttranslational regulation. Detailed reviews on the different regulatory levels are given in (Nestler and Hyman 2002, Jackson, Hellen et al. 2010, Coulon, Chow et al. 2013). Upon transcription of a gene, the transcribed RNA is a target of many regulators until it is finally processed to perform its action. This work focuses on posttranscriptional gene regulation, which generally includes splicing, transportation,

editing, translational control and more. Some of the regulators, which are relevant for this thesis, are outlined in more detail below.

1.2.1 RNA-binding proteins

Part of this thesis is a detailed review of RNA-binding proteins (RBPs) and experimental as well as bioinformatic methods to elucidate protein-RNA interactions (Kloetgen, Münch et al. 2015). RBPs have many functional roles from splicing, transportation, modification, degradation to translation of transcribed RNAs (Glisovic, Bachorik et al. 2008). They provide additional genome diversity by alternative splicing and modification (e.g. RNA editing), which increases the number of translated proteins from a single gene locus. Binding of RBPs to RNAs is either dependent on the sequence composition and/or the secondary structure of the respective RNA or happens unselectively (Gupta and Gribskov 2011).

Because of their wide range of activities, it is not surprising that many RBPs have functional roles in diseases including cancer. An example for an oncogenic RBP is *FUS*, which is affected by genomic rearrangements mainly in sarcomas and leukemias (Ichikawa, Shimizu et al. 1994, Singer, Socci et al. 2007). In addition, two different mutations within its RNA-binding domain, which alter the preferred RNA binding motif, were reported to cause amyotrophic lateral sclerosis (Kwiatkowski, Bosco et al. 2009, Vance, Rogelj et al. 2009).

The information about the RNA network in which a particular RBP operates is important to understand its cellular functions. To identify the RNA binding network and a potential binding sequence pattern of a given RBP, multiple experimental protocols were established. These include SELEX, RIP-CHIP (nowadays also termed RIP-Seq), and several crosslinking and immunoprecipation (CLIP) protocols. This thesis focuses on CLIP methods and their data analysis. A detailed description of a specific CLIP method and its pitfalls in data analysis can be found in Sections 1.3 and 1.4.2.

1.2.2 microRNAs

Another group of posttranscriptional gene regulators are the so-called microRNAs (miRNAs), which belong to the class of non-coding RNAs (ncRNAs), reviewed in (Barrett, Fletcher et al. 2012). The transcribed pri-miRNA is a double-stranded RNA of length roughly between 500 and 3000bp. In short, the pri-miRNA is processed by different RPBs (e.g. *DGCR8* and *DICER1*) in the nucleus and forms a 70-80bp long pre-miRNA (Denli, Tops et al. 2004). After transport into the cytoplasm by exportin-5, the pre-miRNA is further processed and the mature miRNA is excised. The mature miRNA, located in the cytoplasm, is a single-stranded RNA usually of a length around 18-23 bases and acts as a posttranscriptional gene regulator (Lee, Jeon et al. 2002). The miRNA-mediated regulation is carried out within the RNA-induced silencing complex (RISC), which consists of different members of the *AGO* protein family (also RBPs) and miRNAs. The mature miRNAs are non-specifically bound by *AGO* and further proteins to form a miRNA-containing ribonucleoprotein complex. In mammals, brought together by the RISC, miRNA seed regions (i.e. positions 2-8) are complementary binding to mRNAs, which are subsequently destabilized. This often results in the degradation of the respective mRNA. There is no evidence yet for a translational inhibition of the targeted mRNAs as reported for other species than mammals. The preferred binding site of miRNAs lies within the 3' untranslated region (3' UTR) of the respective mRNAs, but targets in coding regions are also reported, albeit less frequently (Holoch and Moazed 2015). A curated database of annotated miRNAs is miRBase (Kozomara and Griffiths-Jones 2010), which lists 2,588 mature human miRNAs in its current release (V21).
Apart from curated databases for mRNA targets of certain miRNAs (e.g. miRTarBase (Chou, Chang et al. 2015)), bioinformatic tools were developed for the prediction of potential miRNA binding sites in coding genes. This includes the most prominent algorithms miRanda (Enright, John et al. 2004) and TargetScan (Garcia, Baek et al. 2011). The algorithms of these tools are mainly based on identifying sequence complementarity between the miRNA seed region and the 3' UTR of an mRNA.

Not only coding genes can act as oncogenes or tumor suppressors, but also non-coding RNAs, such as miRNAs, can cause driving changes to genome stability upon genetic alterations (Chen 2005, Farazi, Spitzer et al. 2011). The analysis for driving mutations should thus not be limited to coding genes, but be extended to ncRNAs by genome-wide classifications of cancer-relevant ncRNAs. The dysregulation of miRNAs might be another diagnostic tool for the classification of cancer. Based on either their up- or downregulation in certain cancers, they are considered oncogenes or tumor suppressors, respectively. For example, the *hsa-miR-17-92* cluster was reported as oncogenic in lymphoproliferative disorders, because of its high upregulation in *MYC*-driven B-cell lymphomas (He, Thomson et al. 2005, Xiao, Srinivasan et al. 2008). As for coding tumor suppressors and oncogenes, miRNAs were also considered as biomarkers and therapeutic targets in specific tumors. Hence, many studies also focused on defining miRNA expression classifiers acting as biomarkers for certain cancers (Lu, Getz et al. 2005). For example, miRNAs enabled the accurate classification of non-small cell lung carcinomas in 95% of the tested cases (Bishop, Benjamin et al. 2010). Also, a classifier containing 15 miRNAs revealed better prognostic power for squamous cell carcinoma with an accuracy of approximately 78% than an expression signature containing 50 genes (Raponi, Dossey et al. 2009).

1.3 Biological background of RNA sequencing and PAR-CLIP

Experimental background of RNA sequencing

Recent progress in sequencing technologies enabled high-throughput analyses of a cell's expressed transcriptome (known as RNA-Seq). The resulting sequencing reads are each representing parts of an RNA molecule expressed in the biological sample. RNA-Seq overcomes many disadvantages of previous methods, such as microarray or qRT-PCR based studies. RNA-Seq includes, besides the large-scale assessments of expression levels, the analysis of unknown and unannotated transcripts (Necsulea and Kaessmann 2014). Many companies provide sequencing machines and protocols that highly differ in sample preparation and the sequencing procedure. This thesis is only based on

sequencing data obtained from an Illumina HiSeq 2500 machine and its respective protocols. Sequencing with the Illumina platform was reviewed in detail in (Metzker 2010). Further prominent platforms include PacBio, 454 and IonTorrent.

Experimental background of PAR-CLIP

The identification of the RNA network in which a certain RBP operates is important to understand its cellular functions. A promising method for the genome-wide identification of RNA binding sites of a given RBP is called photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP). A detailed review on PAR-CLIP and a comparison to similar methods as well as a discussion of a common analysis pipelines and its pitfalls is part of this thesis (Kloetgen, Münch et al. 2015). Experimentally, cells are supplied with 4-thiouridine (4-SU), which replaces the natural uridine to a certain extent during transcription (Figure 2). Further photo-reactive nucleosides are also available, such as 6-thioguanine (6-TG). Next, cells are irradiated with UV light at 365nm to crosslink amino acids of the RBP with nucleotides of the bound RNAs. The RBP of interest is then immunoprecipitated and digested with proteinase K. The remaining RNA is thought to represent the bound RNAs of the immunoprecipitated RBP. During reverse transcription of the purified RNAs, which is a necessary step for sequencing, all crosslinked 4-SUs result in a conversion to a cytidine on sequence level. This conversion is called a “T–C conversion”. Contaminations with unbound but highly abundant RNA fragments commonly happen. The reads resulting from contaminations do not contain T–C conversions (except for rare sequencing errors), so these specific conversions can be used to identify specifically bound RNAs.

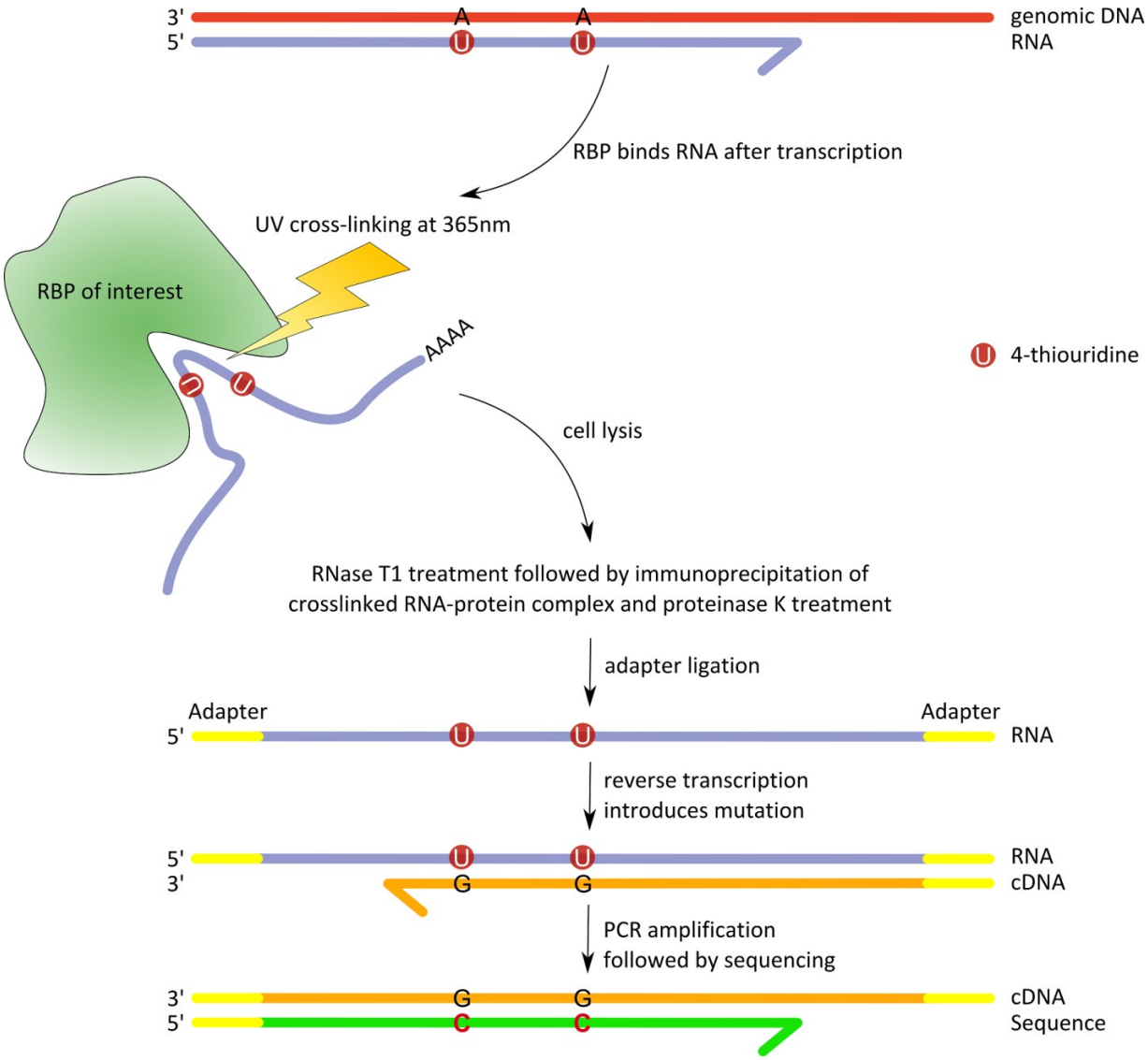


Figure 2: Experimental steps of PAR-CLIP. This Figure is taken from (Kloetgen, Münch et al. 2015) without any modification.

PAR-CLIP analysis used to reveal miRNA-mRNA interactions

Common miRNA-mRNA correlation analyses are based on curated or computationally predicted miRNA-mRNA pairs or its functions, which are oftentimes reported independently of the respective cell type or tissue (Eisenberg, Eran et al. 2007, Wang and Li 2009, Gutiérrez, Sarasquete et al. 2010). The computational target predictions are often only based on sequence complementarity between the miRNA seed region and any region in the 3' UTR of a coding gene, irrespective of further effectors such as

secondary RNA structures. This can lead to large numbers of 1,828,274 (in the case of miRDB (Wong and Wang 2014)) or 1,264,046 (in the case of TarBase (Vergoulis, Vlachos et al. 2012)) for predicted human miRNA-mRNA pairs. Although expression correlation scores obtained from sequencing data might identify a real miRNA-mRNA interaction on the basis of the potential interactions, many other posttranscriptional mechanisms might explain the expression level changes of the respective miRNA and mRNA. Therefore, the disadvantage of the common approach is that the databases often lack experimental validity as well as information about cell type and tissue. But recent applications of PAR-CLIP to the *AGO2* protein, the most frequent member of the RISC, revealed physical miRNA-mRNA interaction pairs on a genome-wide scale (Hafner, Landthaler et al. 2010, Farazi, Ten Hoeve et al. 2014). This approach identified both, the potential of miRNA-binding by sequence complementarity to mRNAs, and the physical interaction of the respective mRNA with the RISC. Thus this experiment represents a genome-wide analysis of experimentally validated miRNA-targeted mRNAs for a certain cell type.

1.4 Bioinformatic analysis of sequencing data

1.4.1 Sample pipeline for RNA-Seq data processing

Commonly, the first step of analyzing sequencing reads is the removal of adapter sequences and low quality ends (Figure 3). Adapters are short RNA fragments, which were ligated to the bound RNA fragment for sequencing purposes. Low quality ends occur on both sides of the reads and often relate to wrongly called bases by the sequencer. Different algorithms are available for removing adapters and low quality ends, e.g. cutadapt (Martin 2011) or trimmomatic (Bolger, Lohse et al. 2014). Then, the remainder of the reads can be either aligned against an existing reference genome sequence or newly assembled to reconstruct the expressed genes on the basis of the obtained sequencing reads. As this thesis only deals with human data for which a reference sequence is available (GRCh38 is the most recent human genome sequence

release (Yates, Akanni et al. 2016)), only sequencing read alignment against a genome sequence is discussed. Depending on the properties of the reads, different read alignment algorithms are available to identify the origin of a particular read within a reference genome sequence. A wide range of properties affects the read alignment process. Compared to DNA-Seq, RNA-Seq reads might span exon–exon junctions, which means that parts of the read are separated by sometimes thousands of bases within the reference sequence. Another difference to DNA-Seq is the read coverage, which is not uniform across the entire genome. Additionally, the number of sequencing cycles or the decision for single-end or paired-end sequencing has a profound impact on the choice of the read aligner. Short read aligners, including the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009), Bowtie (Langmead, Trapnell et al. 2009) or Bowtie2 (Langmead and Salzberg 2012), were developed to align short reads of approximately 30–70 bases in a continuous stretch to the reference sequence. Newer algorithms, such as TopHat2 (Kim, Pertea et al. 2013), STAR (Dobin, Davis et al. 2013) or Subjunc (Liao, Smyth et al. 2013), are capable of aligning longer reads to the genome sequence, which may also span exon–exon junctions in case of RNA-Seq reads. After the reads are aligned against the reference sequence, any downstream analyses can be performed. A common setting is to estimate gene expression from an RNA-Seq dataset by counting reads mapping to annotated genes. These annotations can be downloaded from resource databases, such as UCSC (Rosenbloom, Armstrong et al. 2015) or Ensembl (Cunningham, Amode et al. 2015). Examples for algorithms for read counting are HTSeq (Anders, Pyl et al. 2014) or ngsutils (Breese and Liu 2013). The counts per annotated gene give a rough approximation of the gene expression in relation to the sequencing depth of the particular sequencing run. For most applications, these counts have to be inter-sample normalized based on the sequencing depth, leading for instance to counts per millions calculated as $CPM(gene_A) = \frac{\text{counts}(gene_A) \cdot 1{,}000{,}000}{\sum_{i \in G} \text{counts}(gene_i)}$, with G being the set of all annotated genes and $gene_A \in G$. The CPM values can then be used for a differential gene expression analysis, employed by algorithms such as edgeR (Robinson, McCarthy et al. 2010) or DESeq (Anders and Huber 2010). These compare CPMs between two (or

more) groups of samples to identify transcriptomic differences between the cohorts. A common setup in cancer research is to compare gene expression between healthy tissue and cancer tissue. The differentially expressed genes explain the pathophysiological background to a certain extent. They can further act as clinical biomarkers and might be considered as molecular targets for cancer therapeutics.

Another downstream analysis of aligned RNA-Seq data is the detection of indels and SNVs encoded in the genome. Therefore, the aligned reads have to be locally realigned at indel positions to account for false alignments in the close vicinity of splice sites or repetitive genomic regions. Otherwise, these could result in falsely reported indels. Indel realignment can be handled e.g. by GATK (DePristo, Banks et al. 2011). Next, SNP/indel calling and annotation can also be handled by GATK. The reported variation calls can be further examined for effects on protein-coding regions, for which tools also already exist (Variant Effect Predictor (McLaren, Pritchard et al. 2010), PolyPhen2 (Adzhubei, Schmidt et al. 2010) and SIFT (Kumar, Henikoff et al. 2009)).

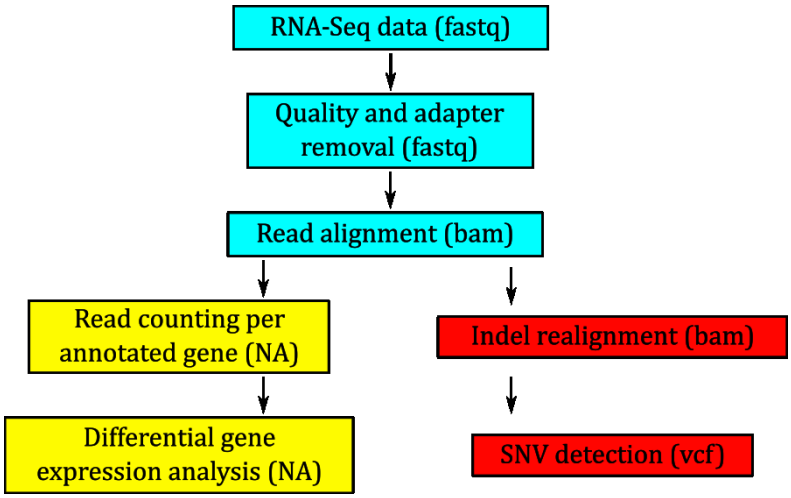


Figure 3: Sample pipeline for RNA-Seq data processing, including two possible downstream analyses. Cyan fields are preprocessing steps, yellow fields are differential gene expression analysis and red fields are SNP/indel detection. Names in brackets show commonly used data-types for each step, where NA depicts not applicable.

1.4.2 Bioinformatics analysis of PAR-CLIP data

The first step of analyzing PAR-CLIP sequencing data is to trim low quality ends and adapter sequences (Figure 4). As PAR-CLIP reads are generally short, all reads shorter than a certain threshold (e.g. 14 bases after trimming) might be excluded, as they cannot be uniquely aligned to the genome. Due to more mismatches per sequencing read compared to a reference genome sequence caused by the T–C conversions, the alignment step is crucial and has to allow for more errors than the alignment of normal RNA-Seq reads. Commonly employed read aligners for PAR-CLIP data analysis are BWA or Bowtie, each allowing for either one or two mismatches between a single read sequence and the reference sequence (Mukherjee, Corcoran et al. 2011, Ascano, Mukherjee et al. 2012, Sievers, Schlumpf et al. 2012, Mukherjee, Jacobs et al. 2014). But these approaches do not distinguish between different mismatch types and treat the PAR-CLIP specific T–C conversions in the same way as sequencing errors. Although they can thus deal with the frequent T–C conversions in a knowledgeable way, it will also give an advantage for reads with any type of mismatches. For PAR-CLIP data, this may result in falsely aligned reads mapping to multiple positions per sequencing read, making the decision for the correct alignment position complicated. An extension of BWA called BWA PSSM (Kerpedjiev, Frellsen et al. 2014) takes a position-specific scoring matrix (PSSM) into account to accommodate for specific types of mismatches occurring more frequently than others. The authors of BWA PSSM provide the user with a PSSM for PAR-CLIP data, which favors T–C mismatches over all others during alignment. However, a drawback of this approach is that the PSSM has to be specified by the user. Thus the underlying errors (and its expected rate of occurrence in the dataset) must be known prior to read alignment to specify the PSSM. Also, the PSSM is limited to nucleotide conversions and does not cover insertions or deletions. The high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP, sometimes referred to as CLIP-Seq) procedure, however, introduces deletions rather than single nucleotide conversions (Zhang and Darnell 2011, Sugimoto, König et al.

2012). These errors are not covered by BWA PSSM and it thus has a limited range of applications.

After read alignment of PAR-CLIP data, RBP binding sites can be detected with further software, which in general apply a clustering algorithm on the basis of the reads' mapping positions. The resulting clusters represent short transcriptomic stretches that might be bound by the RBP. Subsequently, a filtering for reliable T–C conversions for each identified cluster is applied. As mentioned before, the T–C conversions only occur at binding sites that were crosslinked to the particular RBP. This information is used for filtering non-specifically purified RNAs from truly bound RNAs. The filtered clusters are considered the binding sites of the analyzed RBP. Examples for software capable of identifying binding sites are PARalyzer (Corcoran, Georgiev et al. 2011) or BMix (Golumbeanu, Mohammadi et al. 2015). BMix uses a maximum likelihood approach to distinguish between T–C conversions, which are introduced by the crosslinking, and lowly and highly frequent erroneous alterations within the detected clusters. After binding site detection, further annotation information can be loaded for the identified binding sites, e.g. targeted genes and targeted gene regions (3' UTR, exon, intron etc.). A possible follow-up for RBPs binding a specific RNA sequence motif is the inference of the actual binding motif. On the basis of the detected binding sites, the sequence motif of the particular RBP can be inferred by searching for statistically significantly enriched short sequences of about four to eight bases. Commonly applied algorithms for this step are CERMIT (Georgiev, Boyle et al. 2010) or MEME (Bailey and Elkan 1994).

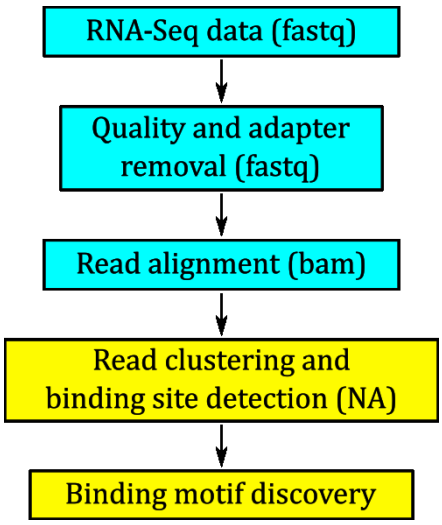


Figure 4: Sample pipeline for PAR-CLIP data processing. Cyan fields show preprocessing steps and yellow fields show an example for a downstream analysis of binding sites. The common data output format is depicted in parenthesis, where NA stands for not applicable.

1.4.3 Sequence read alignment and its evaluation

As explained in the previous section, the sequence read alignment against a reference genome sequence is an important step in analyzing RNA-Seq data. The general algorithm of BWA is outlined in the following, as it was modified for the purpose of this thesis for improved PAR-CLIP data analysis. First, a suffix array of all possible suffixes within the reference sequence is created and lexicographically sorted. Because of the sorting of the suffix array, all exact matches of a certain sequence read within the reference sequence will occur in an interval of the suffix array. The following description and formulas are taken from (Li and Durbin 2009): Let W be the sequence read, X the reference sequence and S the sorted suffix array of X. X_i is the suffix of X starting at position i. Then,

$$\underline{R}(W) = \min\{k: W \text{ is a prefix of } X_{S(k)}\}$$

and

$$\bar{R}(W) = \max\{k: W \text{ is a prefix of } X_{S(k)}\}$$

are the first and last occurrences of W in X, respectively, indicated by the suffix array S. So the set of positions $\{S(k): \underline{R}(W) \leq k \leq \overline{R}(W)\}$ represents all occurrences of W in X. The Burrows-Wheeler transform (Burrows and Wheeler 1994) is applied to the suffix array to create an index, the so called Ferragina and Manzini (FM) index (Ferragina and Manzini 2000). It compresses the suffix array to be linear in space requirements, which is important for large genomes (Lam, Sung et al. 2008). For each suffix of the reference sequence, the FM index points to its positions in the reference genome. It thus acts as a lookup table during the actual read alignment process. Another feature of the FM index is that it can also be represented by a suffix trie, a tree like data structure, which is an easy representation technique for the lookup process.

The actual alignment process is individually performed for each sequencing read, but has to accommodate for inexact matches between the read and the reference sequence. Hence, the alignment of a particular read starts with its last base proceeding to its front, a process considered a backward search. For the last base, the algorithm checks all possible positions within the reference sequence by querying the FM index/suffix trie. Next, only paths within the suffix trie are considered which show the same predecessor base as within the read. This process is recursively carried out until the first base of the sequencing read is reached. This identifies all possible mapping positions of the sequencing read within the reference sequence. Also, mismatches and indels are allowed until a predefined maximal threshold of mismatches and indels is reached, which increases the number of possible downstream paths within the suffix trie. If no further predecessor base can be aligned by checking the suffix trie (considering the mismatch threshold to be reached), the read is discarded as not aligned to the genome sequence.

An important aspect of software development is the performance evaluation of the implemented method. Recent advantages in read alignment algorithms were evaluated on simulated read datasets produced with for example ART (Huang, Li et al. 2012) or GemSim (McElroy, Luciani et al. 2012). These mimic regular RNA-Seq data, as either single-end or paired-end data covering entire transcripts with simulated sequencing

errors. Recent software specific for PAR-CLIP analysis was evaluated on reads produced with standard RNA-Seq read simulators (Kerpedjiev, Frellsen et al. 2014). However, these simulators do not include the PAR-CLIP specific read properties and are thus of limited use for the evaluation of PAR-CLIP specific analysis software.

2 Manuscripts

2.1 Publication I: *The PARA-suite: processing and aligning error-prone CLIP sequencing reads with empirical error model inference*

A. Kloetgen[1,2,3], A. Borkhardt[2], J. I. Hoell[2,*], A. C. McHardy[1,3,*]

[1]Department of Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany

[2]Department of Pediatric Oncology, Hematology and Clinical Immunology, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany

[3]Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany

[*]These authors have contributed equally to the work

Authorship:	first author
Contributed part:	85%
Contribution:	study design, implementation, data analysis, interpretation of results, writing of the manuscript
Journal:	PeerJ
Impact factor:	2.18
Status of publication:	published
DOI:	10.7717/peerj.2619
PubMed-ID:	27812418

2.2 Publication II: *Alterations of miRNAs and miRNA-regulated mRNA expression in GC B cell lymphomas determined by integrative sequencing analysis*

Hezaveh K[*,1], Kloetgen A[*,1,2], Bernhart SH[*,3,4,5], et al., Siebert R[§,7], Borkhardt A[§,1], Hummel M[§,6], Hoell JI[§,1] on behalf of the ICGC MMML-Seq Project[‡]

[1]Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Medical Faculty, Düsseldorf, Germany

[2]Department of Algorithmic Bioinformatics, Heinrich-Heine University, Duesseldorf, Germany

[3]Transcriptome Bioinformatics Group, LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany

[4]Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig, Germany

[5]Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany

[6]Institute of Pathology, Charité – University Medicine Berlin, Berlin, Germany

[7]Institute of Human Genetics, University Hospital Schleswig-Holstein Campus Kiel/ Christian-Albrechts University Kiel, Kiel, Germany

[*]These authors contributed equally to this work.

[‡]A list of all authors and affiliations appears in the supplementary information.

[§]These authors contributed equally to this work.

Authorship:	joint first author
Contributed part:	30%
Contribution:	bioinformatic analysis, interpretation of results, writing of the manuscript
Journal:	Haematologica
Impact factor:	6.67
Status of publication:	published
DOI:	10.3324/haematol.2016.143891
PubMed-ID:	27390358

2.3 Publication III: *T-cell acute lymphoblastic leukemia in infants has distinct genetic and epigenetic features compared to childhood cases*

Doerrenberg M[1], Kloetgen A[1,2], Wössmann W[3], Stanulla M[4], McHardy AC[2], Borkhardt A[1], Hoell JI[1]

[1]Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Medical Faculty, Düsseldorf, Germany

[2]Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany

[3]Department of Pediatric Hematology and Oncology, University Hospital Gießen and Marburg, Gießen, Germany

[4]Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany

Authorship:	co-author
Contributed part:	30%
Contribution:	bioinformatics analysis, interpretation of results, writing of the manuscript
Journal:	Genes, Chromosomes and Cancer
Impact factor:	3.96
Status of publication:	published
DOI:	10.1002/gcc.22423
PubMed-ID:	27717083

2.4 Other Publications

- Kloetgen A, Münch PC, Borkhardt A, Hoell JI, McHardy AC (2015). *Biochemical and bioinformatic methods for elucidating the role of RNA–protein interactions in posttranscriptional regulation*. Brief Funct Genomics. 2015 Mar;14(2):102-14.
- Shinde P, Xu HC, Maney SK, Kloetgen A, Namineni S, Bellora N, Trilling M, Pozdeev VI, van Rooijen N, Pfeffer K, Duggimpudi S, Höll JI, Borkhardt A, Knolle P, Heikenwalder M, Ruland J, Mak TW, Brenner D, Pandyra AA, Häussinger D, Lang KS, Lang PA. *TNF mediated survival of CD169+ cells mediate innate and adaptive immune activation during viral infection*. Under review.

3 Concluding remarks

Cancer is a complex disease and aberrations of multiple layers in the genome contribute to tumor formation, progression and more. Posttranscriptional gene regulation is involved in establishing cell diversity, but as a drawback makes the investigation of complex diseases, such as cancer, even more difficult. Although not all regulators for posttranscriptional gene regulation are fully understood, recent advantages of experimental protocols brought deeper insight into the functionality and importance of these regulators. This thesis contains the following contributions for advancing PAR-CLIP data analysis and applying posttranscriptional gene regulation analysis in certain cancer types:

a. Assessment of systematically induced T–C conversions in PAR-CLIP experiments, improvements of the read alignment of PAR-CLIP datasets implemented in the PARA-suite and a performance evaluation based on specifically simulated PAR-CLIP data and real datasets (Publication I)
b. Application of the PARA-suite on *AGO2* PAR-CLIPs performed in Non-Hodgkin lymphomas (NHL) to integrate knowledge on physically validated miRNA-mRNA interactions with RNA-Seq and miRNA-Seq data for an NHL patient cohort (Publication II)
c. Analysis of the mutational landscape, transcriptome and miRnome expression in infant T-cell acute lymphoblastic leukemia (iT-ALL) compared to childhood T-ALL (Publication III)

This thesis sheds light on current pitfalls in the analysis of PAR-CLIP data, which is a technique to reveal mRNAs that are posttranscriptionally regulated by a certain RBP, summarized in a review related to this thesis (Kloetgen, Münch et al. 2015). The here presented PARA-suite contributes to the improved analysis of PAR-CLIP data. The PARA-suite takes the unique properties of PAR-CLIP data into account, enabling a proper simulation of PAR-CLIP data for the evaluation of read aligners and binding site

detection algorithms, and offering an accurate read alignment pipeline. Compared to previous approaches, the PARA-suite alignment pipeline shows increased accuracy on simulated and real PAR-CLIP data by empirical error profile estimation for individual datasets.

PAR-CLIP data has specific properties compared to RNA-Seq data. Thus PAR-CLIP reads cannot be simulated by regular RNA-Seq read simulators such as ART or GemSim. The PARA-suite contains a read simulator specific for PAR-CLIP data, which mimics the properties of PAR-CLIP reads. It simulates reads that cover only short binding sites, and introduces T–C conversions at much higher rates at specific T–C sites, rather than randomly introducing T–C conversions similar to sequencing errors. The PARA-suite's read simulator is of general use for the scientific community to evaluate all aspects of PAR-CLIP data analysis, from read alignment to binding site detection. It returns reads in a FASTQ-format for read alignment evaluation and the respective binding sites for the evaluation of binding site detection algorithms.

The alignment method of the PARA-suite, which is based on the implementation of BWA, makes use of an error profile calculated from the actual sequencing run to accommodate for data type specific errors. It is not limited to single nucleotide conversions but can also estimate increased rates of insertions and deletions. Hence, the application of the PARA-suite to a HITS-CLIP dataset also revealed additional promising binding sites than previous approaches. Sequencing data in general show varying rates of sequencing errors and systematic errors, depending on the sample preparation or the sequencing protocol (Laehnemann, Borkhardt et al. 2015, Schirmer, Ijaz et al. 2015). The PAR-CLIP datasets analyzed for the purposes of this work showed that this phenomenon is generally important for PAR-CLIP data, as different RBPs show slight variations in the frequencies of T–C conversions per dataset. The read aligner BWA PSSM always uses a fixed PSSM for the respective sequencing datatype irrespective of varying qualities and error rates. However, the PARA-suite automatically adjusts to the sequencing-dependent changes by the inference of an error profile per each sequencing dataset. The evaluation of the PARA-suite alignment pipeline showed an increased accuracy on simulated PAR-CLIP data compared to commonly employed

read aligners. These aligners (including BWA, Bowtie and BWA PSSM) were assessed using different parameter settings, allowing for different numbers of mismatches per application. However, they were still outperformed by the PARA-suite on the simulated datasets. Further assessments of the PARA-suite on real PAR-CLIP data showed its potential for highly accurate data analysis. More examples for the advantage of the adaptive behavior of the PARA-suite on different data types are applications to bisulphite sequencing, which introduces C–T conversions in methylated CpG islands (Frommer, McDonald et al. 1992), or low-quality ancient DNA sequencing data (Briggs, Stenzel et al. 2007).

The PARA-suite was successfully applied to PAR-CLIP data on *AGO2* within a project of the ICGC MMML-Seq showing the importance of posttranscriptional gene regulation in distinct lymphoma subtypes. The subsequent data integration granted deep insights into tumor biology of miRNA-mediated regulations in NHLs: The AGO-PAR-CLIP provided experimental evidence for miRNA-mRNA interaction pairs and the RNA-Seq provided expression measures for both miRNAs and mRNAs. In addition, the proposed miRNA expression classifier, consisting of new biomarkers, is useful for distinguishing the heterogenic NHLs for clinical purposes.

The AGO-PAR-CLIP was applied to endogenously expressed *AGO2* in two BL and two DLBCL cell lines, to reveal mRNAs under miRNA-mediated regulation in these cell types. After identifying the *AGO2* binding sites within mRNAs with the full use of the PARA-suite, these were considered to be regulated by expressed miRNAs in the cells. For the identification of miRNA-mRNA pairs currently under regulation in the cells, a similar method as recently described was applied (Farazi, Ten Hoeve et al. 2014). The detected binding sites of *AGO2* were searched for complementary seed sequences of all annotated and expressed miRNAs. In contrast to recent approaches, this revealed experimentally validated and regulated miRNA-mRNA pairs in BL and DLBCL cell lines. The combination of the PAR-CLIP results with RNA-Seq and miRNA-Seq data revealed insights into active regulations that would have been missed with any of the datasets alone. For further downstream analyses, only those miRNA-mRNA interaction pairs

were used which showed a negative correlation in their expression pattern. A functional analysis of the interaction pairs found in BL and DLBCL patients identified them to be highly associated with lymphomagenesis and tumor formation in general.
The miRNA-Seq data obtained in the ICGC MMML-Seq project was further used for an analysis of potential oncogenic and tumor suppressive miRNAs in BL, DLBCL and FL. A differential gene expression analysis for miRNAs between BL and the combined DLBCL/FL cases resulted in a classifier consisting of 22 distinct miRNAs clearly separating the subtypes. An external validation on microarray data for 150 BL and DLBCL cases confirmed the validity of the classifier, which showed an accuracy of 84%.

The last part of this thesis dealt with the differences in the transcriptome and miRnome between iT-ALL patients and childhood T-ALL patients. RNA-Seq, miRNA-Seq and whole exome sequencing data of a total of three infant T-ALL patients and six childhood T-ALL patients were obtained. The gene expression analysis of this data revealed epigenetic regulations as potential biomarkers for T-ALL in infancy. The analyses on the combined miRNA and mRNA expression patterns showed that the pathophysiology of iT-ALL seems to differ compared to childhood T-ALL. This includes expressional differences in miRNA-mRNA pairs and pathways important for immune system responses. However, whether these differences account for the worse outcome of iT-ALL patients remains to be clinically validated.
RNA-Seq and miRNA-Seq data were used to identify significant expressional differences between iT-ALL and childhood T-ALL cases. To predict transcriptomic changes mediated by miRNAs, information from a total of five public databases for miRNA-mRNA interactions were downloaded and combined. A miRNA-mRNA interaction pair was accepted for downstream analysis, if it was either experimentally validated or predicted by at least two databases. This overcomes issues of a single database lacking completeness or showing not reliably high numbers of interactions. The analysis revealed 62 miRNA-mRNA pairs having a negative correlation in their expression pattern between iT-ALL and childhood T-ALL. Many of these negatively correlating miRNA-mRNA pairs were already associated with tumorigenesis. For instance, the

analysis revealed that *hsa-let-7b* is negatively correlating with *GREB1* and *IGF2BP1*, both associated with different cancers including ALL when upregulated (Stoskus, Gineikiene et al. 2011, Mohammed, D'Santos et al. 2013).

The characterization of the iT-ALL patients also included the analysis of the mutational landscape in iT-ALL by whole exome sequencing. It revealed mutations (SNPs and indels) to be associated with the rare iT-ALL cases, which are also recurrently mutated in T-ALL in general. The most important mutations in *NOTCH2*, *IL7R*, *KRAS* and *PTEN* were validated by Sanger sequencing.

4 References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* **7**(4): 248-249.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T. and Yu, X. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**(6769): 503-511.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11**(10): R106.

Anders, S., Pyl, P. T. and Huber, W. (2014). HTSeq–A Python framework to work with high-throughput sequencing data. *bioRxiv*.

Anderson, K. N., Schwab, R. B. and Martinez, M. E. (2014). Reproductive risk factors and breast cancer subtypes: a review of the literature. *Breast Cancer Res. Treat.* **144**(1): 1-10.

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., Langlois, C., Munschauer, M., Dewell, S. and Hafner, M. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**(7429): 382-386.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Barrett, L. W., Fletcher, S. and Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.* **69**(21): 3613-3634.

Bergsagel, P. L. and Kuehl, W. M. (2001). Chromosome translocations in multiple myeloma. *Oncogene* **20**(40).

Bishop, J. A., Benjamin, H., Cholakh, H., Chajut, A., Clark, D. P. and Westra, W. H. (2010). Accurate classification of non–small cell lung carcinoma using a novel microRNA-based approach. *Clin. Cancer Res.* **16**(2): 610-619.

Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*: btu170.

Bos, J. L. (1989). Ras oncogenes in human cancer: a review. *Cancer Res.* **49**(17): 4682-4689.

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

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T. and Lachmann, M. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U S A* **104**(37): 14616-14621.

Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. *Digital Equipment Corporation: Palo Alto, CA* (Technical Report 124).

Campo, E., Swerdlow, S. H., Harris, N. L., Pileri, S., Stein, H. and Jaffe, E. S. (2011). The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood* **117**(19): 5019-5032.

Chen, C. (2005). MicroRNAs as oncogenes and tumor suppressors. *N. Engl. J. Med.* **353**(17): 1768.

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y. and Tu, S.-J. (2015). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*: gkv1258.

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D. and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.* **12**(8): R79.

Coulon, A., Chow, C. C., Singer, R. H. and Larson, D. R. (2013). Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat Rev Genet.* **14**(8): 572-584.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T.,

Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R. and Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Res.* **43**(Database issue): D662-669.

De Bont, R. and Van Larebeke, N. (2004). Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**(3): 169-185.

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. and Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**(7014): 231-235.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A. and Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**(5): 491-498.

Dietlein, F., Thelen, L. and Reinhardt, H. C. (2014). Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. *Trends Genet.* **30**(8): 326-339.

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

Eisenberg, I., Eran, A., Nishino, I., Moggio, M., Lamperti, C., Amato, A. A., Lidov, H. G., Kang, P. B., North, K. N. and Mitrani-Rosenbaum, S. (2007). Distinctive patterns of microRNA expression in primary muscular disorders. *Proc. Natl. Acad. Sci. U S A* **104**(43): 17016-17021.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S. (2004). MicroRNA targets in Drosophila. *Genome Biol.* **5**(1): R1-R1.

Farazi, T. A., Spitzer, J. I., Morozov, P. and Tuschl, T. (2011). miRNAs in human cancer. *The Journal of pathology* **223**(2): 102-115.

Farazi, T. A., Ten Hoeve, J. J., Brown, M., Mihailovic, A., Horlings, H. M., van de Vijver, M. J., Tuschl, T. and Wessels, L. F. (2014). Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets. *Genome Biol.* **15**(1): R9.

Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. *Proceedings of the 41st Symposium on Foundations of Computer Science*: 390-398.

Friedberg, E. C., McDaniel, L. D. and Schultz, R. A. (2004). The role of endogenous and exogenous DNA damage and mutagenesis. *Curr. Opin. Genet. Dev.* **14**(1): 5-10.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U S A* **89**(5): 1827-1831.

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A. and Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.* **18**(10): 1139-1146.

GenomesProjectConsortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

GenomesProjectConsortium (2015). A global reference for human genetic variation. *Nature* **526**(7571): 68-74.

Georgiev, S., Boyle, A. P., Jayasurya, K., Ding, X., Mukherjee, S. and Ohler, U. (2010). Evidence-ranked motif identification. *Genome Biol.* **11**(2): R19.

Glisovic, T., Bachorik, J. L., Yong, J. and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**(14): 1977-1986.

Goldberg, J. M., Silverman, L. B., Levy, D. E., Dalton, V. K., Gelber, R. D., Lehmann, L., Cohen, H. J., Sallan, S. E. and Asselin, B. L. (2003). Childhood T-cell acute lymphoblastic leukemia: the Dana-Farber Cancer Institute acute lymphoblastic leukemia consortium experience. *J. Clin. Oncol.* **21**(19): 3616-3622.

Golumbeanu, M., Mohammadi, P. and Beerenwinkel, N. (2015). BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data. *Bioinformatics*: btv520.

Greaves, M. F. and Wiemels, J. (2003). Origins of chromosome translocations in childhood leukaemia. *Nature Reviews Cancer* **3**(9): 639-649.

Gupta, A. and Gribskov, M. (2011). The role of RNA sequence and structure in RNA–protein interactions. *J. Mol. Biol.* **409**(4): 574-587.

Gutiérrez, N. C., Sarasquete, M. E., Misiewicz-Krzeminska, I., Delgado, M., De Las Rivas, J., Ticona, F., Ferminan, E., Martin-Jimenez, P., Chillon, C. and Risueno, A. (2010). Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. *Leukemia* **24**(3): 629-637.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C. and Munschauer, M. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**(1): 129-141.

Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *N. Engl. J. Med.* **363**(4): 301-304.

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

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

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W. and Hannon, G. J. (2005). A microRNA polycistron as a potential human oncogene. *Nature* **435**(7043): 828-833.

Henry, N. L. and Hayes, D. F. (2012). Cancer biomarkers. *Mol. Oncol.* **6**(2): 140-146.

Hilden, J. M., Dinndorf, P. A., Meerbaum, S. O., Sather, H., Villaluna, D., Heerema, N. A., McGlennen, R., Smith, F. O., Woods, W. G. and Salzer, W. L. (2006). Analysis of prognostic factors of acute lymphoblastic leukemia in infants: report on CCG 1953 from the Children's Oncology Group. *Blood* **108**(2): 441-451.

Hoeijmakers, J. H. (2001). Genome maintenance mechanisms for preventing cancer. *Nature* **411**(6835): 366-374.

Holoch, D. and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet.* **16**(2): 71-84.

Huang, W., Li, L., Myers, J. R. and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* **28**(4): 593-594.

Ichikawa, H., Shimizu, K., Hayashi, Y. and Ohki, M. (1994). An RNA-binding protein gene, TLS/FUS, is fused to ERG in human myeloid leukemia with t (16; 21) chromosomal translocation. *Cancer Res.* **54**(11): 2865-2868.

Jackson, R. J., Hellen, C. U. and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature reviews Molecular cell biology* **11**(2): 113-127.

Jackson, S. P. and Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature* **461**(7267): 1071-1078.

Kerpedjiev, P., Frellsen, J., Lindgreen, S. and Krogh, A. (2014). Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics* **15**(1): 100.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**(4): R36.

Kloetgen, A., Münch, P. C., Borkhardt, A., Hoell, J. I. and McHardy, A. C. (2015). Biochemical and bioinformatic methods for elucidating the role of RNA-protein interactions in posttranscriptional regulation. *Brief. Funct. Genomics* **14**(2): 102-114.

Kozomara, A. and Griffiths-Jones, S. (2010). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*: gkq1027.

Kumar, P., Henikoff, S. and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**(7): 1073-1081.

Kwiatkowski, T. J., Bosco, D., Leclerc, A., Tamrazian, E., Vanderburg, C., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E. and Munsat, T. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* **323**(5918): 1205-1208.

Laehnemann, D., Borkhardt, A. and McHardy, A. C. (2015). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Brief. Bioinform.*: bbv029.

Lam, T. W., Sung, W.-K., Tam, S.-L., Wong, C.-K. and Yiu, S.-M. (2008). Compressed indexing and local alignment of DNA. *Bioinformatics* **24**(6): 791-797.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**(4): 357-359.

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

Lee, Y., Jeon, K., Lee, J. T., Kim, S. and Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO journal* **21**(17): 4663-4670.

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

Liao, Y., Smyth, G. K. and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**(10): e108.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H. and Ferrando, A. A. (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**(7043): 834-838.

Mansur, M. B., Delft, F. W., Colman, S. M., Furness, C. L., Gibson, J., Emerenciano, M., Kempski, H., Clappier, E., Cave, H. and Soulier, J. (2015). Distinctive genotypes in infants with T-cell acute lymphoblastic leukaemia. *Br. J. Haematol.* **171**(4): 574-584.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**(1): 10-12.

McElroy, K. E., Luciani, F. and Thomas, T. (2012). GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* **13**(1): 74.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**(16): 2069-2070.

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

Metzler, M., Wilda, M., Busch, K., Viehmann, S. and Borkhardt, A. (2004). High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma. *Genes Chromosomes Cancer* **39**(2): 167-169.

Mohammed, H., D'Santos, C., Serandour, A. A., Ali, H. R., Brown, G. D., Atkins, A., Rueda, O. M., Holmes, K. A., Theodorou, V. and Robinson, J. L. (2013). Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor. *Cell reports* **3**(2): 342-349.

Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, M., Tuschl, T., Ohler, U. and Keene, J. D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell* **43**(3): 327-339.

Mukherjee, N., Jacobs, N. C., Hafner, M., Kennington, E. A., Nusbaum, J. D., Tuschl, T., Blackshear, P. J. and Ohler, U. (2014). Global target mRNA specification and regulation by the RNA-binding protein ZFP36. *Genome Biol.* **15**(1): R12.

Necsulea, A. and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet.* **15**(11): 734-748.

Nesbit, C. E., Tersak, J. M. and Prochownik, E. V. (1999). MYC oncogenes and human neoplastic disease. *Oncogene* **18**(19).

Nestler, E. J. and Hyman, S. E. (2002). Regulation of gene expression. *Neuropsychopharmacology: The Fifth Generation of Progress*: 217-228.

Pui, C.-H., Carroll, W. L., Meshinchi, S. and Arceci, R. J. (2011). Biology, risk stratification, and therapy of pediatric acute leukemias: an update. *J. Clin. Oncol.* **29**(5): 551-565.

Quackenbush, J. (2006). Microarray analysis and tumor classification. *N. Engl. J. Med.* **354**(23): 2463-2472.

Raponi, M., Dossey, L., Jatkoe, T., Wu, X., Chen, G., Fan, H. and Beer, D. G. (2009). MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. *Cancer Res.* **69**(14): 5776-5783.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1): 139-140.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L. and Haeussler, M. (2015). The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* **43**(D1): D670-D681.

Rowley, J. D. (1998). The critical role of chromosome translocations in human leukemias. *Annu. Rev. Genet.* **32**(1): 495-519.

Sampson, V. B., Rong, N. H., Han, J., Yang, Q., Aris, V., Soteropoulos, P., Petrelli, N. J., Dunn, S. P. and Krueger, L. J. (2007). MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells. *Cancer Res.* **67**(20): 9762-9770.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*: gku1341.

Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012). Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.* **40**(20): e160.

Singer, S., Socci, N. D., Ambrosini, G., Sambol, E., Decarolis, P., Wu, Y., O'Connor, R., Maki, R., Viale, A. and Sander, C. (2007). Gene expression profiling of liposarcoma identifies distinct biological types/subtypes and potential therapeutic targets in well-differentiated and dedifferentiated liposarcoma. *Cancer Res.* **67**(14): 6626-6636.

Sinha, R., Nastoupil, L. and Flowers, C. R. (2012). Treatment strategies for patients with diffuse large B-cell lymphoma: past, present, and future. *Blood and lymphatic cancer: targets and therapy* **2012**(2): 87.

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J. and Pegram, M. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344**(11): 783-792.

Stoskus, M., Gineikiene, E., Valceckiene, V., Valatkaite, B., Pileckyte, R. and Griskevicius, L. (2011). Identification of characteristic IGF2BP expression patterns in distinct B-ALL entities. *Blood Cells Mol. Dis.* **46**(4): 321-326.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M. and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.* **13**(8): R67.

Vance, C., Rogelj, B., Hortobágyi, T., De Vos, K. J., Nishimura, A. L., Sreedharan, J., Hu, X., Smith, B., Ruddy, D. and Wright, P. (2009). Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* **323**(5918): 1208-1211.

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T. and Hatzigeorgiou, A. G. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.* **40**(D1): D222-D229.

Wang, Y.-P. and Li, K.-B. (2009). Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics* **10**(1): 218.

Weinberg, R. (2013). *The biology of cancer*, Garland science.

Weinberg, R. A. (1995). The retinoblastoma protein and cell cycle control. *Cell* **81**(3): 323-330.

Wong, N. and Wang, X. (2014). miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.*: gku1104.

Xiao, C., Srinivasan, L., Calado, D. P., Patterson, H. C., Zhang, B., Wang, J., Henderson, J. M., Kutok, J. L. and Rajewsky, K. (2008). Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes. *Nat. Immunol.* **9**(4): 405-414.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S. and Gil, L. (2016). Ensembl 2016. *Nucleic Acids Res.* **44**(D1): D710-D716.

Zhang, C. and Darnell, R. B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29**(7): 607-614.

Acknowledgements

I want to acknowledge all the help from all the people who helped me realizing my work during the last years. It was an outstanding experience for me to work in such an interdisciplinary environment, and without the support of my colleagues, my family and my friends this thesis would not have been possible.

In particular, I would like to thank my supervisors Alice McHardy, Arndt Borkhardt and Jessica Höll, who supported me very well with their expertise but also motivated me all the time. I would also like to thank the Düsseldorf School of Oncology (DSO), in particular Cornelia Höner, for the educational and financial support during that time. I further would like to thank all my colleagues from the group for Computational Biology of Infection Research at the Helmholtz Center for Infection Research Braunschweig and all members of the Department for Pediatric Oncology, Hematology and Clinical Immunology at HHU Düsseldorf.

Most importantly, I am greatly indebted to my parents, Carmen and Hans Udo Klötgen, as well as to my brother Marcel and his wife Natalie Klötgen for their encouragement and help during this time.

Of course, many more people helped me with professional and personal discussions making me having a great time in Düsseldorf. So I can only say thanks to everyone!

Appendix

Publication I

The PARA-suite: PAR-CLIP specific sequence read simulation and processing

Andreas Kloetgen[1,2,3], Arndt Borkhardt[2], Jessica I. Hoell[2,§], Alice C. McHardy[1,3,§,*]

[1]Department of Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany.
[2]Department of Pediatric Oncology, Hematology and Clinical Immunology, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany.
[3]Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany.
[§]These authors contributed equally to the work.
[*]To whom correspondence should be addressed.

Running head: Simulating and processing PAR-CLIP data

Abstract

Background: Next-generation sequencing technologies have profoundly impacted biology over recent years. Experimental protocols, such as photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP), which identifies protein–RNA interactions on a genome-wide scale, commonly employ deep sequencing. With PAR-CLIP, the incorporation of photoactivatable nucleosides into nascent transcripts leads to high rates of specific nucleotide conversions during reverse transcription. So far, the specific properties of PAR-CLIP-derived sequencing reads have not been assessed in depth.

Methods: Here, we compared PAR-CLIP sequencing reads to regular transcriptome sequencing reads (RNA-Seq) to identify distinctive properties that are relevant for reference-based read alignment of PAR-CLIP datasets. We describe a set of freely available tools for this purpose, called the PAR-CLIP analyzer suite (PARA-suite). The PARA-suite includes error model inference, PAR-CLIP read simulation based on PAR-CLIP specific properties, a full read alignment pipeline with a modified Burrows–Wheeler Aligner algorithm and CLIP read clustering for binding site detection.

Results: We show that differences in the error profiles of PAR-CLIP reads relative to regular transcriptome sequencing reads (RNA-Seq) make a distinct processing advantageous. We examine the alignment accuracy of commonly applied read aligners on 10 simulated PAR-CLIP datasets using different parameter settings and identified the most accurate setup among those read aligners. We demonstrate the performance of the PARA-suite in conjunction with different binding site detection algorithms on several real PAR-CLIP and HITS-CLIP datasets. Our processing pipeline allowed the improvement of both alignment and binding site detection accuracy.

Availability: The PARA-suite toolkit and the PARA-suite aligner are available at https://github.com/akloetgen/PARA-suite and https://github.com/akloetgen/PARA-suite_aligner, respectively, under the GNU GPLv3 license.

Background

RNAs play a crucial role in cell survival and viability. Coding messenger RNAs (mRNAs), which are translated into proteins, and many other RNA species, such as small and long non-coding RNAs, ribosomal RNAs and transfer RNAs, are essential for the survival and proper functioning of the cells (Eddy 2001). Most RNAs maintain their function by working together with the so-called RNA-binding proteins (RBPs) (Glisovic, Bachorik et al. 2008). RBPs are involved in virtually all steps of the mRNA lifecycle, from polyadenylation, translocation and modification to translation (Hieronymus and Silver 2004). Thus it is not surprising that many RBPs that show aberrant functions or changes in expression patterns have been associated with disease progression or even with carcinogenesis (Lukong, Chang et al. 2008). For instance, the FET protein family, which consists of the three RBPs FUS, EWSR1 and TAF15, is ubiquitously expressed and widely conserved in mammals. Genomic rearrangements, leading to mutant forms of these RBPs in humans, have been described as key players in sarcomas and leukemia (Tan and Manley 2009). More recently, two mutants of FUS causing amyotrophic lateral sclerosis have shown different RNA-binding patterns compared to their wild-type counterparts, supporting the importance of the function of FUS in mRNA processing (Hoell, Larsson et al. 2011).

Experimental protocols have been developed to analyze the functional network in which a particular RBP interacts. A promising method for this purpose is the photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) technique (Hafner, Landthaler et al. 2010). When coupled with deep sequencing, it identifies the bound RNAs for a particular RBP on a genome-wide scale. First, the cells are supplied with a specific photoactivatable nucleoside, such as 4-thiouridine (4-SU), which is incorporated as an alternative to the respective nucleoside into nascent mRNA transcripts. Afterwards, the cells are treated with ultraviolet light at 365 nm to cross-link the amino acids of RBPs to the nucleotides of their bound RNA molecules. The incorporation of 4-SU instead of uridine results in nucleotide

conversions from uridine to cytidine at all cross-linked sites containing a 4-SU during reverse transcription (a necessary step for preparing cDNA libraries for sequencing). This specific replacement is called a 'T–C conversion'. T–C conversions can be used to distinguish between non-specifically bound RNA fragments (considered as contaminations) and those that are specifically bound and cross-linked to the RBP of interest (Ascano, Hafner et al. 2012, Golumbeanu, Mohammadi et al. 2015). We recently published a detailed protocol for the PAR-CLIP procedure (Hoell, Hafner et al. 2014). Other CLIP protocols for the genome-wide identification of RBP targets are also frequently used, such as high-throughput sequencing of RNAs isolated by cross-linking and immunoprecipitation (HITS-CLIP, sometimes also called CLIP-seq) or the iCLIP protocol (Chi, Zang et al. 2009, König, Zarnack et al. 2010). The procedures, experimental designs and bioinformatic analysis of these different CLIP methods differ greatly and are still evolving. Recent reviews compare the strengths and weaknesses of the three methods in detail (Wang, Xiao et al. 2015, Danan, Manickavel et al. 2016). HITS-CLIP, for example, mainly introduces deletions of a single base at the cross-linked sites, whereas single nucleotide conversions do not seem to occur at a significant frequency (Zhang and Darnell 2011, Sugimoto, König et al. 2012).

Current sequencing platforms allow for the sequencing of mammalian transcriptome libraries with high coverage. Nowadays, the most commonly used next-generation sequencing (NGS) platforms are 454, Illumina, IonTorrent and PacBio (van Dijk, Auger et al. 2014). Depending on the sequencing platform and the sample type, sequencing errors vary in type and frequency. The errors that most commonly occur are substitution errors and indels of a few bases between the sequencing read and the reference sequence (large rearrangements, such as those leading to chimeras, are also possible errors but are not discussed here) (Laehnemann, Borkhardt et al. 2015). In an RNA-Seq dataset, a single transcript will be covered by sequencing reads in all its expressed coding exons (apart from, for example, amplification errors or alternative splicing variants). For common sequencing data types, such as RNA-Seq and DNA-Seq, designated read aligners have recently been developed. These include short read

aligners, such as BWA (Li and Durbin 2009) or Bowtie (Langmead, Trapnell et al. 2009), and read aligners such as TopHat (Trapnell, Pachter et al. 2009), STAR (Dobin, Davis et al. 2013) or Subjunc (Liao, Smyth et al. 2013), which can also handle longer sequencing reads spanning exon–exon junctions. Specific software for the evaluation and analysis of the PAR- and HITS-CLIP sequencing data is needed to accommodate their unique error profiles (Kloetgen, Münch et al. 2015). For instance, the read aligner BWA PSSM (Kerpedjiev, Frellsen et al. 2014) makes use of a pre-defined position-specific scoring matrix to process the error-prone PAR-CLIP reads.

In general, the sequencing error profiles of RNA-Seq datasets, including PAR-CLIP data, can vary between different sequencing runs, depending on the sequencing machine, the experimental conditions and the biological properties of the sample (Laehnemann, Borkhardt et al. 2015, Schirmer, Ijaz et al. 2015). Here, we describe the PAR-CLIP analyzer suite (PARA-suite), which includes a PAR-CLIP read simulator, an error estimation tool for CLIP datasets and an alignment pipeline based on a novel alignment algorithm performing on-the-fly dataset-specific error estimation. The alignment pipeline thus automatically adjusts to the quality and error profiles of individual sequencing datasets. We compare PAR-CLIP sequencing reads to regular transcriptome sequencing reads (RNA-Seq) to identify the distinctive properties that are relevant for reference-based read alignment and RBP binding site detection from PAR-CLIP datasets. Generation of simulated PAR-CLIP datasets can be performed with the PARA-suite´s read simulator. The PARA-suite toolkit is available at https://github.com/akloetgen/PARA-suite and https://github.com/akloetgen/PARA-suite aligner, implemented as an extension of BWA (henceforth referred to as BWA PARA). It is licensed under GNU GPLv3, and can be implemented in the programming languages Java and C.

Methods

Datasets and read aligners

We downloaded PAR-CLIP data for the FET family (EWSR1, FUS and TAF15) from the DRASearch database (https://trace.ddbj.nig.ac.jp/DRASearch/) with the accession number SRA025082 (Hoell, Larsson et al. 2011), the HuR dataset with the accession number SRR248532, the MOV10 dataset with the accession number SRR490650 and the HITS-CLIP data on the Argonaute2 protein (AGO2) (Chi, Zang et al. 2009) from http://ago.rockefeller.edu/. For estimating the error profiles of regular RNA-Seq runs, we downloaded two sequencing lanes from an NGS quality assessment study with the accession numbers SRR896663 and SRR896664 (SEQC/MAQC-III-Consortium 2014) from DRASearch and pooled the data. An overview of the analyzed datasets can be found in Table 1.

Table 1: Overview of the analyzed RNA-Seq and CLIP datasets.

Dataset	**Published (year)**	**Sequencing method**	**Platform**	**Accession number/website**
EWSR1	2011	PAR-CLIP	Illumina Genome Analyzer II	SRA025082
FUS	2011	PAR-CLIP	Illumina Genome Analyzer II	SRA025082
TAF15	2011	PAR-CLIP	Illumina Genome Analyzer II	SRA025082
HuR	2011	PAR-CLIP	Illumina Genome Analyzer	SRR248532
MOV10	2012	PAR-CLIP	Illumina Genome Analyzer II	SRR490650
AGO2	2009	HITS-CLIP	Illumina Genome Analyzer II	http://ago.rockefeller.edu/

Human reference RNA	2014	RNA-Seq	Illumina HiSeq 2000	SRR896663, SRR896664

We used the following read aligners and versions, shown in alphabetic order: Bowtie, version 0.12.7 (Langmead, Trapnell et al. 2009), Bowtie2, version 2.2.3 (Langmead and Salzberg 2012), BWA, version 0.7.8 (Li and Durbin 2009), BWA PSSM, initial release version (Kerpedjiev, Frellsen et al. 2014), MOSAIK, version 2.2.3 (Lee, Stromberg et al. 2014), STAR, version 2.3.0 (Dobin, Davis et al. 2013), Subjunc, version 1.4.2 (Liao, Smyth et al. 2013) and TopHat, version 2.0.13 (Trapnell, Pachter et al. 2009).

PAR-CLIP read simulator and hierarchical clustering

We developed a PAR-CLIP read simulator (Figure 1) that creates short RNA reads which mimic important PAR-CLIP specific properties (Section 3.1). First, the following probability distributions are obtained from real PAR-CLIP data: (a) a probability matrix ε representing the background error profile of sequencing errors, (b) a probability vector of T–C conversion frequencies α for ranked T–C conversion sites, (c) a probability vector β for the preferred read positions of T–C conversion sites within binding sites, (d) a probability vector μ for indel frequencies per read position and (e) a probability vector δ for the base-calling quality score distribution per read position. The probability matrix ε contains a probability distribution for each DNA base over the DNA bases {A, C, G, T}. For this purpose, a PAR-CLIP dataset is aligned against a reference genome sequence with an appropriate read aligner.

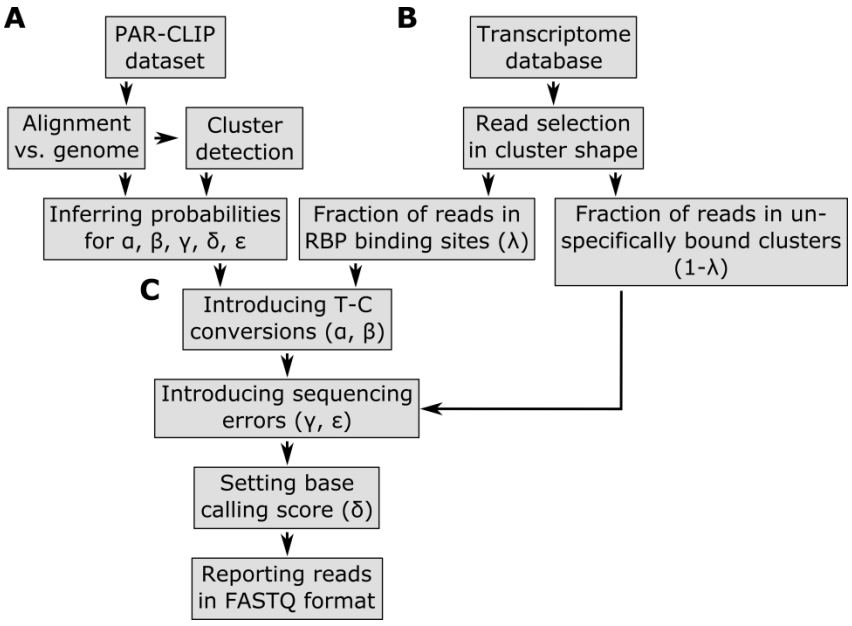


Figure 1: Pipeline of the PAR-CLIP read simulator implemented in the PARA-suite. Part A describes the process of generating the error profile and other parameters learned from a real PAR-CLIP dataset. Part B starts to generate reads mapping to RBP binding sites (clusters) on transcript regions from a given transcript database (e.g. Ensembl genes). In Part C, the pre-calculated profiles are used to introduce T–C conversions, sequencing errors, indels and base-calling quality scores to the defined reads.

Based on these alignments, the sequencing error profile ε is estimated from the observed frequencies of all single nucleotide substitutions, except for T–C errors, as these include PAR-CLIP specific T–C conversions. Standard T–C sequencing errors are approximated by the average over all the other sequencing error frequencies. The probability vectors μ and δ are also inferred from these alignments. Next, all aligned reads of the real dataset are clustered (stacked) using single-linkage hierarchical clustering based on their genomic mapping positions, using a 5-base overlap of the genomic mapping positions as the clustering threshold. To identify high confidence clusters (sometimes referred to as binding sites) as defined in the literature (Hafner, Landthaler et al. 2010), clusters that contain less than 10 reads, less than 25% T–C

conversions per cluster, are longer than 75 bases and include only T–C conversion sites that are reported as single nucleotide polymorphism loci in the dbSNP database (version 142) (Sherry, Ward et al. 2001) are discarded. This implementation of hierarchical clustering is part of the PARA-suite and will later be used for binding site detection. For the subsequent simulation, the positions and frequencies of highly mutated T–C sites within reads are determined to estimate α and β from the high confidence clusters (Figure S1A-B).

Next, the PAR-CLIP read simulation starts with the random selection of transcripts from a pre-selected database of annotated transcripts. One to at most three clusters (the number of clusters is randomly chosen from a uniform distribution) containing several reads are created for a selected transcript sequence. The starting positions of the clusters are randomly selected from a uniform distribution within the entire range of a transcript. The number of reads simulated for a single cluster is drawn from a normal distribution with a mean of 16 and a standard deviation of 10. This enables the simulation of a wide range of read coverages throughout the clusters. Furthermore, small shifts of the start and end site of each read leading to distinctive alignment position shifts in the shape of a cluster are randomly introduced at this step (normal distribution, standard deviation = 1). A user-defined parameter $\lambda \in [0,1]$ specifies the fraction of clusters that are considered to be binding-sites, whereas the remaining clusters mimic contaminations of unbound RNAs that occur in all PAR-CLIP experiments. We recommend values in the range of 0.5–0.7 (50–70%), as we observed this range of aligned sequencing reads stacking into clusters after hierarchical clustering and filtering (Table S1; similar values were previously reported by (Ascano, Hafner et al. 2012)). If more than one T–C site is simulated for a single cluster, a major T–C conversion site is selected according to the site-specific T–C conversion profile β and T–C conversion probabilities are drawn from α. Subsequently, background sequencing errors are introduced on the basis of the pre-computed probability matrix ε and the frequency vector μ for substitutions and indels, respectively. In the last step, every base receives a base-calling quality score, as specified by the position-specific

quality score distribution δ. All generated reads are stored in the universal FASTQ format (Cock, Fields et al. 2010). The PAR-CLIP read simulator is available through the PARA-suite.

The PARA-suite: tools for error profile inference, read simulation, multiple database mapping and more

The PARA-suite is a toolkit for processing and aligning short and error-prone sequencing reads. It is implemented in Java using HTSjdk, a Java API for high-throughput sequencing data formats (https://github.com/samtools/htsjdk). The PARA-suite allows the user to estimate a sequencing run-specific error profile, combine the results of multiple reference database alignments, cluster an aligned sequencing read dataset (Section 2.2), run the PAR-CLIP read simulator, benchmark an alignment of simulated PAR-CLIP sequencing reads and run a full processing pipeline for error-prone short read alignments (Figure 2A). The alignment pipeline of the PARA-suite includes the calculation of an error profile for a particular sequencing run, applying the alignment algorithm described in the following section, and optionally combines the results of read mappings against multiple databases (Figure 2B–D). First, a read alignment against a reference sequence is performed with a fast short read aligner. By default, this is carried out with BWA, as our evaluations have demonstrated this to be a fast and accurate aligner (Section 3.3) on PAR-CLIP reads. However, other read aligners can also be used to produce the reference-based read alignment. This initial read alignment is used to estimate the underlying mismatch and indel probabilities M, I and D (as described in the next section) of the sequencing run. Once the error profile has been estimated, all sequencing reads can be aligned with BWA PARA (Section 2.4) against the reference sequence(s). All aligned reads are reported in a BAM file.

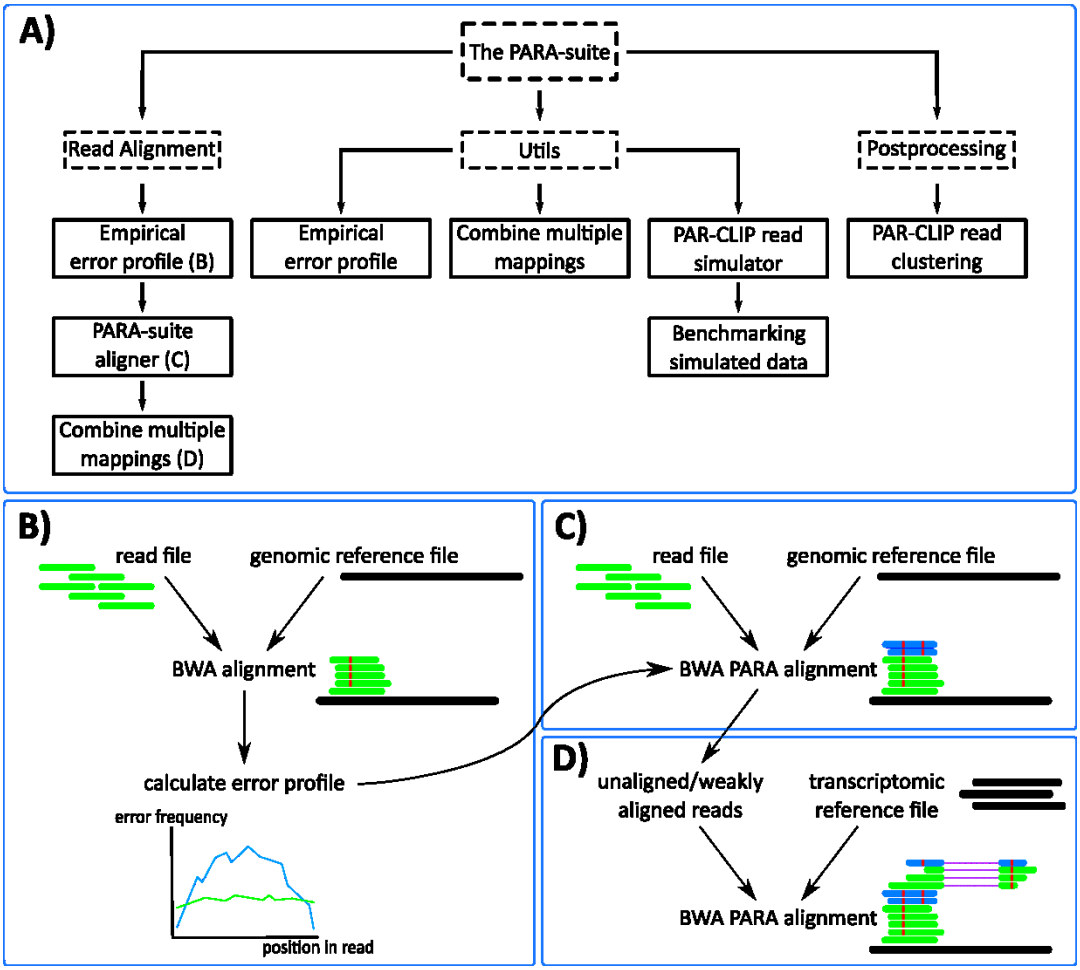


Figure 2: The PARA-suite. (A) The PARA-suite. Dashed boxes represent software packages; all other boxes represent executable programs. The Utils package includes tools for working with error-prone sequencing data and the postprocessing package contains a tool for clustering an aligned PAR-CLIP dataset to identify RBP-bound genomic regions. (B) Read alignment by a fast read aligner is necessary to infer the error profile for a particular read dataset (we selected BWA). (C) BWA PARA is applied to the entire dataset to map error-prone reads, indicated here by the additional mapping of the two reads (shown in blue). (D) An optional alignment versus a transcriptome reference database can be executed using BWA PARA to identify previously unmapped reads.

Algorithm of the PARA-suite aligner BWA PARA

The general BWA algorithm uses a Burrows–Wheeler transform (BWT) (Burrows and Wheeler 1994) to create an index for a reference genome sequence and applies a backward search to identify possible mapping positions in the genome for every single sequencing read. The backward search starts with the last base of a read proceeding to its front, searching the partly decompressed suffix trie using the auxiliary Ferragina and Manzini index (Ferragina and Manzini 2000) for a matching predecessor base of the read´s bases compared so far. Even if a match can be found for a single comparison, mismatches are introduced and all possible downstream paths within the suffix trie are considered until a pre-defined threshold of maximal mismatches is exceeded in a single path (Figure 3, red dotted line).

The principal idea of BWA PARA is the introduction of a probability estimate for each comparison of the backward search. This enables mismatches to be weighted according to their probabilities that they occur in the analyzed dataset. A sequencing run is initially characterized according to its underlying error probabilities. This allows us to determine specific error-profiles for experimental techniques, such as the frequent T–C conversions in PAR-CLIP data, which are more common than sequencing errors. The error profile M is a 4×4 probability matrix specifying substitution probabilities values $\in [0..1]$ for each reference base $\in \{A, C, G, T\}$ to the read bases {A, C, G, T} (Figure 4A). Indels are introduced during the alignment step separately, using the estimated probabilities $I \in [0,1]$ for insertions and $D \in [0,1]$ for deletions.

For each comparison between a read base at read position i (read[i]) and a reference base at position j (ref[j]) in the reference sequence, the algorithm recursively calculates a joint probability value p, which is used to examine the chance of incorporating a matching base or a suitable error, including indels, at the respective read positions (Figure 4D):

$$p_i = \begin{cases} p_{i+1} \cdot D, & if\ ref[j]\ is\ deleted \\ p_{i+1} \cdot I, & if\ read[i]\ is\ inserted \\ p_{i+1} \cdot M(read[i], ref[j]), & otherwise \end{cases}$$

with $p_{|\text{read}|} = 1$, starting with $i = |\text{read}| - 1$ and decreasing i at each step, except in the case of a deletion (where i is left unchanged), for $i \geq 0$.

Before the alignment of a particular read, a minimal threshold T for the probability p is needed to decide whether a read is accepted as aligned or rejected. The calculation for T depends on a parameter X for the average number of mismatches. Note that this is not a maximal threshold in terms of absolute mismatches, as the number of the more frequent errors per aligned read can exceed X. The parameter X can be pre-defined by the user or is by default estimated as the expected number of mismatches for different read lengths based on the error profile M for a sequencing run. Next, the minimal threshold T is computed (Figure 4B&C):

$$T = avg(\text{match})^{|read|-X} \cdot avg(\text{mismatch})^{X},$$

where $avg(\text{match}) = \frac{1}{5}\left[\sum_{i\in\{0..3\}} M_{i,i} + \left(1 - (I + D)\right)\right]$ and $avg(\text{mismatch}) = \frac{1}{14}\left[\sum_{i,j\in\{0..3\};i\neq j} M_{i,j} + I + D\right]$.

Both *avg(*match*)* and *avg(*mismatch*)* are normalized by the number of elements (four matches plus one for no indel occurring, and 12 mismatches plus 2 for either a insertion or a deletion). If p falls below the pre-calculated threshold T during read alignment, the path within the suffix trie is assumed not to match the read and is rejected (Figure 3, blue dashed line). The algorithm thus penalizes rare types of mismatches according to M, whereas frequent errors, such as T–C errors in PAR-CLIP reads, are the most favored substitutions in the alignment process (Figure 4B–D).

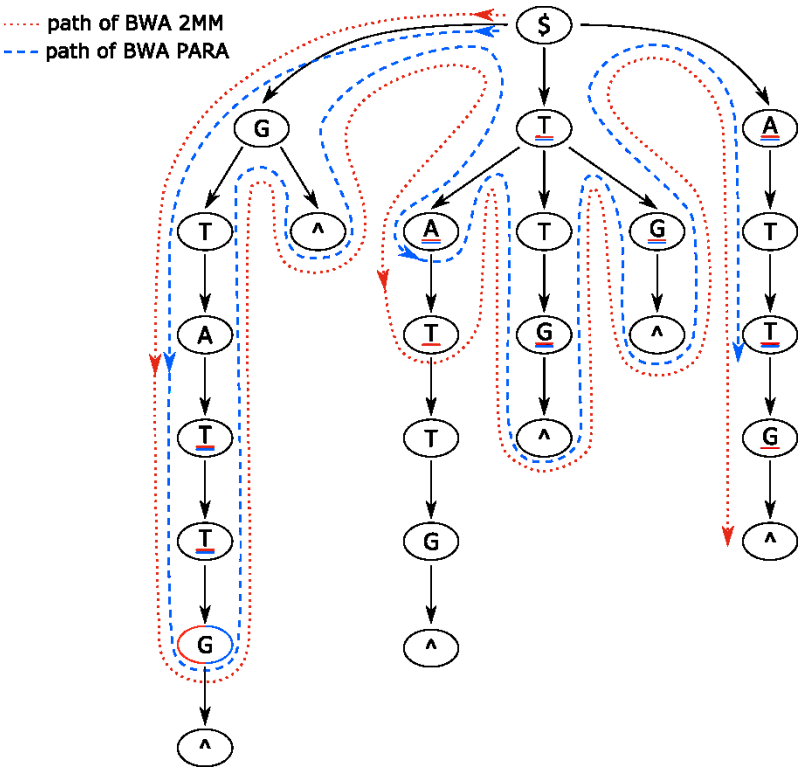


Figure 3: Suffix trie paths for the BWA and PARA-suite. Paths of the algorithms through the suffix trie aligning the read sequence GCCATG$ against the reference sequence GTTATG$ (where $ means the end of a sequence). The red dotted line represents the algorithm of the BWA aligner, allowing for two mismatches; the blue dashed line indicates the BWA PARA algorithm. The underlined bases represent positions where the respective aligner introduces a mismatch. The example shows that BWA PARA needs 14 comparisons but the basic BWA needs 16 comparisons. Indels are not shown for simplicity.

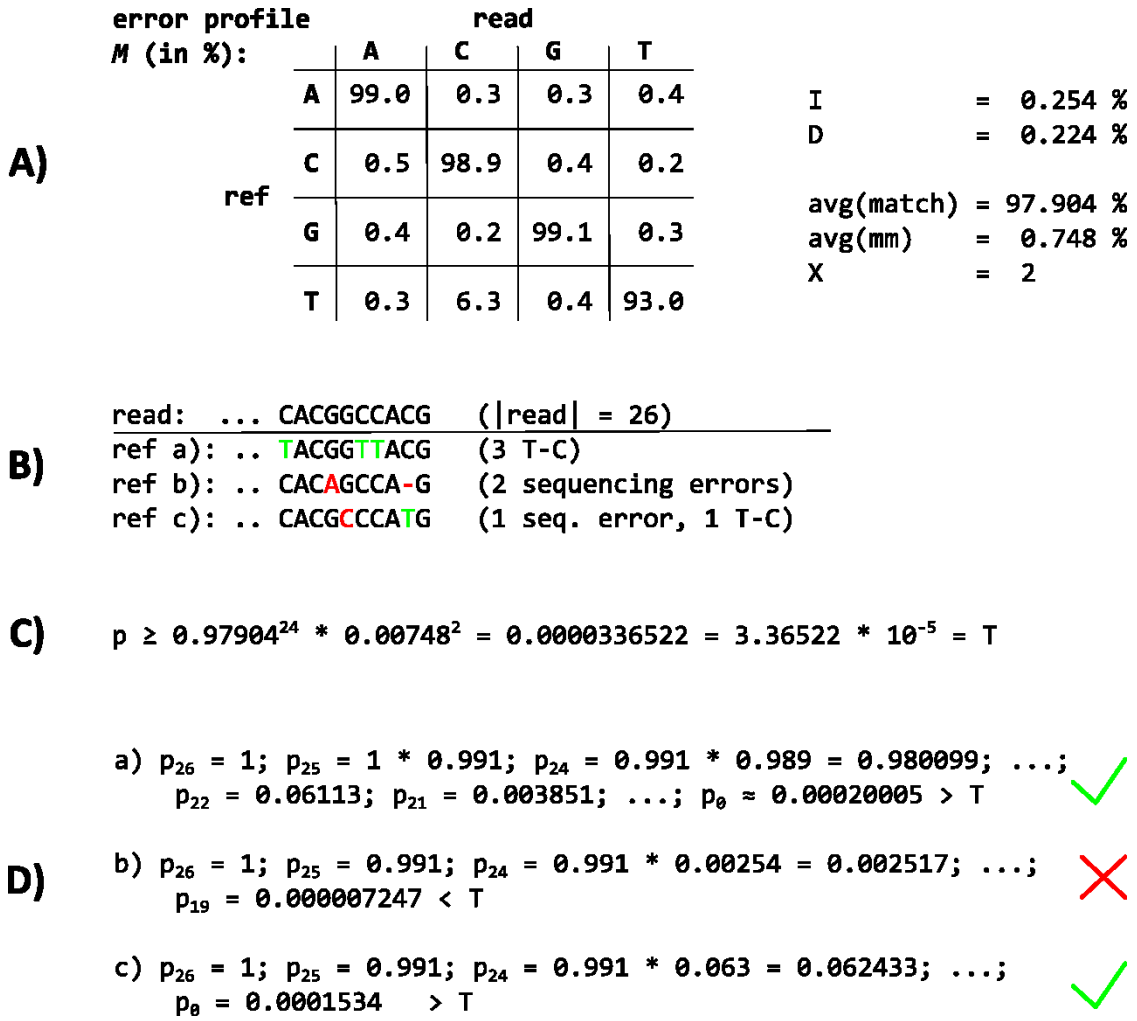


Figure 4: The BWA PARA alignment approach. (A) The error profile probability matrix M and the indel probabilities I and D, which are used as input for the BWA PARA algorithm, as well as exemplary results of the intermediate calculations of the BWA PARA algorithm. In M, only T–C conversions have a higher probability (6.3%) than sequencing errors and indels. (B) The last characters of a particular read and three examples of mapping positions within a reference, called ref a–c. (C) The calculation of a maximum threshold T for the mapping probability p (see the Equation 2 in the main text, and values from (A) in this image). (D) The mapping probability calculation of the read when mapped to References a–c. The read fails to map against ref b with two sequencing errors, whereas ref a and ref c are suitable mapping positions, where the probability p is higher than the threshold T. For implementation, we worked with the

open-source read aligner BWA (version 0.7.8) to extend its algorithm for the alignment of short and error-prone reads.

Results

Properties of PAR-CLIP reads

To assess the most important properties of the PAR-CLIP sequencing reads for read alignment, we systematically compared PAR-CLIP datasets for the three RBPs EWSR1, FUS and TAF15 (the FET protein family) (Hoell, Larsson et al. 2011) to a recently published RNA-Seq run on human reference RNA (SEQC/MAQC-III-Consortium 2014). The 10 outermost bases of the SEQC/MAQC reads showed error rates with peaks at 1.5 and 2.2 errors per 100 reads (EPR). In contrast, the middle read length range showed an average of about 0.3 EPR (Figure S2A, red line). As the short reads of the FET PAR-CLIP datasets consisted only of these outermost bases, they exhibited a two- to threefold higher average sequencing error rate (about 0.7 EPR or even higher) than the SEQC/MAQC reads (Figure S2B, green line). When considering the T–C conversions only, we observed 1.319 EPR for EWSR1, 1.477 EPR for FUS and 1.051 EPR for TAF15 on average. This is an approximately 20- to 30-fold increase in comparison to the SEQC/MAQC dataset with 0.051 EPR for T–C conversions on average (Figure S2). Moreover, we analyzed data from two further PAR-CLIP studies performed on the RBPs HuR (Mukherjee, Corcoran et al. 2011) and MOV10 (Sievers, Schlumpf et al. 2012), which showed similar error profiles and EPRs to the FET PAR-CLIPs for T–C conversions (Figure S3).

Further analyses of the PAR-CLIP read datasets for EWSR1, FUS, TAF15, MOV10 and HuR showed the PAR-CLIP reads (a) to be shorter than 30 bases, (b) to cover only short stretches of an expressed gene rather than the entire expressed RNA (these stretches are henceforth called clusters), (c) to exhibit a specific nucleotide conversion pattern with a strong enrichment of T–C conversions, where (d) such conversions occur in specific ‘conversion sites’ in the clusters. The two properties (a) and (b) are determined by treating the cells with RNAse T1 or the lysate during the PAR-CLIP experimental

protocol. As only short RNA fragments that are not digested by the endonuclease (these are probably protected by the binding pocket of the RBP) are sequenced, the lengths of those fragments are usually short. However, the nucleotide composition of those reads is strongly affected by the digestion enzyme and can vary among different digestion enzymes (Kishore, Jaskiewicz et al. 2011). After quality trimming and adapter trimming of the five PAR-CLIP datasets, the average read lengths were 25.67 bases (EWSR1), 25.60 bases (FUS), 24.21 bases (TAF15), 25.20 bases (HuR) and 23.36 bases (MOV10). As the transcript regions outside the bound RNA fragment are digested by the endonuclease, these are removed during immunoprecipitation and not sequenced, except for additional binding sites on the same transcript further up- or downstream. Thus the sequencing reads are stacked into short clusters covering short stretches of the gene and representing the RBP-bound regions of the transcripts (Figure S4A).

The two properties (c) and (d) were determined by incorporating photoactivatable nucleosides into the nascent transcripts during transcription. In the case of 4-SU, T–C conversions occur in the sequencing reads at all crosslinked sites, where the 4-SU is incorporated instead of the native uridine. These conversions can reach high rates in specific conversion sites within a cluster (Hafner, Landthaler et al. 2010). In the analyzed datasets, we observed an average frequency of about 70% T–C conversions in the main T–C conversion site (Figure S1A). This emphasizes that simulated read datasets with specific properties are necessary for the evaluation of common short read aligners for analyzing PAR-CLIP read data. However, this cannot be created by common sequencing read simulators, such as ART (Huang, Li et al. 2012) or GemSIM (McElroy, Luciani et al. 2012). These produce simulated reads with a continuous coverage over the entire transcript range and the introduced mutations are distributed randomly throughout the simulated reads. This is not the case for PAR-CLIP sequencing reads.

PAR-CLIP read simulation for performance evaluation

We simulated a total of 10 PAR-CLIP read datasets based on information learned from three previously published PAR-CLIP datasets of the FET protein family (Hoell, Larsson et al. 2011) (Table S2). We imitated Illumina GenomeAnalyzer II sequence data

according to the real datasets used. The respective sequencing error and T–C conversion profiles were generated on the basis of alignments of all three datasets against the human reference genome sequence version 38 (GRCh38) (Lander, Linton et al. 2001). The error profile and additionally estimated distributions were similar to the ones from PAR-CLIP data on the two RBPs HuR and MOV10, indicating that these profiles represented a reasonable approximation for PAR-CLIP data in general. We selected human transcript sequences downloaded from Ensembl Genes version 77 (Cunningham, Amode et al. 2015) as our sequence database to simulate human transcript read sequences. We set λ, the parameter for the fraction of sequencing reads that stacked into clusters bound by the RBP, to 65%. These true RBP binding sites showed high T–C conversion frequencies in different T–C conversion sites. The remaining 35% of the simulated sequencing reads were designated to represent non-specifically bound transcripts without an elevated T–C conversion rate, except for a few T–C sequencing errors. These reflected RNA contaminations that can occur during the PAR-CLIP experiment.

To assess the quality of the simulation, we then compared PAR-CLIP-specific properties between the 10 simulated datasets and the FET PAR-CLIP data. Within a cluster detected in a simulated dataset, shifts in the alignment positions of a few nucleotides at the beginning and the end of the simulated cluster could be seen between the reads (Figure S4B). According to the position-wise T–C conversion profile used, a T–C conversion site with a high conversion rate, as well as a few sites with lower conversion rates, were usually present in the detected clusters (e.g. Figure 1B). We compared the error profiles between one of the simulated datasets and the real datasets, and distinguished between T–C errors and all other errors; the latter represent all sequencing errors other than the T–C sequencing errors (Figure S2C). Similar to the real data, the distribution of the sequencing errors in the simulated dataset peaked at the beginning of the reads and dropped to a mean error rate of 0.6 EPR in the middle read length range. Error rates were slightly underestimated in the simulated data compared to the real PAR-CLIP data, presumably because of a small percentage of multiple

mutations that occurred at individual sites. Apart from this, the simulated datasets appeared to be representative of real PAR-CLIP data in the relevant aspects.

Accuracy of common read aligners and the PARA-suite on simulated PAR-CLIP data

Using the simulated PAR-CLIP datasets, we analyzed the accuracy of state-of-the-art read aligners and common binding site detection algorithms, and compared these to the PARA-suite alignment pipeline. The aligners BWA and Bowtie have often been used in CLIP studies (Lebedeva, Jens et al. 2011, Ascano, Mukherjee et al. 2012, Sievers, Schlumpf et al. 2012). BWA PSSM was applied with the PSSM for PAR-CLIP provided by its authors because a PSSM estimated from the sequencing dataset revealed worse accuracy (data not shown). MOSAIK was executed, reporting only unique mappings, allowing for up to three mismatches between the read and the reference sequence, and using a Smith–Waterman bandwidth of 5. The read aligners were used to align the simulated datasets to the reference sequence GRCh38. We also executed the PARA-suite on the Ensembl Genes transcriptome database (version 77) and combined the results with the genomic reference sequence alignments. These results are henceforth referred to as those of the "PARA-suite pipeline", whereas the results of the genomic alignment step using the PARA-suite only are referred to as those of "BWA PARA". For BWA PARA, the sequencing error and T–C conversion profiles for the simulated datasets were obtained on the basis of the BWA alignments, allowing for two mismatches (BWA 2MMs) for each of the simulated datasets separately (execution commands are outlined in the Supplementary Methods). For an overview of the performance, we estimated the average of the recall, precision and accuracy for each aligner over the 10 simulated datasets (our calculations are described in the Supplementary Methods). Unfortunately, BMix does not report negative clusters (contaminations) and thus we were able to neither calculate the recall nor the accuracy, but only the precision.

In terms of overall performance, the PARA-suite performed best, with an accuracy of 69.74% for BWA PARA and 73.14% for the entire pipeline, showing performance gains of 1.57% and 4.97% compared to the second-best aligner (BWA 2MM), respectively (Table 2, Table S3). Many prominent PAR-CLIP studies have used Bowtie 1MM or BWA

2MM for the read alignment step (Lebedeva, Jens et al. 2011, Mukherjee, Corcoran et al. 2011, Ascano, Mukherjee et al. 2012, Sievers, Schlumpf et al. 2012, Mukherjee, Jacobs et al. 2014). When we compared the PARA-suite pipeline with these two aligners, the PARA-suite pipeline showed an increase of 16.95% and 4.97% in the overall accuracy, respectively. Notably, 1.56% of the reads aligned by the PARA-suite pipeline on average spanned an exon–exon junction. These were not identified by the genomic reference mapping step but instead required alignment against the transcriptome reference sequences. Additionally, we compared the recall (the fraction of correctly aligned reads out of all simulated reads) and the precision (the fraction of correctly aligned reads out of all aligned reads) to assess the mapping ability of the read aligners (Table 2, Figure S5). Here, the PARA-suite pipeline and BWA PARA were ranked first and third regarding recall, and first and second regarding precision, respectively, out of 10 analyzed alignment scenarios (Table 2). Hence, the PARA-suite pipeline and BWA PARA offer notable performance increases over commonly applied alignment setups.

We then tested the accuracy of the binding site detection algorithms BMix, PARalyzer and the hierarchical clustering of the PARA-suite using the read alignments of BWA PARA (Table S4). The hierarchical clustering identified the most correct binding sites: 3.26% more correct sites than BMix and 5.54% more correct binding sites than PARalyzer. However, BMix identified fewer false binding sites than the hierarchical clustering (20.30% fewer) and PARalyzer (69.85% fewer). Furthermore, we investigated whether BWA PARA increased the number of binding sites detected, irrespective of the detection algorithm used. In conjunction with BMix, BWA 2MM (the second-best aligner) identified 7.17% fewer correct binding sites than BWA PARA. With PARalyzer, BWA 2MM identified 2.97% fewer correct binding sites than BWA PARA. Finally, the hierarchical clustering identified 7.52% more correct binding sites for BWA PARA than for BWA 2MM. Overall, the combination of BMix and BWA PARA provided the most accurate results on our simulated data.

Table 2: Alignment accuracy on simulated PAR-CLIP data. The most accurate alignment results were obtained for different parameter settings for each read aligner on 10 simulated PAR-CLIP datasets. The results are averaged per read aligner over all 10 datasets and are sorted by accuracy.

Aligner	Accuracy (in %)	Variance	Recall (in %)	Precision (in %)	Mapped overall	Mapped correctly	Real time (s)
PARA-suite pipeline	73.14	1.37E-06	84.49	71.85	1,024,792	969,948	396.8
BWA PARA	69.74	1.38E-06	82.16	68.24	975,672	924,802	153.7
BWA 2MMs	68.17	1.37E-06	82.31	64.98	959,171	904,034	359.2
Bowtie 2MMs	63.38	1.10E-06	77.91	60.93	886,512	840,540	120.6
BWA PSSM	59.80	1.18E-06	74.04	58.72	818,895	793,007	25.4
TopHat	59.69	8.35E-07	76.10	55.35	844,902	791,549	282.9
Bowtie2	56.22	1.11E-06	73.23	51.43	763,893	745,531	13.4
STAR	50.74	9.10E-07	69.57	43.02	826,871	672,920	248.6
MOSAIK	44.88	2.18E-04	62.83	37.16	897,679	595,220	12,128.18
Subjunc	35.42	9.03E-07	50.61	26.09	597,400	469,751	64.2

Analysis of FET PAR-CLIP datasets

To investigate the performance of the PARA-suite on real PAR-CLIP datasets, we applied it to the three FET PAR-CLIP datasets (Hoell, Larsson et al. 2011). The sequencing reads were preprocessed similarly to the method given in the original publication, and low quality ends and adapter sequences were trimmed using Cutadapt (Martin 2011).

Afterwards, all remaining reads longer than 18 bases were aligned against GRCh38 with Bowtie2, Bowtie 2MM, BWA 2MMs, BWA PSSM and BWA PARA (without the transcriptome mapping step to achieve comparable results). Selection of the read aligners (i.e. Bowtie2, Bowtie 2MM, BWA 2MM, BWA PSSM and BWA PARA) was based on the results of the previous section, as these represent the most accurate read aligners on PAR-CLIP data. We measured the fraction of aligned reads for all the aligners on the three datasets (Table S5). BWA PARA generated the largest fraction of aligned reads over all three datasets in comparison to BWA 2MM and BWA PSSM. Next, we stacked (clustered) all the aligned reads using BMix and the hierarchical clustering tool of the PARA-suite (Table 3). BWA 2MM identified fewer binding sites than BWA PSSM or BWA PARA for read alignments prior to either BMix or hierarchical clustering. Using the hierarchical clustering, BWA PARA reported the largest number of binding sites for two out of the three datasets. BWA PSSM identified 6.90% more clusters than BWA PARA for the FUS dataset whereas BWA PARA identified 3.98% more clusters for the EWSR1 dataset and 19.21% more clusters for the TAF15 dataset than BWA PSSM. In comparison to the values reported in the original publication, the use of BWA PARA and hierarchical clustering increased the number of binding sites by 33.71% for EWSR1 and 16.77% for FUS, and decreased them by 12.56% for TAF15. After extracting distinct genes from all binding sites identified by BWA PARA (10,631 genes in total), 26.90% additional genes were found for all three RBPs, in comparison to the original publication (7,771 genes in total). As expected for three RBPs from the same family, there was a substantial overlap in terms of the identified genes, with 2,702 genes targeted by all three RBPs (Figure S6).

Table 3: Binding sites detected for the *FET* protein family. The number of binding sites for the *FET* protein family identified by the aligners BWA 2MM, BWA PSSM, BWA PARA, Bowtie 2MM and Bowite2 in combination with BMix and the hierarchical clustering of the PARA-suite. Filters were applied according to Section 2.2.

	EWSR1	***FUS***	***TAF15***
BWA 2MM BMix	20,703	14,768	5,086
BWA 2MM Clustering	22,760	36,861	5,810
BWA PSSM BMix	24,639	19,628	5,238
BWA PSSM Clustering	27,550	51,606	6,130
BWA PARA BMix	25,478	19,006	5,862
BWA PARA Clustering	28,692	48,042	7,588
Bowtie 2MM BMix	19,173	13,902	4,582
Bowtie 2MM Clustering	21,082	35,490	5,254
Bowtie2 BMix	12,384	8,078	3,558
Bowtie2 Clustering	13,338	20,398	3,710

Analysis of PAR-CLIP data on HuR

We next applied the PARA-suite to a PAR-CLIP dataset on HuR, an RBP promoting RNA stabilization (Mukherjee, Corcoran et al. 2011). Adapters and low-quality ends within the HuR dataset were trimmed using Cutadapt and reads shorter than 14 bases were discarded. The binding motif of HuR is well-studied and is AU-rich, with a consensus motif described as AUUUA, AUUUUA or AUUUUUA (Nabors, Suswam et al. 2003, Lebedeva, Jens et al. 2011), showing potentially more T–C conversions within each binding site than other RBPs. As the generated error profile of the dataset was similar to those of the FET PAR-CLIP data (Section 3.1), the data quality seemed comparable. However, we noted a slight increase in T–C conversions (Figure S3). The AU-rich binding motif might explain the higher T–C conversion rate of 1.684 EPR compared to the conversion rate of 1.477 EPR e.g. for FUS.

We used the same read aligners as described in the previous section (Bowtie2, Bowtie 2MM, BWA 2MM, BWA PSSM and BWA PARA) to align the pre-processed dataset against the human genome reference GRCh38. We applied BMix and the hierarchical clustering of the PARA-suite to determine the binding sites of HuR derived by using the different read aligners. BWA PSSM, in conjunction with BMix, identified the most RBP binding sites within the genome, which was 3.69% more than BWA PARA (Table 4). When we compared the binding sites detected by BMix and the PARA-suite hierarchical clustering for alignments created by BWA PARA (binding site positions overlapping by at least 13 bases), the difference was only marginal, with an overlap of more than 98.25% for the two methods. A recent study of this dataset reported binding sites using Bowtie 2MM for the alignment step and PARalyzer for the binding site detection. We found that the use of either BWA PSSM or BWA PARA in conjunction with either BMix or hierarchical clustering increased the number of binding sites detected by 2.87–7.84%.

Table 4: Binding sites detected by BMix and the hierarchical clustering based on read alignments performed by BWA 2MM, BWA PSSM, BWA PARA, Bowtie 2MM and Bowtie2 on the *HuR* dataset.

	BMix	**Hierarchical clustering**
BWA 2MM	136,775	137,697
BWA PSSM	147,883	148,985
BWA PARA	141,365	141,867
Bowtie 2MM	125,592	125,067
Bowtie2	88,369	87,400

We searched for the exact binding motifs of HuR (ATTTA, ATTTTA and ATTTTTA) within the binding sites detected by BMix within 3' untranslated region (UTR) or introns for all the read aligners tested. We found that all aligners performed comparably, with motifs present in 42–44% of all binding sites detected. The largest fraction was achieved using read alignments with BWA PSSM (44.33%), whereas BWA

PARA in combination with BMix found 42.53% of the binding sites that were most likely correct. Bowtie 2MM in combination with BMix had the lowest fraction of binding sites containing the reported binding motif (42.44%). We also compared the previously reported HuR binding sites to the binding sites determined by the full PARA-suite pipeline with BMix for clustering and detected 13 out of 15 sites, namely 3' UTR PTGS2, 3'UTR CDKN1A, 3'UTR VEGFA, 3'UTR TNF, 3' UTR SLC7A1, 3'UTR CCND1, 3'UTR MYC, 3' UTR XIAP, 3'UTR CELF1, TTS CSF2, 3'UTR CCNB1, intron NCL and 3' UTR KRAS. The binding information for this comparison was taken from the Ingenuity knowledge base (Calvano, Xiao et al. 2005). The original study on the HuR dataset (Mukherjee, Corcoran et al. 2011) only reported 12 out of these 15 genes having confirmed binding site.

Discussion

We provided a detailed characterization of the error profiles of PAR-CLIP reads and an in-depth performance assessment of short read aligners in combination with binding site detection tools. We characterized some of the unique properties of PAR-CLIP sequence datasets, including the preferred read positions for T–C conversion sites and their frequencies per read position. We observed higher frequencies of sequencing errors in PAR-CLIP data than in the human reference RNA-Seq data. A likely reason for this behavior could be that PAR-CLIP reads are much shorter than common RNA-Seq reads, which reach lengths of 200 bases and show high-quality regions in the middle read length range (Laehnemann, Borkhardt et al. 2015, Schirmer, Ijaz et al. 2015). We used these observations for the design of a PAR-CLIP read simulator that embeds PAR-CLIP specific information within the simulation process and the PARA-suite pipeline for error-aware read alignment and processing. The read simulator mimics PAR-CLIP datasets with error profiles drawn from real PAR-CLIP datasets.

Based on the simulated PAR-CLIP datasets, we determined the parameter settings that delivered the best performance for commonly used aligners (Mukherjee, Corcoran et al. 2011, Ascano, Mukherjee et al. 2012, Sievers, Schlumpf et al. 2012, Mukherjee, Jacobs et

al. 2014). Our analysis showed that the read alignment was crucial for detecting RBP binding sites in PAR-CLIP datasets. However, the PAR-CLIP specific read properties make it nearly impossible to identify splice junctions covered by PAR-CLIP reads with RNA-Seq read aligners such as TopHat, STAR or Subjunc, as their algorithms are based on unmet assumptions, such as a similar read coverage across all exons or long reads, in order to achieve high confidence k-mer spectra. Accordingly, these three aligners were outperformed by the other methods (Table S3–4). nterestingly, MOSAIK, an error-aware aligner based on hash queries that has been shown to be more robust on RNA-Seq reads than BWT-based aligners (Lee, Stromberg et al. 2014), was also outperformed by most of the other tested methods. Although it is robust on longer RNA-Seq reads, MOSAIK seemed to struggle with the very short PAR-CLIP reads. The PARA-suite alignment pipeline allowed us to increase the fraction of aligned reads in comparison to other aligners, including the alignment of reads spanning exon–exon junctions, both for PAR-CLIP datasets and data from a HITS-CLIP study (Supplementary Results). We observed this improvement irrespective of the binding site detection algorithm applied downstream. Importantly, unlike the error-aware short read aligner BWA PSSM, our short read alignment algorithm does not need the manual input of an error profile, which is instead inferred de novo within individual sequencing runs. The aligner thus automatically adapts to varying qualities of individual (PAR-)CLIP sequencing runs and is specifically adjusted to each sequence dataset. To our knowledge, it is the first tool for simultaneous de novo error model inference and short read alignment based on the BWA algorithm. Another difference from the BWA PSSM algorithm is that the latter introduces mismatches while considering the base calling quality scores and a probabilistic background model for matching bases in addition to the input error profile. In contrast, the generic error profile estimation of the PARA-suite is not limited to any specific input profile. Further applications of our software could thus be used to analyze other types of error-prone sequencing data such as bisulphite sequencing data, which introduces a high amount of C–T mutations (Frommer, McDonald et al. 1992) or data from low-quality ancient DNA samples (Briggs, Stenzel et al. 2007).

Common read simulators such as ART or GemSim do not allow simulating PAR-CLIP reads with their specific error profiles. When comparing our PAR-CLIP read simulator with the recently developed CSeq simulator for CLIP data (Kassuhn, Ohler et al. 2016), both have different strengths. CSeq takes an exact binding motif and T–C conversion profile that is specific for the respective binding motif as input, thus restricting the read's base composition and T–C conversion sites. This allows to mimic PAR-CLIP reads for a specific RBP, but not to generalize evaluations on these datasets to all kinds of RBPs. In comparison, the PAR-CLIP reads simulated with the PARA-suite are based on data that have been inferred from three different PAR-CLIP datasets to simulate heterogenic reads, which represent a broader spectrum of RBP binding sites. In addition, the read selection is not restricted to sequences containing the actual RBP binding motif. Thus CSeq and the PARA-suite's read simulator have slightly different applications: CSeq allows one to simulate reads to optimize parameters for a specific dataset and the PARA-suite allows one to simulate reads for general tool evaluation and algorithmic improvements.

Our analysis of combinations of read aligners and binding site detection algorithms on simulated and real datasets indicated that no single software performed best in terms of detecting binding sites on the available PAR-CLIP datasets. This observation was recently also made on other datasets (Kassuhn, Ohler et al. 2016). Our analysis of the HuR and FUS datasets revealed that U-rich binding sites tended to show higher rates of T–C conversions per read and were best aligned by BWA PSSM. RBPs with a more heterogeneous nucleotide distribution within the binding site (e.g. EWSR1 and TAF15) are better assessed by BWA PARA. This is supported by an analysis of uridylate-rich sequences from our simulated data aligned by BWA PSSM and BWA PARA (Supplementary Results and Supplementary Table S6). Therefore, a preliminary analysis of the error profile using the PARA-suite error profiler could allow one to determine the best approach to analyze sequencing data of a novel yet uncharacterized RBP.

References

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. and Tuschl, T. (2012). Identification of RNA–protein interaction networks using PAR-CLIP. Wiley Interdiscip. Rev. RNA 3(2): 159-177.

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., Langlois, C., Munschauer, M., Dewell, S. and Hafner, M. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature 492(7429): 382-386.

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T. and Lachmann, M. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. U S A 104(37): 14616-14621.

Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. Digital Equipment Corporation: Palo Alto, CA (Technical Report 124).

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P. and Tschoeke, S. K. (2005). A network-based analysis of systemic inflammation in humans. Nature 437(7061): 1032-1037.

Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature 460(7254): 479-486.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38(6): 1767-1771.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J.,

Yates, A., Zerbino, D. R. and Flicek, P. (2015). Ensembl 2015. Nucleic Acids Res. 43(Database issue): D662-669.

Danan, C., Manickavel, S. and Hafner, M. (2016). PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites. Post-Transcriptional Gene Regulation: 153-173.

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

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. Nature reviews. Genetics 2(12): 919-929.

Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. Proceedings of the 41st Symposium on Foundations of Computer Science: 390-398.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U S A 89(5): 1827-1831.

Glisovic, T., Bachorik, J. L., Yong, J. and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett. 582(14): 1977-1986.

Golumbeanu, M., Mohammadi, P. and Beerenwinkel, N. (2015). BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data. Bioinformatics: btv520.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141(1): 129-141.

Hieronymus, H. and Silver, P. A. (2004). A systems view of mRNP biology. Genes Dev. 18(23): 2845-2860.

Hoell, J. I., Hafner, M., Landthaler, M., Ascano, M., Farazi, T. A., Wardle, G., Nusbaum, J., Cekan, P., Khorshid, M. and Burger, L. (2014). Transcriptome-Wide Identification

of Protein Binding Sites on RNA by PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation). Handbook of RNA Biochemistry: Second, Completely Revised and Enlarged Edition. A. B. R.K. Hartmann, A. Schön, and E. Westhof. Weinheim, Wiley-VCH Verlag GmbH & Co. KGaA. II: 877-898.

Hoell, J. I., Larsson, E., Runge, S., Nusbaum, J. D., Duggimpudi, S., Farazi, T. A., Hafner, M., Borkhardt, A., Sander, C. and Tuschl, T. (2011). RNA targets of wild-type and mutant FET family proteins. Nat. Struct. Mol. Biol. 18(12): 1428-1431.

Huang, W., Li, L., Myers, J. R. and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics 28(4): 593-594.

Kassuhn, W., Ohler, U. and Drewe, P. (2016). Cseq-simulator: a data simulator for CLIP-Seq experiments. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.

Kerpedjiev, P., Frellsen, J., Lindgreen, S. and Krogh, A. (2014). Adaptable probabilistic mapping of short reads using position specific scoring matrices. BMC Bioinformatics 15(1): 100.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat. Methods 8(7): 559-564.

Kloetgen, A., Münch, P. C., Borkhardt, A., Hoell, J. I. and McHardy, A. C. (2015). Biochemical and bioinformatic methods for elucidating the role of RNA-protein interactions in posttranscriptional regulation. Brief. Funct. Genomics 14(2): 102-114.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M. and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. 17(7): 909-915.

Laehnemann, D., Borkhardt, A. and McHardy, A. C. (2015). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. Brief. Bioinform.: bbv029.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. and FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9(4): 357-359.

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

Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M. and Rajewsky, N. (2011). Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Mol. Cell 43(3): 340-352.

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

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

Liao, Y., Smyth, G. K. and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41(10): e108.

Lukong, K. E., Chang, K.-w., Khandjian, E. W. and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet. 24(8): 416-425.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17(1): 10-12.

McElroy, K. E., Luciani, F. and Thomas, T. (2012). GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics 13(1): 74.

Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, M., Tuschl, T., Ohler, U. and Keene, J. D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. Mol. Cell 43(3): 327-339.

Mukherjee, N., Jacobs, N. C., Hafner, M., Kennington, E. A., Nusbaum, J. D., Tuschl, T., Blackshear, P. J. and Ohler, U. (2014). Global target mRNA specification and regulation by the RNA-binding protein ZFP36. Genome Biol. 15(1): R12.

Nabors, L. B., Suswam, E., Huang, Y., Yang, X., Johnson, M. J. and King, P. H. (2003). Tumor Necrosis Factor α Induces Angiogenic Factor Up-Regulation in Malignant Glioma Cells A Role for RNA Stabilization and HuR. Cancer Res. 63(14): 4181-4187.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res.: gku1341.

SEQC/MAQC-III-Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. 32(9): 903-914.

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

Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012). Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. Nucleic Acids Res. 40(20): e160.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M. and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol. 13(8): R67.

Tan, A. Y. and Manley, J. L. (2009). The TET family of proteins: functions and roles in disease. J. Mol. Cell. Biol. 1(2): 82-92.

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

van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. 30(9): 418-426.

Wang, T., Xiao, G., Chu, Y., Zhang, M. Q., Corey, D. R. and Xie, Y. (2015). Design and bioinformatics analysis of genome-wide CLIP experiments. Nucleic Acids Res. 43(11): 5263-5274.

Zhang, C. and Darnell, R. B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat. Biotechnol. 29(7): 607-614.

Supplementary Materials & Methods

Supplementary Methods

For an overview of the performance of different read aligners and binding site detection algorithms on 10 simulated PAR-CLIP datasets, we calculated the precision, recall and accuracy for each. We considered all reads originating from simulated RBP-binding sites (with T–C conversions) as positives and those originating from other areas of the reference (simulated contaminations) as negatives. True positive and negative reads are those which are aligned correctly, whereas false positive and negative reads are those which are wrongly or not aligned (Table 1; Supplementary Table 3). We used BMix, PARalyzer and our hierarchical clustering to obtain the read clusters. Filtering of the clusters generated with the hierarchical clustering was performed as described in Section 2.2. A correctly reported binding site was considered a true positive, a falsely reported cluster (simulated contamination without elevated T–C conversions) as a false positive, an unreported binding site as a false negative and an unreported cluster (without T–C conversions) as a true negative (Supplementary Table 4). Unfortunately, BMix does not report false negative clusters (contaminations) and thus we were not able to calculate the recall nor the accuracy, but only the precision.

Execution commands

Quality and adapter trimming:

```
cutadapt -e 0.05 -q 28 -m 18 -b $adapter -f fastq -o $output $input
```

Alignment:

bwa aln -n $n $hg38_reference $trimmed_input > $output.sai ($n in {1, 2, 0.01, 0.02, 0.04})

bwa samse $hg38_reference $output.sai $trimmed_input > $output.sam

bowtie -S -v 1 --best -m $n --strata $hg38_reference -q $trimmed_input $output.sam ($n in {1, 2})

bowtie2 -x $hg38_reference -U $trimmed_input -S $output.sam

parasuite map --refine -q $trimmed_input -r $hg38_reference -t $hg38_transcriptome -o $output --parasuite-mm $X ($X in {1, 2, 3, -1})

STAR --genomeDir $hg38_reference --readFilesIn $trimmed_input --outFileNamePrefix $output

subjunc -u -n -i $hg38_reference -r $trimmed_input -o $output.sam

tophat -o $output $hg38_reference $trimmed_input

MosaikBuild -q $trimmed_input -out $mosaik_input -st illumina -ga hg38

MosaikAligner -ia $hg38_reference -in $mosaik_input -out $output -mm 3 -annse ./mosaik-2.2.3/network_files/2.1.78.se.ann -annpe ./mosaik-2.2.3/network_files/2.1.78.pe.ann -m unique -bw 5

RBP binding site detection:

PARalyzer config file:

BANDWIDTH=3

CONVERSION=T>C

MINIMUM_READ_COUNT_PER_CLUSTER=5

MINIMUM_READ_COUNT_FOR_KDE=3

MINIMUM_CLUSTER_SIZE=14

MINIMUM_CONVERSION_LOCATIONS_FOR_CLUSTER=1

MINIMUM_CONVERSION_COUNT_FOR_CLUSTER=1

MINIMUM_READ_COUNT_FOR_CLUSTER_INCLUSION=5

MINIMUM_READ_LENGTH=13

MAXIMUM_NUMBER_OF_NON_CONVERSION_MISMATCHES=0

MINIMUM_READ_COUNT_PER_GROUP=5

EXTEND_BY_READ

BMix config file:

COV_MIN=5

REFINE_COV=1

CONFIDENCE_PER=0.95

SEPARATE_STRANDS=1

PARA-suite clustering:

parasuite clust $alignment.bam $hg38_reference $output $dbsnp_142 5

Annotation:

annotatePeaks.pl $clusters.peak hg38 -norevopp -strand "+" > $clusters.annotated

Supplementary Results

Simulation of uridylate-rich and homopolymeric PAR-CLIP reads

To measure the accuracy of the PARA-suite aligner for special types of data (uridylate-rich sequences, which are common in PAR-CLIP and homopolymeric sequences), we generated subsets of our simulated data that contained either >35% T (uridylate-rich sequences) or homopolymeric sequences with stretches of five or more bases of a particular nucleotide.

For the uridylate-rich PAR-CLIP reads, we observed an increase of 1.37% for PARA-suite alignments and an increase of 2.35% in the accuracy for BWA PSSM alignments compared to our basic simulated data (Supplementary Table 5). The accuracy for the PARA-suite decreased by 1.53% but the accuracy was unchanged for BWA PSSM when the PARA-suite was applied to the homopolymeric PAR-CLIP reads (Supplementary Table 5).

Application of the PARA-suite to HITS-CLIP data

Besides PAR-CLIP, other CLIP protocols are also used widely. Therefore, we chose a previously published Argonaute protein HITS-CLIP dataset generated from mouse brain samples (Chi, Zang et al. 2009) to assess the PARA-suite on a different type of CLIP data.

To allow a comparison to previous results on the same dataset, we excluded all sequencing reads that were shorter than 25 bases after quality trimming using cutadapt. Next, we determined the error profile for the pooled replicates of the HITS-CLIP dataset using the respective PARA-suite tool to train its alignment pipeline. Here, we could already verify the high rate of deletions in contrast to insertions or single nucleotide substitutions compared to the mouse reference genome sequence GRCm38 (Chinwalla, Cook et al. 2002). Next, we applied the alignment pipeline to the pooled sequencing reads to align them against GRCm38 and against the transcript database of Ensembl genes Version 77 for the mouse genome assembly, and combined the results. Again, the transcriptomic mapping step revealed 79,658 additional aligned reads spanning exon–exon junctions out of 15,145,095 aligned reads in total (0.526 %). To achieve comparable results for RBP-bound transcribed regions in the mouse genome, we used PIPE-CLIP (Chen, Yun et al. 2014), which is a web-based program for cluster enrichment analysis of CLIP sequencing data. We compared our results with the number of cross-linked regions reported in the PIPE-CLIP publication analyzing the same dataset. The filtering criteria were the same as those in the PIPE-CLIP publication with an enriched cluster length of ≥25 bases and exclusion of duplicated sequencing reads by mapping position. After filtering the entire list of cross-linked regions for those that were supported by deletions in the cross-linked sites, we found 1450 significantly enriched regions by applying false discovery rate (FDR) ≤0.01 filtering. This number was substantially larger than what was found by the initial PIPE-CLIP analysis based on read alignments using Novoalign (http://www.novocraft.com) with 1232 cross-linked regions that were supported by deletions, an increase of 17.69% identified regions in total.

We also applied FDR ≤0.001 filtering to compare our results with the first in-depth analysis of the same data (Zhang and Darnell 2011), which used a cross-linking-induced mutation sites (CIMS) analysis. We identified 984 cross-linked regions showing a reliable deletion, whereas the CIMS analysis applied to the read alignments performed by Novoalign identified only 886 cross-linked regions (Zhang and Darnell 2011).

Supplementary Tables and Figures

Supplementary Table S1: Statistics of FET PAR-CLIP reads (Hoell, Larsson et al. 2011) before and after filtering for confident clusters.

Dataset	Reads in clusters	Reads in confident clusters	% reads passing the filter
EWSR1	1,375,517	700,936	50.96
FUS	1,249,406	923,904	73.95
TAF15	1,310,291	761,710	58.13

Supplementary Table S2: Average numbers for 10 simulated PAR-CLIP datasets.

Simulated reads	1,326,151
Mean read length	23
Clusters	85,691
T–C conversions	624,737
Sequencing errors	367,325
Indels	7324

Supplementary Table S3: Average performance of short read aligners on 10 simulated PAR-CLIP datasets sorted by accuracy. The runtime for BWA PARA was determined without error profile estimation, whereas the runtime for the entire PARA-suite pipeline includes error profile estimation, and alignment against genomic and transcriptomic reference sequences and both of these in combination. The results for "PARAsuite pipeline" refer to an execution where the parameter X was automatically evaluated (default). The results for "PARAsuite X1", "X2" and "X3" refer to executions with fixed values for X (i.e. X = 1, X = 2 and X = 3; see section "execution commands" for further information).

Aligner	Accuracy (in %)	Variance	Recall (in %)	Precision (in %)	Mapped overall	Mapped correctly	CPU time (in s)	Real time (in s)	Memory (in GB)
PARAsuite pipeline	73.14	1.37E-06	84.49	71.85	1,024,792	969,948	2287.3	396.8	6.27
PARAsuite X3 pipeline	72.61	1.26E-06	84.57	70.76	1,057,149	962,901	1365.9	307.7	6.21
PARAsuite X2 pipeline	71.63	1.31E-06	83.39	70.35	993,244	949,870	3786.6	539.2	6.33
PARAsuite	69.74	1.38E-06	82.16	68.24	975,672	924,802	1189.7	153.7	4.42
PARAsuite X3	68.57	1.46E-06	81.85	66.36	995,213	909,390	356.6	73.0	4.42
PARAsuite X2	68.26	1.33E-06	81.04	66.79	945,035	905,293	2405.1	265.1	4.42
BWA 002	68.17	1.38E-06	82.32	64.98	959,235	904,090	3621.9	359.2	4.42
BWA 004	68.17	1.37E-06	82.31	64.98	959,171	904,034	3981.5	390.7	4.42
BWA 2MM	68.17	1.37E-06	82.31	64.98	959,171	904,034	795.5	109.5	4.42
BWA 001	66.73	1.46E-06	80.61	64.26	958,919	884,964	797.2	109.5	4.42
Bowtie 2MM	63.38	1.10E-06	77.91	60.93	886,512	840,540	713.2	120.6	4.46
BWA PSSM	59.80	1.18E-06	74.04	58.72	818,895	793,007	232.4	25.4	2.26
TopHat	59.69	8.35E-07	76.10	55.35	844,902	791,549	592.9	282.9	-
BWA 1MM	59.29	8.68E-07	77.01	53.26	808,033	786,330	76.8	13.4	3.32
Bowtie2	56.22	1.11E-06	73.23	51.43	763,893	745,531	93.8	45.8	4.41
Bowtie 1mm	56.19	1.11E-06	73.20	51.42	763,631	745,227	1016.3	268.0	6.12
PARAsuite X1 pipeline	53.02	8.44E-07	68.55	51.20	716,838	703,161	54.0	10.8	2.26
PARAsuite X1	50.85	9.15E-07	66.52	49.08	685,788	674,399	75.0	43.7	4.41
STAR	50.74	9.10E-07	69.57	43.02	826,871	672,920	133.5	248.6	28.39
MOSAIK	44.88	2.18E-04	62.83	37.16	897,679	595,220	18,125.54	12,128.18	194.16
Subjunc	35.42	9.03E-07	50.61	26.09	597,400	469,751	24.3	64.2	6.65

Supplementary TableS 4: Binding sites detected by BMix, PARalyzer and the hierarchical clustering applied to read alignments of 10 simulated PAR-CLIP datasets. Recall and accuracy cannot be calculated for BMix because it does not provide a list of negative (discarded) clusters.

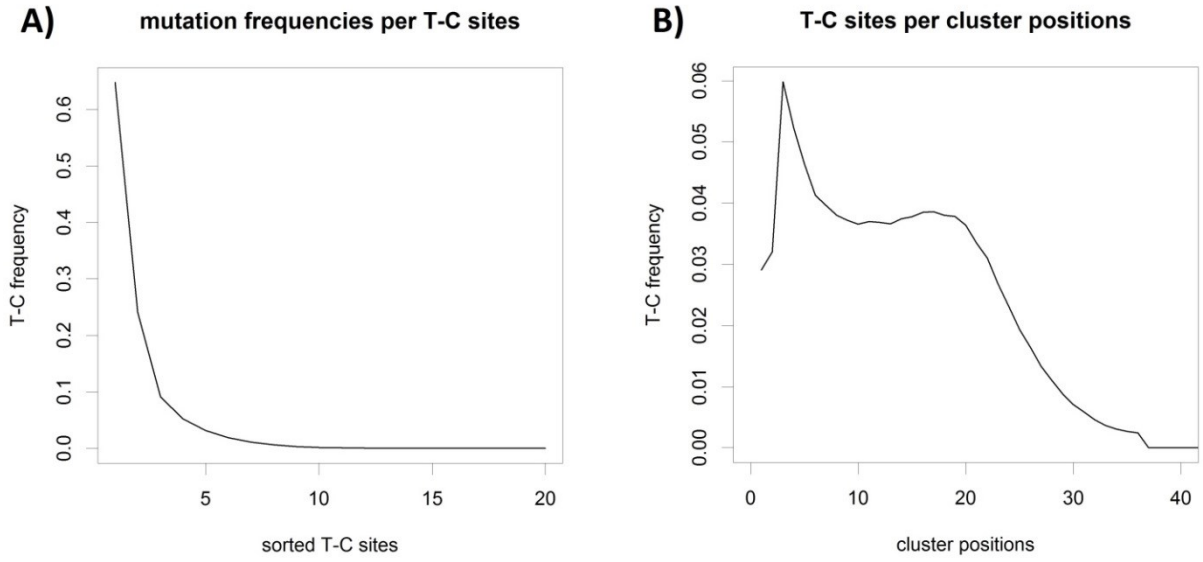
Aligner	True positives	True negatives	False positives	False negatives	Precision (in %)
BWA 2mm BMix	29,631	0	1456	0	95.32
BWA 2mm clustering	30,516	17,587	1795	5229	94.45
BWA 2mm paralyzer	29,255	12,184	5684	1575	83.73
BWA PSSM BMix	28,440	0	1470	0	95.09
BWA PSSM clustering	29,130	15,993	1837	2222	94.07
BWA PSSM paralyzer	28,396	11,172	5663	952	83.37
Bowtie 1mm BMix	26,824	0	969	0	96.51
Bowtie 1mm clustering	27,234	16,230	1137	3605	95.99
Bowtie 1mm paralyzer	27,464	11,252	5223	1299	84.02
Bowtie 2mm BMix	28,061	0	1375	0	95.33
Bowtie 2mm clustering	28,911	16,359	1691	4491	94.47
Bowtie 2mm paralyzer	27,979	11,218	5303	1280	84.07
Bowtie2 BMix	26,832	0	969	0	96.52
Bowtie2 clustering	27,231	16,239	1138	3611	95.99
Bowtie2 paralyzer	29,631	0	1456	0	84.03
PARA-suite BMix	31,918	0	1908	0	94.36
PARA-suite clustering	32,995	17,940	2394	4065	93.23
PARA-suite paralyzer	30,149	12,448	6329	2176	82.65

Supplementary Table S5: Alignment fractions of selected short read aligners applied to the PAR-CLIP results of the *FET* protein family. The PARA-suite aligner outperformed BWA 2MMs and BWA PSSM for all three datasets.

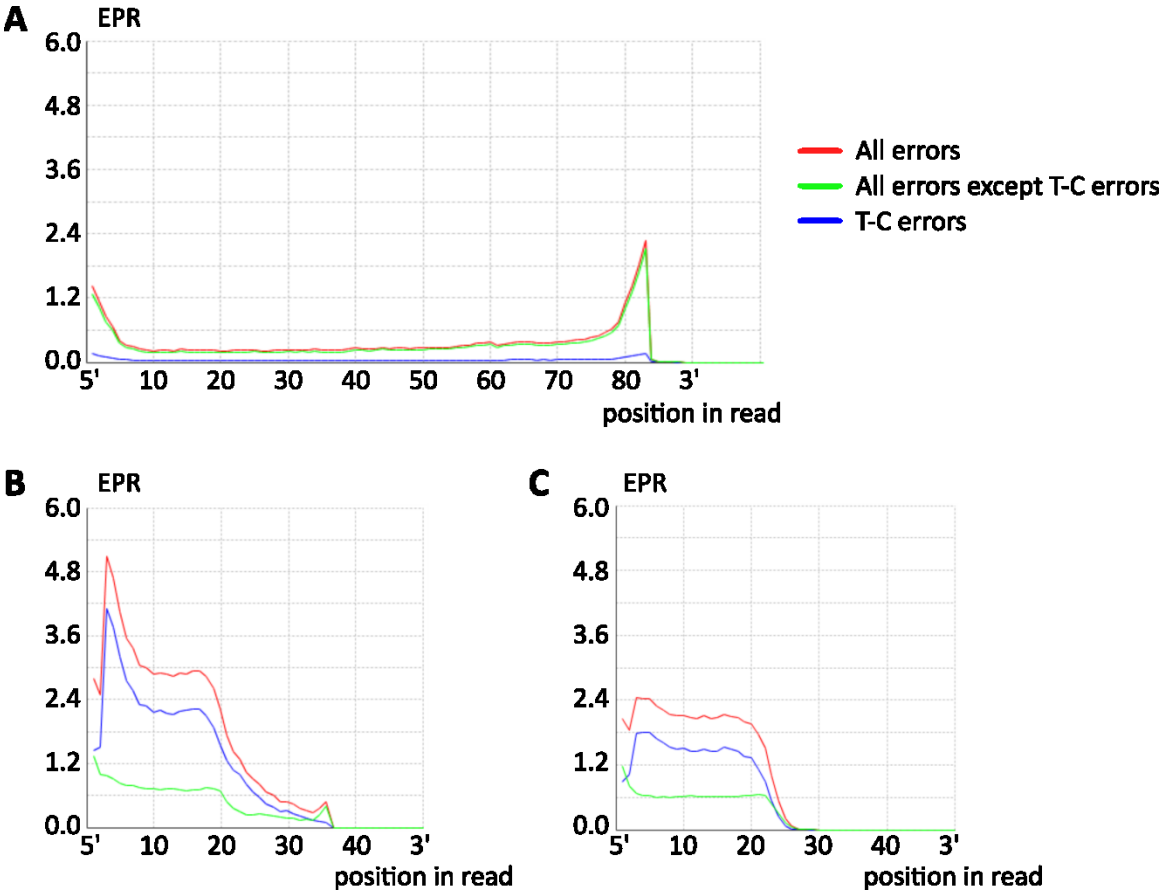
Dataset	**Reads after trimming**	**PARA-suite aligner**	**PARA-suite aligner fraction**	**BWA PSSM**	**BWA PSSM fraction**	**BWA 2MMs**	**BWA 2MMs fraction**
EWSR1	14,557,174	3,193,140	21.94%	2,350,935	16.15%	2,870,884	19.72%
FUS	10,981,718	3,571,035	32.70%	3,161,867	28.79%	3,083,820	28.08%
TAF15	10,611,969	2,457,585	23.16%	1,605,642	15.13%	2,326,287	21.92%

Supplementary Table S6: Accuracy of the PARA-suite and BWA PSSM on uridylate-rich and homopolymeric simulated PAR-CLIP data.

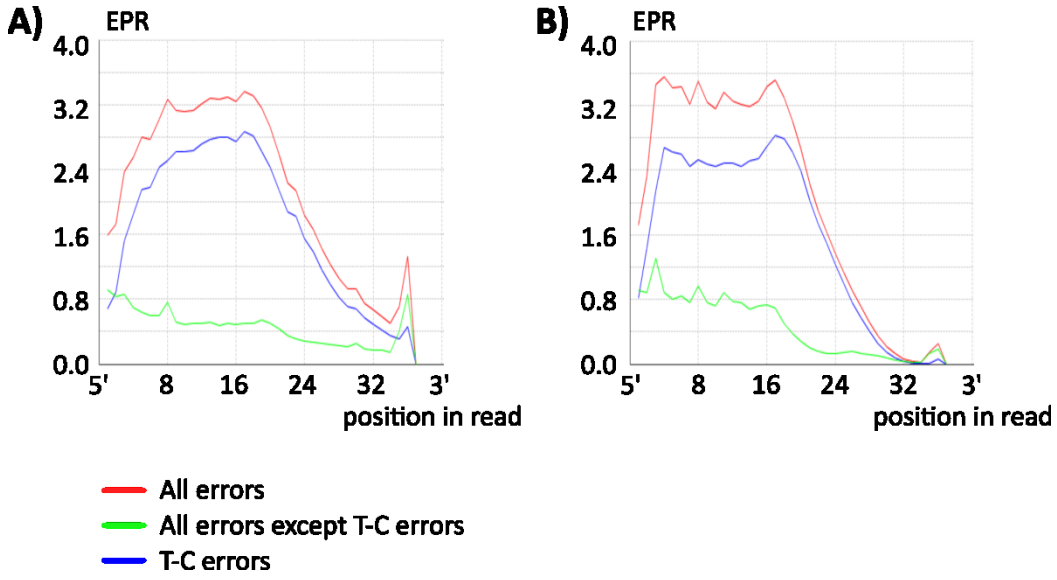
Aligner	**Accuracy**	
	Uridylate-rich	**Homopolymers**
PARA-suite	71.11	68.21
BWA PSSM	62.15	59.80



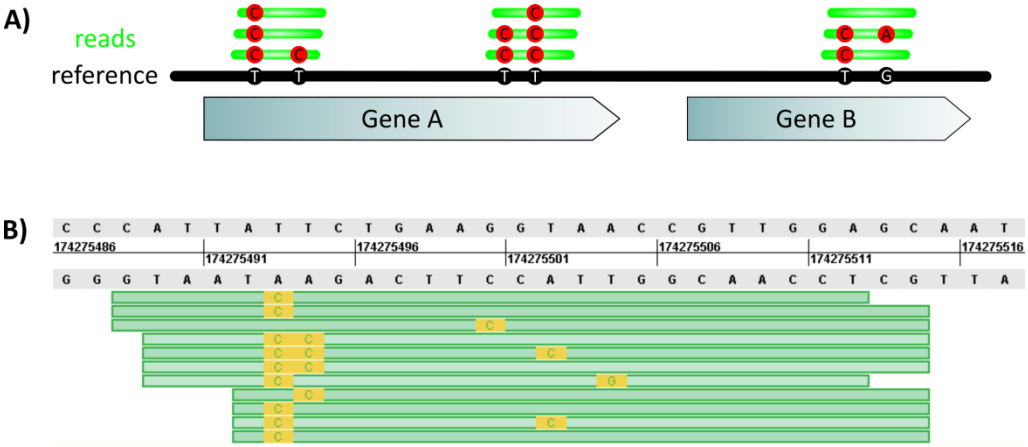
Supplementary Figure S1: (A) T–C conversion frequencies (α) in real PAR-CLIP data (summarized over all FET PAR-CLIPs (Hoell, Larsson et al. 2011)) and sorted by T–C sites within highly confident clusters. (B) Probabilities (β) for the preferred read positions of T–C conversion sites within confident clusters. This graph shows a peak at the beginning of the clusters where the majority of T–C conversions occurred.



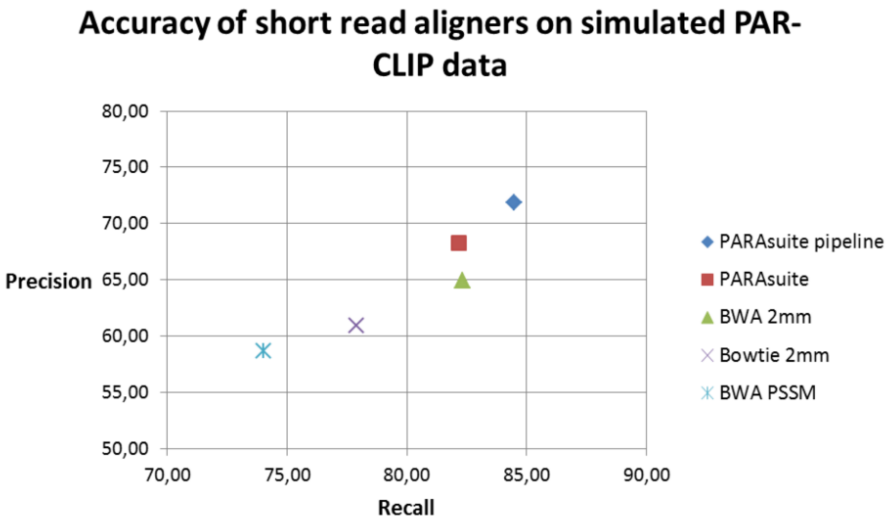
Supplementary Figure S2: Error profiles for (A) human reference RNA-Seq, (B) FUS PAR-CLIP and (C) simulated PAR-CLIP data (averaged over 10 simulated datasets) showing position-wise errors per reads × 100 (EPR). The RNA-Seq profile in (A) has higher sequencing error rates in the outermost bases and a very low average in the mid-range of the reads. The two PAR-CLIP error-profiles in (B) and (C) show a high increase in T–C errors between the read sequences and the reference sequence.



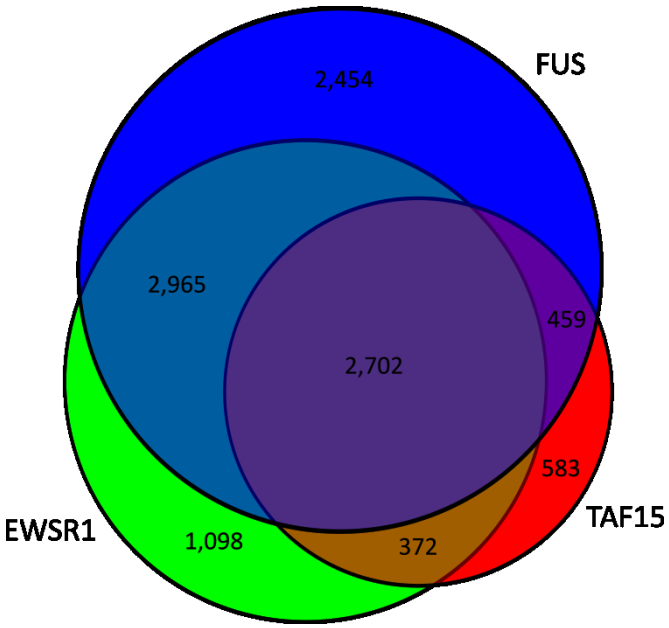
Supplementary Figure S3: Error profiles for (A) HuR (Mukherjee, Corcoran et al. 2011) and (B) MOV10 (Sievers, Schlumpf et al. 2012). Both error profiles lack a peak in the error rate for the first bases but show nearly the same average T–C conversion frequencies as the FET PAR-CLIP dataset with 1.684 errors per reads × 100 (EPR) for HuR and 1.561 EPR for MOV10 as compared to 1.477 EPR for, say, FUS.



Supplementary Figure S4: (A) Schematic view of PAR-CLIP reads aligned against a reference sequence. All reads are stacked into three clusters covering only small parts of the respective genes. Furthermore, T–C conversion sites with high and low mutation frequencies as well as a G–A sequencing errors are shown. (B) Modified representation of a cluster of simulated PAR-CLIP sequencing reads, produced by GenomeView version 2350 (http://genomeview.org/). The cluster shows three T–C conversion sites, one of which has a very high amount of T–C conversions, and A–G and G–C sequencing errors.



Supplementary Figure S5: Average accuracy of short read aligners on 10 simulated PAR-CLIP datasets. Bowtie and BWA were run allowing for two mismatches (Bowtie 2MMs and BWA 2MMs). The PARA-suite, including the transcriptome alignment (called the PARA-suite pipeline), outperformed all other aligners in recall and precision. The performance values obtained for additional aligners are listed in Supplementary Table 2.



Supplementary Figure S6: Overlaps of genes targeted by the FET family identified by the cross-linked regions after cluster filtering. P-values for the Pairwise enrichments are as follows using Fisher's exact test: EWSR1–FUS enrichment = 2.1 (p-value < 0.000); FUS–TAF15 enrichment = 2.0 (p-value < 0.000); EWSR1–TAF15 enrichment = 2.4 (p-value < 0.000). The largest fraction of 2702 distinct genes is covered by all three datasets, which correlates with the results of the initial study.

Supplementary References

Chen, B., Yun, J., Kim, M. S., Mendell, J. T. and Xie, Y. (2014). PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome Biol. 15: R18.

Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature 460(7254): 479-486.

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R. and McPherson, J. D. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915): 520-562.

Hoell, J. I., Larsson, E., Runge, S., Nusbaum, J. D., Duggimpudi, S., Farazi, T. A., Hafner, M., Borkhardt, A., Sander, C. and Tuschl, T. (2011). RNA targets of wild-type and mutant FET family proteins. Nat. Struct. Mol. Biol. 18(12): 1428-1431.

Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, M., Tuschl, T., Ohler, U. and Keene, J. D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. Mol. Cell 43(3): 327-339.

Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012). Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. Nucleic Acids Res. 40(20): e160.

Zhang, C. and Darnell, R. B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat. Biotechnol. 29(7): 607-614.

Publication II

Alterations of miRNAs and miRNA-regulated mRNA expression in GC B cell lymphomas determined by integrative sequencing analysis

Hezaveh K[*,1], Kloetgen A[*,1,2], Bernhart SH[*,3,4,5], et al., Siebert R[§,7], Borkhardt A[§,1], Hummel M[§,6], Hoell JI[§,1] on behalf of the ICGC MMML-Seq Project[‡]

[1]Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Medical Faculty, Düsseldorf, Germany

[2]Department of Algorithmic Bioinformatics, Heinrich-Heine University, Duesseldorf, Germany

[3]Transcriptome Bioinformatics Group, LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany

[4]Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig, Germany

[5]Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany

[6]Institute of Pathology, Charité – University Medicine Berlin, Berlin, Germany

[7]Institute of Human Genetics, University Hospital Schleswig-Holstein Campus Kiel/ Christian-Albrechts University Kiel, Kiel, Germany

[*]These authors contributed equally to this work.

[‡]A list of all authors and affiliations appears in the supplementary information.

[§]These authors contributed equally to this work.

Running head: miRNA-mRNA target regulation in B-cell lymphomas

Abstract

MicroRNAs are well-established players in posttranscriptional gene regulation. However, information on the effects of microRNA deregulation mainly relies on bioinformatic prediction of potential targets, whereas proof of the direct physical microRNAs/target mRNAs interaction is mostly lacking. Within the International Cancer Genome Consortium Project "Determining Molecular Mechanisms in Malignant Lymphoma by Sequencing" (ICGC MMML-Seq), we performed miRnome sequencing from 16 Burkitt lymphomas, 19 diffuse large B-cell lymphomas, and 21 follicular lymphomas. Twenty-two miRNAs separated Burkitt lymphomas from diffuse large B-cell lymphomas/follicular lymphomas, of which 13 have shown regulation by MYC. Moreover, we show expression of three hitherto unreported microRNAs. Additionally, we detect recurrent mutations of hsa-miR-142 in diffuse large B-cell lymphomas and follicular lymphomas, and editing of the hsa-miR-376 cluster, providing evidence for microRNA editing in lymphomagenesis. To interrogate the direct physical interactions of microRNAs with mRNAs, we performed Argonaute-2 photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation experiments. MicroRNAs directly targeted 208 mRNAs in the Burkitt lymphomas and 328 mRNAs in the non-Burkitt lymphoma models. This integrative analysis discovered several regulatory pathways of relevance in lymphomagenesis including Ras, PI3K-Akt and MAPK signaling pathways, also recurrently deregulated in lymphomas by mutations. Our dataset uncovers in detail the mRNA deregulation through microRNAs as a highly relevant mechanism in lymphomagenesis.

Introduction

B-cell lymphomas account for approximately 85% of all lymphomas and form a heterogeneous group of lymphoid neoplasms arising at different stages of B-cell development (Lenz and Staudt 2010). They are classified according to morphological and immunophenotypical features, supplemented by characteristic genomic

translocations (WHO 2008). Although these allow the diagnosis of histological B-cell lymphoma subtypes, molecular subtypes remain largely indistinguishable (Campo, Swerdlow et al. 2011). Presumably due to this molecular heterogeneity, many patients do not respond well to common therapy regimens (Sinha, Nastoupil et al. 2012). Identifying new therapeutic targets and biomarkers is therefore required to improve the accuracy of lymphoma diagnosis and subsequent therapy selection.

One potential class of biomarkers and/or therapeutic targets are a subset of RNA molecules named microRNAs (miRNAs). These are small non-coding RNAs (17–25 nucleotides in length) that bind mostly to target sequences within the 3' UTR of mRNAs. MiRNAs regulate the expression of thousands of mRNAs including those with key roles in cell differentiation and cancer pathogenesis (Farazi, Spitzer et al. 2011). MiRNAs influence immune cell differentiation and play crucial roles in both early and late B-cell differentiation (Di Lisio, Martinez et al. 2012) and lymphomagenesis (Musilova and Mraz 2015). Mechanisms of miRNA dysregulation in lymphomas include copy number alterations (e.g. the miR-17~92 polycistron (He, Thomson et al. 2005)), chromosomal translocation (e.g. hsa-miRNA-125 (Enomoto, Kitaura et al. 2011)) and mutations (e.g. hsa-miR-142 (Kwanhian, Lenze et al. 2012)). Several molecular profiling studies have tried to assess differential miRNA expression in B-cell lymphomas (recently reviewed in (Di Lisio, Martinez et al. 2012, Lim, Trinh et al. 2015, Musilova and Mraz 2015)). Lately, a signature of 38 miRNAs containing MYC-regulated and nuclear factor-□B pathway-associated miRNAs was published, which differentiated Burkitt lymphoma (BL) from diffuse large B-cell lymphoma (DLBCL) (Lenze, Leoncini et al. 2011).

Available data on miRNA expression profiling in B-cell lymphomas is, however, still preliminary, as published profiles are either mostly not derived from large sample collections, do not compare subtypes or originate from either qRT-PCR-based approaches or microarrays. Next generation sequencing (NGS) is able to overcome the disadvantages of previous methods such as probe cross-hybridization (Creighton, Reid et al. 2009) and the limitations of qRT-PCR, such as restricting the analysis to

previously known miRNAs. Furthermore, sequencing-based approaches allow for the discovery of novel miRNAs and large-scale identification of mutated miRNAs.

The present study was performed within the framework of the International Cancer Genome Consortium Project "Determining Molecular Mechanisms in Malignant Lymphoma by Sequencing" (ICGC MMML-Seq). Our aim was to identify NGS-based miRNA signatures in three common subtypes of B-cell lymphomas, i.e BL, DLBCL and follicular lymphoma (FL) and to correlate these to mRNA expression and genomic mutations. Moreover, by performing photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) experiments (Hafner, Landthaler et al. 2010) and intersecting the results with the patient-derived m(i)RNA expression profiles, we aimed at identifying specific miRNA-mRNA target pairs in BL and DLBCL.

Methods

Patient samples

The ICGC MMML-Seq project was approved by the IRB of the Medical Faculty of Kiel University (A150/10) and by the recruiting centers. Informed consent was obtained from all patients (in case of children from their legal guardians). Histopathologic, immunophenotypic and genetic characterization of the initial diagnosis (tumor cell content ≥ 60%) tumor samples was performed as described recently (Richter, Schlesner et al. 2012).

Next generation sequencing

Nucleic acid extraction was performed as previously detailed (Richter, Schlesner et al. 2012). Libraries for miRNA sequencing were prepared using TruSeq Small RNA sample prep kit (Illumina, San Diego, California, USA) according to the manufacturer's instructions. with 100 ng - 1 µg total RNA as input. Libraries were size-fractionated on 6 x TBE gels (Life Technologies, Carlsbad, California, USA). DNA concentration and sizes were analyzed on a 2100 Bioanalyzer (Agilent, Santa Clara, California, USA). 7 pmol of

DNA of each library were loaded onto a flow cell (multiplexing up to four libraries per lane), 50 cycle sequencing was performed using the TruSeq SBS Kit v3 on the HiSeq 2500 (Illumina).

Whole genome sequencing data of tumors and matched controls and transcriptome sequencing data of tumors were generated by the ICGC MMML-Seq project as previously described (Richter, Schlesner et al. 2012). All sequencing data have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/, accession number EGAS00001001394).

Quantitative real-time (qRT)-PCR

Undiluted RT reactions (20 ng of RNA per sample) were combined with TaqMan Universal Master Mix II (no UNG) (Life Technologies) and amplified (7500HT Real-Time PCR System, Life Technologies) with RNU24 and RNU48 as housekeeping genes. Experiments were performed in triplicate and analyzed using the 2-ΔΔCT method.

AGO2-PAR-CLIP

AGO2-PAR-CLIP was carried out as previously described (Hafner, Landthaler et al. 2010) with modifications (mostly relating to the washing steps during immunoprecipitation) owing to the use of monoclonal anti-AGO2 antibody (#4-642, EMD Millipore, Billerica, Massachusetts, USA) (Farazi, Ten Hoeve et al. 2014). In brief, following the addition of 4-thiouridine, an immunoprecipitation using a monoclonal anti-AGO2 antibody isolated the RNA-protein complexes. After protein digestion, sequencing adapters were ligated to the purified RNA fragments. Following reverse transcription, PAR-CLIP libraries were sequenced on a HiSeq2500 (Illumina) (Spitzer, Hafner et al. 2014).

Bioinformatic methods

Bioinformatic analyses of the genome and transcriptome data were performed as described recently, employing the various pipelines established in the ICGC MMML-Seq (Richter, Schlesner et al. 2012) (also Supplementary Methods).

MiRNA and PAR-CLIP analysis

Following adapter removal, reads were mapped onto the human genome (1000 genomes project, hs37d5100) using segemehl (Hoffmann, Otto et al. 2009). Novel miRNA prediction was performed using miRanalyzer 0.3 (Hackenberg, Sturm et al. 2009) (default parameters), target prediction using miRanda (Enright, John et al. 2004) (miRsvr-score < -1.2).

After filtering and trimming the PAR-CLIP reads, we obtained a total of 62 281 382 single-end reads, which were aligned with BWA(Li and Durbin 2009) with up to two mismatches between a read sequence and the reference sequence (hg19). All reads failing this mapping were aligned against the transcriptome database (Ensembl Genes 75). Aligned reads were piled into clusters by PARA-suite (Kloetgen et al., submitted). As PAR-CLIP reads contain thymidine to cytidine (T-C) conversions at the sites of crosslinks, all identified clusters were filtered to receive the most confident target regions. Excluding clusters containing < 5 reads and < 25% T-C conversions (excluding 100% T-C conversion sites as these might result from SNVs) resulted in (prior to pooling) 1 329 clusters for SU-DHL-4, 1 517 clusters for SU-DHL-6, 1 209 clusters for Namalwa and 425 clusters for Raji. Please refer to Supplementary Methods for more details (including miRNA-mRNA correlation analyses).

Results

Molecular classification of BL versus DLBCL/FL using a 25 miRNA classifier

We profiled tumor samples from 56 patients including 16 BL (based on a molecular classifier; all patients ≤ 18 years), 19 DLBCL (including 7 GCB DLBCLs, 10 ABC DLBCLs and 2 type III DLBCLs) and 21 FL (mainly grade 1/2) (Supplementary Table 1). We obtained 1 169 752 727 sequencing reads in total (average of 20 888 442 reads per sample, Supplementary Table 2). Following normalization of miRNA reads, we performed an unsupervised hierarchical clustering. Unexpectedly (and different to what we observed on the transcriptomic level, data not shown), no clear distinction between BL, DLBCL and FL was achieved based on miRNA expression profiles

(Supplementary Figure 1A). We then ranked the miRNAs by mean expression and, discarding those that showed little expression variability, chose the top ten miRNAs for validating our NGS data by qRT-PCR. A correlation analysis showed the consistency of miRNA expression levels regardless of the employed method of quantification (spearman's rank correlation test, 10/10, high correlation (R>0.7), p-values for the correlation between qRT-PCR expression and NGS expression ≤ 0.05 in 7/10 cases; please refer to supplementary bioinformatic methods for details on all p-value calculations) (Supplementary Figure 1B, Supplementary Table 3).

To recognize subtler molecular differences that escape unsupervised clustering approaches, we performed a differential gene expression analysis between BL versus DLBCL, BL versus FL and DLBCL versus FL using edgeR (Supplementary Table 4 & supplementary bioinformatic methods). Clustering of the top 25 differentially expressed miRNAs between each two lymphoma subtypes (BL/DLBCL, BL/FL, and FL/DLBCL) revealed separation according to the subtypes (Figure 1A). Employing this approach, BL and FL separated clearly, whereas the discrimination between BL/DLBCL and FL/DLBCL was less pronounced, most likely due to the molecular heterogeneity of DLBCL (Alizadeh, Eisen et al. 2000, Iqbal, Shen et al. 2015). As there were no dual hit patients and no DLBCL cases with MYC breaks as single events in our cohort, we were not able to test, whether our classifier was able to single out those cases.

Interestingly, 7/25 miRNAs differentially expressed between BL/DLBCL (hsa-miRs-23a/29b/130b/146a/155/196b/222) were also part of a recently published, 27-miRNA qRT-PCR derived classifier for the differentiation of those both subtypes (Iqbal, Shen et al. 2015) (6.9 fold enrichment, one-sided Fisher's exact test, p-value for the overlap of the two classifiers 2.322E-05). In a previous array-based study, we established a classifier consisting of 38 miRNAs, which differentiated BL from DLBCL (Lenze, Leoncini et al. 2011). From the 25 miRNAs top differentially expressed herein, eight overlapped with these 38 miRNAs (hsa-miRs-23a/29b/146a/155/193a/221/222/339) (5.1 fold, p-value for this overlap 5.900E-05). In summary, five miRNAs (hsa-miR-23a/29b/146a/155/222) seem to be robustly able to differentiate BL from DLBCL irrespective of the collection of cases and the method

used for analysis. We additionally analyzed previously published microarray data (Lenze, Leoncini et al. 2011) for 64 BL cases and 86 DLBCL cases to validate the predictive power of our classifiers on an independent dataset (see Supplementary Methods for further details). We predicted correct class labels for 126/128 cases with a majority vote of at least 80% (recall = 98.44%; 57/58 BL cases and 69/70 DLBCL cases; overall accuracy = 84.0%) on our 25 miRNA-classifier for BL vs. DLBCL.

To address the question on how to distinguish BL from the other investigated histological subtypes, we merged DLBCL and FL and clustered the top 25 differentially expressed miRNAs between BL and DLBCL/FL inferred with edgeR. This resulted - with the exception of two BL cases - in a clear separation between BL and DLBCL/FL (Figure 1B). Of those top 25 differentially expressed miRNAs, 14 were up- and 11 were downregulated in BL compared to DLBCL/FL (Table 1). As our analysis takes both "5p" and "3p" versions (previously referred to as mature miRNA and star strand) of each miRNA into account, our classifier consists of 22 unique miRNAs. Interestingly, for a total of 13 of these miRNAs a regulation by MYC was reported in the literature (Chang, Yu et al. 2008, Di Lisio, Sanchez-Beato et al. 2012, Zhao, Bai et al. 2012, Xiong, Jiang et al. 2013, Liu, Mai et al. 2014, Tao and Zhao 2014).

Table 1: 25 miRNA classifier separating BL from DLBCL/FL. MiRNAs for whom regulation by MYC has been shown are shaded in gray. cpm indicates counts per million; FDR, false discovery rate.

miRNA	p-value	FDR	cpm BL	cpm DLBCL/FL
hsa-miR-17-3p	1.4 E-14	1.6 E-12	2177.0	279.2
hsa-miR-18a-3p	4.8 E-12	2.7 E-10	88.9	16.1
hsa-miR-19a-3p	5.7 E-12	3.0 E-10	1852.6	349.2
hsa-miR-20a-3p	7.2 E-28	4.1 E-25	28.9	3.9
hsa-miR-25-5p	1.7 E-21	5.0 E-19	79.2	11.2
hsa-miR-29c-5p	6.6 E-12	3.1 E-10	12.6	51.0
hsa-miR-93-3p	4.7 E-10	1.1 E-08	153.4	26.0

hsa-miR-106b-3p	7.6 E-11	2.3 E-09	1010.8	381.5
hsa-miR-106b-5p	3.9 E-11	1.4 E-09	617.7	211.0
hsa-miR-130b-3p	1.1 E-18	2.1 E-16	701.2	148.6
hsa-miR-150-3p	9.3 E-12	4.1 E-10	3.5	36.5
hsa-miR-150-5p	8.4 E-13	6.0 E-11	649.6	7980.4
hsa-miR-155-5p	3.2 E-10	7.9 E-09	1152.3	10989.1
hsa-miR-184	1.6 E-10	4.4 E-09	0.9	123.6
hsa-miR-196b-5p	2.9 E-10	7.6 E-09	4.5	53.7
hsa-miR-151b	6.1 E-11	1.9 E-09	15.5	146.0
hsa-miR-211-5p	1.4 E-11	5.9 E-10	0.1	1.7
hsa-miR-221-3p	9.6 E-15	1.4 E-12	461.7	3018.5
hsa-miR-296-3p	2.1 E-12	1.3 E-10	6.6	1.7
hsa-miR-335-3p	1.9 E-11	7.2 E-10	566.5	100.8
hsa-miR-339-5p	1.9 E-13	1.6 E-11	98.8	20.3
hsa-miR-664-3p	1.1 E-10	3.1 E-09	14.7	97.9
hsa-miR-664-5p	4.1 E-10	9.8 E-09	4.2	29.1
hsa-miR-573	5.3 E-11	1.8 E-09	3.3	0.4
hsa-miR-4521	1.9 E-13	1.6 E-11	35.6	4.6

Figure 1

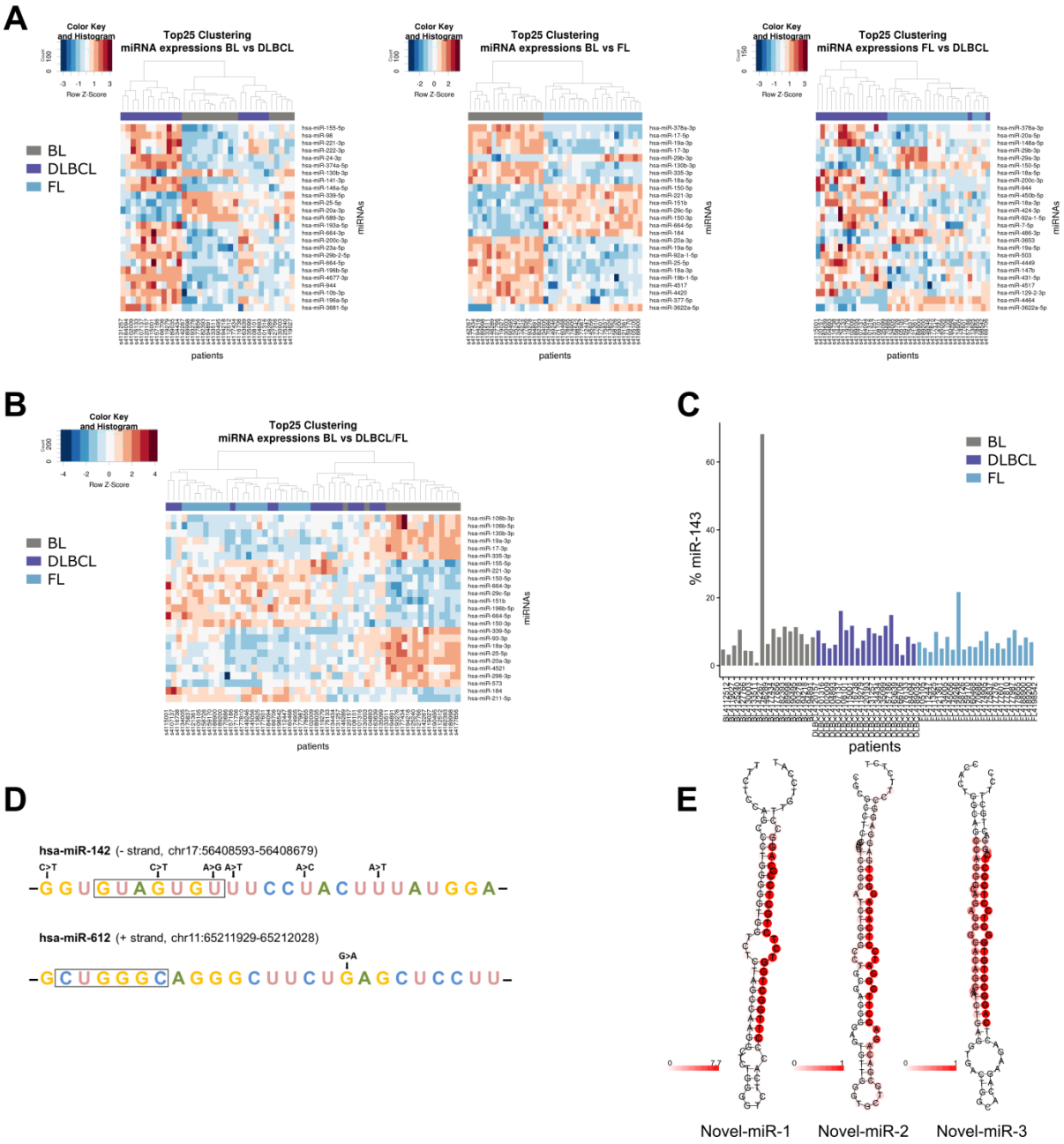


Figure 1: The miRnome of B-cell lymphomas. A) Clustering according to the top 25 differentially expressed miRNAs inferred with edgeR between FL (light blue), DLBCL (blue), and BL (gray), in pairwise comparisons. B) Clustering according to the top 25 differentially expressed miRNAs inferred with edgeR between BL and DLBCL/FL. C) Hsa-miR-143 expression across all patient samples. D) Visualization of the genomic mutations of those miRNAs, which show alterations in their mature sequences. Shown

are the mature sequences of the respective miRNAs, the seed sequences are highlighted by black boxes. The positions of the mutations are also indicated. E) Predicted folding of the three biochemically validated novel miRNAs.

Hsa-miRNA-143 is highly abundant in GC B cell lymphomas

Contrary to earlier reports (Akao, Nakagawa et al. 2007), hsa-miR-143 showed a very high expression across most lymphoma samples (Figure 1C). Expression ranged from 0.8% to 68.2% (mean 8.9%) of all reads mapping to miRNAs for this miRNA alone with no significant differences between subtypes (means 10.8%, 7.4% and 8.9% for BL/FL/DLBCL, respectively). The extremely high expression of this miRNA (68.2%) in patient 4146289 (BL) was confirmed by qRT-PCR as was the lower expression (0.8%) in patient 4142267 (BL) (Supplementary Figure 1B). As hsa-miR-143 forms a bicistronic cluster on chromosomal region 5q33.1 with hsa-miR-145, we also investigated the latter's expression. MiRNAs in bicistronic clusters are transcribed simultaneously and thus show similar expression patterns. The correlation analysis (p-value 0.0034 for the correlation between hsa-miR-143 and hsa-miR-145 expression, R=0,39) confirmed the validity of the hsa-miR-143 expression with similar expression patterns (Supplementary Figure 1C). Whole-genome derived copy number analysis of all patient samples revealed no relevant alterations in either the promoter or the genomic region of hsa-miR-143/145. The reason for the observed high expression of the hsa-miR-143/145 cluster thus remained unclear.

To identify molecular pathways associated with the high expression of hsa-miR-143, we performed a target prediction and investigated, which of the predicted targets were downregulated in the respective RNASeq data. This resulted in 186 predicted hsa-miR-143/mRNA interaction pairs (Supplementary Table 5). Gene Ontology analysis employing Gorilla (Eden, Navon et al. 2009) revealed that the GO term "ubiquitin-protein transferase activity" (GO:0004842) showed the highest enrichment (5.33 fold, p-value <0.001). The associated target genes are listed in Supplementary Table 6, the entire GO output in Supplementary Table 7.

Hsa-miR-142 is recurrently mutated in its mature sequence in DLBCL and FL

Next, we searched for mutated miRNAs, which were detectable on both DNA as well as RNA level. Mutations in miRNAs located in the IGH gene locus were excluded. We identified 10 mutations (Table 2) in 8 patients (6 mutations in 5 DLBCL patients, 4 mutations in 3 FL patients) with a total of 4 miRNAs affected (hsa-miR-142/-612/-3655/-4322). In two miRNAs (hsa-miR-142/-612), the mutations were within the mature sequence (Figure 1D).

Hsa-miR-142 was the most frequently mutated miRNA with six different mutations in 5/40 DLBCL/FL patients. Two of those were located within the seed sequence. Looking at the subgroups, this broke up into 3/19 in DLBCL and 2/21 in FL. A recent publication (Kwanhian, Lenze et al. 2012) reported a mutation frequency of hsa-miR-142 in 11/56 DLBCL cases. Our data therefore confirms the mutation frequency in DLBCL and extends this finding to FL.

Table 2: Genomically mutated miRNAs. chr indicates chromosome; gen. pos., genomic position (hg19); mature, whether or not the sequenced alteration is located within the mature miRNA sequence; ref/alt, reference/alternative; PID, personal identifier.

miRNA	**chr**	**gen. pos.**	**mature**	**ref/alt**	**PID**	**Subtype**
hsa-mir-142	chr17	56408624	y	C>T	4102009	DLBCL
hsa-mir-142	chr17	56408616	y	A>C	4112447	FL
hsa-mir-142	chr17	56408630	n	C>T	4120193	DLBCL
hsa-mir-142	chr17	56408620	y	A>T	4160468	FL
hsa-mir-142	chr17	56408621	y	A>G	4160468	FL
hsa-mir-142	chr17	56408612	y	A>T	4176133	DLBCL
hsa-mir-612	chr11	65211962	y	G>A	4135099	DLBCL
hsa-mir-3655	chr5	140027478	n	A>G	4177376	FL
hsa-mir-4322	chr19	10341090	n	C>A	4134434	DLBCL
hsa-mir-4322	chr19	10341109	n	C>T	4135099	DLBCL

The hsa-miR-376 cluster is recurrently edited in GC B cell lymphoma subtypes

RNA editing is a process in which (most commonly) adenosine deaminases perform the site-specific hydrolytic deamination of adenosine to inosine (Tomaselli, Locatelli et al. 2014). When an RNA molecule contains an inosine, the sequenced change usually is A-to-G. We searched for mutations exclusive to the miRNA data (not seen on genomic level), which thus represented bona fide miRNA editing events. Starting with all SNVs, we restricted our search to those in the seed regions and discarded known SNVs as reported in dbSNP_135 including rare variants. The remaining 40 candidates were manually evaluated (correct position of SNVs in sequence reads, A-to-G change, sequencing quality of errors), narrowing the list to four SNVs (Table 3). These mapped to hsa-miR-1260b, hsa-miR-376a1/2, and hsa-miR-376c, with the hsa-miR-376 family belonging to the same genomic cluster on 14q32. Editing frequencies (edited reads versus all reads) ranged from 35-86% across miRNAs in the lymphoma samples showing this phenomenon. The editing "efficiency" (percent alternative base) and the expression of ADAR, one of the main enzymes responsible for RNA editing (Tomaselli, Locatelli et al. 2014), per case (with observed editing) showed a weak correlation (p-value 0.044; R=0.30), possibly pointing at the mechanism behind the observed miRNA editing.

Table 3: RNA editing events across lymphoma subtypes. Numbers of samples showing the editing events at the indicated genomic positions with at least 10 sequenced reads at this position are given. Chr indicates chromosome; gen. pos., genomic position (hg19); ref/alt, reference/alternative; mean % alternative, mean % of reads differing from the reference sequence

miRNA	**chr**	**gen. pos.**	**ref/alt**	**mean % alternative**	**# samples with editing**
hsa-miR-376a1/2	chr14	101506460	A>G	86.2%	39
hsa-miR-376c	chr14	101506074	A>G	45.2%	19
hsa-miR-1260b	chr11	96074619	A>G	35.3%	11

Discovery of three hitherto unreported miRNAs expressed in GC B cell lymphomas

We employed miRanalyzer to predict hitherto unreported miRNAs (Hackenberg, Sturm et al. 2009), then choosing a subset of 20 (Supplementary Table 8) and observed the correct processing of three candidates (Table 4) by Northern blotting (Supplementary Figure 2A). Secondary structures of these three hitherto unreported miRNAs as predicted by RNAfold (Denman 1993) are depicted in Figure 1E.

Novel-miR-1 was moderately expressed in SU-DHL-4 and weakly expressed in Namalwa and Raji. Novel-miR-2 was expressed in Raji and SU-DHL-4, novel-miR-3 in Raji and Namalwa (Supplementary Figure 2A). We next assessed publicly available RNASeq data of 16 cell lines (details in Supplementary Material) across a variety of tissues/diseases for expression of our three novel miRNAs. Novel-miR-2 and novel-miR-3 were broadly expressed (16/16 cell lines, 12/16 cell lines, respectively), novel-miR-1 showed a restricted expression, and was only detected in the B-lymphoblastoid cell line GM12878 (data not shown). We then focused on novel-miR-1 (restricted expression) and novel-miR-2 (broad expression) for further experiments.

We performed overexpression/knockdown studies in SU-DHL-4 (novel-miR-1) and Raji (novel-miR-2) followed by RNASeq (Supplementary Figure 2B). To only identify mRNAs whose differential expression was due to direct targeting effects, we searched for mRNAs that carried the respective seed sequence, had a significant miRanda score and were inversely regulated (FDR for all further calculations < 0.05).

Downregulation of novel-miR-1 and novel-miR-2 resulted in two (EIF3C, MPEG1) and 3 (HLA-DRB5, PFKFB4, PPP1R35) upregulated mRNAs, respectively. Upregulation of novel-miR-1 led to the downregulation of 55 coding mRNAs (Supplementary Table 9), whereas overexpression of novel-miR-2 only resulted in two downregulated mRNAs (SLCO2B1, UPP1). Interestingly, there were many genes previously reported in the context of lymphomagenesis among those mRNAs, which carried novel-miR-1 seed sequences. These genes represent its bona fide direct targets and included CARD11, E2F1, MCM2 and MCM7. Novel-miR-1 thus potentially represents a new player in lymphomagenesis. Sequences of novel-miR-1/-2/-3 have been submitted to miRBase.

Table 4: Novel miRNAs in B-cell lymphomas. chr indicates chromosome; gen. pos., genomic position (hg19).

Northern Blot	probe	chr	gen. pos.	mature miRNA sequence
Positive (novel-miR-1)	NB-5	10	50035510-50035603	GCACACTGACACAGAGAGAGAGA
Positive (novel-miR-2)	NB-19	M	3363-3463	CCAACGTTGTAGGCCCCTACGGGCTACT
Positive (novel-miR-3)	NB-20	12	52453530-52453613	TCACTGCAGGGCCCTAGCAATA

AGO2-PAR-CLIP identifies direct mRNA-miRNA interactions in BL and DLBCL/FL

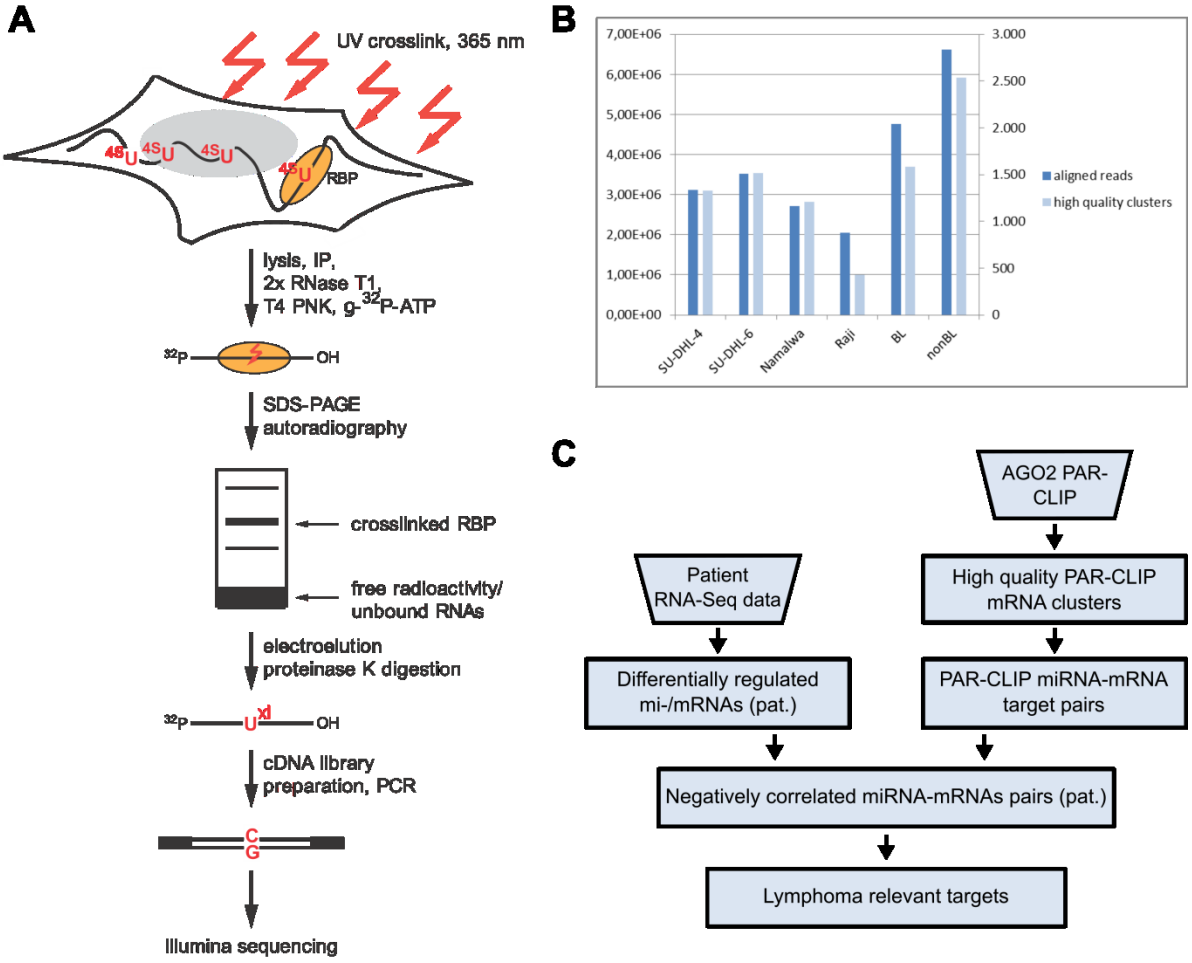
To identify those mRNAs, which were physically targeted by miRNAs in Argonaute-miRNA-mRNA complexes (versus performing bioinformatic predictions to identify putative interactions only), we performed PAR-CLIP experiments of endogenous AGO2 (Hafner, Landthaler et al. 2010) (Figure 2A) in two BL cell lines (Namalwa, Raji) and two non-BL cell lines (SU-DHL-4, SU-DHL-6; both t(14;18) positive). Merging the BL and the non-BL sequencing reads resulted in 1 587 and 2 532 clusters, respectively (individual read numbers see Figure 2B). Combining these miRNA-target sites with the transcriptome data also available for each patient (Figure 2C) led to 302 (BL) / 540 (non-BL) miRNA-mRNA interactions with negative correlations, with several genes being targeted by more than one miRNA (Supplementary Table 10). On individual gene level the numbers were 208 (BL) and 328 (non-BL).

Many of the genes showing direct regulation by miRNAs have well-known roles in lymphomagenesis (Figure 2D). These genes fell into different functional categories, some for which expression was correlated to prognosis (B2M (Hagberg, Killander et al. 1983) (targeted by hsa-miR-106b), MDM2 (Solenthaler, Matutes et al. 2002) (hsa-miR-361)), for which differential expression was shown (CCR6 (Durig, Schmucker et al. 2001) (hsa-miR-296) or were correlated to treatment resistance (e.g. THY1 (Ishiura, Kotani et al. 2010) (hsa-miR-149)). For other targeted genes, mutations (ID3 (Richter, Schlesner et al. 2012) (hsa-miR-4424), NPAT (Kuppers 2011) (hsa-miR-4518),

SMARCA4 (Love, Sun et al. 2012) (hsa-miR-2467), TCF3 (Schmitz, Young et al. 2012) (hsa-miR-184)) or translocations (e.g. CDK6 (Chen, Law et al. 2009) (hsa-miR-148b)) have been described in several types of lymphomas.

Significantly enriched and lymphoma-relevant targeted KEGG pathways (Table 5) showing a differential expression between BL and non-BL included “miRNAs in cancer” (hsa05206, 10 genes, p-value 7.56E-07, enrichment 7.7), “MAPK signaling” (hsa04010, 11 genes, p-value 1.79E-08, enrichment 9.7), “Ras signaling” (hsa04014, 8 genes, 7.56E-06, enrichment 8.0), and “PI3K-Akt signaling” (hsa04151, 8 genes, p-value 1.48E-04, enrichment 5.2). Total numbers of genomically detected mutations in the four mentioned pathways were in that order 122 (297 genes in the pathway), 111 (257 genes), 83 (228 genes) and 124 (347 genes). As these overlaps were not statistically significant (p-values 0.221 to 0.409), this suggests that the respective pathways are targeted and deregulated either by virtue of miRNA interference or by mutations.

Figure 2



D

Prognosis	Diff. expression	Mutations	Translocations	Treatment resistance	Other
B2M	CCR6	ID3	CDK6	TCF4	DOCK2
MDM2		NPAT	IRF8	THY1	
		SMARCA4		PAK1	
		TCF3			

Figure 2: Direct miRNA-mRNA regulation in B-cell lymphomas. A) PAR-CLIP principle. Following the addition of 4-thiouridine, an immunoprecipitation with subsequent protein digestion is performed. Purified RNA fragments are reverse transcribed and cDNA libraries are sequenced on a HiSeq2500 followed by bioinformatic analysis (adapted from Hafner et al. (Hafner, Landthaler et al. 2010)). B) PAR-CLIP library statistics. The left y-axis shows the number of aligned reads, the right y-axis the number of high quality PAR-CLIP clusters. Employed cell lines are indicated. C) Flow chart of the integrative miRNA-mRNA regulation analysis (adapted from Farazi

et al. (Farazi, Ten Hoeve et al. 2014)). D) List of lymphoma relevant genes for which regulation by distinct miRNAs could be elucidated.

Table 5: Targeted KEGG pathways and associated miRNA-mRNA regulation pairs.

KEGG pathway	gene	targeting miRNA(s)	mutations detected
hsa05206: microRNAs in cancer	APC2	hsa-miR-663b, hsa-miR-3648	-
hsa05206: microRNAs in cancer	CCND1	hsa-miR-27b-5p, hsa-miR-590-5p	-
hsa05206: microRNAs in cancer	E2F3	hsa-miR-141-5p	-
hsa05206: microRNAs in cancer	MDM2	hsa-miR-361-3p	BL4112512
hsa05206: microRNAs in cancer	MMP16	hsa-miR-151a-3p	BL4190495
hsa05206: microRNAs in cancer	NOTCH4	hsa-miR-573	FL4178655
hsa05206: microRNAs in cancer	PAK4	hsa-miR-2355-5p	-
hsa05206: microRNAs in cancer	PDGFA	hsa-miR-181b-3p, hsa-miR-4420	-
hsa05206: microRNAs in cancer	PRKCB	hsa-miR-577	DLBCL4131257
hsa05206: microRNAs in cancer	ZFPM2	hsa-miR-127-5p, hsa-miR-181b-3p, hsa-miR-4420	DLBCL4134434,FL4112447
hsa04014: Ras signaling pathway	FLT4	hsa-miR-17-3p	-
hsa04014: Ras signaling pathway	MRAS	hsa-miR-181b-3p, hsa-miR-1304-3p	-
hsa04014: Ras signaling pathway	PAK1	hsa-miR-424-5p	-
hsa04014: Ras signaling pathway	PAK4	hsa-miR-2355-5p	-
hsa04014: Ras signaling pathway	PAK6	hsa-miR-125a-3p	DLBCL4135099
hsa04014: Ras signaling pathway	PDGFA	hsa-miR-181b-3p, hsa-miR-4420	-
hsa04014: Ras signaling pathway	PLA2G4A	hsa-miR-3940-3p	-
hsa04014: Ras signaling pathway	PRKCB	hsa-miR-577	DLBCL4131257
hsa04151: PI3K-Akt signaling pathway	CCND1	hsa-miR-27b-5p, hsa-miR-590-5p	-
hsa04151: PI3K-Akt signaling pathway	COL6A6	hsa-miR-135b-5p, hsa-miR-140-3p, hsa-miR-4424, hsa-miR-4999-5p	-
hsa04151: PI3K-Akt signaling pathway	FLT4	hsa-miR-17-3p	-
hsa04151: PI3K-Akt signaling pathway	LPAR1	hsa-miR-3194-5p, hsa-miR-3940-3p	-
hsa04151: PI3K-Akt signaling pathway	MDM2	hsa-miR-361-3p	BL4112512
hsa04151: PI3K-Akt signaling pathway	PDGFA	hsa-miR-181b-3p, hsa-miR-4420	-
hsa04151: PI3K-Akt signaling pathway	PPP2R1B	hsa-miR-140-3p	BL4127766
hsa04151: PI3K-Akt signaling pathway	PPP2R3A	hsa-miR-708-5p	-
hsa04010: MAPK signaling pathway	CACNB1	hsa-miR-3622a-5p	-
hsa04010: MAPK signaling pathway	ECSIT	hsa-miR-34a-5p, hsa-miR-3605-3p	-

hsa04010: MAPK signaling pathway	MRAS	hsa-miR-181b-3p, hsa-miR-1304-3p	-
hsa04010: MAPK signaling pathway	PAK1	hsa-miR-424-5p	-
hsa04010: MAPK signaling pathway	PDGFA	hsa-miR-181b-3p, hsa-miR-4420	-
hsa04010: MAPK signaling pathway	PLA2G4A	hsa-miR-3940-3p	-
hsa04010: MAPK signaling pathway	PPM1A	hsa-miR-199a-3p, hsa-miR-199b-3p	-
hsa04010: MAPK signaling pathway	PRKCB	hsa-miR-577	DLBCL4131257
hsa04010: MAPK signaling pathway	RAPGEF2	hsa-miR-641, hsa-miR-3613-3p, hsa-miR-4517	DLBCL4177376
hsa04010: MAPK signaling pathway	TAB1	hsa-miR-361-3p	-
hsa04010: MAPK signaling pathway	TGFBR2	hsa-miR-4487	DLBCL4108101

Discussion

We here report a deep sequencing analysis to identify differences in miRNA expression in BL, FL and DLBCL patient samples collected within the ICGC MMML-Seq Consortium. Comparing our miRNA classifiers separating the three entities to previous array- and qRT-PCR based studies, five miRNAs (hsa-miRs-23a/29b/146a/155/222) were recurrently identified to differentiate BL from DLBCL (Lenze, Leoncini et al. 2011, Iqbal, Shen et al. 2015) and two miRNAs (hsa-miR-92/150) to robustly separate FL from DLBCL (Roehle, Hoefig et al. 2008, Lawrie, Chi et al. 2009). Of note, 13 of those miRNAs differentiating BL/DLBCL were previously reported to be regulated by MYC (Chang, Yu et al. 2008, Di Lisio, Sanchez-Beato et al. 2012, Zhao, Bai et al. 2012, Xiong, Jiang et al. 2013, Liu, Mai et al. 2014, Tao and Zhao 2014), emphasizing the role of MYC in the pathogenesis of BL.

The higher discriminative power between BL, DLBCL and FL based on unsupervised analysis of the RNA-Seq data likely comes from less variation among the patients, which might be partly due to the higher number of analyzed genes when compared to miRNA-Seq as well as overlapping effects of some miRNAs. Supervised analysis based on differentially expressed miRNAs, however, had a similar discriminative power as the supervised analysis of differentially expressed mRNAs.

We identified hsa-miR-143 as highly expressed (compared to all other miRNAs) across all three subtypes. This miRNA has hitherto mostly been discussed as a tumor suppressor in (mainly) epithelial malignancies (Kent, McCall et al. 2014). However, a recent study in colorectal cancer found hsa-miR-143 overexpression correlated to short

overall survival (Schou, Rossi et al. 2014). Earlier publications have also reported a downregulation (mostly associated with its deletion) of the hsa-miR143/145 cluster in some leukemias and lymphomas (Dou, Zheng et al. 2012, Liu, Iqbal et al. 2013). Examples of other miRNAs, which have - depending on the tumor type - been described as both tumor suppressors and oncogenes include hsa-miR-26a, and the hsa-miR-141/200a-cluster (Farazi, Spitzer et al. 2011). The high expression of hsa-miR-143 raises the possibility of a new and more general role for this miRNA in lymphomagenesis.

We describe recurrent mutations in hsa-miR-142 in FL at a frequency of 9.5%. Additionally, we confirm recurrent mutations of hsa-miR-142 at a frequency of 12.5% in DLBCL compared to 19.6% as previously published (Kwanhian, Lenze et al. 2012). Hsa-miR-142 mutations lead to the generation of new target sites as well as abolishing originally canonical ones in lymphoma-relevant genes, suggesting that hsa-miR-142 mutations act as a pathogenic mechanism across lymphoma subtypes. Other - albeit non-recurrent - seed sequence mutations affected hsa-miR-612, which was previously shown to suppress local invasion and distant colonization of hepatocellular carcinoma (Tao, Wan et al. 2013) but has not been linked to lymphoid malignancies yet.

RNA editing as a posttranscriptional modification is the site-specific alteration of an RNA transcript. The most frequently observed form is adenosine to inosine (A-to-I) editing, catalyzed by ADAR enzymes. Both the splicing and the translation machinery recognize inosines as guanosines. RNA editing occurs in a tissue-specific manner and increases the diversity of protein products in the case of mRNA editing. The specific deamination of miRNAs affects the stability of their precursors and thus the processing efficacy as well as results in the generation of novel mRNA targets sites in addition to altering existing ones (Blow, Grocock et al. 2006). Although not yet in lymphoma, the hsa-mir-376 family was previously shown to be subject to miRNA editing in different cancer types. This resulted in an altered mRNA target profile with both the loss of regulation of previous targets as well as the gain of new targets (Mizuguchi, Mishima et al. 2011, Choudhury, Tay et al. 2012). Both aspects promoted the respective cancers.

We provide here evidence of miRNA editing (hsa-miR-1260b, hsa-miR-376a1/2, and hsa-miR-376c) in lymphomas.

Only sequencing data allows the larger scale identification of novel miRNAs. The current release (21) of miRBase lists 1 881 human miRNAs. Similar to previous studies (Lim, Trinh et al. 2015), we identified hundreds of putative novel miRNA candidates. By Northern blot experiments, we provide experimental evidence of the correct processing of three novel miRNAs. Novel-miR-1 emerged as the most interesting candidate, only being detectable in SU-DHL-4, Namalwa and a B-lymphoblastoid cell line. Our analysis showed that it regulates many well-known lymphoma genes including CARD11, E2F1, MCM2 and MCM7, thus presenting itself as a potential novel player in lymphomagenesis.

Through our integrative analysis of miRNA and mRNA patient profiles in combination with AGO2 PAR-CLIP data, it is for the first time possible to pinpoint individual, biochemically defined miRNA/mRNA target interactions in lymphomas as well as functional consequences of miRNA dysregulation. We focused our analysis on those target pairs (208 in BL, 328 in DLBCL/FL) with consistent expression changes (presumably due to aberrant miRNA expression) in the respective patient RNASeq data. Just performing a correlation analysis between differentially expressed miRNAs and mRNAs in patient samples coupled with a miRanda target prediction would have resulted in a much greater number of predicted interaction pairs (2151 predicted pairs, data not shown). We described associated regulatory pathways including “Ras signaling”, “PI3K-Akt signaling”, and “MAPK signaling”. As there was very little overlap between those mRNAs that are targeted by miRNAs and those genes for which genomic mutations were detected (in those pathways), we suggest miRNA-mRNA targeting with subsequent deregulation as an additional oncogenic mechanism. We also provide evidence of miRNA regulation of many genes with already established roles in lymphomagenesis including ID3, CDK6, MDM2, SMARCA4, and TCF3.

Our miRNA expression profiles uncovered subtype-specific differences in miRNA expression, evidence of recurrent hsa-miR-142 mutations in FL and DLBCL as well as

miRNA editing and revealed distinct miRNA/mRNA target interaction pairs with roles in lymphomagenesis. Thus, we confirm and extend the important role that miRNAs play in lymphomagenesis.

References

Akao, Y., Nakagawa, Y., Kitade, Y., Kinoshita, T. and Naoe, T. (2007). Downregulation of microRNAs-143 and -145 in B-cell malignancies. Cancer Sci 98(12): 1914-1920.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769): 503-511.

Blow, M. J., Grocock, R. J., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., Wooster, R. and Stratton, M. R. (2006). RNA editing of human microRNAs. Genome Biol 7(4): R27.

Campo, E., Swerdlow, S. H., Harris, N. L., Pileri, S., Stein, H. and Jaffe, E. S. (2011). The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. Blood 117(19): 5019-5032.

Chang, T. C., Yu, D., Lee, Y. S., Wentzel, E. A., Arking, D. E., West, K. M., Dang, C. V., Thomas-Tikhonenko, A. and Mendell, J. T. (2008). Widespread microRNA repression by Myc contributes to tumorigenesis. Nat Genet 40(1): 43-50.

Chen, D., Law, M. E., Theis, J. D., Gamez, J. D., Caron, L. B., Vrana, J. A. and Dogan, A. (2009). Clinicopathologic features of CDK6 translocation-associated B-cell lymphoproliferative disorders. The American journal of surgical pathology 33(5): 720-729.

Choudhury, Y., Tay, F. C., Lam, D. H., Sandanaraj, E., Tang, C., Ang, B. T. and Wang, S. (2012). Attenuated adenosine-to-inosine editing of microRNA-376a* promotes invasiveness of glioblastoma cells. J Clin Invest 122(11): 4059-4076.

Creighton, C. J., Reid, J. G. and Gunaratne, P. H. (2009). Expression profiling of microRNAs by deep sequencing. Brief Bioinform 10(5): 490-497.

Denman, R. B. (1993). Using RNAFOLD to predict the activity of small catalytic RNAs. Biotechniques 15(6): 1090-1095.

Di Lisio, L., Martinez, N., Montes-Moreno, S., Piris-Villaespesa, M., Sanchez-Beato, M. and Piris, M. A. (2012). The role of miRNAs in the pathogenesis and diagnosis of B-cell lymphomas. Blood 120(9): 1782-1790.

Di Lisio, L., Sanchez-Beato, M., Gomez-Lopez, G., Rodriguez, M. E., Montes-Moreno, S., Mollejo, M., Menarguez, J., Martinez, M. A., Alves, F. J., Pisano, D. G., Piris, M. A. and Martinez, N. (2012). MicroRNA signatures in B-cell lymphomas. Blood Cancer J 2(2): e57.

Dou, L., Zheng, D., Li, J., Li, Y., Gao, L., Wang, L. and Yu, L. (2012). Methylation-mediated repression of microRNA-143 enhances MLL-AF4 oncogene expression. Oncogene 31(4): 507-517.

Durig, J., Schmucker, U. and Duhrsen, U. (2001). Differential expression of chemokine receptors in B cell malignancies. Leukemia 15(5): 752-756.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48.

Enomoto, Y., Kitaura, J., Hatakeyama, K., Watanuki, J., Akasaka, T., Kato, N., Shimanuki, M., Nishimura, K., Takahashi, M., Taniwaki, M., Haferlach, C., Siebert, R., Dyer, M. J., Asou, N., Aburatani, H., Nakakuma, H., Kitamura, T. and Sonoki, T. (2011). Emu/miR-125b transgenic mice develop lethal B-cell malignancies. Leukemia 25(12): 1849-1856.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S. (2004). MicroRNA targets in Drosophila. Genome Biol 5(1): R1.

Farazi, T. A., Spitzer, J. I., Morozov, P. and Tuschl, T. (2011). miRNAs in human cancer. J Pathol 223(2): 102-115.

Farazi, T. A., Ten Hoeve, J. J., Brown, M., Mihailovic, A., Horlings, H. M., van de Vijver, M. J., Tuschl, T. and Wessels, L. F. (2014). Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets. Genome Biol 15(1): R9.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J. M. and Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 37(Web Server issue): W68-76.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141(1): 129-141.

Hagberg, H., Killander, A. and Simonsson, B. (1983). Serum beta 2-microglobulin in malignant lymphoma. Cancer 51(12): 2220-2225.

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J. and Hammond, S. M. (2005). A microRNA polycistron as a potential human oncogene. Nature 435(7043): 828-833.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F. and Hackermuller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5(9): e1000502.

Iqbal, J., Shen, Y., Huang, X., Liu, Y., Wake, L., Liu, C., Deffenbacher, K., Lachel, C. M., Wang, C., Rohr, J., Guo, S., Smith, L. M., Wright, G., Bhagavathi, S., Dybkaer, K., Fu, K., Greiner, T. C., Vose, J. M., Jaffe, E., Rimsza, L., Rosenwald, A., Ott, G., Delabie, J., Campo, E., Braziel, R. M., Cook, J. R., Tubbs, R. R., Armitage, J. O., Weisenburger, D. D., Staudt, L. M., Gascoyne, R. D., McKeithan, T. W. and Chan, W. C. (2015). Global

microRNA expression profiling uncovers molecular markers for classification and prognosis in aggressive B-cell lymphoma. Blood 125(7): 1137-1145.

Ishiura, Y., Kotani, N., Yamashita, R., Yamamoto, H., Kozutsumi, Y. and Honke, K. (2010). Anomalous expression of Thy1 (CD90) in B-cell lymphoma cells and proliferation inhibition by anti-Thy1 antibody treatment. Biochemical and biophysical research communications 396(2): 329-334.

Kent, O. A., McCall, M. N., Cornish, T. C. and Halushka, M. K. (2014). Lessons from miR-143/145: the importance of cell-type localization of miRNAs. Nucleic Acids Res 42(12): 7528-7538.

Kuppers, R. (2011). NPAT mutations in Hodgkin lymphoma. Blood 118(3): 484-485.

Kwanhian, W., Lenze, D., Alles, J., Motsch, N., Barth, S., Doll, C., Imig, J., Hummel, M., Tinguely, M., Trivedi, P., Lulitanond, V., Meister, G., Renner, C. and Grasser, F. A. (2012). MicroRNA-142 is mutated in about 20% of diffuse large B-cell lymphoma. Cancer Med 1(2): 141-155.

Lawrie, C. H., Chi, J., Taylor, S., Tramonti, D., Ballabio, E., Palazzo, S., Saunders, N. J., Pezzella, F., Boultwood, J., Wainscoat, J. S. and Hatton, C. S. (2009). Expression of microRNAs in diffuse large B cell lymphoma is associated with immunophenotype, survival and transformation from follicular lymphoma. J Cell Mol Med 13(7): 1248-1260.

Lenz, G. and Staudt, L. M. (2010). Aggressive lymphomas. N Engl J Med 362(15): 1417-1429.

Lenze, D., Leoncini, L., Hummel, M., Volinia, S., Liu, C. G., Amato, T., De Falco, G., Githanga, J., Horn, H., Nyagol, J., Ott, G., Palatini, J., Pfreundschuh, M., Rogena, E., Rosenwald, A., Siebert, R., Croce, C. M. and Stein, H. (2011). The different epidemiologic subtypes of Burkitt lymphoma share a homogenous micro RNA profile distinct from diffuse large B-cell lymphoma. Leukemia 25(12): 1869-1876.

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

Lim, E. L., Trinh, D. L., Scott, D. W., Chu, A., Krzywinski, M., Zhao, Y., Robertson, A. G., Mungall, A. J., Schein, J., Boyle, M., Mottok, A., Ennishi, D., Johnson, N. A., Steidl, C.,

Connors, J. M., Morin, R. D., Gascoyne, R. D. and Marra, M. A. (2015). Comprehensive miRNA sequence analysis reveals survival differences in diffuse large B-cell lymphoma patients. Genome Biol 16: 18.

Liu, C., Iqbal, J., Teruya-Feldstein, J., Shen, Y., Dabrowska, M. J., Dybkaer, K., Lim, M. S., Piva, R., Barreca, A., Pellegrino, E., Spaccarotella, E., Lachel, C. M., Kucuk, C., Jiang, C. S., Hu, X., Bhagavathi, S., Greiner, T. C., Weisenburger, D. D., Aoun, P., Perkins, S. L., McKeithan, T. W., Inghirami, G. and Chan, W. C. (2013). MicroRNA expression profiling identifies molecular signatures associated with anaplastic large cell lymphoma. Blood 122(12): 2083-2092.

Liu, Z., Mai, C., Yang, H., Zhen, Y., Yu, X., Hua, S., Wu, Q., Jiang, Q., Zhang, Y., Song, X. and Fang, W. (2014). Candidate tumour suppressor CCDC19 regulates miR-184 direct targeting of C-Myc thereby suppressing cell growth in non-small cell lung cancers. J Cell Mol Med 18(8): 1667-1679.

Love, C., Sun, Z., Jima, D., Li, G., Zhang, J., Miles, R., Richards, K. L., Dunphy, C. H., Choi, W. W., Srivastava, G., Lugar, P. L., Rizzieri, D. A., Lagoo, A. S., Bernal-Mizrachi, L., Mann, K. P., Flowers, C. R., Naresh, K. N., Evens, A. M., Chadburn, A., Gordon, L. I., Czader, M. B., Gill, J. I., Hsi, E. D., Greenough, A., Moffitt, A. B., McKinney, M., Banerjee, A., Grubor, V., Levy, S., Dunson, D. B. and Dave, S. S. (2012). The genetic landscape of mutations in Burkitt lymphoma. Nat Genet 44(12): 1321-1325.

Mizuguchi, Y., Mishima, T., Yokomuro, S., Arima, Y., Kawahigashi, Y., Shigehara, K., Kanda, T., Yoshida, H., Uchida, E., Tajiri, T. and Takizawa, T. (2011). Sequencing and bioinformatics-based analyses of the microRNA transcriptome in hepatitis B-related hepatocellular carcinoma. PLoS One 6(1): e15304.

Musilova, K. and Mraz, M. (2015). MicroRNAs in B-cell lymphomas: how a complex biology gets more complex. Leukemia 29(5): 1004-1017.

Richter, J., Schlesner, M., Hoffmann, S., Kreuz, M., Leich, E., Burkhardt, B., Rosolowski, M., Ammerpohl, O., Wagener, R., Bernhart, S. H., Lenze, D., Szczepanowski, M., Paulsen, M., Lipinski, S., Russell, R. B., Adam-Klages, S., Apic, G., Claviez, A., Hasenclever, D., Hovestadt, V., Hornig, N., Korbel, J. O., Kube, D., Langenberger, D., Lawerenz, C., Lisfeld, J., Meyer, K., Picelli, S., Pischimarov, J., Radlwimmer, B.,

Rausch, T., Rohde, M., Schilhabel, M., Scholtysik, R., Spang, R., Trautmann, H., Zenz, T., Borkhardt, A., Drexler, H. G., Moller, P., MacLeod, R. A., Pott, C., Schreiber, S., Trumper, L., Loeffler, M., Stadler, P. F., Lichter, P., Eils, R., Kuppers, R., Hummel, M., Klapper, W., Rosenstiel, P., Rosenwald, A., Brors, B., Siebert, R. and Project, I. M.-S. (2012). Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat Genet 44(12): 1316-1320.

Roehle, A., Hoefig, K. P., Repsilber, D., Thorns, C., Ziepert, M., Wesche, K. O., Thiere, M., Loeffler, M., Klapper, W., Pfreundschuh, M., Matolcsy, A., Bernd, H. W., Reiniger, L., Merz, H. and Feller, A. C. (2008). MicroRNA signatures characterize diffuse large B-cell lymphomas and follicular lymphomas. Br J Haematol 142(5): 732-744.

Schmitz, R., Young, R. M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Shaffer, A. L., Hodson, D. J., Buras, E., Liu, X., Powell, J., Yang, Y., Xu, W., Zhao, H., Kohlhammer, H., Rosenwald, A., Kluin, P., Muller-Hermelink, H. K., Ott, G., Gascoyne, R. D., Connors, J. M., Rimsza, L. M., Campo, E., Jaffe, E. S., Delabie, J., Smeland, E. B., Ogwang, M. D., Reynolds, S. J., Fisher, R. I., Braziel, R. M., Tubbs, R. R., Cook, J. R., Weisenburger, D. D., Chan, W. C., Pittaluga, S., Wilson, W., Waldmann, T. A., Rowe, M., Mbulaiteye, S. M., Rickinson, A. B. and Staudt, L. M. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature 490(7418): 116-120.

Schou, J. V., Rossi, S., Jensen, B. V., Nielsen, D. L., Pfeiffer, P., Hogdall, E., Yilmaz, M., Tejpar, S., Delorenzi, M., Kruhoffer, M. and Johansen, J. S. (2014). miR-345 in Metastatic Colorectal Cancer: A Non-Invasive Biomarker for Clinical Outcome in Non-KRAS Mutant Patients Treated with 3rd Line Cetuximab and Irinotecan. PLoS One 9(6): e99886.

Sinha, R., Nastoupil, L. and Flowers, C. R. (2012). Treatment Strategies for Patients with Diffuse Large B-Cell Lymphoma: Past, Present, and Future. Blood Lymphat Cancer 2012(2): 87-98.

Solenthaler, M., Matutes, E., Brito-Babapulle, V., Morilla, R. and Catovsky, D. (2002). p53 and mdm2 in mantle cell lymphoma in leukemic phase. Haematologica 87(11): 1141-1150.

Spitzer, J., Hafner, M., Landthaler, M., Ascano, M., Farazi, T., Wardle, G., Nusbaum, J., Khorshid, M., Burger, L., Zavolan, M. and Tuschl, T. (2014). PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. Methods Enzymol 539: 113-161.

Tao, J. and Zhao, X. (2014). c-MYC-miRNA circuitry: a central regulator of aggressive B-cell malignancies. Cell Cycle 13(2): 191-198.

Tao, Z. H., Wan, J. L., Zeng, L. Y., Xie, L., Sun, H. C., Qin, L. X., Wang, L., Zhou, J., Ren, Z. G., Li, Y. X., Fan, J. and Wu, W. Z. (2013). miR-612 suppresses the invasive-metastatic cascade in hepatocellular carcinoma. J Exp Med 210(4): 789-803.

Tomaselli, S., Locatelli, F. and Gallo, A. (2014). The RNA editing enzymes ADARs: mechanism of action and human disease. Cell Tissue Res 356(3): 527-532.

Xiong, L., Jiang, W., Zhou, R., Mao, C. and Guo, Z. (2013). Identification and analysis of the regulatory network of Myc and microRNAs from high-throughput experimental data. Comput Biol Med 43(9): 1252-1260.

Zhao, Z. N., Bai, J. X., Zhou, Q., Yan, B., Qin, W. W., Jia, L. T., Meng, Y. L., Jin, B. Q., Yao, L. B., Wang, T. and Yang, A. G. (2012). TSA suppresses miR-106b-93-25 cluster expression through downregulation of MYC and inhibits proliferation and induces apoptosis in human EMC. PLoS One 7(9): e45133.

Supplementary Materials & Methods

Supplementary Methods

Genome and transcriptome analysis

RNA-seq reads (average of 11,320,300 / sample) were aligned with TopHat (Trapnell, Pachter et al. 2009) (version 2.0.12). Differential gene expression was measured with

edgeR (Robinson, McCarthy et al. 2010) (version 3.8.6) using inter-sample normalized counts per millions (CPM) and applying multiple testing corrections using FDR. CPMs are normalizing the actual read counts mapping to a respective gene by the entire number of sequenced reads per sample times 1,000,000 (Robinson, McCarthy et al. 2010). Thereby, CPMs are representing a measure of expression counts that are comparable across multiple samples.

Read pairs obtained by whole genome paired-end sequencing were mapped to the human reference genome (hg19) using BWA (Li and Durbin 2009) version 0.5.9-r16 (maximum insert size of 1 kb). SAMtools (Li, Handsaker et al. 2009) was used to generate a coordinate-sorted BAM file, and Picard (version 1.48) was used to merge BAM files from one sample and remove PCR duplicates. Detection of somatic SNVs from tumor and matched control whole genome sequencing data was performed as described previously (Jones, Jager et al. 2012). SNVs were functionally annotated using Annovar (Wang, Li et al. 2010) and annotated for overlaps with SNPs (dbSNP build 135 and 1000 Genome project data) using BEDTools (Quinlan and Hall 2010). Allele-specific copy-number alterations were detected as described in Richter et al (Richter, Schlesner et al. 2012).

MiRNA sequencing data processing and novel miRNA prediction

Adapter sequences were removed from raw reads using fastx_clipper (http://hannonlab.cshl.edu/fastx_toolkit/index.html). A total of 1,169,752,727 clipped reads were mapped onto the human genome (1000 genomes project, hs37d5100) using segemehl(Hoffmann, Otto et al. 2009), with a minimum accuracy of 90% (average of 20,888,442 reads per sample, see Supplementary Table 2). Quantification of annotated microRNAs from miRBase version 19 was performed using ngsutils (Breese and Liu 2013), partially counting multimapped reads. Read counts were inter-sample normalized leading to CPM values per annotated miRNA. Differential expression was achieved by applying edgeR and subsequent calculation of the FDR for multiple testing correction.

Novel miRNA prediction was performed using miRanalyzer (Hackenberg, Sturm et al. 2009) (version 0.3) using default parameters.

MiRNA target prediction

We used miRanda (Enright, John et al. 2004) (version August 2010) to predict potential interaction targets of miRNAs in the transcriptome. For narrowing down the list of miRNA-mRNA target correlations, we filtered for a miRsvr-score < -1.2 per correlation pair. We selected this score as it results in the 5% most significant miRNA-mRNA correlations (Betel, Koppal et al. 2010).

Differential expression analysis on RNA-Seq data – novel miRNA overexpression/knockdown

In all sequencing reads, adapters and low quality ends were trimmed using seqtk (https://github.com/lh3/seqtk) and cutadapt (Martin 2011). All reads shorter than 25 bases after trimming were discarded, leading to 11,320,300 reads on average per sample. Next, the preprocessed reads were aligned against the human genome sequence (hg19) with TopHat (Trapnell, Pachter et al. 2009), which is capable of aligning RNA-Seq data because of the identification of splice junctions spanned by individual reads. On average, more than 80% of the preprocessed reads have been aligned against the reference sequence and were used for further analysis. To measure the transcript abundances representing an estimate of the gene expression levels in the samples, HTSeq (Anders, Pyl et al. 2014) (version 0.5.4) was employed using gene annotations downloaded from Ensembl Genes 75(Flicek, Amode et al. 2013). Differential gene expression was measured with edgeR (Robinson, McCarthy et al. 2010) using inter-sample normalized counts per millions (CPM) and applying multiple testing corrections using FDR.

PAR-CLIP analysis

The preprocessing of PAR-CLIP reads was similar to that of RNA-Seq reads, however, because short reads align multiple times to the genome, we selected a cutoff of 17 bases

per sequencing read after quality and adapter trimming using cutadapt. On average, this resulted in about 15,570,345 reads per sequencing sample. Next, reads were aligned with BWA (Li and Durbin 2009) allowing for up to two mismatches between a single read sequence and the reference sequence of hg19. All reads that failed in this mapping (mapping quality <10) were aligned against the transcriptome database Ensembl Genes 75 using BWA allowing for up to two mismatches between both sequences (Kloetgen, Münch et al. 2014). Next, the aligned reads were piled up into clusters by the PARA-suite (Kloetgen, Borkhardt et al., submitted) which applies a hierarchical clustering algorithm, where reads overlapping by at least 5 bases in their genomic mapping positions are stacked into a single cluster. To identify high confidence RNA-binding protein-bound regions, clusters having <5 reads and <25% T-C conversion frequency were excluded (for further details on analyzing PAR-CLIP datasets see (Hafner, Landthaler et al. 2010)).

To equalize cluster lengths for subsequent mRNA-miRNA correlation analysis, crosslink-centered regions (CCRs) consisting of 20 nucleotides up- and downstream of the major T-C conversion site within a cluster (i.e. the T-C conversion site with the highest conversion frequency per cluster), were generated to calculate all possible 7-mers within the CCRs (Farazi, Ten Hoeve et al. 2014) (Figure 2C). The most significantly enriched (compared to random sequences of the same dinucleotide compositions) corresponded to the reverse complement of the miRNA seed region. The miRNA-mRNA correlation was achieved by matching miRNA seed positions 2-8 as well as miRNA seed positions 1-7 to the reverse complement of the enriched CCR 7-mers for each cluster. As we also had information on the expression levels of miRNAs and mRNAs, we kept only miRNA-mRNA pairs, which showed a differential expression between BL versus non-BL patient samples (differential expression FDR ≤0.05). All miRNAs meeting this cutoff (FDR ≤0.05) showed log2 fold expression changes of at least ± 0.6.

Enrichment analysis

We have performed enrichment analyses by applying a one-sided Fisher's exact test to the number of overlapping genes/miRNAs between two lists, thereby calculating a significance value for the respective enrichment between the two lists. On the one hand, this was applied to identify enrichments between our reported miRNA classifier and recently published miRNA classifiers for the respective lymphoma subtypes. Therefore, the total number of tested miRNAs was taken as a basis (667 in [28] and 602 in [10]). On the other hand, enriched genes of the negatively correlated miRNAs-mRNAs pairs were identified within KEGG pathways. The gene-basis was set to 22,525 genes, which were accessible for differential gene expression analysis based on RNA-Seq data after filtering lowly expressed genes (logCPM < 1 in all patients of all subgroups). Enrichment scores were calculated as follows: Enrichment = #overlap / (#DE genes * #pathway genes / #population genes).

Expressional correlation analysis

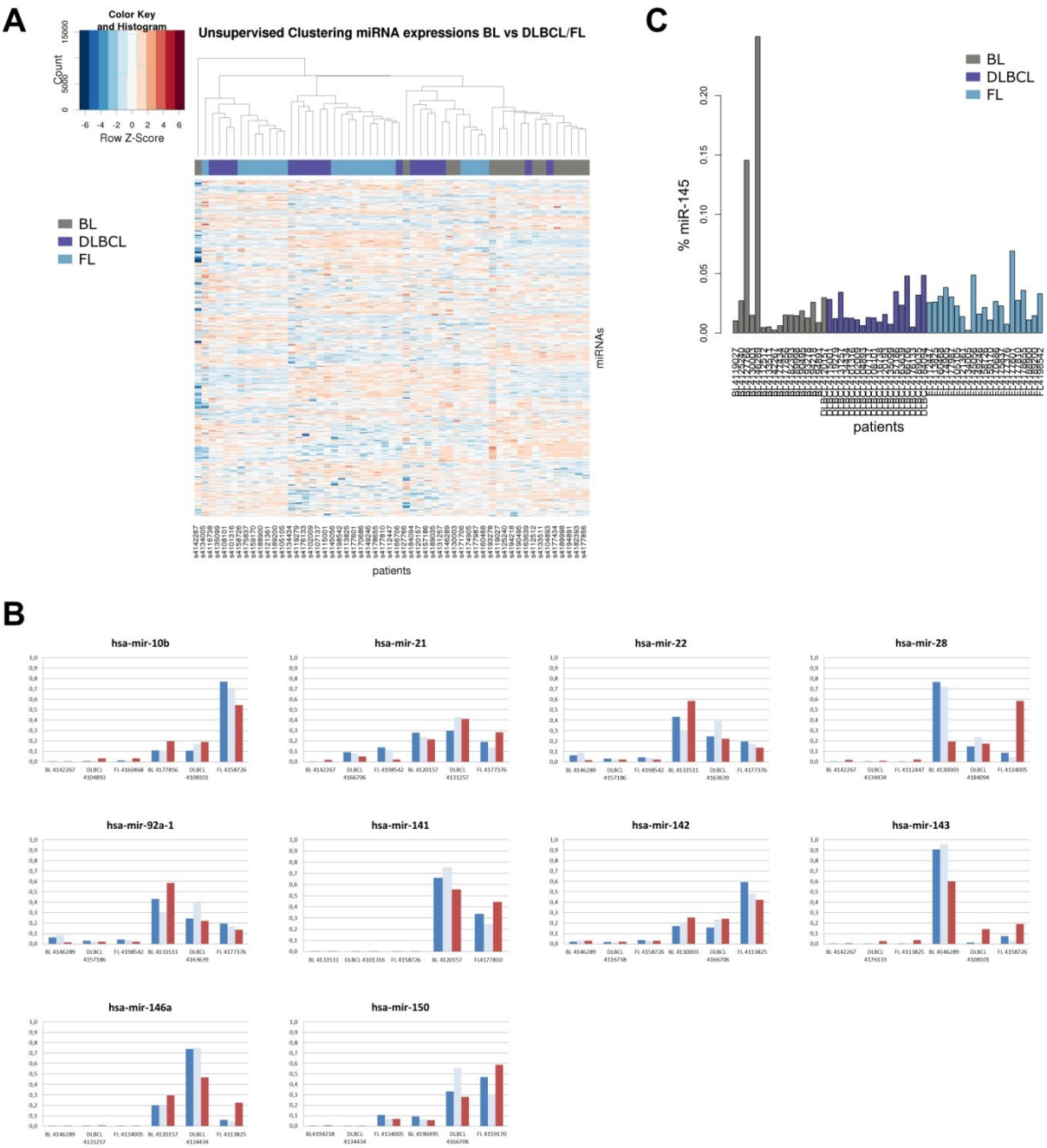
We have used the spearman's rank correlation to test for significant similarity between two expression patterns. This was applied to check for similar expressions of miRNAs between different platforms, i.e. between NGS derived expression values and qPCR derived expression values. We also used the spearman's rank correlation to check whether the expressional pattern of hsa-miR-143 and hsa-miR-145 were similar across all patient samples.

Leave-one-out cross-validation

To test the validity of our reported miRNA-classifiers, we have performed leave-one-out cross-validation (LOO-CV) on an independent dataset downloaded from NCBI GEO (Accession GSE22420) (Lenze, Leoncini et al. 2011). Because these expression values were obtained from microarray experiments, we used LOO-CV to train our classifier on microarray expression values rather than on cpm values obtained from our RNA-Seq data. Additionally, we had to exclude miRNAs from the classifier which were not covered by the microarray (7 miRNAs each for the BL vs. DLBCL classifier and the BL vs.

DLBCL/FL classifier). The class prediction for the left-out sample was performed using the K-nearest-neighbours algorithm with k=33 (half the size of the smaller BL group containing 64 samples). All cases showing less than 80% majority vote during the class prediction were excluded as not classified. Recall was calculated as the number of remaining predictions per classified cases and overall accuracy was calculated as the number of correct assignments after majority vote exclusion per all tested cases.

Supplementary Figure 1



Supplementary Figure 1. Validation of NGS data. A) For unsupervised clustering, miRNAs were discarded if less than 16 patients showed a base-line expression of > 0 log2 cpm after normalization. In total, 573 mature miRNAs were used for unsupervised clustering. FL (light blue), DLBCL (dark blue), and BL (grey). B) Validation of NGS

miRNA expression by qRT-PCR. Light blue = qRT-PCR expression (normalized to RNU48 as housekeeping gene), dark blue = qRT-PCR expression (normalized to RNU24 as housekeeping gene), red: expression according to NGS analysis. To allow for comparison across platforms, the expression levels were set to add up to 100 per experiment and are shown as % total expression. Also see Supplementary Table 3 for statistical analysis. C) Hsa-miR-145 expression across all patient samples. Color code as in A).

Supplementary Tables

Supplementary Table 1: Patient information.

PID	Diagnosis	Classification 2	Age at first diagnosis	Gender	MYC_STATUS	BCL2_BREAK	BCL6_BREAK	IGH_status
4112512	BL	NA	18	female	IG-MYC pos	negative	negative	IGH-MYC pos
4119027	BL	NA	12	male	IG-MYC pos	negative	negative	IGH-MYC pos
4125240	BL	NA	4	male	IG-MYC pos	negative	negative	IGH-MYC pos
4127766	BL	NA	8	male	IG-MYC pos	negative	negative	IGH-MYC pos
4130003	BL	NA	6	male	IG-MYC pos	negative	negative	IGH neg
4133511	BL	NA	5	male	IG-MYC pos	negative	negative	IGH-MYC pos
4142267	BL	NA	5	male	IG-MYC pos	negative	negative	IGH neg
4146289	BL	NA	14	male	IG-MYC pos	negative	negative	IGH-MYC pos
4177434	BL	NA	16	female	IG-MYC pos	negative	negative	IGH-MYC pos
4177856	BL	NA	10	male	IG-MYC pos	negative	negative	IGH-MYC pos
4182393	BL	NA	10	male	IG-MYC pos	negative	negative	IGH-MYC pos
4189998	BL	NA	13	male	IG-MYC pos	negative	negative	IGH-MYC pos
4190495	BL	NA	15	male	IG-MYC pos	negative	negative	IGH neg
4193278	BL	NA	17	male	IG-MYC pos	negative	negative	IGH neg
4194218	BL	NA	4	male	IG-MYC pos	negative	negative	IGH-MYC pos
4194891	BL	NA	4	male	IG-MYC pos	negative	negative	IGH-MYC pos
4101316	DLBCL	ABC	74	female	MYC neg	negative	negative	IGH neg
4102009	DLBCL	ABC	64	male	MYC neg	negative	negative	IGH neg
4104893	DLBCL	TypeIII	16	male	MYC neg	negative	negative	IGH-IRF4 pos
4107137	DLBCL	GCB	59	male	MYC neg	negative	positive	IGH pos
4108101	DLBCL	ABC	66	male	MYC neg	negative	negative	IGH neg
4115001	DLBCL	GCB	70	female	MYC neg	positive	negative	IGH-BCL2 pos
4116738	DLBCL	TypeIII	15	male	MYC neg	negative	positive	IGH neg
4119279	DLBCL	ABC	62	female	MYC neg	negative	negative	IGH neg
4120157	DLBCL	GCB	46	male	MYC neg	negative	negative	IGH pos
4120193	DLBCL	ABC	41	female	MYC neg	negative	negative	IGH neg
4131257	DLBCL	ABC	72	male	MYC neg	negative	positive	IGH pos
4134434	DLBCL	GCB	84	male	MYC neg	positive	positive	IGH pos
4135099	DLBCL	ABC	49	male	MYC neg	negative	positive	IGH pos
4157186	DLBCL	ABC	74	male	MYC neg	negative	negative	IGH neg
4163639	DLBCL	GCB	75	female	MYC neg	negative	negative	IGH neg
4166706	DLBCL	GCB	62	male	MYC neg	negative	positive	IGH pos
4176133	DLBCL	ABC	61	female	MYC neg	negative	positive	IGH pos
4184094	DLBCL	GCB	57	female	MYC neg	positive	negative	IGH-BCL2 pos
4189035	DLBCL	ABC	46	male	MYC neg	negative	negative	IGH neg
4105105	FL	grade 1/2	40	female	MYC neg	positive	negative	IGH-BCL2 pos
4112447	FL	grade 1/2	46	male	MYC neg	positive	positive	IGH-BCL2 pos
4113825	FL	grade 1/2	74	female	MYC neg	negative	negative	IGH neg
4121361	FL	grade 1/2	74	male	MYC neg	positive	negative	IGH-BCL2 pos
4134005	FL	grade 1/2	67	male	MYC neg	positive	negative	IGH-BCL2 pos
4145056	FL	grade 1/2	67	female	MYC neg	positive	negative	IGH-BCL2 pos
4149246	FL	grade 1/2	41	male	MYC neg	positive	positive	IGH-BCL2 pos
4158726	FL	grade 1/2	48	male	MYC neg	positive	negative	IGH-BCL2 pos
4159170	FL	grade 1/2	43	male	MYC neg	positive	negative	IGH-BCL2 pos
4160468	FL	grade 1/2	62	male	MYC neg	positive	negative	IGH-BCL2 pos
4170686	FL	grade 1/2	56	male	MYC neg	positive	negative	IGH-BCL2 pos
4174905	FL	grade 1/2	72	male	MYC neg	positive	negative	IGH-BCL2 pos
4175837	FL	grade 1/2	74	female	MYC neg	positive	positive	IGH-BCL2 pos
4177376	FL	grade 3A	73	female	MYC neg	positive	negative	IGH-BCL2 pos
4177601	FL	grade 2/3A	52	female	MYC neg	positive	negative	IGH-BCL2 pos
4177810	FL	grade 1/2	47	female	MYC neg	positive	negative	IGH-BCL2 pos
4177987	FL	grade 1/2	71	male	MYC neg	positive	negative	IGH-BCL2 pos
4178655	FL	grade 1/2	50	male	MYC neg	positive	negative	IGH-BCL2 pos
4188900	FL	grade 1/2	76	male	MYC neg	positive	negative	IGH neg
4189200	FL	grade 1/2	51	female	MYC neg	positive	negative	IGH-BCL2 pos
4198542	FL	grade 1/2/3A	68	female	MYC neg	negative	negative	IGH neg

Supplementary Table 2: miRNA sequencing library statistics.

Diagnosis	PID	raw reads	clipped reads	% clipped reads	mapped reads	% mapped reads	reads on primary miRs	% reads on primary miRs
BL	4112512	111,835,821	90,656,876	81.06%	83,739,794	92.37%	30,310,990	36.20%
BL	4119027	96,943,814	80,372,849	82.91%	74,224,275	92.35%	27,687,024	37.30%
BL	4125240	64,001,808	56,406,459	88.13%	53,503,949	94.85%	10,026,892	18.74%
BL	4127766	35,157,062	34,510,881	98.16%	33,693,807	97.63%	1,706,671	5.07%
BL	4130003	25,502,310	24,828,949	97.36%	23,122,148	93.13%	12,787,211	55.30%
BL	4133511	71,261,813	64,018,378	89.84%	59,083,671	92.29%	18,112,535	30.66%
BL	4142267	69,108,952	65,274,277	94.45%	56,674,054	86.82%	10,070,922	17.77%
BL	4146289	41,192,177	40,108,479	97.37%	37,541,386	93.60%	34,377,156	91.57%
BL	4177434	62,332,955	53,529,191	85.88%	46,581,276	87.02%	23,865,056	51.23%
BL	4177856	75,642,487	68,181,918	90.14%	61,370,463	90.01%	24,861,850	40.51%
BL	4182393	84,780,972	72,038,555	84.97%	66,059,591	91.70%	21,585,793	32.68%
BL	4189998	102,523,454	91,607,098	89.35%	83,325,274	90.96%	22,585,978	27.11%
BL	4190495	156,613,394	137,235,359	87.63%	130,397,032	95.02%	34,428,505	26.40%
BL	4193278	72,506,909	65,405,761	90.21%	57,367,702	87.71%	18,460,967	32.18%
BL	4194218	76,405,749	73,447,437	96.13%	67,907,053	92.46%	5,543,923	8.16%
BL	4194891	90,762,811	68,540,713	75.52%	63,179,485	92.18%	24,980,361	39.54%
DLBCL	4101316	67,519,903	64,199,209	95.08%	61,007,246	95.03%	28,775,304	47.17%
DLBCL	4102009	42,086,479	39,944,756	94.91%	36,818,259	92.17%	11,789,929	32.02%
DLBCL	4104893	95,311,645	85,124,518	89.31%	76,526,308	89.90%	21,888,192	28.60%
DLBCL	4107137	40,477,513	34,734,623	85.81%	31,436,185	90.50%	19,242,992	61.21%
DLBCL	4108101	57,832,471	55,071,105	95.23%	51,979,749	94.39%	27,543,402	52.99%
DLBCL	4115001	27,198,492	26,148,020	96.14%	25,497,799	97.51%	7,386,128	28.97%
DLBCL	4116738	78,232,596	68,884,140	88.05%	60,645,188	88.04%	31,000,823	51.12%
DLBCL	4119279	39,034,533	34,535,595	88.47%	32,847,597	95.11%	8,154,515	24.83%
DLBCL	4120157	45,688,615	41,860,450	91.62%	38,598,518	92.21%	8,676,661	22.48%
DLBCL	4120193	40,566,871	39,713,411	97.90%	37,901,161	95.44%	5,340,938	14.09%
DLBCL	4131257	31,446,420	30,697,771	97.62%	29,248,137	95.28%	10,648,290	36.41%
DLBCL	4134434	34,646,386	28,750,213	82.98%	25,728,439	89.49%	9,642,469	37.48%
DLBCL	4135099	82,046,323	75,673,959	92.23%	70,326,721	92.93%	42,315,069	60.17%
DLBCL	4157186	20,558,444	18,165,585	88.36%	17,261,128	95.02%	7,126,511	41.29%
DLBCL	4163639	82,468,882	76,639,372	92.93%	71,664,118	93.51%	22,990,794	32.08%
DLBCL	4166706	49,112,501	46,502,593	94.69%	43,144,375	92.78%	26,491,474	61.40%
DLBCL	4176133	27,241,860	26,650,486	97.83%	25,782,166	96.74%	4,973,146	19.29%
DLBCL	4184094	37,960,574	35,916,452	94.62%	33,999,067	94.66%	4,891,289	14.39%
DLBCL	4189035	39,456,122	33,743,677	85.52%	30,162,878	89.39%	18,823,323	62.41%
FL	4105105	66,137,979	64,339,942	97.28%	62,248,882	96.75%	33,016,425	53.04%
FL	4112447	45,773,088	42,548,261	92.95%	40,644,074	95.52%	10,268,194	25.26%
FL	4113825	41,550,848	31,506,429	75.83%	26,934,801	85.49%	15,915,215	59.09%
FL	4121361	73,759,960	67,969,941	92.15%	63,948,878	94.08%	49,654,552	77.65%
FL	4134005	67,942,851	61,098,231	89.93%	58,991,880	96.55%	12,599,478	21.36%
FL	4145056	26,282,929	23,842,066	90.71%	20,609,766	86.44%	9,084,189	44.08%
FL	4149246	20,825,007	15,932,481	76.51%	14,785,326	92.80%	12,274,678	83.02%
FL	4158726	80,798,657	77,575,725	96.01%	72,581,907	93.56%	56,127,947	77.33%
FL	4159170	90,316,042	87,915,910	97.34%	84,580,056	96.21%	48,098,750	56.87%
FL	4160468	66,934,137	64,234,781	95.97%	57,676,380	89.79%	45,480,635	78.85%
FL	4170686	28,225,201	28,047,528	99.37%	27,224,450	97.07%	6,080,646	22.34%
FL	4174905	40,688,848	39,370,043	96.76%	36,779,632	93.42%	17,879,064	48.61%
FL	4175837	145,487,505	142,602,072	98.02%	138,805,355	97.34%	47,162,234	33.98%
FL	4177376	59,033,653	57,041,478	96.63%	54,405,387	95.38%	29,629,182	54.46%
FL	4177601	39,845,682	39,113,638	98.16%	37,491,676	95.85%	5,775,674	15.41%
FL	4177810	23,339,328	22,495,874	96.39%	21,827,428	97.03%	9,958,437	45.62%
FL	4177987	38,144,137	34,176,587	89.60%	31,030,702	90.80%	21,971,916	70.81%
FL	4178655	33,166,391	29,713,962	89.59%	27,842,516	93.70%	21,914,993	78.71%
FL	4188900	87,727,261	78,502,242	89.48%	71,530,316	91.12%	40,556,165	56.70%
FL	4189200	62,264,380	29,431,654	47.27%	26,122,247	88.76%	23,250,761	89.01%
FL	4198542	41,391,420	38,304,878	92.54%	36,352,845	94.90%	13,960,509	38.40%

Supplementary Table 3: Validation of NGS expression via qRT-PCR of select miRNAs

miRNA	qRT-PCR RNU24 vs. NGS	qRT-PCR RNU48 vs. NGS
hsa-mir-10b	**0.003**	**0.017**
hsa-mir-21	**0.033**	**0.033**
hsa-mir-22	**0.017**	**0.017**
hsa-mir-28	0.058	0.058
hsa-mir-92a-1	0.103	0.136
hsa-mir-141	0.083	0.083
hsa-mir-142	**0.003**	**0.033**
hsa-mir-143	**0.003**	**0.003**
hsa-mir-146a	**0.003**	**0.003**
hsa-mir-150	**0.003**	**0.017**

P-values of the correlation analyses between qRT-PCRs with two housekeeping genes (RNU24 and RNu48) versus expression as determined by NGS are indicated. P-values ≤ 0.05 are highlighted.

Supplementary Table 4: 25 miRNA classifiers separating B-cell lymphoma subtypes

miRNA	logFC	logCPM	p-value	FDR	cpm BL	cpm DLBCL
hsa-miR-20a-3p	-2.50	4.07	1.92E-12	9.71E-10	29.57	5.22
hsa-miR-221-3p	2.89	11.04	8.01E-12	2.02E-09	478.24	3539.57
hsa-miR-146a-5p	4.22	16.55	1.81E-11	3.05E-09	9293.79	172652.84
hsa-miR-141-3p	8.14	12.63	9.70E-11	1.02E-08	42.20	11905.84
hsa-miR-155-5p	3.87	13.27	1.03E-10	1.02E-08	1197.20	17531.21
hsa-miR-25-5p	-2.44	5.51	1.21E-10	1.02E-08	79.98	14.70
hsa-miR-196b-5p	3.76	5.16	7.07E-10	4.46E-08	4.64	63.07
hsa-miR-3681-5p	5.32	5.31	6.54E-10	4.46E-08	1.83	73.42
hsa-miR-200c-3p	4.70	7.67	1.33E-09	7.45E-08	14.33	370.91
hsa-miR-24-3p	2.66	8.88	2.58E-09	1.30E-07	123.21	780.97
hsa-miR-196a-5p	4.03	2.91	2.91E-09	1.34E-07	0.81	13.30
hsa-miR-664-3p	2.79	5.99	4.82E-09	2.03E-07	15.34	106.32
hsa-miR-130b-3p	-1.69	8.82	7.99E-09	3.04E-07	713.64	220.80
hsa-miR-664-5p	2.94	4.31	8.43E-09	3.04E-07	4.35	33.58
hsa-miR-10b-3p	2.93	2.79	1.18E-08	3.92E-07	1.49	11.62
hsa-miR-29b-2-5p	1.93	2.24	1.24E-08	3.92E-07	1.82	7.14
hsa-miR-23a-5p	2.24	2.76	1.62E-08	4.82E-07	2.24	10.60
hsa-miR-944	3.95	6.01	3.24E-08	9.08E-07	7.40	114.75
hsa-miR-193a-5p	2.09	6.56	3.87E-08	1.03E-06	34.58	147.85
hsa-miR-4677-3p	2.70	4.79	4.96E-08	1.25E-06	7.08	45.92
hsa-miR-339-5p	-2.26	5.89	5.64E-08	1.35E-06	101.54	21.14
hsa-miR-589-3p	-2.05	3.46	5.86E-08	1.35E-06	18.22	4.31
hsa-miR-98	2.09	11.99	6.24E-08	1.37E-06	1492.12	6346.58
hsa-miR-222-3p	2.12	9.93	7.27E-08	1.51E-06	351.22	1525.88
hsa-miR-374a-5p	2.50	8.15	7.47E-08	1.51E-06	81.97	464.14

Separating BL and DLBCL/FL. Table is sorted by FDR. logFC indicates log fold change; logCPM, log counts per million, FDR, false discovery rate.

miRNA	logFC	logCPM	p-value	FDR	cpm BL	cpm FL
hsa-miR-150-3p	3.75	4.97	4.45E-26	1.04E-23	3.85	52.05
hsa-miR-150-5p	4.12	12.86	6.01E-26	1.04E-23	722.73	12575.15
hsa-miR-20a-3p	-3.15	3.94	3.55E-26	1.04E-23	30.87	3.41
hsa-miR-19a-5p	-3.13	3.70	2.30E-23	2.97E-21	25.92	2.97
hsa-miR-18a-3p	-3.49	5.48	3.79E-22	3.92E-20	92.00	8.13
hsa-miR-335-3p	-3.58	8.20	7.95E-22	6.85E-20	614.46	51.34
hsa-miR-378a-3p	-2.43	10.47	1.01E-21	7.49E-20	2641.93	489.06
hsa-miR-184	4.86	3.88	8.06E-21	5.21E-19	0.85	25.05
hsa-miR-130b-3p	-2.79	8.61	1.25E-19	7.16E-18	760.44	110.29
hsa-miR-18a-5p	-3.01	7.05	9.84E-19	4.62E-17	263.04	32.68
hsa-miR-25-5p	-3.09	5.38	9.76E-19	4.62E-17	83.37	9.80
hsa-miR-19b-1-5p	-2.77	1.39	2.07E-18	8.94E-17	4.84	0.73
hsa-miR-151b	3.67	7.05	4.25E-17	1.69E-15	17.20	220.12
hsa-miR-17-5p	-2.32	9.52	1.91E-16	7.04E-15	1348.29	269.90
hsa-miR-92a-1-5p	-2.59	4.43	4.09E-16	1.41E-14	40.81	6.82
hsa-miR-221-3p	2.60	10.96	6.85E-16	2.21E-14	516.96	3123.54
hsa-miR-17-3p	-3.53	10.08	1.38E-15	4.20E-14	2252.58	194.47
hsa-miR-4517	-3.19	1.51	4.84E-15	1.39E-13	5.68	0.61
hsa-miR-29c-5p	2.32	5.53	1.32E-14	3.60E-13	14.05	70.51
hsa-miR-664-5p	2.55	4.12	1.62E-14	4.19E-13	4.61	26.97
hsa-miR-3622a-5p	3.63	1.02	4.44E-14	1.09E-12	0.27	3.20
hsa-miR-4420	-3.99	2.35	2.22E-13	5.22E-12	10.97	0.66
hsa-miR-377-5p	-3.44	1.75	1.75E-12	3.93E-11	6.65	0.58
hsa-miR-29b-3p	2.57	9.29	9.31E-12	2.01E-10	164.06	976.85
hsa-miR-19a-3p	-2.73	10.01	1.07E-11	2.20E-10	1987.69	300.17

Separating BL and DLBCL. Table is sorted by FDR. logFC indicates log fold change; logCPM, log counts per million, FDR, false discovery rate.

miRNA	logFC	logCPM	p-value	FDR	cpm FL	cpm DLBCL
hsa-miR-200c-3p	-3.89	7.26	8.36E-11	4.16E-08	20.77	307.99
hsa-miR-424-3p	-2.12	3.98	2.60E-10	6.48E-08	6.16	26.89
hsa-miR-147b	-3.41	2.08	2.76E-09	4.58E-07	0.76	8.11
hsa-miR-3622a-5p	2.65	0.84	9.35E-09	7.76E-07	2.75	0.42
hsa-miR-4517	-2.34	0.64	7.56E-09	7.76E-07	0.52	2.64
hsa-miR-503	-1.81	1.81	8.73E-09	7.76E-07	1.60	5.53
hsa-miR-486-3p	2.64	5.25	1.34E-08	9.55E-07	62.03	9.95
hsa-miR-148a-5p	-2.64	9.30	3.26E-08	2.03E-06	184.70	1148.17
hsa-miR-378a-3p	-1.48	9.55	4.12E-08	2.28E-06	411.38	1146.34
hsa-miR-18a-3p	-1.80	3.84	5.79E-07	2.62E-05	6.62	23.09
hsa-miR-4449	-2.68	2.42	5.26E-07	2.62E-05	1.53	9.72
hsa-miR-92a-1-5p	-2.22	3.99	6.35E-07	2.64E-05	5.93	27.53
hsa-miR-129-2-3p	-4.30	3.72	9.54E-07	3.45E-05	1.34	26.66
hsa-miR-3653	2.93	4.98	9.69E-07	3.45E-05	52.34	6.81
hsa-miR-19a-5p	-2.09	2.71	1.97E-06	6.13E-05	2.59	11.04
hsa-miR-29b-3p	1.69	9.11	1.86E-06	6.13E-05	809.33	250.37
hsa-miR-4464	3.47	2.53	3.12E-06	9.13E-05	9.73	0.87
hsa-miR-450b-5p	-1.52	4.12	3.76E-06	1.04E-04	9.28	26.54
hsa-miR-18a-5p	-1.65	5.78	3.97E-06	1.04E-04	27.74	86.86
hsa-miR-29a-3p	1.62	13.09	4.69E-06	1.16E-04	12651.18	4113.90
hsa-miR-431-5p	-1.83	1.21	4.88E-06	1.16E-04	1.00	3.62
hsa-miR-944	-2.57	5.67	5.93E-06	1.34E-04	15.53	92.18
hsa-miR-150-5p	1.90	12.79	6.55E-06	1.39E-04	10730.09	2878.09
hsa-miR-20a-5p	-1.93	10.01	6.70E-06	1.39E-04	449.16	1706.57
hsa-miR-7-5p	-2.22	7.18	8.53E-06	1.70E-04	53.95	251.76

Separating BL and FL. Table is sorted by FDR. logFC indicates log fold change; logCPM, log counts per million, FDR, false discovery rate.

Supplementary References

Anders, S., Pyl, P. T. and Huber, W. (2014). HTSeq–A Python framework to work with high-throughput sequencing data. bioRxiv.

Betel, D., Koppal, A., Agius, P., Sander, C. and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 11(8): R90.

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

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414): 57-74.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S. (2004). MicroRNA targets in Drosophila. Genome biology 5(1): R1-R1.

Farazi, T. A., Ten Hoeve, J. J., Brown, M., Mihailovic, A., Horlings, H. M., van de Vijver, M. J., Tuschl, T. and Wessels, L. F. (2014). Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets. Genome Biol 15(1): R9.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G. and Fitzgerald, S. (2013). Ensembl 2014. Nucleic Acids Res.: gkt1196.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J. M. and Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 37(Web Server issue): W68-76.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141(1): 129-141.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F. and Hackermuller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5(9): e1000502.

Jones, D. T., Jager, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y. J., Pugh, T. J., Hovestadt, V., Stutz, A. M., Rausch, T., Warnatz, H. J., Ryzhova, M., Bender, S., Sturm, D., Pleier, S., Cin, H., Pfaff, E., Sieber, L., Wittmann, A., Remke, M., Witt, H.,

Hutter, S., Tzaridis, T., Weischenfeldt, J., Raeder, B., Avci, M., Amstislavskiy, V., Zapatka, M., Weber, U. D., Wang, Q., Lasitschka, B., Bartholomae, C. C., Schmidt, M., von Kalle, C., Ast, V., Lawerenz, C., Eils, J., Kabbe, R., Benes, V., van Sluis, P., Koster, J., Volckmann, R., Shih, D., Betts, M. J., Russell, R. B., Coco, S., Tonini, G. P., Schuller, U., Hans, V., Graf, N., Kim, Y. J., Monoranu, C., Roggendorf, W., Unterberg, A., Herold-Mende, C., Milde, T., Kulozik, A. E., von Deimling, A., Witt, O., Maass, E., Rossler, J., Ebinger, M., Schuhmann, M. U., Fruhwald, M. C., Hasselblatt, M., Jabado, N., Rutkowski, S., von Bueren, A. O., Williamson, D., Clifford, S. C., McCabe, M. G., Collins, V. P., Wolf, S., Wiemann, S., Lehrach, H., Brors, B., Scheurlen, W., Felsberg, J., Reifenberger, G., Northcott, P. A., Taylor, M. D., Meyerson, M., Pomeroy, S. L., Yaspo, M. L., Korbel, J. O., Korshunov, A., Eils, R., Pfister, S. M. and Lichter, P. (2012). Dissecting the genomic complexity underlying medulloblastoma. Nature 488(7409): 100-105.

Kloetgen, A., Münch, P. C., Borkhardt, A., Hoell, J. I. and McHardy, A. C. (2014). Biochemical and bioinformatic methods for elucidating the role of RNA–protein interactions in posttranscriptional regulation. Briefings in functional genomics: elu020.

Lenze, D., Leoncini, L., Hummel, M., Volinia, S., Liu, C., Amato, T., De Falco, G., Githanga, J., Horn, H. and Nyagol, J. (2011). The different epidemiologic subtypes of Burkitt lymphoma share a homogenous micro RNA profile distinct from diffuse large B-cell lymphoma. Leukemia 25(12): 1869-1876.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078-2079.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17(1): 10-12.

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

Richter, J., Schlesner, M., Hoffmann, S., Kreuz, M., Leich, E., Burkhardt, B., Rosolowski, M., Ammerpohl, O., Wagener, R., Bernhart, S. H., Lenze, D., Szczepanowski, M., Paulsen, M., Lipinski, S., Russell, R. B., Adam-Klages, S., Apic, G., Claviez, A., Hasenclever, D., Hovestadt, V., Hornig, N., Korbel, J. O., Kube, D., Langenberger, D., Lawerenz, C., Lisfeld, J., Meyer, K., Picelli, S., Pischimarov, J., Radlwimmer, B., Rausch, T., Rohde, M., Schilhabel, M., Scholtysik, R., Spang, R., Trautmann, H., Zenz, T., Borkhardt, A., Drexler, H. G., Moller, P., MacLeod, R. A., Pott, C., Schreiber, S., Trumper, L., Loeffler, M., Stadler, P. F., Lichter, P., Eils, R., Kuppers, R., Hummel, M., Klapper, W., Rosenstiel, P., Rosenwald, A., Brors, B., Siebert, R. and Project, I. M.-S. (2012). Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat Genet 44(12): 1316-1320.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1): 139-140.

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

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

Publication III

T-cell acute lymphoblastic leukemia in infants has distinct genetic and epigenetic features compared to childhood cases

Doerrenberg M[1], Kloetgen A[1,2], Wössmann W[3], Stanulla M[4], McHardy AC[2], Borkhardt A[1], Hoell JI[1]

[1]Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Medical Faculty, Düsseldorf, Germany
[2]Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany
[3]Department of Pediatric Hematology and Oncology, University Hospital Gießen and Marburg, Gießen, Germany
[4]Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany

Running head: Genetic and epigenetic features in infant T-ALL

Abstract

For reasons not yet understood, nearly all infants with acute lymphoblastic leukemia (ALL) are diagnosed with the B-cell type, with T-ALL in infancy representing the very rare exception. Clinical and molecular knowledge about infant T-ALL is still nearly completely lacking and it is also still unclear, whether it represents a distinct disease compared to childhood T-ALL.

To address this, we performed exome sequencing of three infant cases, which enabled the detection of mutations in NOTCH2, NOTCH3, PTEN and KRAS. When analyzing the transcriptomes and miRNomes of the three infant and an additional six childhood T-ALL samples, we found 760 differentially expressed mRNAs and 58 differentially expressed miRNAs between these two cohorts. Correlation analysis for differentially expressed miRNA-mRNA target pairs revealed 47 miRNA-mRNA pairs, with numerous of them already described to be aberrantly expressed in leukemia and cancer. Pathway analysis revealed differentially expressed pathways and upstream regulators related to the immune system or cancerogenesis such as the ERK5 pathway, which was activated in infant T-ALL. In summary, there are distinct molecular features in infant compared to childhood T-ALL on a transcriptomic and epigenetic level, which potentially have an impact on the development and course of the disease.

Introduction

Leukemia is an aggressive malignant disease of the hematopoietic system and the most frequent pediatric cancer type, with about one in every third cancer patient diagnosed with acute lymphoblastic leukemia (ALL). Next generation sequencing brought deeper insights into the tumor biology of ALL by identifying novel and recurrent genomic mutations affecting genes with roles in cell proliferation, differentiation, apoptosis, and drug resistance (Pui, Carroll et al. 2011). ALL can be subdivided into distinct subtypes (B- and T-ALL) and risk classes according to genomic aberrations. This enables the administration of risk-directed, in some instances even molecular-targeted, therapeutic

strategies. Nowadays, approximately 80% of all treated ALL patients in developed countries achieve an event-free survival (EFS) of ≥ 5 years (Pui, Carroll et al. 2011). However, treatment failure rates in infants (≤ 1 year of age) remain high at 60% (Hilden, Dinndorf et al. 2006), mostly because of a higher drug resistance, especially to prednisolone and asparaginase (Pieters, den Boer et al. 1998).

About 10–15% of all childhood ALL patients are diagnosed with T-ALL (Goldberg, Silverman et al. 2003), which represents a more aggressive leukemia subtype compared to B-ALL with a higher risk of relapse (Uckun, Gaynon et al. 1997, Aifantis, Raetz et al. 2008) . T-ALL patients are always classified high risk and thus receive a more intensive chemotherapy regimen. The development of T-ALL in infancy (iT-ALL) is extremely rare. From what is known, children older than one year – similar to B-cell ALL – seem to have a better outcome than infants suffering from T-ALL (Goldberg, Silverman et al. 2003, Mansur, van Delft et al. 2015). At present, it is not clear whether iT-ALL represents a distinct disease to childhood T-ALL.

We thus aimed to analyze the genetic and epigenetic differences between infant and childhood T-ALL. We used next generation sequencing to discover molecular aberrations on different levels, which promote the development of iT-ALL. We analyzed the exomes of three infant patients to uncover distinct mutations. We also compared the transcriptomes and miRNomes of these iT-ALL cases with six childhood T-ALL cases. Our findings show that infant and childhood T-ALL differ strongly on a genetic and epigenetic level.

Methods

Patient samples

Mononuclear cells of bone marrow or blood samples from patients diagnosed with T-ALL were obtained at initial diagnosis. The study was approved by the local ethics committee (study numbers 3432, 4769, 5036) and written informed consent was obtained from the parents.

High throughput sequencing

Total RNA and genomic DNA from cell pellets frozen in PBS were extracted with AllPrep DNA/RNA/miRNA Universal Kit (Qiagen, Hilden, #80224). Quality and quantity of RNA samples were measured on 2100 Bioanalyzer (Agilent, Amstelveen). Integrity numbers ranged from 8.4 to 10. For RNA sequencing, a total of 1 µg total RNA was used for library preparation with TruSeq Small RNA Sample Prep Kit (Illumina, San Diego, CA, USA, #RS-200-0012/0024). Between 2.7 and 10 µg of total RNA was used for library preparation with TruSeq Stranded Total RNA Sample Prep Kit with Ribo-Zero Gold (Illumina, San Diego, #RS-122-2301). Size and DNA concentration of prepared cDNA libraries were also analyzed on the 2100 Bioanalyzer. 7 pM of template DNA were loaded per flow cell, high-throughput sequencing was performed on the Illumina HiSeq 2500 with 50 cycles (miRNAs) or 100 cycles (mRNAs).

Quality and quantity of genomic DNA samples were measured on Nanodrop 1000 (VWR, Darmstadt). 1 µg of DNA from each of the infant patient samples was used for whole exome sequencing. Exome capturing of extracted DNA was performed with the SureSelectV5+UTR kit (Agilent, Santa Clara) and sequenced on Illumina HiSeq 2500 with 100 cycles (paired-end). In total, we obtained 466,376,392 sequencing reads, out of which 465,870,438 (99.89%) were mapped and paired uniquely to the genome. We achieved a coverage of ≥ 30x for about 95% of all captured features.

Sanger sequencing

Sanger sequencing was performed to validate mutations and chromosomal translocation found by exome sequencing and RNA-seq. For validation of the mutations, genomic DNA of patient material was amplified with REPLI-g Ultra Fast Mini Kit (Qiagen, Hilden #150035). For validation of chromosomal translocations RNA of patient material was reverse transcribed into cDNA by SuperScript III Reverse Transcriptase (Thermo Fisher, Braunschweig #18080-044). Genomic DNA and cDNA were used for amplification of mutated or translocated regions. PCR primers for PCR amplification of the mutations were flanking the mutated sites (Supplementary Table 1). PCR was performed with Phusion High Fidelity DNA Polymerase (NEB, Ipswich #M0530L) in

50µl reactions with HF buffer. Cycling for validation mutations was performed 30 seconds at 98°, a 30 times repetition of 5 seconds at 98°, 30 seconds at 60°, 12 seconds at 72° and a final extension of 10 minutes at 72°.

Bioinformatic methods

For a detailed overview of the bioinformatics methods for sequencing data analysis, pathway analysis, correlation analysis and exome data analysis, please refer to Supplementary Methods.

Results

Oncogenes and tumor suppressors important for leukomogenesis are frequently mutated in iT-ALL

Blast material of three infant patients from 0 to 12 months of age and of six childhood patients from one to 16 years with T-ALL was analyzed (Table 1). The median age of the infant patients was 9 months. The median age of the childhood patients was 11 years.

To analyze the mutational spectra of iT-ALL, we performed WES on the three infant patients. In total, 4,504 mutations in 1,595 genes were detected in three infant patients, 1,305 recurrent mutations in 798 genes in at least two patients, and 557 recurrent mutations in 426 genes in all three patients. As we did not have germline material available, these numbers refer to all detected mutations (SNPs and indels) and not only to leukemia-specific alterations. Based on recent studies on genetic mutations in T-ALL, we focused on genes recurrently mutated in T-ALL to ascertain, whether these genes were mutated also in iT-ALL. We validated a total of 19 of the here reported mutations by Sanger sequencing (Table 2).

We found mutations in NOTCH2 and NOTCH3. One of the three infant patients (#102) had a heterozygous NOTCH2 mutation, which was predicted as deleterious as it causes an aminoacid change from phenylalanin to valin in the extracellular EGF-like domain in the NOTCH2 protein, which is needed for Ca2+-dependent ligand binding (Rao, Handford et al. 1995). The same patient exhibited a homozygous 10 bp deletion in NOTCH3. All infant patients had a 4 bp deletion in NOTCH3; however, this is located in a

repetitive genome region. In the patients 101 and 102, the aforementioned mutation was heterozygous and in patient 103 it was homozygous.

We also found several mutations in PTEN. One patient (103) had a heterozygous 2 bp insertion in PTEN leading to an elongated frameshift variant. In patient 102, we found two heterozygous mutations in PTEN. The first was a 1 bp deletion leading to a feature truncation , the second was a 2 bp deletion. Patient 101 harbored six heterozygous PTEN mutations – a 1 bp deletion, a 3 bp deletion, two 1 bp insertions, one 1 bp insertion, and one 3 bp insertion. In addition, patients 101 (heterozygous) and 102 (homozygous) had a 5 bp insertion. All of the infant patients harbored KRAS mutations. Patient 103 had two homozygous 1 bp insertions. Both mutations occurred, in a heterozygous form, in patient 101. Patient 102 carried one of these insertions heterozygously and the other homozygously. Patient 103 (homozygous) and patient 101 (heterozygous) had also a 2 bp deletion in KRAS.

Table 1: Patient information for three infant (101, 102 and 103) and six childhood T-ALL cases (201, 202, 203, 204, 205, 206) analyzed in this study**.** m: male, f: female, PB: peripheral blood, BM: bone marrow.

Patient ID	**Age at diagnosis**	**sex**	**material**	**outcome**	**BCR/ABL**	**MLL/AF4**	**TEL/AML1**	**MLL/ENL**
101	5 months	m	PB	deceased	neg	neg	neg	n.d.
102	9 months	f	BM/PB	deceased	neg	neg	neg	n.d.
103	10 months	f	PB	deceased	neg	neg	neg	Neg
201	15 years	m	PB	remission	neg	neg	neg	Neg
202	14 years	m	BM	remission	neg	n.d.	neg	n.d.
203	5 years	m	PB	remission	neg	neg	neg	Neg
204	16 months	w	BM	remission	neg	neg	neg	Neg
205	12 years	m	PB	remission	neg	neg	neg	Neg
206	10 years	w	PB	remission	n.d.	n.d.	n.d.	n.d.

Table 2: Validated genetic alterations (SNVs and indels) in iT-ALL samples.

Coordinates	Gene affected	Type of alteration	Patient	Genotype
1:120478125-120478125	NOTCH2	SNV	102	heterozygous
19:15285382-15285386	NOTCH3	4 bp deleteion	103	homozygous
19:15285382-15285386	NOTCH3	4 bp deleteion	101	heterozygous
19:15285382-15285386	NOTCH3	4 bp deleteion	102	heterozygous
5:35857308-35857309	IL7R	1 bp insertion	103	homozygous
5:35857308-35857309	IL7R	1 bp insertion	101	heterozygous
10:89653620-89653621	PTEN	1 bp deletion	101	heterozygous
10:89690952-89690957	PTEN	5 bp ins	102	homozygous
10:89690952-89690957	PTEN	5 bp ins	101	heterozygous
10:89717674-89717676	PTEN	2 bp ins	103	heterozygous
10:89725886-89725887	PTEN	1 bp ins	101	heterozygous
10:89728633-89728634	PTEN	1 bp deletion	102	heterozygous
10:89731315-89731317	PTEN	2 bp ins	101	heterozygous
12:25358662-25358664	KRAS	2 bp del	101	heterozygous
12:25358662-25358664	KRAS	2 bp del	103	homozygous
4:154626317-154626317	TLR2	SNV	102	heterozygous
9:120466929-120466930	TLR4	1 bp deletion	103	heterozygous
9:120466929-120466930	TLR4	1 bp deletion	101	heterozygous
9:120466929-120466930	TLR4	1 bp deletion	102	heterozygous

We also found several mutations in PTEN. One patient (103) had a heterozygous 2 bp insertion in PTEN leading to an elongated frameshift variant. In patient 102, we found two heterozygous mutations in PTEN. The first was a 1 bp deletion leading to a feature truncation, the second was a 2 bp deletion. Patient 101 harbored six heterozygous PTEN mutations – a 1 bp deletion, a 3 bp deletion, two 1 bp insertions, one 1 bp insertion and one 3 bp insertion. In addition, patients 101 (heterozygous) and 102 (homozygous) had a 5 bp insertion. All of the infant patients harbored KRAS mutations. Patient 103 had two homozygous 1 bp insertions. Both mutations also occurred heterozygous in patient 101. Patient 102 carried one of these insertions heterozygous and the other homozygous. Patient 103 (homozygous) and patient 101 (heterozygous) had also a 2 bp deletion in KRAS.

RNA-Seq enabled discrimination of iT-ALL on gene and pathway level

To analyze the differences in expressed mRNAs between infant and childhood cases, we sequenced the transcriptomes of the six childhood and three infant samples. We obtained a total of 135,216,180 sequencing reads mapping to coding genes. We found 760 differentially expressed genes ($|\log FC| \geq 1$, p-value ≤ 0.01). Out of these, 207 were downregulated in the infants and 553 were upregulated. Clustering of the significantly differentially expressed genes clearly separated childhood from infant samples indicating the genetic difference between iT-ALL and childhood T-ALL (Figure 1A).

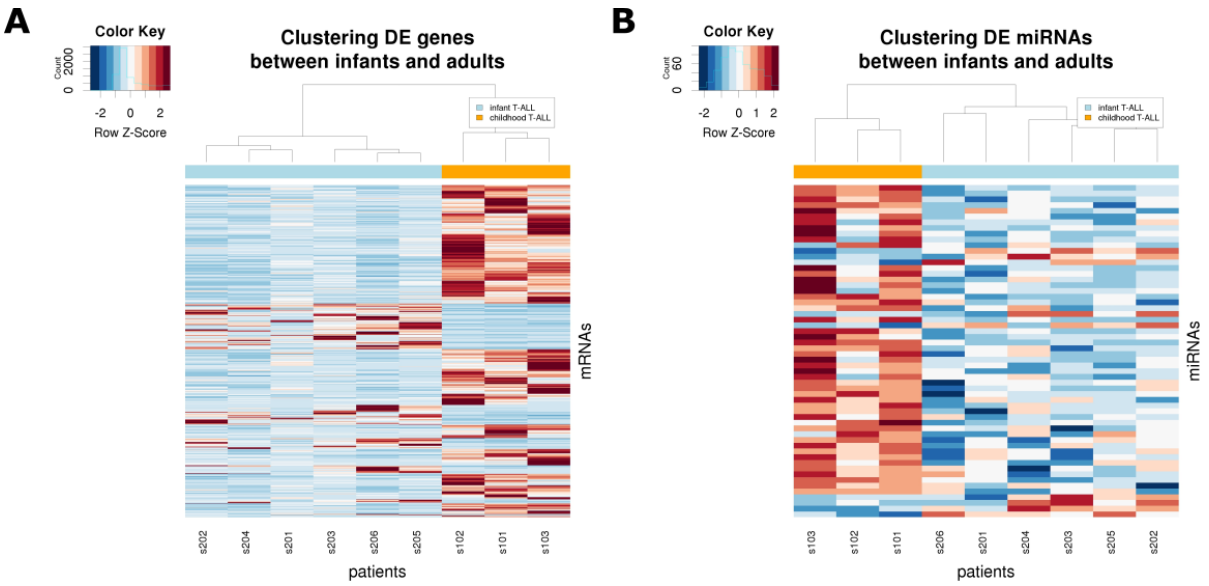


Figure 1: Hierarchical clustering of differentially expressed mRNAs (A) and miRNAs (B) between iT-ALL and childhood T-ALL.

To get a better understanding of the biological processes in which the identified differentially expressed genes are involved, a pathway analysis with Ingenuity Pathway Analysis (IPA) was performed. We found 9 pathways (Supplementary Table 2), which were differentially regulated between infant and childhood patients (p-value < 0.05, z-score ≥ ± 1). Most perturbed pathways were related to immune functions or cancer, including differentiation, proliferation, apoptosis or cell survival signaling. Among the identified pathways we found the ERK5 (mitogen-activated protein kinase 7) pathway

to be activated in iT-ALL (z-score = 1.134; Supplementary Figure 1). The ERK5 pathway can be regulated by a variety of stimuli such as growth factors, G-Protein coupled receptors or cellular stress factors. The pathway map shows an activation of ERK5 via TRKA (Tyrosine Kinase Receptor A) in infant samples, which further promotes the activation of transcription factors such as MYC (cellular myelocytomatosis viral oncogene homolog), SAP1 (putative AAA family ATPase SAP1) or FRA1 (FOSL1, FOS like antigen 1).

We also performed an Upstream Regulator Analysis with IPA and found the toll-like receptors 2 (TLR2) and 4 (TLR4) being inhibited based on the aberrant expression of their downstream targets. TLR2 itself was not significantly downregulated across all iT-ALL samples with a logFC of -0.98 (p-value = 0.29; Supplementary Figure 2A). TLR4 was downregulated in iT-ALL samples with a logFC of -2.07 (p-value = 0.03; Supplementary Figure 2B). However, patient 102 additionally harbored two heterozygous, deleterious TLR2 mutations and all three infant patients harbored a heterozygous 1 bp deletion in TLR4 (Table 2). The differentially expressed downstream targets for TLR2 include SELP (selectin P), ITGA2B (integrin subunit alpha 2b), IL1B (interleukin 1 beta), CD86 (cluster of differentiation 86) and IL6 (interleukin 6), which were all significantly downregulated in infant samples compared to childhood T-ALL. The downregulated targets of TLR4 were CD86, IL1B, IL6, CCR7 (C-C motif chemokine receptor 7) and CCL5 (C-C motif chemokine ligand 5).

Furthermore, we checked upregulated genes in iT-ALL cases for approved therapeutics and for those that are currently being tested in clinical trials. We found a total of six genes to be significantly upregulated in iT-ALL compared to childhood T-ALL; these are potentially targetable and are thus interesting targets for individualized therapy (Supplementary Table 3), possibly as an add-on to standard therapy regimens. This included KIT (logFC = 2.03), a receptor tyrosine-kinase frequently associated with different cancers, for which four approved receptor tyrosine-kinase inhibitors are available and of potential therapeutic interest: Imatinib (Debiec-Rychter, Sciot et al. 2006), Dasatinib (Antonescu, Busam et al. 2007), Sorafenib (Bisagni, Rossi et al. 2009), and Pazopanib (Sloan and Scheinfeld 2008).

miRNA-mRNA correlations reveal further mechanisms specific to iT-ALL

To analyze differences between infant and childhood T-ALL samples on an epigenetic level, we performed miRNA sequencing. We obtained a total of 31,761,705 sequencing reads mapping to miRNAs annotated in miRBase V21. We found 58 miRNAs that were differentially expressed between infant and childhood T-ALL samples ($|\text{logFC}| \geq 1$ and P-value ≤ 0.01; Supplementary Table 4). Nine of these miRNAs were downregulated in iT-ALL and 49 were upregulated. Hierarchical clustering of the most significant differentially expressed genes again showed a clear separation between childhood and infant cases (Fig. 1B). This illustrates differences between infant and childhood T-ALL not only on the transcriptomic but also on the epigenetic level.

To identify mRNAs whose aberrant expression pattern might be explained by differentially expressed miRNAs, we performed a correlation analysis based on five public miRNA target databases. This correlation analysis for differentially expressed miRNA-mRNA target pairs revealed 47 miRNA-mRNA pairs (Spearman's Rho ≤ -0.6 and P-value ≤ 0.05; Table 3). MiRNA hsa-let-7b was downregulated in the infant samples. Hsa-let7b was previously described to be downregulated in infant B-ALL with MLL-(lysine methyltransferase 2A) rearrangements (MLL-r) (Nishi, Eguchi-Ishimae et al. 2013, Wu, Eguchi-Ishimae et al. 2015). We could not detect any MLL-r in the infant samples, but the low expression of hsa-let7b might promote the expression of its target genes. For hsa-let-7b, we found six potential target genes showing a negative correlation (Supplementary Fig. 3). This includes an upregulation of IGF2BP1 (insulin like growth factor 2 mRNA binding protein 1) and an upregulation of GREB1 (Figs. 2A-B). We also found hsa-miR-31 to be downregulated in iT-ALL samples (logFC = -3.86, P-value = 0.004), which may act as a tumor suppressor. Our findings suggest that hsa-miR-31 likely targets RPA1 (replication protein A1) (spearman's Rho: -0.62 and P-value = 0.04; Fig. 2C).

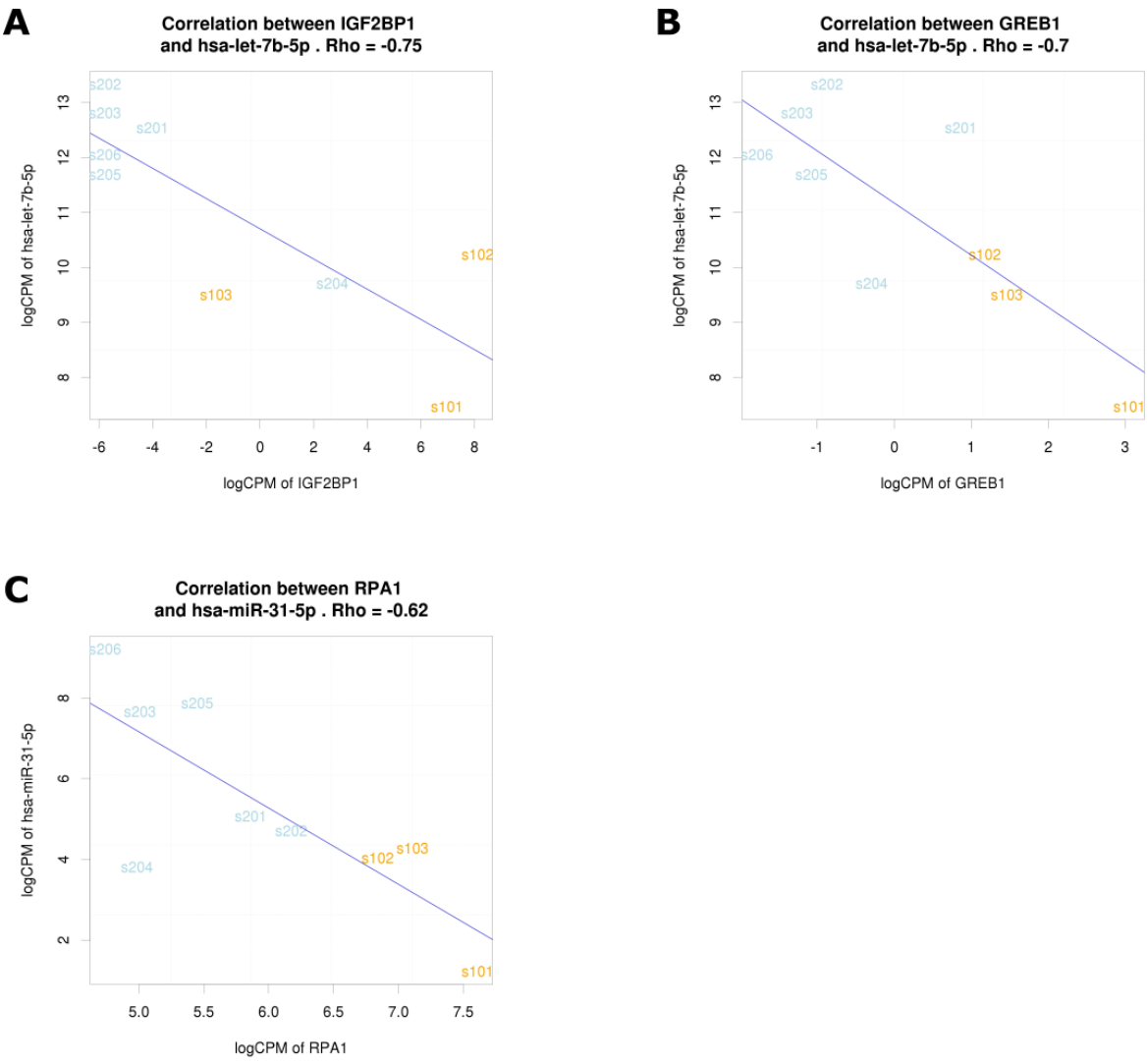


Figure 2: Negative correlation between hsa-let-7b and IGF2BP1 (A) and GREB1 (B) and hsa-miR-31 and RPA1 (C).

Table 3: Negatively correlated miRNA-mRNA interaction pairs. Information on interactions was taken from publicly available databases (miRanda, miRDB, TarBase, TargetScan and miRTarBase). For each such a correlation, a spearman correlation was calculated. The list is filtered for spearmans Rho ≤ -0.6 and correlation p-value ≤ 0.05. logFC specifies the logarithmic fold-change of expression between infant and childhood cases. All p-values for differential expression of mRNAs and miRNAs were statistically significant (< 0.01).

miRNA	miRNA logFC	targeted mRNA	mRNA logFC	Spearman Rho	Correlation p-value
hsa-miR-5683	4.0	**HLF**	-3.0	-0.73	0.02
hsa-miR-205-5p	-3.7	**NFAT5**	1.4	-0.75	0.01
hsa-miR-421	1.9	**HOXA9**	-5.0	-0.6	0.05
hsa-miR-3909	2.2	**NOG**	-2.6	-0.62	0.04
hsa-let-7b-5p	-2.9	**SCAF4**	1.4	-0.68	0.03
hsa-let-7b-5p	-2.9	**GREB1**	2.6	-0.7	0.02
hsa-miR-766-3p	3.3	**NR3C2**	-4.5	-0.78	0.01
hsa-miR-5683	4.0	**CDK14**	-2.3	-0.93	0.00
hsa-miR-183-5p	2.9	**CR1**	-2.9	-0.63	0.04
hsa-miR-3909	2.2	**CD300E**	-2.9	-0.6	0.05
hsa-miR-3143	2.0	**ABCD2**	-4.0	-0.72	0.02
hsa-miR-18a-5p	2.4	**MEF2C**	-2.0	-0.63	0.04
hsa-let-7f-1-3p	-2.5	**ZFAND3**	1.8	-0.78	0.01
hsa-let-7b-5p	-2.9	**PDPR**	1.4	-0.6	0.05
hsa-miR-148b-3p	2.5	**NR3C2**	-4.5	-0.67	0.03
hsa-let-7b-5p	-2.9	**COL1A2**	2.9	-0.64	0.03
hsa-miR-421	1.9	**MEF2C**	-2.0	-0.67	0.03
hsa-let-7f-1-3p	-2.5	**PRPF8**	1.5	-0.6	0.05
hsa-let-7b-3p	-3.6	**WNK1**	1.5	-0.72	0.02

hsa-miR-5683	4.0	**ZBTB38**	-2.9	-0.62	0.04
hsa-miR-183-5p	2.9	**KDELC2**	-2.3	-0.83	0.00
hsa-let-7b-5p	-2.9	**PRDM2**	1.4	-0.7	0.02
hsa-miR-5683	4.0	**MEF2C**	-2.0	-0.87	0.00
hsa-let-7b-5p	-2.9	**FRAS1**	2.4	-0.82	0.01
hsa-miR-1276	2.5	**ADCY9**	-2.6	-0.69	0.02
hsa-miR-3143	2.0	**NOG**	-2.6	-0.67	0.03
hsa-miR-5581-3p	2.4	**MEIS1**	-3.5	-0.8	0.01
hsa-let-7f-1-3p	-2.5	**NBEA**	1.5	-0.73	0.02
hsa-let-7b-5p	-2.9	**IGF2BP1**	6.9	-0.75	0.01
hsa-miR-18a-5p	2.4	**ERRFI1**	-3.3	-0.75	0.01
hsa-miR-671-5p	1.8	**ALDH3A2**	-2.1	-0.62	0.04
hsa-miR-421	1.9	**PTGER2**	-1.9	-0.72	0.02
hsa-let-7b-5p	-2.9	**WNK1**	1.5	-0.73	0.02
hsa-miR-331-3p	2.7	**MEIS1**	-3.5	-0.87	0.00
hsa-miR-148b-3p	2.5	**PTGER2**	-1.9	-0.67	0.03
hsa-miR-205-5p	-3.7	**RUNX2**	3.6	-0.6	0.05
hsa-miR-148b-3p	2.5	**MYBL1**	-3.4	-0.68	0.03
hsa-let-7b-5p	-2.9	**COL4A6**	4.6	-0.83	0.00
hsa-miR-5581-3p	2.4	**FOS**	-3.0	-0.85	0.00
hsa-let-7b-3p	-3.6	**JUP**	2.9	-0.7	0.02
hsa-miR-31-5p	-3.9	**RPA1**	1.8	-0.62	0.04
hsa-miR-31-5p	-3.9	**WNK1**	1.5	-0.72	0.02
hsa-let-7b-3p	-3.6	**SLC18A2**	2.7	-0.63	0.04
hsa-miR-5581-3p	2.4	**MYBL1**	-3.4	-0.87	0.00
hsa-let-7b-3p	-3.6	**ZFAND3**	1.8	-0.75	0.01
hsa-miR-148b-3p	2.5	**RAB34**	-3.8	-0.63	0.04
hsa-miR-421	1.9	**NR3C2**	-4.5	-0.73	0.02

Discussion

By exome sequencing we identified multiple mutations in oncogenes in the iT-ALL samples including NOTCH2 and NOTCH3. The latter was mutated in all three infant cases. However, we did not detect any NOTCH1 mutation, which are the most frequent mutations in childhood T-ALL with a frequency around 60% in all T-ALL patients (Weng, Ferrando et al. 2004). We did not detect any FBXW7 (F-box and WD repeat domain containing 7) mutation, which is frequently mutated in childhood T-ALL (Park, Taki et al. 2009) and was also described to be mutated in some iT-ALL patients (Mansur, van Delft et al. 2015). Recent studies reported non-synonymous mutations in the C2 domain of PTEN to cause C-terminal truncations of the protein in T-ALL patients (Gutierrez, Sanda et al. 2009). Some deletions in PTEN are also associated with treatment failure in T-ALL, having an impact on patient outcome. We identified short deletions in PTEN in two out of the three infant patients and short insertions in all patients. Also, KRAS mutations that were found in the iT-ALL cases are frequently associated with treatment response. The KRAS mutations prevent responses to cetuximab especially in colorectal cancer (Di Fiore, Blanchard et al. 2007).

We found distinct mechanisms acting in infant and childhood T-ALL by multiple sequencing analyses. Hierarchical clustering of both differentially expressed mRNAs and miRNAs separated infant from childhood samples. Our findings hence indicate that infant and childhood T-ALL differ strongly on a genetic and epigenetic level. We identified multiple differentially affected signaling pathways between infant and childhood T-ALL. By upstream regulator analysis we found multiple downstream targets of TLR2 and TLR4 to be significantly downregulated in iT-ALL. TLR2 and TLR4 themselves were not significantly downregulated in all iT-ALL patients, respectively, but predicted to be inhibited. However, we found mutations in TLR2/4 which might explain the downregulation of their targets without TLR2/4 being differentially expressed in all cases. TLR2/4 are pattern-recognition receptors, which are activated by binding to conserved molecules on pathogens leading to an activation of signaling pathways promoting inflammatory immune responses (Mahla, Reddy et al. 2013). The

detected TLR2 SNP has been annotated as a deleterious, missense mutation (rs5743708). A loss of TLR2 was described to promote liver cancer development in mice (Lin, Yan et al. 2013). The TLR4 deletion lies within an intron (rs5900307). TLR2/4 are also downregulated in patients with hepatocarcinoma (Soares, Pimentel-Nunes et al. 2012). Additionally, TLRs initiate signaling pathways leading to cell proliferation and chemoresistance (Chen, Alvero et al. 2008). However, knowledge on the function of TLR2/4 in T-cells or T-cell blasts is sparse, apart from its physiologic low expression (Muzio, Bosisio et al. 2000, Hornung, Rothenfusser et al. 2002).

When searching for potentially druggable targets in the infant cases, we identified the receptor tyrosine kinase KIT, for which four approved drugs are available (Imatinib, Dasatinib, Sorafenib, and Pazopanib). KIT is activated by its ligand SCF (stem cell factor), which is important during the differentiation of hematopoietic cells (Ashman 1999). However, patients carrying or acquiring certain are resistant to those TKIs (tyrosine kinase inhibitors) such as Imatinib (Heinrich, Corless et al. 2003, McLean, Gana-Weisz et al. 2005, Roberts, Odell et al. 2007). We could not identify any KIT mutations in the infant cases, but a higher expression compared to the childhood cases. Thus the drugs targeting KIT are of potential interest as add-on to standard therapy regimens.

Target prediction and correlation analysis between mRNAs and miRNAs showed 47 differentially expressed and negatively correlated miRNA-mRNA pairs. Most of these genes and miRNAs have previously been described to influence cancer development and progression. That high expression of hsa-let-7b targets GREB1 has been described in breast (Liu, Wang et al. 2012, Mohammed, D'Santos et al. 2013), ovarian (Bauerschlag, Ammerpohl et al. 2011), and prostate cancer (Antunes, Leite et al. 2012). A high expression of GREB1 was experimentally shown to increase proliferation of breast cancer cells (Rae, Johnson et al. 2005, Liu, Wang et al. 2012) and knockdown of GREB1 decreased tumor growth in ovarian cancer cells in vitro and in mouse xenografts (Laviolette, Hodgkinson et al. 2014). Another target, IGF2BP1, has been shown to be overexpressed in ETV6 (ETS variant 6)-RUNX1 (runt related transcription factor) ALL-subtype (Stoskus, Gineikiene et al. 2011) and aberrantly expressed by fusion with IGH

(immunoglobulin heavy locus) locus (Gu, Sederberg et al. 2014, Jeffries, Jones et al. 2014). In adult T-ALL cases, downregulation of hsa-miR-31 has been reported to activate NF-kB-signaling and promote apoptosis resistance (Yamagishi, Nakano et al. 2012). The here regulated target of hsa-miR-31, RPA1, is important for DNA damage response during DNA replication (Lin, Shivji et al. 1998). A significant upregulation of this gene has also been described in CLL (chronic lymphocytic leukemia) (Poncet, Belleville et al. 2008, Hoxha, Fabris et al. 2014).

In summary, we identified the landscape of genetic alterations in a limited set of iT-ALL by WES, transcriptome, and miRNome sequencing. We identified multiple differentially expressed signaling pathways between iT-ALL and childhood T-ALL and showed that iT-ALL differs substantially from childhood T-ALL on the transcriptomic and epigenetic levels.

References

Aifantis, I., Raetz, E. and Buonamici, S. (2008). Molecular pathogenesis of T-cell leukaemia and lymphoma. Nat. Rev. Immunol. 8(5): 380-390.

Antonescu, C. R., Busam, K. J., Francone, T. D., Wong, G. C., Guo, T., Agaram, N. P., Besmer, P., Jungbluth, A., Gimbel, M. and Chen, C. T. (2007). L576P KIT mutation in anal melanomas correlates with KIT protein expression and is sensitive to specific kinase inhibition. Int. J. Cancer 121(2): 257-264.

Antunes, A. A., Leite, K. R., Reis, S. T., Sousa-Canavez, J. M., Camara-Lopes, L. H., Dall'oglio, M. F. and Srougi, M. (2012). GREB1 tissue expression is associated with organ-confined prostate cancer. Urol. Oncol. 30(1): 16-20.

Ashman, L. K. (1999). The biology of stem cell factor and its receptor C-kit. The international journal of biochemistry & cell biology 31(10): 1037-1051.

Bauerschlag, D. O., Ammerpohl, O., Brautigam, K., Schem, C., Lin, Q., Weigel, M. T., Hilpert, F., Arnold, N., Maass, N., Meinhold-Heerlein, I. and Wagner, W. (2011). Progression-free survival in ovarian cancer is reflected in epigenetic DNA methylation profiles. Oncology 80(1-2): 12-20.

Bisagni, G., Rossi, G., Cavazza, A., Sartori, G., Gardini, G. and Boni, C. (2009). Long lasting response to the multikinase inhibitor bay 43-9006 (Sorafenib) in a heavily pretreated metastatic thymic carcinoma. J. Thorac. Oncol. 4(6): 773-775.

Chen, R., Alvero, A. B., Silasi, D. A., Steffensen, K. D. and Mor, G. (2008). Cancers take their Toll--the function and regulation of Toll-like receptors in cancer cells. Oncogene 27(2): 225-233.

Debiec-Rychter, M., Sciot, R., Le Cesne, A., Schlemmer, M., Hohenberger, P., van Oosterom, A. T., Blay, J.-Y., Leyvraz, S., Stul, M. and Casali, P. G. (2006). KIT mutations and dose selection for imatinib in patients with advanced gastrointestinal stromal tumours. Eur. J. Cancer 42(8): 1093-1103.

Di Fiore, F., Blanchard, F., Charbonnier, F., Le Pessot, F., Lamy, A., Galais, M., Bastit, L., Killian, A., Sesboüé, R. and Tuech, J. (2007). Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by Cetuximab plus chemotherapy. Br. J. Cancer 96(8): 1166-1169.

Goldberg, J. M., Silverman, L. B., Levy, D. E., Dalton, V. K., Gelber, R. D., Lehmann, L., Cohen, H. J., Sallan, S. E. and Asselin, B. L. (2003). Childhood T-cell acute lymphoblastic leukemia: the Dana-Farber Cancer Institute acute lymphoblastic leukemia consortium experience. J. Clin. Oncol. 21(19): 3616-3622.

Gu, G., Sederberg, M. C., Drachenberg, M. R. and South, S. T. (2014). IGF2BP1: a novel IGH translocation partner in B acute lymphoblastic leukemia. Cancer Genet. 207(7-8): 332-334.

Gutierrez, A., Sanda, T., Grebliunaite, R., Carracedo, A., Salmena, L., Ahn, Y., Dahlberg, S., Neuberg, D., Moreau, L. A. and Winter, S. S. (2009). High frequency of PTEN, PI3K, and AKT abnormalities in T-cell acute lymphoblastic leukemia. Blood 114(3): 647-650.

Heinrich, M. C., Corless, C. L., Demetri, G. D., Blanke, C. D., von Mehren, M., Joensuu, H., McGreevey, L. S., Chen, C. J., Van den Abbeele, A. D., Druker, B. J., Kiese, B., Eisenberg, B., Roberts, P. J., Singer, S., Fletcher, C. D. M., Silberman, S., Dimitrijevic, S. and Fletcher, J. A. (2003). Kinase mutations and imatinib

response in patients with metastatic gastrointestinal stromal tumor. J Clin Oncol 21(23): 4342-4349.

Hilden, J. M., Dinndorf, P. A., Meerbaum, S. O., Sather, H., Villaluna, D., Heerema, N. A., McGlennen, R., Smith, F. O., Woods, W. G., Salzer, W. L., Johnstone, H. S., Dreyer, Z., Reaman, G. H. and Children's Oncology, G. (2006). Analysis of prognostic factors of acute lymphoblastic leukemia in infants: report on CCG 1953 from the Children's Oncology Group. Blood 108(2): 441-451.

Hornung, V., Rothenfusser, S., Britsch, S., Krug, A., Jahrsdorfer, B., Giese, T., Endres, S. and Hartmann, G. (2002). Quantitative expression of toll-like receptor 1-10 mRNA in cellular subsets of human peripheral blood mononuclear cells and sensitivity to CpG oligodeoxynucleotides. J. Immunol. 168(9): 4531-4537.

Hoxha, M., Fabris, S., Agnelli, L., Bollati, V., Cutrona, G., Matis, S., Recchia, A. G., Gentile, M., Cortelezzi, A., Morabito, F., Bertazzi, P. A., Ferrarini, M. and Neri, A. (2014). Relevance of telomere/telomerase system impairment in early stage chronic lymphocytic leukemia. Genes Chromosomes Cancer 53(7): 612-621.

Jeffries, S. J., Jones, L., Harrison, C. J. and Russell, L. J. (2014). IGH@ translocations co-exist with other primary rearrangements in B-cell precursor acute lymphoblastic leukemia. Haematologica 99(8): 1334-1342.

Laviolette, L. A., Hodgkinson, K. M., Minhas, N., Perez-Iratxeta, C. and Vanderhyden, B. C. (2014). 17beta-estradiol upregulates GREB1 and accelerates ovarian tumor progression in vivo. Int. J. Cancer 135(5): 1072-1084.

Lin, H., Yan, J., Wang, Z., Hua, F., Yu, J., Sun, W., Li, K., Liu, H., Yang, H., Lv, Q., Xue, J. and Hu, Z. W. (2013). Loss of immunity-supported senescence enhances susceptibility to hepatocellular carcinogenesis and progression in Toll-like receptor 2-deficient mice. Hepatology 57(1): 171-182.

Lin, Y. L., Shivji, M. K., Chen, C., Kolodner, R., Wood, R. D. and Dutta, A. (1998). The evolutionarily conserved zinc finger motif in the largest subunit of human replication protein A is required for DNA replication and mismatch repair but not for nucleotide excision repair. J. Biol. Chem. 273(3): 1453-1461.

Liu, M., Wang, G., Gomez-Fernandez, C. R. and Guo, S. (2012). GREB1 functions as a growth promoter and is modulated by IL6/STAT3 in breast cancer. PLoS One 7(10): e46410.

Mahla, R. S., Reddy, M. C., Prasad, D. V. and Kumar, H. (2013). Sweeten PAMPs: Role of Sugar Complexed PAMPs in Innate Immunity and Vaccine Biology. Front. Immunol. 4: 248.

Mansur, M. B., van Delft, F. W., Colman, S. M., Furness, C. L., Gibson, J., Emerenciano, M., Kempski, H., Clappier, E., Cave, H., Soulier, J., Pombo-de-Oliveira, M. S., Greaves, M. and Ford, A. M. (2015). Distinctive genotypes in infants with T-cell acute lymphoblastic leukaemia. Br. J. Haematol. 171(4): 574-584.

McLean, S. R., Gana-Weisz, M., Hartzoulakis, B., Frow, R., Whelan, J., Selwood, D. and Boshoff, C. (2005). Imatinib binding and cKIT inhibition is abrogated by the cKIT kinase domain I missense mutation Val654Ala. Molecular cancer therapeutics 4(12): 2008-2015.

Mohammed, H., D'Santos, C., Serandour, A. A., Ali, H. R., Brown, G. D., Atkins, A., Rueda, O. M., Holmes, K. A., Theodorou, V., Robinson, J. L., Zwart, W., Saadi, A., Ross-Innes, C. S., Chin, S. F., Menon, S., Stingl, J., Palmieri, C., Caldas, C. and Carroll, J. S. (2013). Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor. Cell Rep 3(2): 342-349.

Muzio, M., Bosisio, D., Polentarutti, N., D'Amico, G., Stoppacciaro, A., Mancinelli, R., van't Veer, C., Penton-Rol, G., Ruco, L. P., Allavena, P. and Mantovani, A. (2000). Differential expression and regulation of toll-like receptors (TLR) in human leukocytes: selective expression of TLR3 in dendritic cells. J. Immunol. 164(11): 5998-6004.

Nishi, M., Eguchi-Ishimae, M., Wu, Z., Gao, W., Iwabuki, H., Kawakami, S., Tauchi, H., Inukai, T., Sugita, K., Hamasaki, Y., Ishii, E. and Eguchi, M. (2013). Suppression of the let-7b microRNA pathway by DNA hypermethylation in infant acute lymphoblastic leukemia with MLL gene rearrangements. Leukemia 27(2): 389-397.

Park, M. J., Taki, T., Oda, M., Watanabe, T., Yumura-Yagi, K., Kobayashi, R., Suzuki, N., Hara, J., Horibe, K. and Hayashi, Y. (2009). FBXW7 and NOTCH1 mutations in childhood T cell acute lymphoblastic leukaemia and T cell non-Hodgkin lymphoma. Br. J. Haematol. 145(2): 198-206.

Pieters, R., den Boer, M. L., Durian, M., Janka, G., Schmiegelow, K., Kaspers, G. J., van Wering, E. R. and Veerman, A. J. (1998). Relation between age, immunophenotype and in vitro drug resistance in 395 children with acute lymphoblastic leukemia--implications for treatment of infants. Leukemia 12(9): 1344-1348.

Poncet, D., Belleville, A., t'kint de Roodenbeke, C., Roborel de Climens, A., Ben Simon, E., Merle-Beral, H., Callet-Bauchu, E., Salles, G., Sabatier, L., Delic, J. and Gilson, E. (2008). Changes in the expression of telomere maintenance genes suggest global telomere dysfunction in B-chronic lymphocytic leukemia. Blood 111(4): 2388-2391.

Pui, C. H., Carroll, W. L., Meshinchi, S. and Arceci, R. J. (2011). Biology, risk stratification, and therapy of pediatric acute leukemias: an update. J. Clin. Oncol. 29(5): 551-565.

Rae, J. M., Johnson, M. D., Scheys, J. O., Cordero, K. E., Larios, J. M. and Lippman, M. E. (2005). GREB 1 is a critical regulator of hormone dependent breast cancer growth. Breast Cancer Res. Treat. 92(2): 141-149.

Rao, Z., Handford, P., Mayhew, M., Knott, V., Brownlee, G. G. and Stuart, D. (1995). The structure of a Ca(2+)-binding epidermal growth factor-like domain: its role in protein-protein interactions. Cell 82(1): 131-141.

Roberts, K. G., Odell, A. F., Byrnes, E. M., Baleato, R. M., Griffith, R., Lyons, A. B. and Ashman, L. K. (2007). Resistance to c-KIT kinase inhibitors conferred by V654A mutation. Molecular cancer therapeutics 6(3): 1159-1166.

Sloan, B. and Scheinfeld, N. S. (2008). Pazopanib, a VEGF receptor tyrosine kinase inhibitor for cancer therapy. Current opinion in investigational drugs (London, England: 2000) 9(12): 1324-1335.

Soares, J. B., Pimentel-Nunes, P., Afonso, L., Rolanda, C., Lopes, P., Roncon-Albuquerque, R., Jr., Goncalves, N., Boal-Carvalho, I., Pardal, F., Lopes, S., Macedo, G., Lara-Santos, L., Henrique, R., Moreira-Dias, L., Goncalves, R., Dinis-Ribeiro, M. and Leite-Moreira, A. F. (2012). Increased hepatic expression of TLR2 and TLR4 in the hepatic inflammation-fibrosis-carcinoma sequence. Innate Immun. 18(5): 700-708.

Stoskus, M., Gineikiene, E., Valceckiene, V., Valatkaite, B., Pileckyte, R. and Griskevicius, L. (2011). Identification of characteristic IGF2BP expression patterns in distinct B-ALL entities. Blood Cells Mol. Dis. 46(4): 321-326.

Uckun, F. M., Gaynon, P. S., Sensel, M. G., Nachman, J., Trigg, M. E., Steinherz, P. G., Hutchinson, R., Bostrom, B. C., Sather, H. N. and Reaman, G. H. (1997). Clinical features and treatment outcome of childhood T-lineage acute lymphoblastic leukemia according to the apparent maturational stage of T-lineage leukemic blasts: a Children's Cancer Group study. J. Clin. Oncol. 15(6): 2214-2221.

Weng, A. P., Ferrando, A. A., Lee, W., Morris, J. P. t., Silverman, L. B., Sanchez-Irizarry, C., Blacklow, S. C., Look, A. T. and Aster, J. C. (2004). Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. Science 306(5694): 269-271.

Wu, Z., Eguchi-Ishimae, M., Yagi, C., Iwabuki, H., Gao, W., Tauchi, H., Inukai, T., Sugita, K., Ishii, E. and Eguchi, M. (2015). HMGA2 as a potential molecular target in KMT2A-AFF1-positive infant acute lymphoblastic leukaemia. Br. J. Haematol. 171(5): 818-829.

Yamagishi, M., Nakano, K., Miyake, A., Yamochi, T., Kagami, Y., Tsutsumi, A., Matsuda, Y., Sato-Otsubo, A., Muto, S., Utsunomiya, A., Yamaguchi, K., Uchimaru, K., Ogawa, S. and Watanabe, T. (2012). Polycomb-mediated loss of miR-31 activates NIK-dependent NF-kappaB pathway in adult T cell leukemia and other cancers. Cancer Cell 21(1): 121-135.

Supplementary Materials & Methods

Supplementary Methods

Analysis of Transcriptome Data

mRNA and miRNA datasets were handled in a similar fashion. First, adapter sequences and low-quality ends were trimmed off the reads using seqtk (https://github.com/lh3/seqtk) and cutadapt (Martin, 2011). Next, all remaining reads were aligned against the human reference genome sequence GRCh38 with STAR (Dobin et al., 2013) and BWA (Li and Durbin, 2009) for regular RNA-Seq and miRNA-Seq, respectively. BWA was applied with -n 0.04 option to accommodate for sequencing errors in the miRNA-Seq. To calculate read counts (relative expression values), HTSeq (Anders et al., 2014) was applied on both types of datasets using annotations from Ensembl Genes V82 (Yates et al., 2016) for mRNAs and annotations from miRBAse V21 (Kozomara and Griffiths-Jones, 2010) for miRNAs. Lastly, edgeR (Robinson et al., 2010) was used to inter-sample normalize expression counts (leading to counts per millions, CPM values) and to perform a differential gene expression between the infant and the childhood T-ALL cases.

Pathway Analysis

Pathway analysis was performed using QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, CA, USA, www.qiagen.com/ingenuity). We also applied an Upstream Regulator Analysis with IPA. This takes differentially expressed genes as input and queries literature entries for upstream regulators of the respective genes. Depending on the regulatory effect of the potential upstream regulator and the actual expression of its targets, it is either considered inhibited or activated.

Correlation Analysis

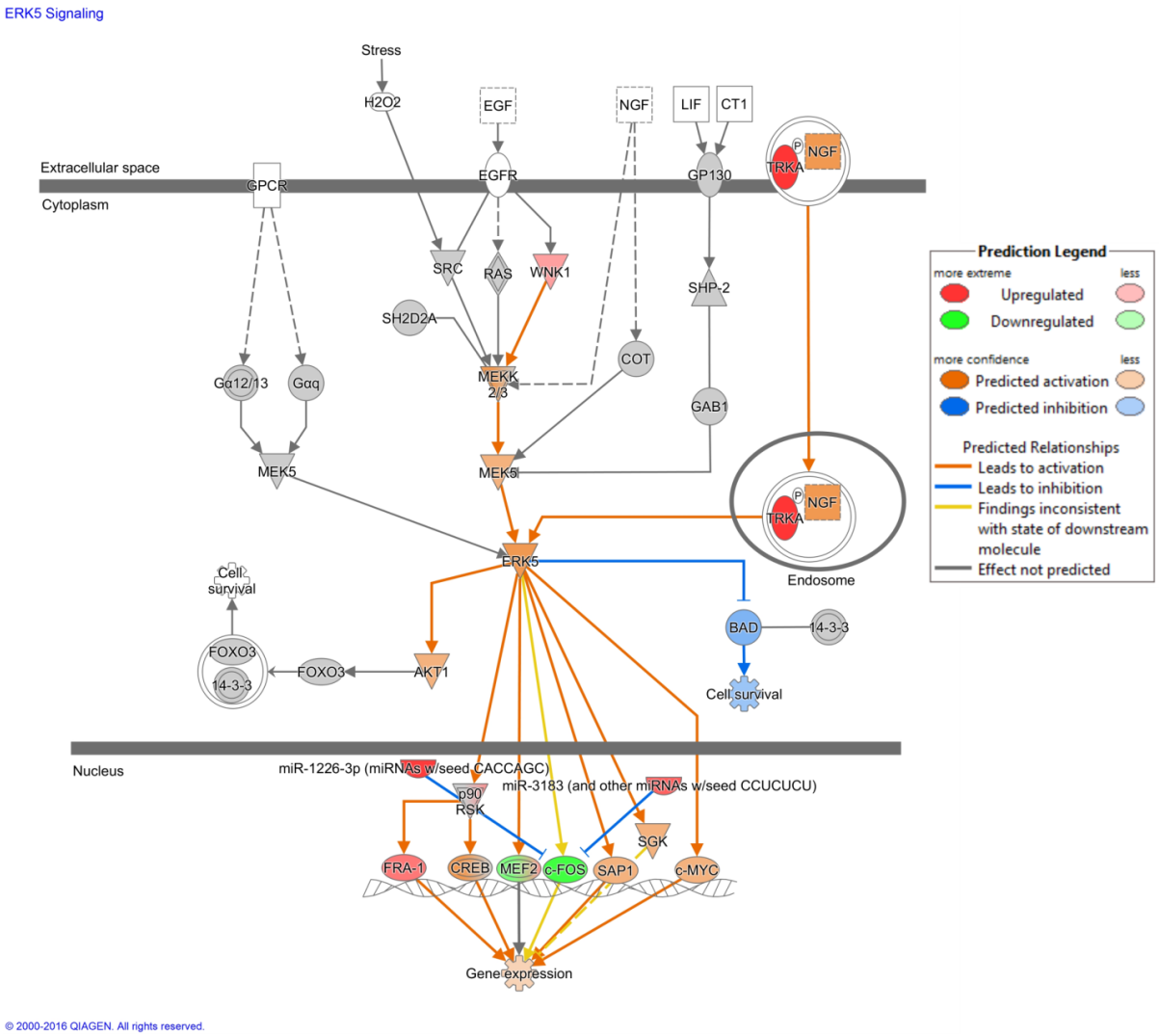
A correlation analysis based on public databases for miRNA-mRNA targeting to reveal miRNA-mRNA pairs expressed in a negative correlation fashion was performed. Information of the following five databases for miRNA-mRNA targeting was

downloaded: miRanda August 2010 (Kozomara and Griffiths-Jones, 2010), miRDB V5 (Wong and Wang, 2014), TarBase V7 (Vergoulis et al., 2012), TargetScan V6.2 (Garcia et al., 2011) (all computationally predicted), and the experimentally validated miRTarBase V4.5 (Chou et al., 2015). A target pair was used for downstream analysis if it was either reported as experimentally validated or predicted by at least two computational predictions. Next, all pairs not showing a significant differential expression for any miRNA or mRNA between infant and childhood cases were discarded. Lastly, for each kept miRNA-mRNA pair the Spearman's correlation was calculated.

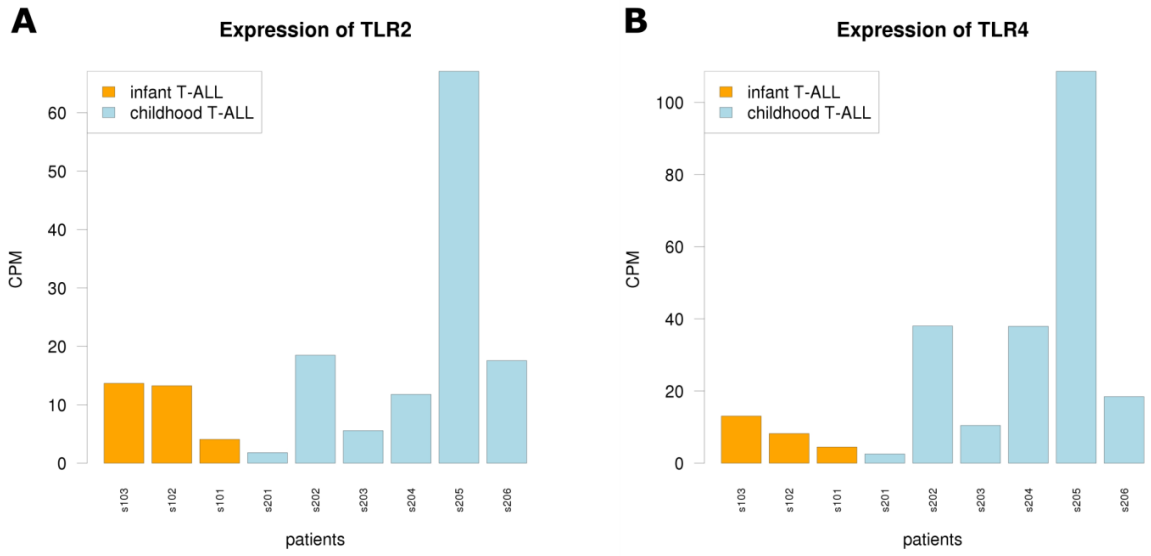
Analysis of Exome Data

BWA version 0.6.1-r104 (Li and Durbin, 2010) was used to align read sequences to the human reference genome (GRCh37). Conversion steps were carried out using Samtools (Li et al., 2009) followed by removal of duplicated reads by mapping positions (Duraku et al., 2013). All reads aligned with at least one insertion/deletion (indels) compared to the genome were locally realigned by GATK (DePristo et al., 2011). All following steps including SNP calling, annotation, and recalibration were also performed by GATK. For recalibration, data from HapMap, OmniArray and dbSNP V135 provided by the Broad Institute were used. The resulting variation calls were annotated by the Variant Effect Predictor (McLaren et al., 2010) using data from Ensembl V70 and imported into an in-house MySQL database to facilitate automatic and manual annotation, reconciliation and data analysis. Predictions for loss of function were provided by PolyPhen2 (Adzhubei et al., 2010) and SIFT (Kumar et al., 2009). Sequence variants within protein coding genes with less than 15% frequency in the 1000 genomes and hapmap project were considered for further analysis.

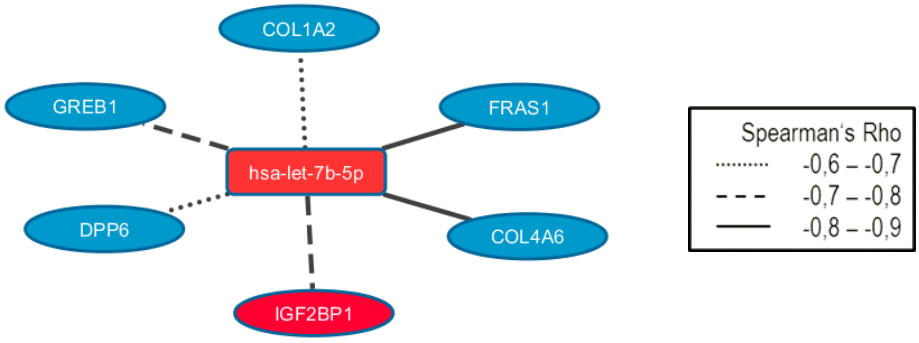
Supplementary Figures



Supplementary Figure 1. Differences in the ERK5 pathway for infant T-ALLs compared with childhood T-ALLs. The pathway map includes differentially expressed genes and miRNAs as well as predictions for activations and inhibitions of genes.



Supplementary Figure 2. Expression of TLR2 (A) and TLR4 (B) in the infant (light blue) and childhood (orange) cases.



Supplementary Figure 3. Regulatory network of hsa-let-7b, negatively correlating with the expression of six target genes.

Supplementary Tables

Supplementary Table 1. PCR Primers for Validation of Mutations (indels or SNVs).

Primer	**Sequence**
NOTCH2 SNP for	CAAGATGCCAAACAACCAGA
NOTCH2 SNP rev	TTTGTTTCCTCATGGCAGTG
NOTCH3 for	GTCAGGGGTCAGAGGAGACA
NOTCH3 rev	AGGTCAGGAGTTCAACAGCA
IL7R for	TGGCTATGCTCAAAATGGTG
IL7R rev	TGAATCCAGTTTGATCTCCTGA
PTEN1for	TCTTTTCAGGCAGGTGTCAA
PTEN1rev	TCTGCAGGAAATCCCATAGC
PTEN2for	ACCGCCAAATTTAATTGCAG
PTEN2rev	GCTTTCTCCCTGTGATTGCT
PTEN3for	TCGTTTTTGACAGTTTGACAGT
PTEN3rev	CACCAATGCCAGAGTAAGCA
PTEN4for	TCCGAAGGGTTTTGCTACAT
PTEN4rev	TCAAGCCCATTCTTTGTTGA
PTEN5for	CCACCCTTTTGACCTTACACA
PTEN5rev	TGCAGTCTGGGCATATCAAA
PTEN6for	GCCTCATCCCAATCAGATGT
PTEN6rev	TGGACTTTTTCAGGACTAGAACG
PTEN7for	CACCTTTAGGATTTTCTGCCTA
PTEN7rev	TGCCAACTTTGGTTTAATGC
KRAS for	GGCACTCAAAGGAAAAATGC
KRAS rev	TGCATTGAGAAACTGAATAGCTG
TLR2delfor	GTAATTCCGGATGGTTGTGC
TLR2delrev	CTTCCTTGGAGAGGCTGATG
TLR2SNPfor	TCCATTGAAAAGAGCCACAA
TLR2SNPrev	TCCTCAAATGACGGTACATCC
TLR4for	TCAGAAACTGCTCGGTCAGA
TLR4rev	GCCCCTGTTAGCACTCAAAA

Supplementary Table 2. Ingenuity Pathway Analysis Showing Significantly Altered Pathways when Searching for Differentially Expressed Genes. The shown pathways are filtered for p-value < 0.05 and are either activated (z-score > 1) or inhibited (z-score < -1) in infant T-ALLs.

Ingenuity canonical pathways	**P-value**	**z-score**
Role of NFAT in Regulation of the Immune Response	0.0005	-1.000
Dendritic Cell Maturation	0.0006	-1.387
ERK5 Signaling	0.0014	1.134
Type I Diabetes Mellitus Signaling	0.0021	-1.342
TREM1 Signaling	0.0026	-1.134
Intrinsic Prothrombin Activation Pathway	0.0062	1.000
Agrin Interactions at Neuromuscular Junction	0.0331	1.000
HMGB1 Signaling	0.0380	-1.134
Gαi Signaling	0.0417	1.134

Supplementary Table 3. Clinically Approved Drugs and Drugs Currently Under Investigation in Clinical Trials for Significantly Upregulated Genes in iT-ALL Cases.

Gene	logFC infants	Drugs
BRD3	1.44	OTX015, CPI-0610, I-BET-762 (GSK525762), TEN-010
KIT	2.03	Imatinib, Dasatinib, Sorafenib, Axitinib, Pazopanib, Cabozantinib, Sunitinib, Ponatinib, Regorafenib, Nilotinib, Cediranib (AZD2171), Dovitinib (TKI-258, CHIR-258), Motesanib (AMG 706), Tandutinib (MLN518), Masitinib (AB1010), Telatinib, Tivozanib, OSI-930, Amuvatinib (MP-470), Midostaurin (PKC412), Quizartinib (AC220), MGCD516, Famitinib
BLK	2.48	Dasatinib, Saracatinib (AZD0530)
FLT1	2.62	Axitinib, Cabozantinib, Nintedanib (BIBF 1120), Regorafenib, Pazopanib, Motesanib (AMG706), MGCD-265, Cediranib (AZD2171), Foretinib (GSK1363089), Lenvatinib (E7080), Tivozanib (AV-951), Lucitanib (E-3810), Famitinib, Linifanib (ABT-869), OSI-930, Cabozantinib
NTRK1	4.07	Ponatinib, Cabozantinib, Entrectinib (RXDX-101), TSR-011, PLX7486, LOXO-101, Dovitinib (TKI-258, CHIR-258), MGCD516, Lestaurtinib (CEP-701), DS-6051b, BMS-754807, Milciclib (PHA-848125), Danusertib (PHA-739358), Lestaurtinib (CEP-701), Cabozantinib
ERBB4	5.96	Lapatinib, Dacomitinib (PF299804, PF299), AC480 (BMS-599626)

Supplementary Table 4. Differentially Expressed miRNAs Between Infant and Childhood T-ALL Cases.

Symbol	logFC	P-value
hsa-let-7b-5p	-2.92	0.0028
hsa-miR-18a-5p	2.38	0.0041
hsa-miR-31-5p	-3.86	0.0036
hsa-miR-30c-5p	2.50	0.0079
hsa-miR-183-5p	2.93	0.0004
hsa-miR-205-5p	-3.71	0.0040
hsa-miR-210-3p	1.62	0.0099
hsa-miR-185-5p	2.15	0.0008
hsa-miR-190a-5p	2.02	0.0019
hsa-miR-200c-3p	1.90	0.0064
hsa-miR-148b-3p	2.49	0.0001
hsa-miR-331-3p	2.65	0.0000
hsa-miR-324-5p	1.96	0.0015
hsa-miR-196b-5p	-5.23	0.0016
hsa-miR-502-5p	2.26	0.0059
hsa-miR-652-3p	2.12	0.0014
hsa-miR-421	1.92	0.0037
hsa-miR-671-5p	1.78	0.0079
hsa-miR-766-3p	3.29	0.0001
hsa-let-7b-3p	-3.61	0.0004
hsa-let-7f-1-3p	-2.52	0.0079
hsa-miR-223-5p	2.35	0.0026
hsa-miR-125b-2-3p	3.53	0.0004
hsa-miR-29c-5p	1.95	0.0092
hsa-miR-874-3p	2.25	0.0015
hsa-miR-1226-3p	3.13	0.0001
hsa-miR-1301-3p	1.83	0.0093
hsa-miR-1180-3p	1.87	0.0033
hsa-miR-1249-3p	1.90	0.0050
hsa-miR-1276	2.47	0.0078
hsa-miR-196b-3p	-5.66	0.0033
hsa-miR-3143	2.04	0.0096
hsa-miR-3186-3p	3.57	0.0001

Supplementary Table 4 cont.

hsa-miR-3620-3p	2.77	0.0033
hsa-miR-3661	2.86	0.0014
hsa-miR-3681-5p	3.36	0.0001
hsa-miR-3909	2.16	0.0014
hsa-miR-3922-3p	2.27	0.0024
hsa-miR-4485-3p	-5.60	0.0008
hsa-miR-4687-5p	2.43	0.0023
hsa-miR-5010-3p	2.34	0.0092
hsa-miR-664b-3p	4.32	0.0000
hsa-miR-5581-3p	2.36	0.0033
hsa-miR-5683	4.02	0.0000
hsa-miR-561-5p	2.90	0.0079
hsa-miR-652-5p	2.18	0.0030
hsa-miR-1306-5p	2.16	0.0038
hsa-miR-6503-5p	-5.17	0.0033
hsa-miR-210-5p	2.61	0.0015
hsa-miR-128-1-5p	2.36	0.0017
hsa-miR-6802-3p	2.51	0.0061
hsa-miR-6803-3p	2.68	0.0006
hsa-miR-6806-3p	2.81	0.0037
hsa-miR-6855-3p	2.26	0.0071
hsa-miR-6769b-3p	2.55	0.0019
hsa-miR-6894-3p	2.57	0.0054
hsa-miR-7706	2.72	0.0001
hsa-miR-128-2-5p	4.42	0.0001

Supplementary References

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, 2010, A method and server for predicting damaging missense mutations: Nat Methods, v. 7, p. 248-9.

Anders, S., P. T. Pyl, and W. Huber, 2014, HTSeq–A Python framework to work with high-throughput sequencing data: bioRxiv.

Chou, C.-H., N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, and S.-J. Tu, 2015, miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database: Nucleic acids research, p. gkv1258.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, 2011, A framework for variation discovery and genotyping using next-generation DNA sequencing data: Nat Genet, v. 43, p. 491-8.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, 2013, STAR: ultrafast universal RNA-seq aligner: Bioinformatics, v. 29, p. 15-21.

Duraku, L. S., M. Hossaini, B. N. Schuttenhelm, J. C. Holstege, M. Baas, T. J. Ruigrok, and E. T. Walbeehm, 2013, Re-innervation patterns by peptidergic Substance-P, non-peptidergic P2X3, and myelinated NF-200 nerve fibers in epidermis and dermis of rats with neuropathic pain: Exp Neurol, v. 241, p. 13-24.

Garcia, D. M., D. Baek, C. Shin, G. W. Bell, A. Grimson, and D. P. Bartel, 2011, Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs: Nature structural & molecular biology, v. 18, p. 1139-1146.

Kozomara, A., and S. Griffiths-Jones, 2010, miRBase: integrating microRNA annotation and deep-sequencing data: Nucleic acids research, p. gkq1027.

Kumar, P., S. Henikoff, and P. C. Ng, 2009, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm: Nat Protoc, v. 4, p. 1073-81.

Li, H., and R. Durbin, 2009, Fast and accurate short read alignment with Burrows-Wheeler transform: Bioinformatics, v. 25, p. 1754-60.

Li, H., and R. Durbin, 2010, Fast and accurate long-read alignment with Burrows-Wheeler transform: Bioinformatics, v. 26, p. 589-95.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing, 2009, The Sequence Alignment/Map format and SAMtools: Bioinformatics, v. 25, p. 2078-9.

Martin, M., 2011, Cutadapt removes adapter sequences from high-throughput sequencing reads: EMBnet. J., v. 17, p. 10-12.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, 2010, Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor: Bioinformatics, v. 26, p. 2069-70.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data: Bioinformatics, v. 26, p. 139-140.

Vergoulis, T., I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, and A. G. Hatzigeorgiou, 2012, TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support: Nucleic acids research, v. 40, p. D222-D229.

Wong, N., and X. Wang, 2014, miRDB: an online resource for microRNA target prediction and functional annotations: Nucleic acids research, p. gku1104.

Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, and L. Gil, 2016, Ensembl 2016: Nucleic acids research, v. 44, p. D710-D716.