



The University  
of Milano  
Bicocca

**THE  
PhD  
PROGRAM  
DIMET**

The PhD Program in  
Translational and Molecular  
Medicine (DIMET)  
is an inter-departmental  
project between the School  
of Medicine and the Faculty  
of Science, organized  
by the University of  
Milano-Bicocca.



**PhD**

**PROGRAM IN TRANSLATIONAL  
AND MOLECULAR MEDICINE**

**DIMET**

**UNIVERSITY OF MILANO-BICOCCA  
SCHOOL OF MEDICINE AND FACULTY OF SCIENCE**

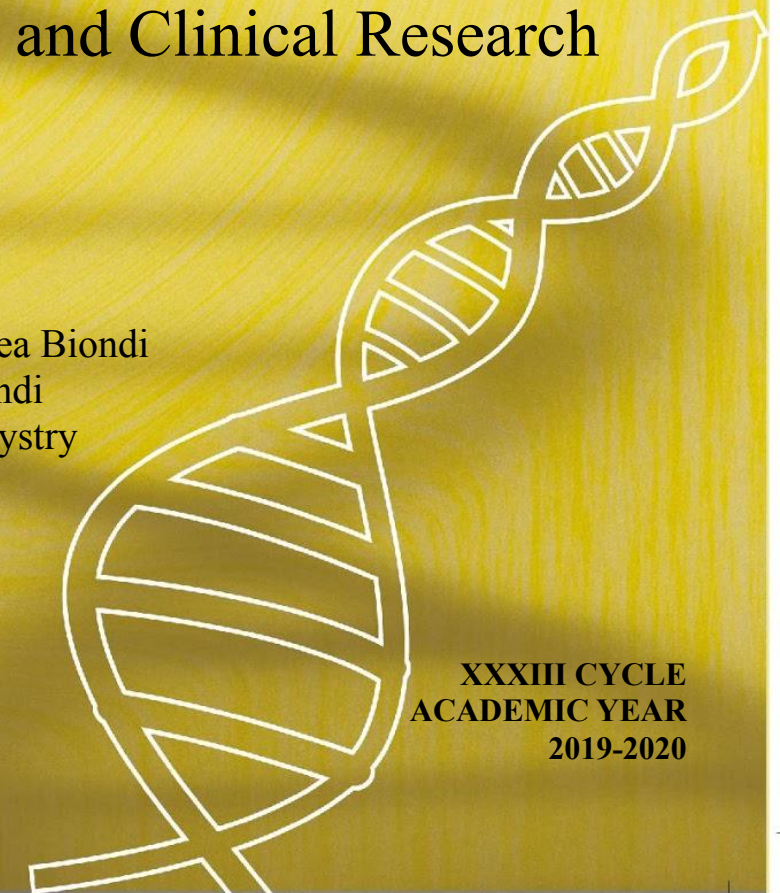
**Application of Modern Data Science to  
Genomics and Clinical Research**

**Coordinator: Prof. Andrea Biondi  
Tutor: Prof. Andrea Biondi  
Co-Tutor: Dr. Vojtech Bystry**

**Dr. Andrea Grioni  
Reg. No. 718282**

**XXXIII CYCLE  
ACADEMIC YEAR  
2019-2020**

**DIMET – XXXIII CYCLE - Dr. Andrea Grioni**





Dimet

Ph.D. Program  
Translational and  
Molecular Medicine



## **PhD Program in Translational and Molecular Medicine**

(XXXIII cycle, academic year 2019/2020)

*University of Milano-Bicocca  
Department of Medicine and Surgery*

### **APPLICATION OF MODERN DATA SCIENCE TO GENOMICS AND CLINICAL RESEARCH**

*Candidate:* Andrea Grioni

*Registration number:* 718282

*Tutor:* Prof. Andrea Biondi, MD

*Co-Tutor:* Dr. Vojtech Bystry, PhD

*Coordinator:* Prof. Andrea Biondi, MD







# Table of Contents

## Contents

Chapter 1: General Introduction.....	7
Part I - Genomics as a Tool for Precision Medicine .....	8
Acute Lymphoblastic Leukemia .....	8
Next-Generation Sequencing .....	11
Bioinformatics .....	14
Informatics.....	18
Part II - Deep Learning for Genomics.....	20
MicroRNA.....	20
Machine Learning.....	21
Machine Learning in Genomics .....	23
Deep Learning .....	24
Convolutional Neural Networks.....	31
Application of Deep Learning in Genomics.....	36
Reference.....	38
Scope of the thesis.....	50
Chapter 2 .....	52
Informatics infrastructure for clinical diagnostics.....	52
Deep Learning for Small RNA analysis.....	54

Chapter 3 .....	58
Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci. ....	58
Chapter 4 .....	69
Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality- NGS validation study. ....	69
Chapter 5 .....	84
Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. ....	84
Chapter 6 .....	96
A novel EP300 mutation associated with Rubinstein-Taybi syndrome type 2 presenting as combined immunodeficiency....	96
Chapter 7 .....	107
First evidence of a paediatric patient with Cornelia de Lange syndrome with acute lymphoblastic leukaemia.....	107
Chapter 8 .....	122
A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia. ....	112
Chapter 9 .....	122

ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data.Bioinformatics. ....	122
Chapter 10 .....	126
High resolution IgH repertoire analysis reveals fetal liver as the likely origin of life-long, innate B lymphopoiesis in humans. ....	126
Conclusion.....	136
Acknowledgements .....	140



## Chapter 1: General Introduction

## Part I - Genomics as a Tool for Precision Medicine

### Acute Lymphoblastic Leukemia

#### Pediatric Acute Lymphoblastic Leukemia

Acute lymphoblastic leukaemia (ALL) is a malignancy characterised by the proliferation of lymphoid cells blocked at an early stage of differentiation that can invade the bone marrow, blood, and extramedullary sites. Approximately 60% of ALL patients are diagnosed before 20 years of age; indeed, ALL is an age-specific malignancy that has the highest incidence in children aged 1–4 years, and then drops sharply through childhood (5–14 years), adolescence, and young adulthood (15–39 years). In the last decades, the 5-year overall survival rate increased from 31% in 1975 to nearly 70% in 2009. This increase can be attributed to the development of fine-tuned clinical protocols and better patient risk stratifications. The introduction of minimal residual disease assessment as part of the clinical diagnostics allowed the evaluation of the effectiveness of the chemotherapy treatment by the quantification of leukaemic cells in the peripheral blood. The degree of the minimal residual disease and genomic biomarkers define the patient's risk group and the specific clinical treatment that should be used. However, the survival rate is still poor for patients who relapse. Analysis of paired diagnosis/relapse ALL samples has shown that the accumulation of new deletions and mutations over time produces new leukaemic clone types (Malard and Mohty, 2020).

The introduction of NGS provided a revolutionary tool for the application of genomics in clinical practice. The application of NGS in clinical diagnostics unveiled new genomic biomarkers, such as the characterisation of new fusion genes and their involvement in the patient's relapse, which can be used to identify new patient subgroups and estimate their survival (Lopes et al., 2019; Mullighan, 2014; Stanulla et al., 2018; Zaliouva et al., 2019). Currently, NGS allows precision medicine, leading to the implementation of personalised treatment for each patient based on their genomic profile (Carrasco-Ramiro et al., 2017; Gulilat et al., 2019; Luh and Yen, 2018; Suwinski et al., 2019).

#### Minima Residual Disease

The introduction of targeted therapies, alongside advances in diagnostic procedures, have improved outcomes for patients with B-ALL (Bassan and Hoelzer, 2011; Hoelzer, 2015). However, despite the substantial proportion (74% to 91%) of patients achieving complete remission (CR), one-third or more will eventually relapse because of the presence of submicroscopic levels of leukaemic cells in the bone marrow (Annino et al., 2002, 2002; Larson et al., 1995). The presence of these remaining cancer cells is known as a minimal residual disease (MRD; alternatively, termed 'measurable residual disease').

MRD is increasingly being used in clinical practice as an independent prognostic marker of the duration of CR and the long-term outcomes of patients with ALL, and also for informing treatment decisions (Bassan et al., 2017; Chen et al., 2015; Gökbuget et al., 2012; Scheuring et al., 2003). In drug development, MRD response has been considered as an

early marker of efficacy in clinical studies, with potential use as a surrogate endpoint in registration studies for accelerated drug Approval (Research, 2020).

### Chromosomal Rearrangements

Fusion genes arise from chromosomal translocations and intrachromosomal rearrangements that mainly disrupt genetic regulators of normal haematopoiesis as well as lymphoid development (e.g., those involving RUNX1 and ETV6) and constitutively activate tyrosine kinases (e.g., ABL1 chimeras) (Hunger and Mullighan, 2015; Inaba et al., 2013).

Fusion genes are hallmarks of ALL that play a pivotal role in leukaemogenesis, and their identification is crucial for patient risk stratification. Common fusion genes in B-lineage ALL include: t(12;21)(p13;q22), encoding ETV6-RUNX1 (TEL-AML); t(1;19)(q23;p13), encoding TCF3-PBX1 (E2A-PBX1); t(9;22)(q34;q11.2), resulting in the formation of the “Philadelphia” chromosome, encoding BCR-ABL1; rearrangements of KMT2A(MLL) at 11q23 to a range of fusion partners; and rearrangements of the cytokine receptor gene *CRLF2* at the pseudoautosomal region 1 (PAR1) at Xp22.3/Yp11.3. Fusion genes correlate with clinical outcome, and are used as biomarkers for patient risk stratification: for example, patients positive for t(12;21)/ETV6-RUNX1 have the most favourable prognosis, whereas t(9;22)/BCR-ABL1, t(1;19)/TCF3-PBX1, and KMT2A-AFF1 correlate with a brief disease latency and have a poor prognosis. Specific drug inhibitors antagonising these fusion proteins provide a more efficient and less toxic tool for disease eradication: for

example, the imatinib tyrosine kinase inhibitor inhibits the oncogenic deregulation caused by the (9;22)/BCR-ABL1 fusion protein.

## Next-Generation Sequencing

### General Workflow

Applications described in this thesis were developed based on the sequencing data generated through the Sequencing by Synthesis (SBS) technique used by the Illumina Sequencing Platforms (see <https://bit.ly/358vk6X>). The Illumina Sequencing workflow is described here:

The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation. Alternatively, “tagmentation” combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified.

Cluster Amplification - For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. When cluster generation is complete, the templates are ready for sequencing (Figure 1.A).

Sequencing - Illumina sequencing uses four fluorescently labelled nucleotides to sequence the tens of millions of clusters on the flow cell surface in parallel. During each sequencing cycle, a single labelled

dNTP is added to the nucleic acid chain. The nucleotide label serves as a “reversible terminator” for polymerisation: after dNTP incorporation, the fluorescent dye is identified through laser excitation and imaging, which is then enzymatically cleaved to allow the next round of incorporation. Base calls are made directly from signal intensity measurements during each cycle (Figure 1.B).

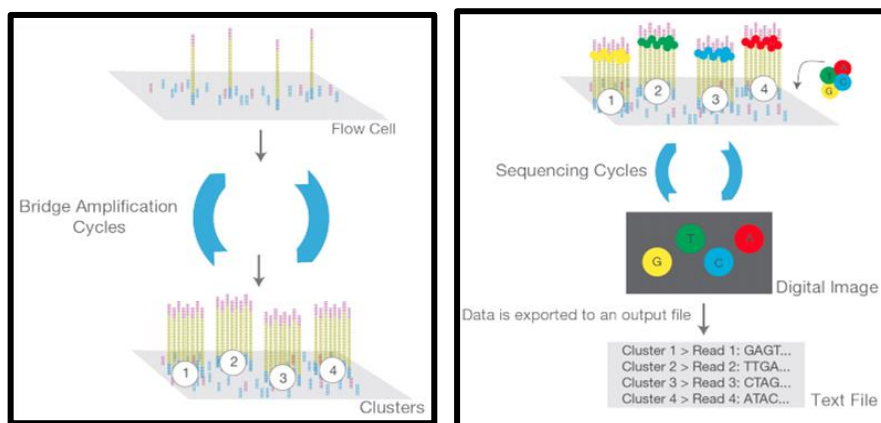


Figure 1. A) Fragments are amplified to generate clusters. B) Sequencing by Synthesis strategy.

### Pair-End Sequencing

A Major advance in NGS technology occurred with the development of paired-end (PE) sequencing (Figure 2). PE sequencing involves the sequencing of both ends of the DNA fragments in a library and the alignment of the forward and reverse sequences as read pairs. In addition to producing twice the number of reads in the same time and with the same effort in library preparation, sequences aligned as read

pairs enable more accurate read alignment and the ability to detect indels, which is not possible with single read data.



*Figure 2. Pair-end sequencing. Information about DNA template nucleotide sequence are provided from sides.*

### Multiplexing

Multiplexing allows large numbers of libraries to be pooled and sequenced simultaneously during a single sequencing run. With multiplexed libraries, unique index sequences are added to each DNA fragment during library preparation such that each read can be identified and sorted before final data analysis. With PE sequencing and multiplexing, NGS has dramatically reduced the sequencing time for multi-sample studies and enabled researchers to go from analysis to data quickly and easily. Multiplexing involves an added layer of complexity, as sequencing reads from pooled libraries need to be identified and sorted computationally in a process called demultiplexing, before final data analysis.

### Library Preparation

In this study, two main next-generation sequencing strategies were used to obtain genomics datasets: amplicon sequencing and target-capture sequencing. Both methodologies have the advantage of restricting the sequencing to pre-selected regions of the genome, thus reducing both

computational analysis time as well as the cost of sequencing while increasing the sequencing depth.

**Amplicon Sequencing** - Amplicon sequencing is a highly targeted approach that enables the analysis of genetic variation in specific genomic regions. This method uses oligonucleotide primers designed to target and amplify pre-selected regions of interest. After amplification, the PCR product is sequenced. This method ensures deep sequencing allowing the identification of rare genomic variants with low abundance.

**Target Capture Sequencing** - Targeted Capture (TC) next-generation sequencing is a type of NGS that focuses on specific areas of the genome. TC relies on the design of biotinylated oligos (baits or probes) whose nucleotide sequence is complementary to the genomic region of interest. Probes bind to the complementary DNA region after DNA fragmentation. Hybrid probes are captured and pooled down by the use of streptavidin beads and magnets, and then sequenced.

## Bioinformatics

### FASTQ file format

Next-generation sequencing output is a set of files containing millions of nucleotide sequences represented by four-letter strings, also known as FASTQ files. By definition, the FASTQ file stores a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity. A FASTQ file containing a single sequence is shown below:



```
@SEQ_ID
TTGGGGTTCAAAGCAGTATCGATCAAATAGTAACATTTGTTCAACTCACAGTTT
+
*((( (**+)) %%%++) (%%%) .1***-+*'! *5CCF>>>>>CCCCCCC65
```

where:

Line1 corresponds to the unique identifier of the read.

Line2 corresponds to the nucleotide sequence of the read.

Line3 corresponds to an empty line that can be used to store read's information.

Line4 corresponds to the base quality of the nucleotide in the same position.

### Preprocessing

FASTQ data preprocessing is a first step in the bioinformatics analysis, which involves the removal of low-quality sequencing data from the original FASTQ file. Low quality sequencing data are represented by: unwanted sequences (e.g. poly-A in RNA sequencing), artificial sequences (e.g. vectors, adapters, primers), join short-overlapping pair-end reads (similar to primer-dimers), low quality reads, PCR duplicates, and contaminations.

The type of FASTQ file preprocessing used is commonly referred to as trimming when it removes low-quality nucleotides at the end of each read or filtering when the entire reads are removed. Preprocessing is a well-established field of research and several open-source software are available online. In this study, preprocessing was performed by the use

of Cutadapt and subsequent quality control was done by FASTQC; the final quality report was built with the software MultiQC (Ewels et al., 2016; Martin, 2011).

### Genome Alignment

Once high-quality data are obtained from preprocessing, the next step is to map the reads to the genome or transcriptome reference of the sequenced organism. In this study, I referred to human and mouse reference genomes, which are freely available on UCSC and ENSEMBL web services. Several algorithms can be used to align reads to a reference genome. The main feature affecting the choice of the aligner is the biological material used for the generation of the FASTQ files, thus RNA or DNA. In the first case, splicing-aware aligners, such as STAR and HISAT2 (Dobin et al., 2013; Kim et al., 2019), are more suitable than non-splicing-aware aligners. On the contrary, non-splicing-aware aligners such as BWA or Minimap2 will perform better for DNA datasets (Li, 2018, p. 2; Li and Durbin, 2009).

### Somatic Single Nucleotide Variant Calling

Next-generation sequencing is by far the most promising technology for *de novo* mutation detections. Theoretically, all mutations, regardless of the variant allele frequency (VAF) or genomic region, can be observed given enough read depth (coverage). However, due to the non-marginal amount of background noise, the bioinformatics process of identification and calling of somatic single nucleotide variants (SNV) is a non-trivial complex task. As described above, preprocessing of FASTQ files highly improves the final results since it helps to partially

reduce the background noise; however, it needs to be paired with proper algorithms for SNV calling. Modern variant calling algorithms such as Mutect2 or strelka2 use multiple methods to differentiate real variants from background noise, such as read local assembly and realignment. The subsequent hitch in somatic SNV detection is to separate germline and somatic variants. Two main strategies exist: matching normal-tumour samples and single-sample variant calling. The first and far superior strategy relies on a pair-wise comparison between variants identified from the analysis of normal and tumour tissues derived from the same patient. This type of analysis depends on the availability of sequenced normal tissue data. This is the case of solid tumours analysis, where it is possible to isolate both tumour and normal tissues from the patient. However, pairwise tissue comparison is not always possible due to the limited availability of patient's samples or the technical difficulties involved in the isolation of tumour tissue. In this scenario, the analysis can be performed by comparing the patient's sample to the human reference genome. The identified variants can be compared with known variants in publicly available databases, such as ClinVar or SNPdb (Landrum et al., 2014; Sherry et al., 2001), filtered based on VAF or common variants in multi-sample studies to identify low frequency variants (<1%) potentially associated with the tumour or disease.

As described in the methods section, we used the best practice of Genome Analysis Toolkit (GATK) for the identification of single-sample variants (McKenna et al., 2010).

## Structural Variants

Genomic structural variations (SVs) are generally defined as deletions (DELs), insertions (INSs), duplications (DUPS), inversions (INVs), and translocations (TRAs) of at least 50 bp in size. SVs are often considered separately from small variants, including single nucleotide variants (SNVs) and short insertions, and deletions (indels), as they are often formed by distinct mechanisms. This dissertation describes the implementation of an analysis pipeline for the identification of chromosomal rearrangements from NGS short-reads pair-end sequencing. Two main strategies exist for the identification of chromosomal rearrangements from whole-genome or RNA sequencing. The first strategy relies on the identification of reads spanning the breakpoint of the chromosomal translocation. The breakpoint can be identified by the detection of soft-clip reads, meaning reads partially aligned to the genome. The second method relies on the identification of discordant pair-end reads. A pair-end read is called discordant when the distance and/or location between the two mates do not match the expected distance based on the insert-size distribution. This is also the case when the mates are aligned on different chromosomes (Kumar et al., 2016).

## Informatics

### Programming Languages for Bioinformatics

Bioinformatics is a wide research field ranging from protein structure prediction to DNA sequence analysis. In the field of genomics and transcriptomics, the most popular programming languages are R and

Python. During this study, I developed high-level expertise in both the programming languages R and Python. R was designed by statisticians and was specialised for statistical computing, and thus is known as the lingua franca of statistics. As technology improves, the data collected by companies or research institutions have become increasingly complex, and R has been adopted by many as the language of choice. R has been used to generate graphs and visualisations with the use of the Tidyverse package. Moreover, the R-Shiny package is used to build interactive web-applications that allow biologists and medical doctors to run and visualise bioinformatics pipelines.

Python is a high-level and versatile language because of its clear syntax and simple text manipulation. The clear syntax of Python has earned it the name executable pseudo-code. The default installation of Python already consists of high-level data types such as tuple, list, sets, and dictionaries. In addition, many machine learning frameworks have recently been developed in Python. Python is also highly popular. Many libraries are available for the analysis and extraction of information from NGS datasets, as well as for the development of DL models with the use of Tensorflow and Keras API.

Bioinformatics pipelines have been written and controlled through Unix-Shell and the programming language Bash. Therefore, during the study, I gained advanced experience as a Linux Server administrator and manager, web-developer, and database maintainer.

## Part II - Deep Learning for Genomics

### MicroRNA

MicroRNAs are short non-coding RNA molecules approximately 22 nucleotides in length that play important regulatory roles in animals and plants by regulating gene expression. MicroRNAs interact with Argonaute (AGO) proteins and guide them to target sites located in the 3' untranslated region (UTR) of messenger RNAs. The miRNA-loaded AGO forms the miRNA-induced silencing complex (miRISC), which promotes translational repression and degradation of targeted mRNAs (O'Brien et al., 2018).

MicroRNAs function in post-transcriptional regulation of target gene expression. One miRNA can simultaneously target several genes. Recent studies have shifted our understanding on how miRNAs interact with their targets, which include not only mRNAs but also long noncoding RNAs (lncRNAs), pseudogenes, and circular RNAs (circRNAs). With the ability to interact with multiple target genes, miRNAs have been shown to influence many important biological processes such as cell growth, tissue differentiation, cell proliferation, embryonic development, and apoptosis. Since the discovery of the first miRNA *lin-4* in 1993, 48,885 mature miRNAs in 271 species have been identified and deposited into the miRBase gold standard central repository (Chen et al., 2019).

MicroRNAs are involved in almost every cellular process from development and cell differentiation to homeostasis; deletions of the fundamental miRNA biogenesis factors *Dicer* and *Drosha* are lethal in

mouse embryos. Dereglulation of miRNA functions is associated with numerous diseases, particularly cancer: miRNAs can be both oncogenes and tumour suppressors, although overall downregulation of miRNA expression is a hallmark of cancer (Bhaskaran and Mohan, 2014).

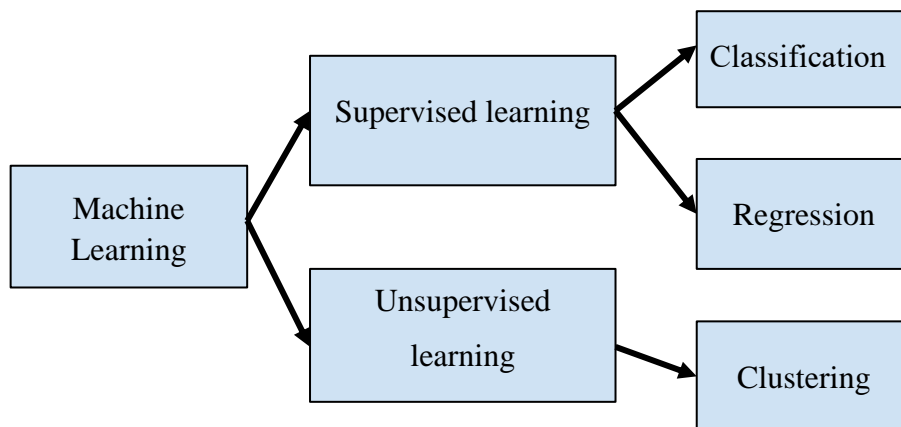
The introduction of NGS allowed a broader identification and study of microRNAs. Several purpose-built assays have been developed for the characterisation of microRNAs and for the identification of their target messenger RNA, such as high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) and Cross-linking, ligation, and sequencing of hybrids (CLASH) (Helwak et al., 2013; Kudla et al., 2011; Licatalosi et al., 2008; Ule et al., 2003). Information collected through experimental biology has been used to implement curated databases, for example, mirBase, TarBase, and TargetScan (Agarwal et al., n.d.; Griffiths-Jones et al., 2006; Karagkouni et al., 2018). Since then, bioinformatics tools have been developed for the analysis and interpretation of these vast amounts of data. In the last decade, the bioinformatics community released dedicated software and pipelines for the study of microRNA biogenesis, functions, target predictions, and microRNA editing. This software can be a set of rules hard-coded through conditional statements or more recently statistical models developed by the application of machine learning approaches, such as linear regression or supported vector machines.

## Machine Learning

Machine learning is actively used today, for example in recommendation systems in e-commerce, in the detection of bank

frauds or spam emails. Machine learning uses statistics to implement computer algorithms (models) that improve performances automatically through experience. In classification problems, the model predicts in which class specific data should fall into. Another task in machine learning is regression. Regression is the prediction of a numeric value. Classification and regression are examples of supervised learning. This set of problems is known as supervised because we are instructing the algorithm what to predict.

The opposite of supervised learning is a set of tasks known as



*Figure 3. Machine learning can be organised in supervised and unsupervised methods. Supervised requires labelled data and it is used to solve classification and regression problems. Unsupervised learning is used to cluster unlabelled data.*

unsupervised learning. In unsupervised learning, there is no label or target value given for the data. A task of unsupervised learning is clustering where group similar items together. In unsupervised learning, we may also want to find statistical values that describe the data. This is known as density estimation. Another task of unsupervised learning is the reduction of the many features of the data to a small number of



them data from many features to a small number in order to properly visualise them in two or three dimensions.

## Machine Learning in Genomics

Genomics data are too large and complex to be mined solely by visual investigation of pairwise correlations. Instead, analytical tools are required to support the discovery of hidden relationships between data and observations. Traditional bioinformatics rely on hard-coded algorithms that require time and a great deal of effort to be developed. Unlike some algorithms, machine learning algorithms are designed to automatically detect patterns in data. Hence, machine learning algorithms are suited to data-driven researches, and in particular, genomics (Eraslan et al., 2019). In theory, it would be possible to model any biological system by the use of proper machine learning techniques. However, the ability of a machine-learning algorithm to model a biological system strongly depends on the quality of the input data and their representation. This preprocessing of the input data consists of manual extraction of features that characterise the biological system to be modelled. For example, the development of a machine learning model to predict pre-micro RNA by genome scanning will require the selection of hand-picked features, such as percentage of GC content, RNA secondary structure, entropy, etc. This process may require the manual selection of hundreds of features for thousands of examples, which is not always feasible. Another example is the classification of a tumour as malignant or benign based on a fluorescent microscopy image; first, a preprocessing algorithm could detect cells, identify the cell type and generate a list of cell counts for each cell type. A machine

learning model would then take these estimated cell counts as input features to classify the tumour. A central issue of these machine learning algorithms is that the classification performance is heavily dependent on the quality and the relevance of the human-selected features. Deep learning, a subdiscipline of machine learning, addresses this issue by embedding the computation of features into the machine learning model itself to yield end-to-end models.

## Deep Learning

Deep learning is a generic name that refers to the recent advances in artificial neural networks (LeCun et al., 2015; McCulloch and Pitts, 1943). The building block of an artificial neural network is an artificial neuron.

### Artificial Neuron

An artificial neuron is a mathematical model that takes as input a vector of real values  $(V_1, V_{n-1})$  and computes the weighted average of these values followed by an activation function. The weights  $(W_1, W_{n-1})$  are the parameters of the model that are learned during the training process.

A schematic representation of an artificial neuron is shown below:

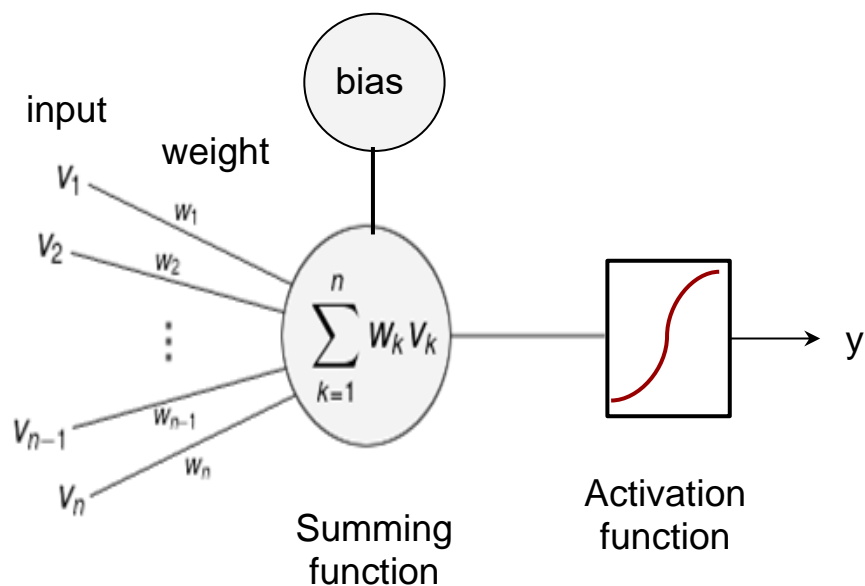


Figure 4. A schematic representation of an artificial neuron. Each input  $V$  is multiplied to its corresponding weight, then the neuron computes the weighted average. The weighted average is input to an activation function.

In this representation,  $v$  indicates input values, and  $w$  the corresponding weight for each input. Each input  $v$  is multiplied with the corresponding weight, and then the sum of these values is provided as input to an activation function. If the value is greater than an established threshold the neuron generates an output (the neuron fires). The bias is an arbitrary value (generally equal to 1) that is subtracted after the summing function and is used to ‘silence’ the neuron. This model is known as perceptron and was proposed by Rosenblatt in 1958 (Rosenblatt, 1958).

#### Train an Artificial Neuron

Training an artificial neuron such as the perceptron is all about finding a set of weights through which the classification task is successful (Figure 5). The weights are known as parameters. To find

the best set of parameters, we iterate the classification task over all samples from the training dataset. Since the training dataset is provided with the correct label for each sample, we can then define a loss function (or cost function) that calculates how distant the model prediction is

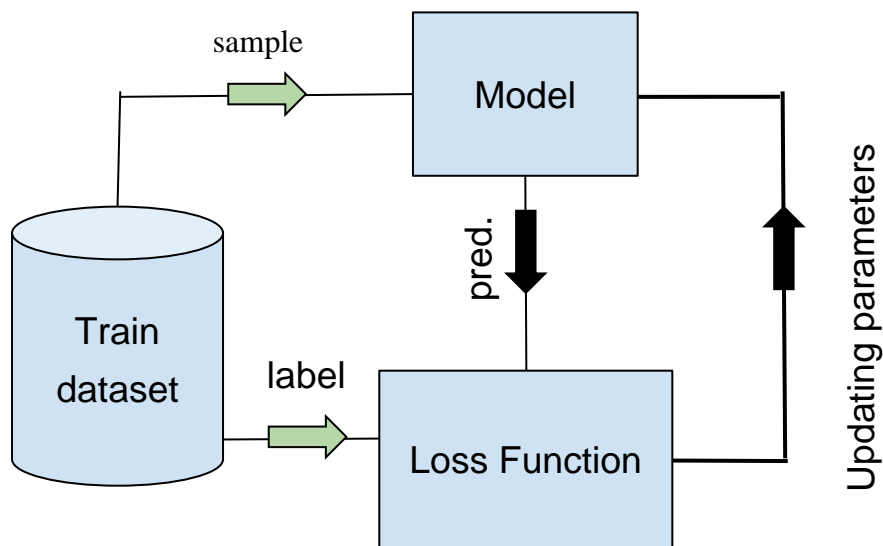


Figure 5. Training scheme: the model evaluates each input sample to generate a prediction. This prediction is compared to the true sample label. A loss function calculates the 'distance' between the real value and the prediction.

from the real label (true value). Based on the loss function, we can update the parameters to improve the prediction.

### Artificial Neural Networks

The output of an artificial neuron can be the input for another. Since artificial neurons are extremely versatile, they can be stacked together into layers, also called artificial neural networks (ANN). A simple ANN is represented in the figure below, where a circle defines a single artificial neuron (Figure 6).

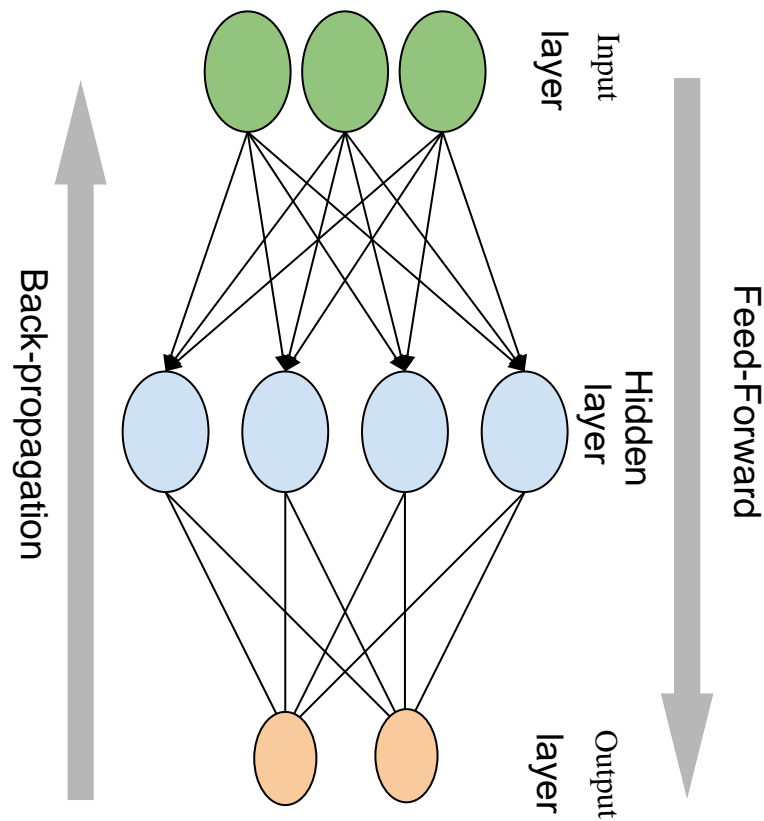


Figure 6. Vanilla Neural Network: the input of the network is the first layer, followed by a hidden layer of 4 neurons, and a final output layer of 2 neurons.

This simple - vanilla - ANN is defined by:

- The input layer is the first layer of an ANN that receives the input information in the form of various texts, numbers, audio files, image pixels, etc.
- In the middle of the ANN model are the hidden layers. There can be a single hidden layer, as in the case of a perceptron, or multiple hidden layers. These hidden layers perform various types of mathematical computation on the input data and recognise the patterns that are part of.
- In the output layer, we received the probability of the prediction.

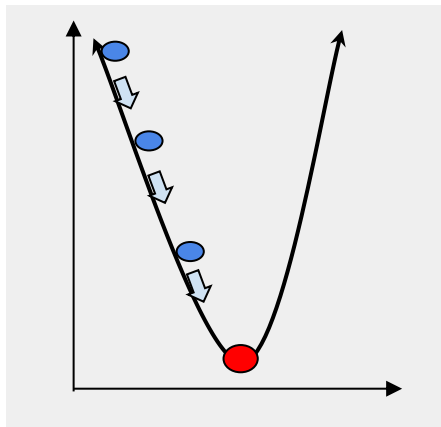
A neural network layer can have an infinite number of neurons per layer as well as an infinite number of hidden layers. This architecture is also known as a dense layer or fully connected layers. The ANN performs two main actions:

- Feed-forward means that the flow of information occurs only in one direction. That is, feed-forward connections represent information flow from one neuron to another where the data being transferred is the computed neuronal activation at the current time step. There are no feedback loops present in this neural network.
- Backpropagation, short for "backward propagation of errors," is an algorithm for supervised learning of artificial neural networks using gradient descent. Gradient descent is an optimisation algorithm used to minimise a function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. Given an artificial neural network and an error function, the backpropagation method calculates the gradient of the error function with respect to the neural network's weights. The "backward" part of the name stems from the fact that calculation of the gradient proceeds backward through the network, with the gradient of the final layer of weights being calculated first and the gradient of the first layer of weights being calculated last. Partial computations of the gradient from one layer are reused in the computation of the gradient for the previous layer. This backward flow of the error information allows for efficient computation of the

gradient at each layer versus the naive approach of calculating the gradient of each layer separately.

### Gradient Descent

Machine learning requires the finding of the correct parameters (weights,  $w$ ) to minimise the loss function (Loss). We can represent the loss and parameters in a cartesian system, where on the y-axis we define the loss and on the x-axis the parameters  $w$  (Figure 7).



*Figure 7. Schematic representation of gradient descent.*

A neural network model is first initialised by randomising the value of parameters. Then, the first iteration of predictions is performed, and a corresponding loss function is calculated. On the cartesian system, our model is now the blue ball on the top-left corner, and we want to reach the bottom of the graph to minimise the loss function (red ball). For this reason, the parameters are updated following the direction opposite to the gradient (negative gradient) and using backpropagation. Next, we recalculate the negative gradient (passing in the coordinates of our new point) and take another step in the direction it specifies. We continue

this process iteratively until we reach the bottom of our graph, or to a point where we can no longer move downhill - a local minimum.

### Neural Networks Architectures

The input of a neural network is typically a matrix of real values. In genomics, the input might be a DNA sequence, in which the nucleotides A, C, G, and T are encoded as  $[1,0,0,0]$ ,  $[0,1,0,0]$ ,  $[0,0,1,0]$ , and  $[0,0,0,1]$ . Neurons that directly read the data input are called the input layer. The following neurons are referred to as hidden layers. The output of a neural network is the prediction of interest, e.g. whether the input DNA is a microRNA. There are three common families of architectures for connecting neurons into a network: feed-forward, convolutional, and recurrent. Feed-forward is the simplest architecture. Every neuron of layer  $i$  is connected only to neurons of layer  $i + 1$ , and all the connection edges can have different weights. In a convolutional neural network (CNN), a neuron is scanned across the input matrix, and at each position of the input, the CNN computes the locally weighted sum and produces an output value. This procedure is highly similar to taking the position weight matrix of a motif and scanning it across the DNA sequence. CNNs are useful in settings in which some spatially invariant patterns in the input are expected. Recurrent neural networks (RNN) are designed for sequential or time-series data. At each point in the sequence, a neural network, which could be feed-forward or convolutional, is applied to generate an internal signal, which is also fed to the next step of the RNN. Hidden layers of the RNN can be viewed as memory states that retain information from the sequence previously observed and are updated at each time step (Zou et al., 2019).



## Convolutional Neural Networks

Convolution is a neural network architecture that takes inspiration from the study of human vision.

### Human Vision

Inspired by how human vision functions, layers of a convolutional network have neurons arranged in three dimensions, so layers have a width, height, and depth. The neurons in a convolutional layer are only connected to a small, local region of the preceding layer. A convolutional layer's function can be expressed simply: it processes a three-dimensional volume of information to produce a new three-dimensional volume of information.

An intuition about human vision came from David Hubel and Torsten Wiesel, who discovered that parts of the visual cortex that are responsible for detecting edges. In 1959, they inserted electrodes into the brain of a cat and projected black-and-white patterns on the screen. They found that some neurons fired only when there were vertical lines, others when there were horizontal lines, and still others when the lines were at angles. Further study determined that the visual cortex was organised in layers (Figure 8). Each layer is responsible for building on the features detected in the previous layers from lines to contours, to shapes, to entire objects.

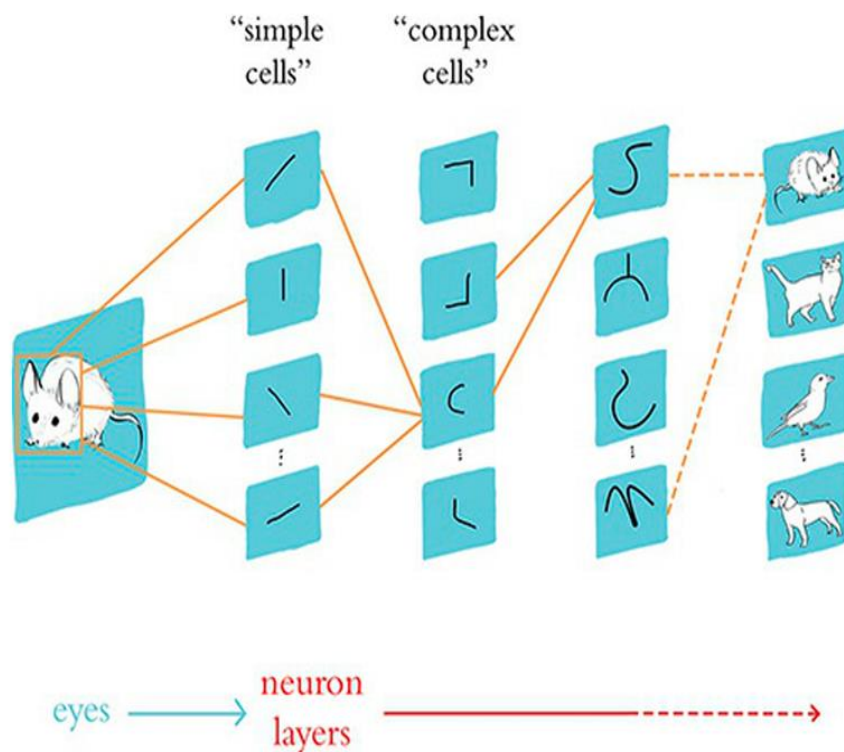


Figure 8. Schematic representation of how consecutive layers of biological neurons represent visual information in the brain. Cells of the primary cortex detect simple features, such as vertical and horizontal lines. Simple features are assembled into more complex representation by cells in the deeper layers (complex cells).

Furthermore, within a layer of the visual cortex, the same feature detectors are replicated over the whole area to detect features in all parts of an image. These ideas significantly impacted the design of convolutional neural networks (“Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence (Addison-Wesley Data & Analytics Series) 1, Krohn, Jon, Beyleveld, Grant, Bassens, Aglaé, eBook - Amazon.com,” n.d.; *Fundamentals of Deep Learning*, 2017).

### Filters and Feature Maps

The first concept was that of a filter. A filter is essentially a feature detector, whose output is called an activation map or feature map. Considering that in the image below, we want to detect vertical lines (Figure 9). An image (left figure) is an array of values (right figure), each of them corresponding to the pixel intensity.

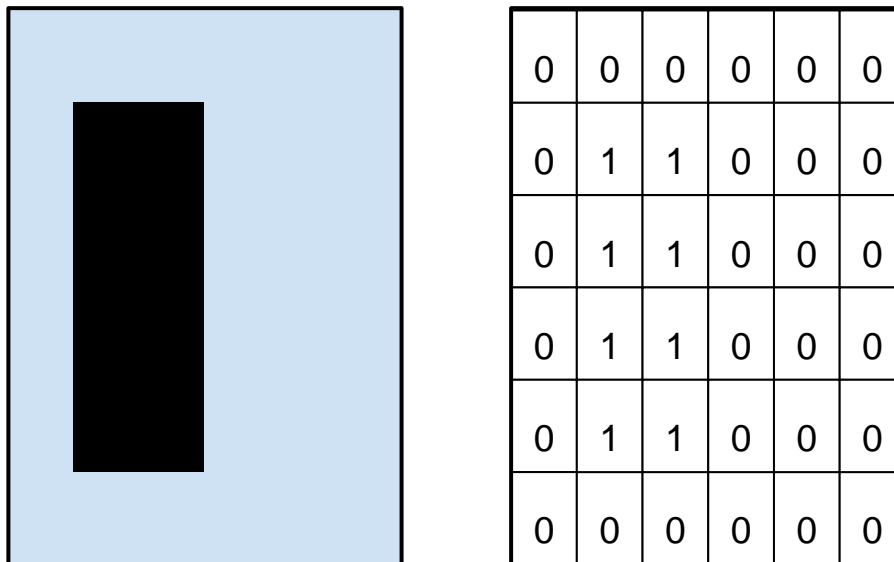


Figure 9. left: real image, right: the same image can be represent as an array of values.

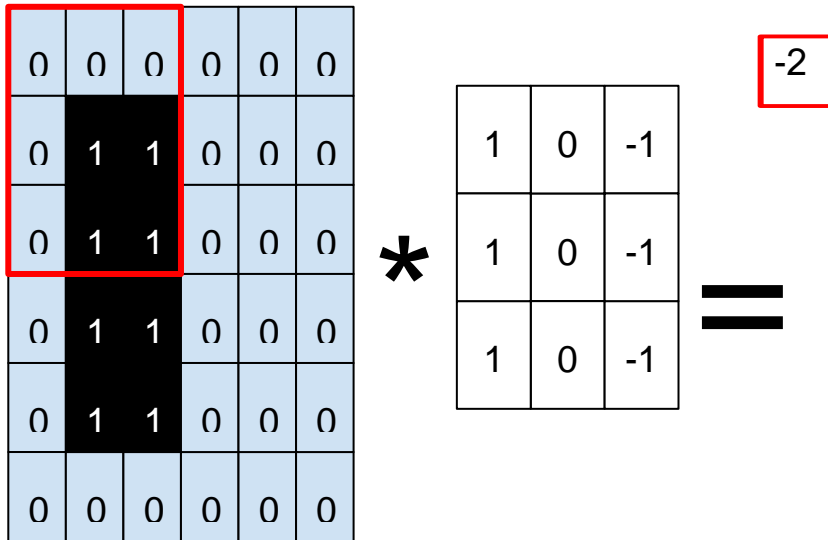
To detect vertical lines, we need to define a filter that maximises its value when overlapping a vertical line (Figure 10).

1	0	-1
1	0	-1
1	0	-1

*Figure 10. A filter that can detect vertical lines.*

We scan the filter along the image to obtain a feature map, or activation map, for the filter (Figure 11). The activation map is generated by the dot product of the image array and the filter. Images are arrays of numbers, in our case black is equal to 1 and white to 0. At the end of the scanning process, a feature map is generated for the corresponding filter.

$$0*1 + 0*1 + 0*1 + 0*0 + 1*0 + 1*0 + 0*-1 + 1*-1 +$$



$$0*1 + 1*1 + 1*1 + 0*0 + 1*0 + 1*0 + 0*-1 + 0*-1 +$$

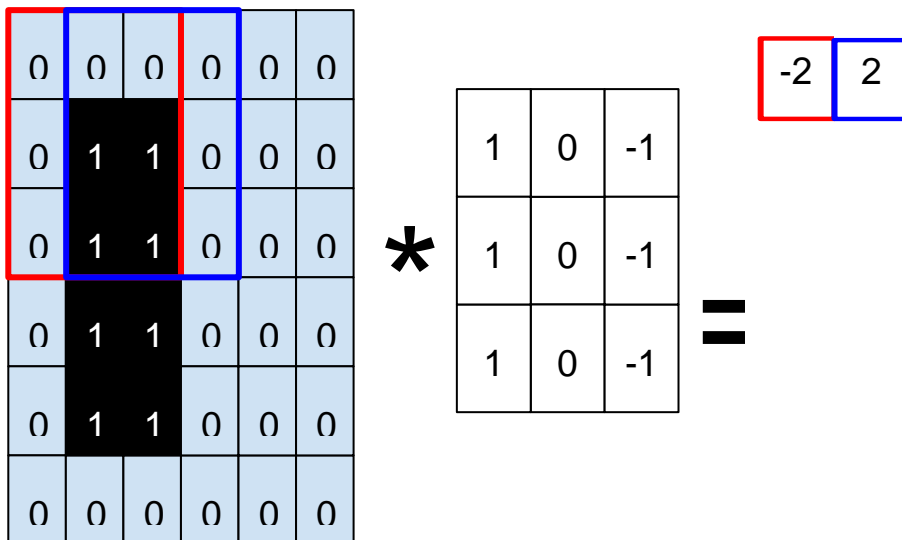


Figure 11. A filter scanning the entire picture. Regions containing vertical lines are maximized.

This operation is a convolution. We take a filter and we multiply it over the entire area of an input image. We can stack several convolution layers one after the other to then provide the convolution net output to a dense layer. The dense layer is known as a fully connected layer, and it is made of neurons that are used for the final classification task.

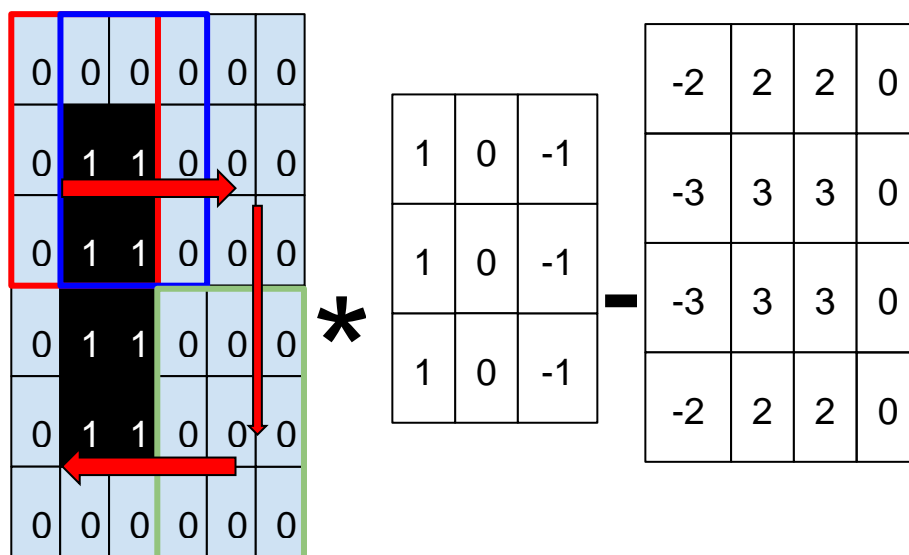


Figure 12. The convolution of the filter along the whole image generates a feature map.

## Application of Deep Learning in Genomics

Deep learning has been successfully implemented in areas such as image recognition or robotics (e.g., self-driving cars) and is most useful when large amounts of data are available. In this respect, using DL as a tool in the field of genomics is entirely apt. Although it is still in somewhat early stages, DL in genomics has the potential to inform

fields such as cancer diagnosis and treatment, clinical genetics, crop improvement, epidemiology and public health, population genetics, evolutionary or phylogenetic analyses, and functional genomics (Ching et al., 2018; Eraslan et al., 2019; Esteva et al., 2019; Jones, 2019). Functional genomics is a field of molecular biology that attempts to describe gene functions and interactions. The application of DL to functional genomics has made the most inroads to date (“Deep learning for genomics,” 2019). The availability of vast troves of data of various types (DNA, RNA, methylation, chromatin accessibility, histone modifications, chromosome interactions, and so forth) ensures that there are enough training datasets to build accurate prediction models related to gene expression, genomic regulation, or variant interpretation. This dissertation is meant to present the development of strategies for the application of DL and convolutional neural network architectures for the identification of small non-coding RNA elements from genome scanning. For this purpose, we used several available database information such as genome sequences, conservation tracks, RNA secondary structure predictors as well as new techniques for the training of neural network models.

## Reference

- 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R., 2015. A global reference for human genetic variation. *Nature* 526, 68–74. <https://doi.org/10.1038/nature15393>
- Agarwal, V., Bell, G.W., Nam, J.-W., Bartel, D.P., n.d. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4. <https://doi.org/10.7554/eLife.05005>
- Annino, L., Vegna, M.L., Camera, A., Specchia, G., Visani, G., Fioritoni, G., Ferrara, F., Peta, A., Ciolli, S., Deplano, W., Fabbiano, F., Sica, S., Di Raimondo, F., Cascavilla, N., Tabilio, A., Leoni, P., Invernizzi, R., Baccarani, M., Rotoli, B., Amadori, S., Mandelli, F., GIMEMA Group, 2002. Treatment of adult acute lymphoblastic leukemia (ALL): long-term follow-up of the GIMEMA ALL 0288 randomized study. *Blood* 99, 863–871. <https://doi.org/10.1182/blood.v99.3.863>
- Bassan, R., Hoelzer, D., 2011. Modern therapy of acute lymphoblastic leukemia. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 29, 532–543. <https://doi.org/10.1200/JCO.2010.30.1382>
- Bassan, R., Intermesoli, T., Scattolin, A., Viero, P., Maino, E., Sancetta, R., Carobolante, F., Gianni, F., Stefanoni, P., Tosi, M., Spinelli, O., Rambaldi, A., 2017. Minimal Residual Disease Assessment and Risk-based Therapy in Acute Lymphoblastic Leukemia. *Clin. Lymphoma Myeloma Leuk.* 17S, S2–S9. <https://doi.org/10.1016/j.clml.2017.02.019>
- Bhaskaran, M., Mohan, M., 2014. MicroRNAs: History, Biogenesis, and Their Evolving Role in Animal Development and Disease. *Vet. Pathol.* 51, 759–774. <https://doi.org/10.1177/0300985813502820>
- Brüggemann, M., Kotrová, M., Knecht, H., Bartram, J., Boudjoghra, M., Bystry, V., Fazio, G., Froňková, E., Giraud, M., Grioni, A., Hancock, J., Herrmann, D., Jiménez, C., Krejci, A., Moppett, J., Reigl, T., Salson, M., Scheijen, B., Schwarz, M., Songia, S., Svaton, M., van Dongen, J.J.M., Villarese, P., Wakeman, S., Wright, G., Cazzaniga, G., Davi, F., García-



- Sanz, R., Gonzalez, D., Groenen, P.J.T.A., Hummel, M., Macintyre, E.A., Stamatopoulos, K., Pott, C., Trka, J., Darzentas, N., Langerak, A.W., EuroClonality-NGS working group, 2019. Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. *Leukemia* 33, 2241–2253. <https://doi.org/10.1038/s41375-019-0496-7>
- Bystry, V., Reigl, T., Krejci, A., Demko, M., Hanakova, B., Grioni, A., Knecht, H., Schlitt, M., Dreger, P., Sellner, L., Herrmann, D., Pingeon, M., Boudjoghra, M., Rijntjes, J., Pott, C., Langerak, A.W., Groenen, P.J.T.A., Davi, F., Brüggemann, M., Darzentas, N., EuroClonality-NGS, 2017. ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. *Bioinforma. Oxf. Engl.* 33, 435–437. <https://doi.org/10.1093/bioinformatics/btw634>
- Carrasco-Ramiro, F., Peiró-Pastor, R., Aguado, B., 2017. Human genomics projects and precision medicine. *Gene Ther.* 24, 551–561. <https://doi.org/10.1038/gt.2017.77>
- Cauchy, M.A., n.d. Méthode générale pour la résolution des systèmes d'équations simultanées 3.
- Chen, L., Heikkinen, L., Wang, C., Yang, Y., Sun, H., Wong, G., 2019. Trends in the development of miRNA bioinformatics tools. *Brief. Bioinform.* 20, 1836–1852. <https://doi.org/10.1093/bib/bby054>
- Chen, X., Xie, H., Wood, B.L., Walter, R.B., Pagel, J.M., Becker, P.S., Sandhu, V.K., Abkowitz, J.L., Appelbaum, F.R., Estey, E.H., 2015. Relation of clinical response and minimal residual disease and their prognostic impact on outcome in acute myeloid leukemia. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 33, 1258–1264. <https://doi.org/10.1200/JCO.2014.58.3518>
- Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., DeCaprio, D., Qi, Y.,

- Kundaje, A., Peng, Y., Wiley, L.K., Segler, M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., Greene, C.S., 2018. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15, 20170387. <https://doi.org/10.1098/rsif.2017.0387>
- Deep learning for genomics, 2019. . *Nat. Genet.* 51, 1–1. <https://doi.org/10.1038/s41588-018-0328-0>
- Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence (Addison-Wesley Data & Analytics Series) 1, Krohn, Jon, Beyleveld, Grant, Bassens, Aglaé, eBook - Amazon.com [WWW Document], n.d. URL [https://www.amazon.com/Deep-Learning-Illustrated-Intelligence-Addison-Wesley-ebook/dp/B07W585JGG/ref=sr\\_1\\_1?crd=15H1Z67TNO1OG&dchild=1&keywords=deep+learning+illustrated&qid=1587918043&s=books&sprefix=deep+learning+ill%2Cstripbooks-intl-ship%2C249&sr=1-1](https://www.amazon.com/Deep-Learning-Illustrated-Intelligence-Addison-Wesley-ebook/dp/B07W585JGG/ref=sr_1_1?crd=15H1Z67TNO1OG&dchild=1&keywords=deep+learning+illustrated&qid=1587918043&s=books&sprefix=deep+learning+ill%2Cstripbooks-intl-ship%2C249&sr=1-1) (accessed 4.26.20).
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* 29, 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Eraslan, G., Avsec, Ž., Gagneur, J., Theis, F.J., 2019. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20, 389–403. <https://doi.org/10.1038/s41576-019-0122-6>
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. <https://doi.org/10.1038/s41591-018-0316-z>
- Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Fazio, G., Massa, V., Grioni, A., Bystry, V., Rigamonti, S., Saitta, C., Galbiati, M., Rizzari, C., Consarino, C., Biondi, A., Selicorni, A., Cazzaniga, G., 2019. First evidence of a paediatric patient with Cornelia de Lange syndrome with acute lymphoblastic leukaemia. *J. Clin. Pathol.* 72, 558–561.

<https://doi.org/10.1136/jclinpath-2019-205707>

- Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms, 1 edition. ed, 2017. . O'Reilly Media, Sebastopol, CA.
- Georgakilas, G.K., Grioni, A., Liakos, K.G., Malanikova, E., Plessas, F.C., Alexiou, P., 2019. MuStARD: Deep Learning for intra- and inter-species scanning of functional genomic patterns (preprint). *Bioinformatics*. <https://doi.org/10.1101/547679>
- Gökbuget, N., Kneba, M., Raff, T., Trautmann, H., Bartram, C.-R., Arnold, R., Fietkau, R., Freund, M., Ganser, A., Ludwig, W.-D., Maschmeyer, G., Rieder, H., Schwartz, S., Serve, H., Thiel, E., Brüggemann, M., Hoelzer, D., German Multicenter Study Group for Adult Acute Lymphoblastic Leukemia, 2012. Adult patients with acute lymphoblastic leukemia and molecular failure display a poor prognosis and are candidates for stem cell transplantation and targeted therapies. *Blood* 120, 1868–1876. <https://doi.org/10.1182/blood-2011-09-377713>
- Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J., 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34, D140–D144. <https://doi.org/10.1093/nar/gkj112>
- Grioni, A., Fazio, G., Rigamonti, S., Bystry, V., Daniele, G., Dostalova, Z., Quadri, M., Saitta, C., Silvestri, D., Songia, S., Storlazzi, C.T., Biondi, A., Darzentas, N., Cazzaniga, G., 2019. A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia. *HemaSphere* 3, e250. <https://doi.org/10.1097/HS9.0000000000000250>
- Gulilat, M., Lamb, T., Teft, W.A., Wang, J., Dron, J.S., Robinson, J.F., Tirona, R.G., Hegele, R.A., Kim, R.B., Schwarz, U.I., 2019. Targeted next generation sequencing as a tool for precision medicine. *BMC Med. Genomics* 12, 81. <https://doi.org/10.1186/s12920-019-0527-2>
- Helwak, A., Kudla, G., Dudnakova, T., Tollervey, D., 2013. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. <https://doi.org/10.1016/j.cell.2013.03.043>

- Hoelzer, D., 2015. Personalized medicine in adult acute lymphoblastic leukemia. *Haematologica* 100, 855–858.  
<https://doi.org/10.3324/haematol.2015.127837>
- Hunger, S.P., Mullighan, C.G., 2015. Redefining ALL classification: toward detecting high-risk ALL and implementing precision medicine. *Blood* 125, 3977–3987.  
<https://doi.org/10.1182/blood-2015-02-580043>
- Inaba, H., Greaves, M., Mullighan, C.G., 2013. Acute lymphoblastic leukaemia. *Lancet Lond. Engl.* 381, 1943–1955.  
[https://doi.org/10.1016/S0140-6736\(12\)62187-4](https://doi.org/10.1016/S0140-6736(12)62187-4)
- Jones, D.T., 2019. Setting the standards for machine learning in biology. *Nat. Rev. Mol. Cell Biol.* 20, 659–660.  
<https://doi.org/10.1038/s41580-019-0176-5>
- Karagkouni, D., Paraskevopoulou, M.D., Chatzopoulos, S., Vlachos, I.S., Tastsoglou, S., Kanellos, I., Papadimitriou, D., Kavakiotis, I., Maniou, S., Skoufos, G., Vergoulis, T., Dalamagas, T., Hatzigeorgiou, A.G., 2018. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res.* 46, D239–D245. <https://doi.org/10.1093/nar/gkx1141>
- Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915.  
<https://doi.org/10.1038/s41587-019-0201-4>
- Knecht, H., Reigl, T., Kotrová, M., Appelt, F., Stewart, P., Bystry, V., Krejci, A., Grioni, A., Pal, K., Stranska, K., Plevova, K., Rijntjes, J., Songia, S., Svatoň, M., Froňková, E., Bartram, J., Scheijen, B., Herrmann, D., García-Sanz, R., Hancock, J., Moppett, J., van Dongen, J.J.M., Cazzaniga, G., Davi, F., Groenen, P.J.T.A., Hummel, M., Macintyre, E.A., Stamatopoulos, K., Trka, J., Langerak, A.W., Gonzalez, D., Pott, C., Brüggemann, M., Darzentas, N., EuroClonality-NGS Working Group, 2019. Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. *Leukemia* 33, 2254–2265.  
<https://doi.org/10.1038/s41375-019-0499-4>
- Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., Tollervey, D., 2011.

Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10010–10015.  
<https://doi.org/10.1073/pnas.1017386108>

Kumar, S., Vo, A.D., Qin, F., Li, H., 2016. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci. Rep.* 6, 21597. <https://doi.org/10.1038/srep21597>

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.,

Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.  
<https://doi.org/10.1038/35057062>

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., Maglott, D.R., 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985.  
<https://doi.org/10.1093/nar/gkt1113>

Larson, R.A., Dodge, R.K., Burns, C.P., Lee, E.J., Stone, R.M., Schulman, P., Duggan, D., Davey, F.R., Sobol, R.E., Frankel, S.R., 1995. A five-drug remission induction regimen with intensive consolidation for adults with acute lymphoblastic leukemia: cancer and leukemia group B study 8811. *Blood* 85, 2025–2037.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521,

436–444. <https://doi.org/10.1038/nature14539>

- Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* 34, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* 25, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., Darnell, R.B., 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469. <https://doi.org/10.1038/nature07488>
- Lopes, B.A., Meyer, C., Barbosa, T.C., Poubel, C.P., Mansur, M.B., Duployez, N., Bashton, M., Harrison, C.J., Zur Stadt, U., Horstmann, M., Pombo-de-Oliveira, M.S., Palmi, C., Cazzaniga, G., Venn, N.C., Sutton, R., Alonso, C.N., Tsauro, G., Gupta, S.K., Bakhshi, S., Marschalek, R., Emerenciano, M., 2019. IKZF1 Deletions with COBL Breakpoints Are Not Driven by RAG-Mediated Recombination Events in Acute Lymphoblastic Leukemia. *Transl. Oncol.* 12, 726–732. <https://doi.org/10.1016/j.tranon.2019.02.002>
- Luh, F., Yen, Y., 2018. FDA guidance for next generation sequencing-based testing: balancing regulation and innovation in precision medicine. *Npj Genomic Med.* 3, 1–3. <https://doi.org/10.1038/s41525-018-0067-2>
- Malard, F., Mohty, M., 2020. Acute lymphoblastic leukaemia. *Lancet Lond. Engl.* 395, 1146–1162. [https://doi.org/10.1016/S0140-6736\(19\)33018-1](https://doi.org/10.1016/S0140-6736(19)33018-1)
- Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10–12. <https://doi.org/10.14806/ej.17.1.200>
- McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. <https://doi.org/10.1007/BF02478259>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a

MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. <https://doi.org/10.1101/gr.107524.110>

Metzker, M.L., 2010. Sequencing technologies — the next generation. *Nat. Rev. Genet.* 11, 31–46. <https://doi.org/10.1038/nrg2626>

Mullighan, C.G., 2014. The genomic landscape of acute lymphoblastic leukemia in children and young adults. *Hematol. Am. Soc. Hematol. Educ. Program* 2014, 174–180. <https://doi.org/10.1182/asheducation-2014.1.174>

O’Brien, J., Hayder, H., Zayed, Y., Peng, C., 2018. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front. Endocrinol.* 9. <https://doi.org/10.3389/fendo.2018.00402>

Raney, B.J., Cline, M.S., Rosenbloom, K.R., Dreszer, T.R., Learned, K., Barber, G.P., Meyer, L.R., Sloan, C.A., Malladi, V.S., Roskin, K.M., Suh, B.B., Hinrichs, A.S., Clawson, H., Zweig, A.S., Kirkup, V., Fujita, P.A., Rhead, B., Smith, K.E., Pohl, A., Kuhn, R.M., Karolchik, D., Haussler, D., Kent, W.J., 2011. ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.* 39, D871-875. <https://doi.org/10.1093/nar/gkq1017>

Research, C. for D.E. and, 2020. Hematologic Malignancies: Regulatory Considerations for Use of Minimal Residual Disease in Development of Drug and Biological Products for Treatment [WWW Document]. US Food Drug Adm. URL <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/hematologic-malignancies-regulatory-considerations-use-minimal-residual-disease-development-drug-and> (accessed 4.26.20).

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. <https://doi.org/10.1037/h0042519>

Roy, A., Bystry, V., Bohn, G., Goudevenou, K., Reigl, T., Papaioannou, M., Krejci, A., O’Byrne, S., Chaidos, A., Gioni, A., Darzentas, N., Roberts, I.A.G., Karadimitris, A., 2017. High resolution IgH repertoire analysis reveals fetal liver as the likely origin of life-long, innate B lymphopoiesis in humans. *Clin. Immunol. Orlando Fla* 183, 8–16.



<https://doi.org/10.1016/j.clim.2017.06.005>

- Saettini, F., Moratto, D., Grioni, A., Maitz, S., Iascone, M., Rizzari, C., Pavan, F., Spinelli, M., Bettini, L.R., Biondi, A., Badolato, R., 2018. A novel EP300 mutation associated with Rubinstein-Taybi syndrome type 2 presenting as combined immunodeficiency. *Pediatr. Allergy Immunol. Off. Publ. Eur. Soc. Pediatr. Allergy Immunol.* 29, 776–781. <https://doi.org/10.1111/pai.12968>
- Scheuring, U.J., Pfeifer, H., Wassmann, B., Brück, P., Gehrke, B., Petershofen, E.K., Gschaidmeier, H., Hoelzer, D., Ottmann, O.G., 2003. Serial minimal residual disease (MRD) analysis as a predictor of response duration in Philadelphia-positive acute lymphoblastic leukemia (Ph + ALL) during imatinib treatment. *Leukemia* 17, 1700–1706. <https://doi.org/10.1038/sj.leu.2403062>
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* 128, 336–359. <https://doi.org/10.1007/s11263-019-01228-7>
- Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. <https://doi.org/10.1093/nar/29.1.308>
- Stanulla, M., Dagdan, E., Zaliova, M., Möricke, A., Palmi, C., Cazzaniga, G., Eckert, C., Te Kronnie, G., Bourquin, J.-P., Bornhauser, B., Koehler, R., Bartram, C.R., Ludwig, W.-D., Bleckmann, K., Groeneveld-Krentz, S., Schewe, D., Junk, S.V., Hinze, L., Klein, N., Kratz, C.P., Biondi, A., Borkhardt, A., Kulozik, A., Muckenthaler, M.U., Basso, G., Valsecchi, M.G., Izraeli, S., Petersen, B.-S., Franke, A., Dörge, P., Steinemann, D., Haas, O.A., Panzer-Grümayer, R., Cavé, H., Houlston, R.S., Cario, G., Schrappe, M., Zimmermann, M., TRANSCALL Consortium, International BFM Study Group, 2018. IKZF1plus Defines a New Minimal Residual Disease-Dependent Very-Poor Prognostic Profile in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 36, 1240–1249. <https://doi.org/10.1200/JCO.2017.74.3617>

- Suwinski, P., Ong, C., Ling, M.H.T., Poh, Y.M., Khan, A.M., Ong, H.S., 2019. Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front. Genet.* 10. <https://doi.org/10.3389/fgene.2019.00049>
- Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., Darnell, R.B., 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302, 1212–1215. <https://doi.org/10.1126/science.1090095>
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nuskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L.,

Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The sequence of the human genome. *Science* 291, 1304–1351. <https://doi.org/10.1126/science.1058040>

Weissenbach, J., 2016. The rise of genomics. *C. R. Biol.* 339, 231–239. <https://doi.org/10.1016/j.crv.2016.05.002>

Zaliova, M., Stuchly, J., Winkowska, L., Musilova, A., Fiser, K., Slamova, M., Starkova, J., Vaskova, M., Hrusak, O., Sramkova, L., Stary, J., Zuna, J., Trka, J., 2019. Genomic landscape of pediatric B-other acute lymphoblastic leukemia in a consecutive European cohort. *Haematologica*. <https://doi.org/10.3324/haematol.2018.204974>

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A., 2019. A primer on deep learning in genomics. *Nat. Genet.* 51, 12–18. <https://doi.org/10.1038/s41588-018-0295-5>

## Scope of the thesis

This study first required the development of bioinformatics strategies and pipelines for the identification of genomic biomarkers. Those pipelines were implemented as an informatics infrastructure for the clinical diagnosis of acute lymphoblastic leukaemia. This study was coordinated through regular meetings and visits between the two institutions, Fondazione Tettamanti at the University of Milan-Bicocca, and CEITEC at Masaryk University.

The continuous increase of genomic information led me to search for new technologies for the analysis of big-data. Therefore, the second part of the thesis was focused on the acquisition of expertise in Machine Learning, and more specifically DL. For this purpose, I have been involved in the development of a DL model for the identification of small non-coding RNA genomic loci from whole genome scanning. This project allowed me to expand my domain knowledge to the research field of non-coding genomes.

**Chapter 2:** this section summarizes the application developed for the clinical diagnostics of Leukemia and implemented at Fondazione Tettamanti. It also describes the application of Deep Learning for the study of small RNA molecules.

**Chapter 3:** In this paper we presented a modern application of Deep Neural Network for the identification of small coding RNA throughout genome scanning.

**Chapter 4:** In this paper we presented the wet lab and in silico approach for the identification of biomarkers from next-generation sequencing.

**Chapter 5:** In this paper we compared the quality control and quantification of IG/TR with next generation sequencing and the bioinformatics tool.

**Chapter 6:** In this case report we identified a novel EP300 mutation potentially causative of Rubinstein-Taybi syndrome type II.

**Chapter 7:** In this paper we presented a genetic analysis of a patient affected by both acute lymphoblastic leukemia and Cornelia de Lange syndrome.

**Chapter 8:** In this paper we presented the in-silico and web laboratory application for the identification of fusion genes from target-capture next-generation sequencing

**Chapter 9:** In this paper we presented a novel tool for the immunoprofiling of IG/TR from next-generation amplicon sequencing as well as IMGT.

**Chapter 10:** In this paper we analysed the evolution of the immunoreportoire throughout the human life stages.

## Chapter 2

### Informatics infrastructure for clinical diagnostics

The application of Next-Generation Sequencing into clinical diagnostics required the automation of bioinformatics processes to provide a fast and autonomous system of analysis to the final user. Under these settings, clinical biologists and medical doctors can overcome the gap between NGS data analysis and clinical reports. The informatics infrastructure supporting the daily clinical diagnostics of acute lymphoblastic leukaemia at the Fondazione Tettamanti - University of Milan - was designed and implemented in collaboration with the Bioinformatics Core Facility of the Central European Institute of Technology (Dr. Vojtech Bystry, Ph.D.).

The infrastructure is used to analyse NGS experiments required for the clinical diagnosis of acute lymphoblastic leukaemia. Users access an interactive web-interface to select the NGS experiment and subsequently the pipeline of analysis (Figure 13).

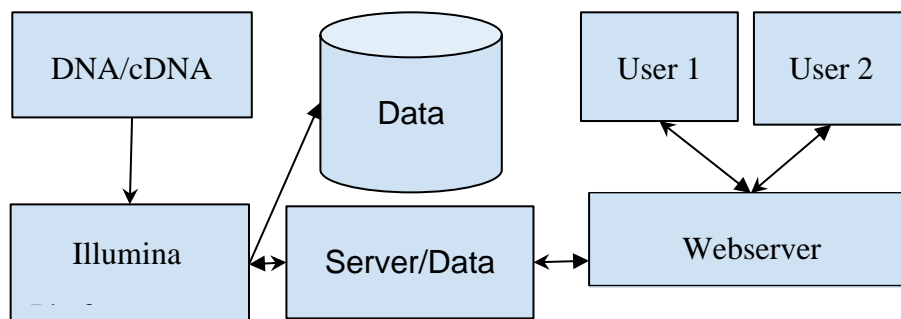


Figure 13. Schematic representation of the informatics infrastructure of Fondazione Tettamanti.

The main bioinformatics analysis pipelines used in ALL clinical diagnostics are the identification of immunoglobulins and T-cell receptor biomarkers for minimal residual disease (Immunoprofile) and the identification of chromosomal rearrangements (capture and capture\_NuGen) (Figure 14). The Immunoprofile pipeline generates the input files for the online platform of analysis ARResT/Interrogate, which is available at the following url: <http://bat.infspire.org/arrest/interrogate/>. Both ARResT/Interrogate and the corresponding NGS strategy have been developed within the European Consortium EuroClonality-NGS (<https://euroclonalityngs.org/usr/pub/pub.php>). I had the opportunity to actively participate in the development of both bioinformatics and wet-laboratory strategies as a member of Dr. Giovanni Cazzaniga's research group (Fondazione Tettamanti - Monza). This collaborative study is described in three main papers (Brüggemann et al., 2019; Bystry et al., 2017; Knecht et al., 2019).

NGS-Tettamanti/stable-pipelines

demultiplex  
bc12fastq manual

Show 10 entries Search:

Date	MachineID	RunNumber	FlowCellID
200430	M01706	0145	000000000-J2FN2
200429	M01706	0144	000000000-J3DFN
200427	M01706	0143	000000000-J3B8L
200424	M01706	0142	000000000-J2RDR
200417	M01706	0141	000000000-J2LKG
200409	M01706	0140	000000000-J2P9C
200401	M01706	0138	000000000-J2R7F
200327	M01706	0137	000000000-J2SV2
200319	M01706	0136	000000000-J2BMV
200312	M01706	0135	000000000-J2J58

Showing 1 to 10 of 180 entries

selected folder: 200430\_M01706\_0145\_000000000-J2FN2

select analysis [?]

- need selection
- need selection
- immunoprofile
- capture
- FASTQ-only
- export
- capture\_NuGen

Previous 1 2 3 4 5 ... 18 Next

Figure 14. Screenshot of the web-based application. The interface allows users to select a specific NGS experiment and run standard pipelines for clinical diagnostics.

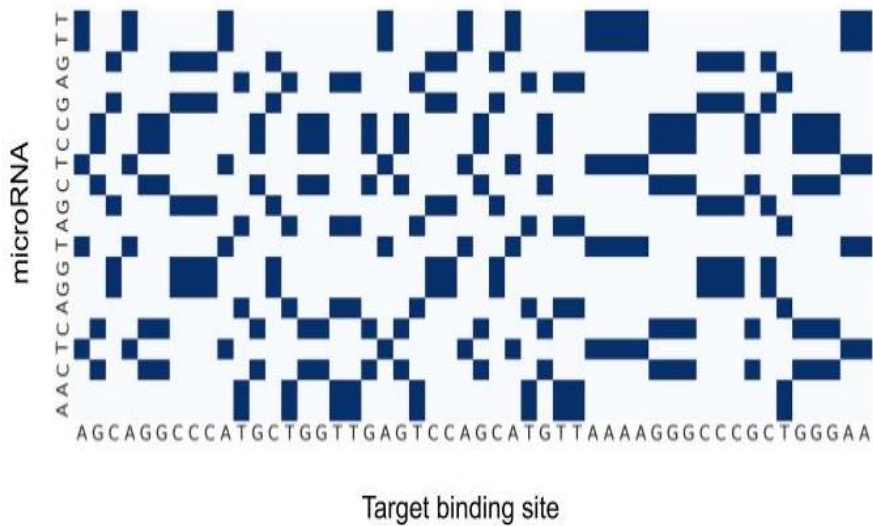
The identification of chromosomal rearrangements is addressed by the use of the Target Capture Sequencing. The “capture\_NuGen” pipeline embedded a preprocessing step to remove specific nucleotide sequences (linkers) used to perform the experiment. Otherwise, both pipelines, “capture” and “capture\_NuGen”, run a purpose-build bioinformatics pipeline that I have developed for the identification of fusion genes from the specific target capture datasets. The strategy for the identification of fusion genes was published as an article in the HemaSphere journal, which is the official journal of the European Hematology Association (Griani et al., 2019)

### Deep Learning for Small RNA analysis

MicroRNA regulates transcription by association with AGOs proteins and binding to target messenger RNA. As a personal project, I developed a neural network model with convolutional architecture to predict microRNA target binding sites by scanning of a genomic nucleotide sequence. The model takes as input a dot matrix (Figure 16), which is a two-dimensional matrix representing the microRNA and target binding site interaction. One dimension is the microRNA nucleotide sequence (20 nt) and the second dimension is the nucleotide sequence of the binding site (50 nt). Watson-Crick interactions between



the two sequences are represented by a positive value of 1, other values are set to 0.



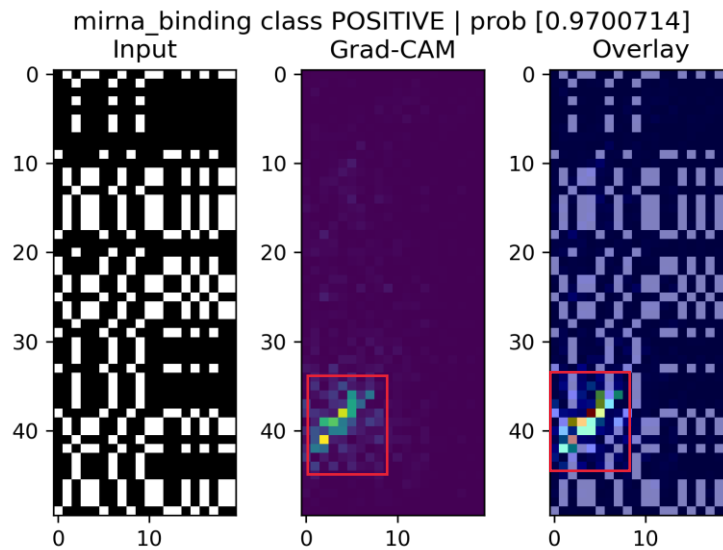
*Figure 15. Example of dot-matrix. Y-axis correspond to the nucleotide sequence of the miRNA (20nt). X-axis correspond to the nucleotide sequence of the genomic binding site. Watson-crick interactions are highlighted in dark blue, while non-Watson-Crick interaction are highlighted in light blue. The dot-matrix is the input of the neural network.*

The model output is a prediction score corresponding to the probability of a real interaction between the microRNA and the target binding site. The score is a continuous number between 0 and 1, where 1 indicates a high probability of a positive interaction. The model facilitates classification of the interaction between a miRNA and a target binding site as positive or negative. However, the model does not localise the nucleotides involved in the binding. For this purpose, I implemented and used the Gradient-weighted Class Activation Mapping (Grad-CAM) method described by Selvaraju et al. (Selvaraju et al., 2020).

The application of Grad-CAM allowed the spatial identification of the regions causing the activation of the neurons in the model. This region

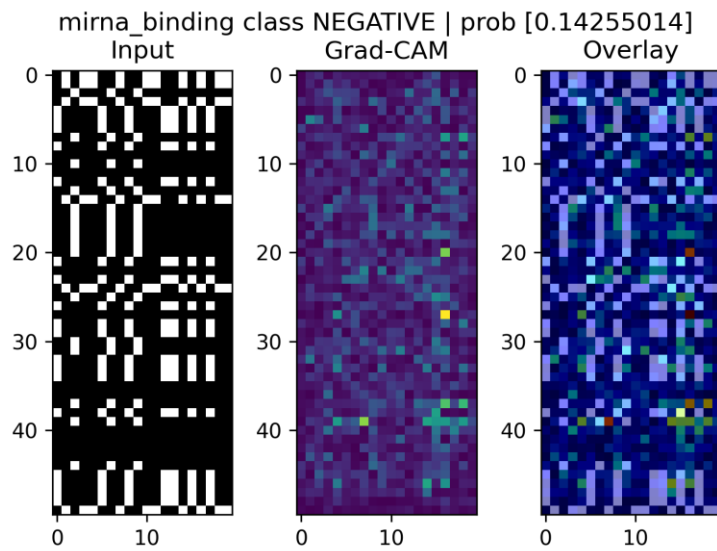
corresponds to the nucleotides involved in the binding of the microRNA with the target binding site. Therefore, through this approach it is possible not only to have qualitative results as positive or negative interaction, but also to identify which nucleotides are responsible for the binding of the microRNA to the target messenger RNA.

Figure 17 shows the activation map of an input dot-matrix. The x and y axes correspond to the miRNA and binding site, respectively. The activation map is represented as a heat-map mirroring the input dot-matrix. Light colours correspond to high values of activation, while dark colours approximate activation values close to 0. The interaction between the miRNA and the candidate binding site involved the first ~8 nucleotides of the miRNA and the nucleotides in position ~35 to ~42 of the binding site.



*Figure 16. Visualization of the activated neurons. The bottom-left corner highlights neurons that recognised the miRNA-Binding site interaction.*

As a comparison, the Figure 18 shows a negative interaction. In this case the model was not able to detect strong and localised features that would indicate a miRNA-binding site interaction. As you can see, the activation is nonspecific and occurred simultaneously in all neurons but with a low intensity.



*Figure 17. Visualization of the negative activation. In this case, all neurons are lightly activated, and it is not possible to distinguish a clear signal. Therefore, the sample is predicted as negative interaction.*

## Chapter 3

Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci.

Paper has been accepted by Nature – Scientific Report and it still needs to be provided with DOI.

Georgios K Georgakilas, **Andrea Grioni**, Konstantinos G Liakos, Eliska Chalupová, Fotis C Plessas and Panagiotis Alexiou. Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci. Scientific Reports – Nature



OPEN

# Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci

Georgios K. Georgakilas<sup>1</sup>, Andrea Grioni<sup>1</sup>, Konstantinos G. Liakos<sup>3</sup>, Eliska Chalupová<sup>2</sup>, Fotis C. Plessas<sup>3</sup> & Panagiotis Alexiou<sup>1</sup>✉

Genomic regions that encode small RNA genes exhibit characteristic patterns in their sequence, secondary structure, and evolutionary conservation. Convolutional Neural Networks are a family of algorithms that can classify data based on learned patterns. Here we present MuStARD an application of Convolutional Neural Networks that can learn patterns associated with user-defined sets of genomic regions, and scan large genomic areas for novel regions exhibiting similar characteristics. We demonstrate that MuStARD is a generic method that can be trained on different classes of human small RNA genomic loci, without need for domain specific knowledge, due to the automated feature and background selection processes built into the model. We also demonstrate the ability of MuStARD for inter-species identification of functional elements by predicting mouse small RNAs (pre-miRNAs and snoRNAs) using models trained on the human genome. MuStARD can be used to filter small RNA-Seq datasets for identification of novel small RNA loci, intra- and inter- species, as demonstrated in three use cases of human, mouse, and fly pre-miRNA prediction. MuStARD is easy to deploy and extend to a variety of genomic classification questions. Code and trained models are freely available at [gitlab.com/RBP\\_Bioinformatics/mustard](https://gitlab.com/RBP_Bioinformatics/mustard).

Since the human genome was first sequenced about two decades ago<sup>1</sup>, our understanding of regulatory and non-coding elements in humans, and other organisms, has been steadily increasing with the identification and cataloguing of a variety of encoded molecule and regulatory region classes<sup>2</sup>. Several small non-coding RNA molecule families such as microRNA (miRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), piwi-interacting RNA (piRNA), short hairpin RNA (shRNA), small interfering RNA (siRNA), promoter-associated short RNAs (PASRs), termini-associated short RNAs (TASRs)<sup>3,4</sup>, transcription initiation RNAs (tiRNAs)<sup>5</sup>, and others, now populate the functional expression map of known genomes. The plethora of functional small non-coding RNA classes supports the idea of a highly interconnected transcriptomic landscape and highlights the necessity of computational approaches that can effectively identify them against the enormous background variability of eukaryotic genomes. Along with our deeper understanding of well-established organisms, the total number of sequenced genomes has been increasing hand in hand with fast pace. NCBI currently lists just over 7,000 eukaryotic sequenced genomes, of which almost 50 have fully assembled genomes, and approximately 1,000 have some assembled chromosomes. The experimental annotation of newly sequenced genomes is a much slower and piecemeal process that benefits greatly from the availability of computational techniques that can guide and assist the annotation.

Q1 Q2 Q3 Q4

Computational methods for genomic annotation have a history at least as long as full genome sequencing, with computational identification of exons and protein coding genes<sup>6</sup> starting in parallel with the sequencing of the first human genome. Small non-coding RNAs, with their shorter length, lack of coding three nucleotide periodicity pattern, and often small number of known examples per class, offer a tougher challenge for computational methods. A common approach for in silico identification of putative small non-coding RNA genomic loci has been the use of sequence homology between molecules from well annotated species, such as humans, and the new species in question. These methods, while efficient when homology is high, are bound to preferentially annotate a subset of loci, biased towards extra-conserved molecules. However, a large number of small non-coding RNAs

<sup>1</sup>Central European Institute of Technology, Brno, Czech Republic. <sup>2</sup>Faculty of Science, National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic. <sup>3</sup>Department of Electrical and Computer Engineering, School of Engineering, University of Thessaly, Volos, Greece. ✉e-mail: [panagiotis.alexiou@ceitec.muni.cz](mailto:panagiotis.alexiou@ceitec.muni.cz)

are more evolutionary constrained. For example, an estimated 40% of human miRNAs have developed recently in evolutionary history and can only be found in other primates<sup>7</sup>.

To avoid the constraints and biases of homology based identification of small non-coding RNA loci, there has been a steady development of algorithms that aim at modelling characteristics of a specific class or subclass, and then evaluating proposed regions of a genome for their potential to encode a small non-coding RNA of this class. For example, over thirty computational methods aiming at pre-microRNA identification have been developed to date, with no tool significantly outperforming all others on benchmarked datasets<sup>8</sup>. A large drawback of such methods is their dependence on expert-defined features and background sets, which tend to produce methods that perform well in evaluations closely matching their training biases, but fail to produce robust classification in more realistic conditions, such as when ‘scanning’ a large genomic region. The second large drawback of these methods is that they are, by design, focused on one specific class or subclass of small non-coding RNA molecules. For example, a method tailored for pre-miRNA prediction is not suitable for snoRNA prediction and vice versa. This issue leads to an unbalanced development of methods towards specific families and ignores others that may not be populous or well-researched enough to warrant the attention of *in silico* method developers. For example, as mentioned above, pre-microRNA prediction is a well-researched field with over thirty computational methods published in the past decade or so, while in contrast snoRNA prediction displays a distinct paucity of options, with methods becoming obsolete and unusable after more than a decade<sup>9,10</sup> and the rate of identification severely slowing down in new species<sup>11</sup>.

Taking into account the limitations and drawbacks of *in silico* methods to date, we have decided to approach the problem of small non-coding RNA identification from a different angle. Here we introduce MuStARD (Machine-learning System for Automated RNA Discovery), a flexible Deep Learning framework that utilizes raw sequence, conservation, and folding data to identify genomic loci with similar characteristics of a given set of regions (Fig. 1a). Instead of hundreds of expertly curated features, we employ a Convolutional Neural Network (CNN) Deep Learning (DL) architecture that can identify important characteristics from raw data directly<sup>12</sup>. Rather than biased background training sets, we opted for a novel iterative background selection process that allows the method itself to identify the background ‘hard cases’ for a specific classification task, and preferentially learn how to avoid them. While other tools focus on one class of small non-coding RNAs, we have developed a framework that can be applied, directly out of the box, on any class of genomic loci. We show the power of this methodology by training models that outperform the state of the art for pre-miRNAs and snoRNAs by scanning large genomic regions. We demonstrate the practical use of our method by performing a cross-species prediction using models trained on human data to accurately identify mouse pre-miRNAs and snoRNAs in numbers well above homology searches. Additionally, we applied MuStARD on small RNA-Seq enriched regions and pre-miRNAs that have been removed from miRBase since version 14, to further highlight the usability spectrum of our algorithm. The source code is available at [https://gitlab.com/RBP\\_Bioinformatics/mustard](https://gitlab.com/RBP_Bioinformatics/mustard) and trained models at [https://gitlab.com/RBP\\_Bioinformatics/mustard\\_paper](https://gitlab.com/RBP_Bioinformatics/mustard_paper).

## Methods

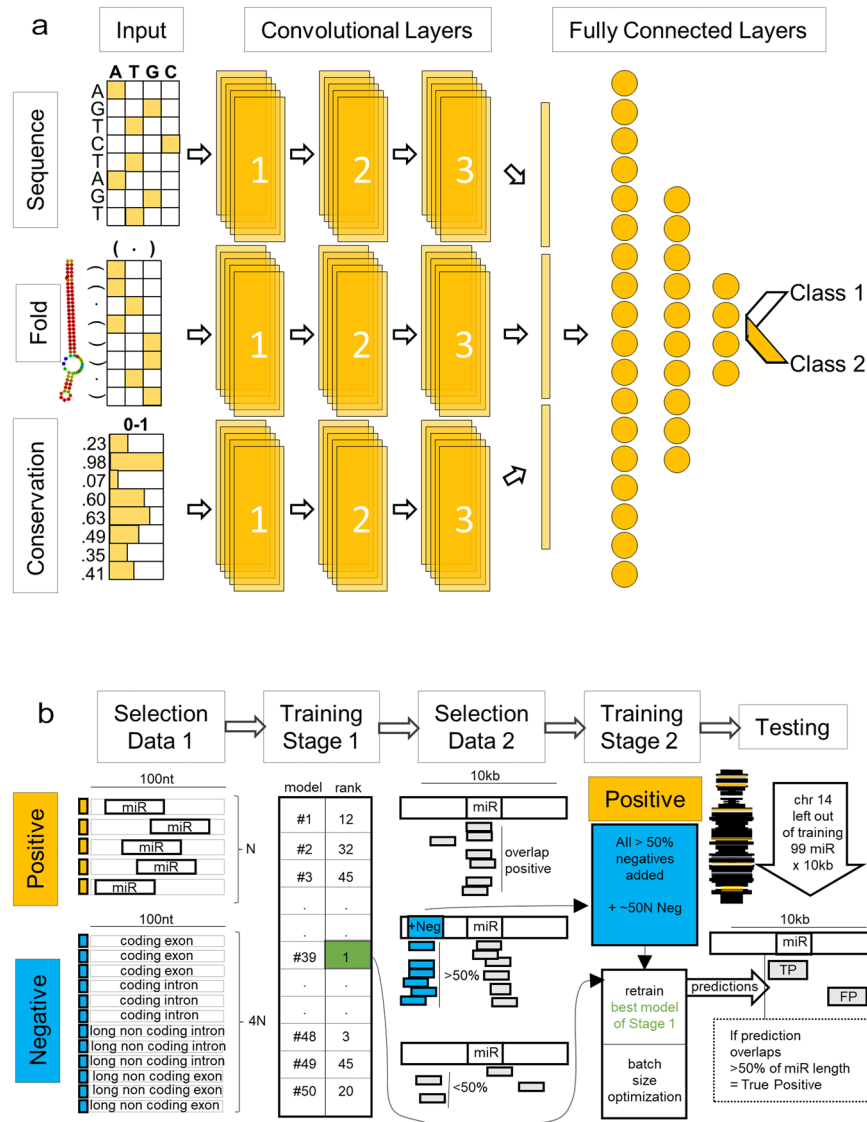
**Network architecture and training scheme.** MuStARD is able to handle any combination of either raw DNA sequences, basewise evolutionary conservation and folding data (Fig. 1a). Each feature category is forwarded to a separate ‘branch’ that consists of three convolutional layers and the computations from all branches are concatenated prior to being forwarded to the fully connected part of the network. The training scheme consists of two steps (Fig. 1b). First 50 models are trained in parallel with random background selection. The 50 trained models are used to scan a large region of the genome. From these 50 scans, the ‘hard cases’ where the majority of models detect a false positive are isolated and a new negative set created. The best performing of the 50 models is then used as the starting point to train the final model on the ‘hard cases’ while keeping the same positive set. This final trained model is then evaluated on targets located in chromosomes completely left out of the whole previous process, thus ensuring no cross-contamination.

This process was repeated 6 times to train pre-miRNA detection models composed of different input combinations; raw sequence with secondary structure and conservation (MuStARD-mirSFC model), raw sequence and conservation (MuStARD-mirSC), raw sequence and secondary structure (MuStARD-mirSF), secondary structure and conservation (MuStARD-mirFC), secondary structure only (MuStARD-mirF) and sequence only (MuStARD-mirS). For the combination of raw sequence, secondary structure and conservation, we have trained an additional model after disabling the class weights option in Keras (MuStARD-mirSFC-U model).

The same pipeline was used to create three snoRNA detection models, one for detecting the C/D box snoRNA subspecies (MuStARD-snoSFC-U-CDbox), one for H/ACA box (MuStARD-snoSFC-U-HACAbbox) and one for detecting all types of snoRNAs (MuStARD-snoSFC-U).

Detailed information related to the network architecture and training scheme can be found in Supplementary Methods.

**Training sets.** Human (GRCh38) and mouse (GRCm38) genomes and corresponding gene and snoRNA annotations were downloaded from Ensembl v93 repository<sup>13</sup>. Fly genome (version 5.32) was downloaded from FlyBase<sup>14</sup>. Pre-miRNA sequences were downloaded from miRBase v22.1<sup>15</sup>. Basewise conservation scores, based on phyloP algorithm, of 99 and 59 vertebrate genomes with human and mouse respectively were downloaded from the UCSC genome repository<sup>16</sup>. For genome scanning tests, targets were extended by  $\pm 5,000$  bp and the resulting regions were merged in the case of strand specific overlaps. The regions were assessed by a moving window of width 100 and step 5. Any prediction overlapping the target by at least 50% was considered a positive. A full explanation of the product of Training Sets, and the Methodology of comparisons can be found in Supplementary Methods. Results of the comparison between MuStARD models using distinct combinations of raw sequence, secondary structure and evolutionary conservation as input, are presented in Supplementary



**Figure 1.** Overview of MuStARD modular architecture and iterative training pipeline. (a) MuStARD is able to handle any combination of either raw DNA sequences, RNAfold derived secondary structure and basewise evolutionary conservation from PhyloP. DNA sequences and RNAfold output are one-hot encoded while PhyloP score is not pre-processed. Each feature category is forwarded to a separate ‘branch’ that consists of three convolutional layers. The computations from all branches are concatenated prior to being forwarded to the fully connected part of the network. (b) The training pipeline of MuStARD consists of two steps. Initially, pre-miRNA sequences are randomly shuffled to exonic and intronic (protein-coding and lincRNA genes) regions of the genome to extract equal sized negative sequences with 1:4 positive to negative ratio. This process is repeated 50 times to facilitate the training of equal number of models. The performance of each model is assessed based on the test set and all false positives that are supported by at least 25 models are extracted. This set of false positives is added to the negative pool of the best performing model to create an enhanced training set. The enhanced set is used to train the final MuStARD model.

Table 1. The evaluation of the final MuStARD models and comparisons to other state of the art programs can be found in Supplementary Tables 2–7. The performance of MuStARD-mir models in the first training iteration can be found in Supplementary Table 8.

**Software and hardware requirements.** MuStARD is developed in python utilizing tensorflow and Keras for the Deep Learning aspect, R for visualizing the performance and Perl for file processing, reformatting and module connectivity. Full list of dependencies can be found on MuStARD’s gitlab page.

MuStARD is able to execute either on CPU or GPU depending on the underlying hardware configuration by taking into advantage tensorflow’s flexibility. The framework has been designed to maintain a minimal memory footprint thus allowing the execution even on personal computers. Running time heavily depends on input dimensionality, number of instances in the training set, learning rate and GPU availability. On a GPU

Algorithms	Homo Sapiens - chr14 pre-labelled dataset			Homo Sapiens - chr14 scanning dataset		
	Precision	Sensitivity	F1	Precision	Sensitivity	F1
MuStARD-mirSFC-U	0.958	0.522	0.675	0.953	0.424	0.587
MiPred	0.128	0.977	0.226	0.069	1	0.130
miRBoost	0.063	0.840	0.118	0.080	0.898	0.146
HuntMi	0.147	1	0.256	0.070	0.979	0.131
microPred	0.114	0.977	0.205	0.197	1	0.330
triplet-SVM	0.194	0.931	0.321	0.061	0.898	0.115
Random	0.051	0.545	0.094	N/A	N/A	N/A

**Table 1.** Performance results based on the human chromosome 14 pre-labelled and scanning datasets.

(NVIDIA GeForce GTX 1050Ti) it took approximately 5 minutes to train a model on 30,000 positive and negative sequences.

MuStARD operates directly on genomic intervals in BED format, in the cases of both the training and prediction modules. For example, regarding small RNA-Seq datasets, MuStARD does not directly process aligned reads. Instead, users need to provide a bed file with small RNA-Seq enriched regions to be scanned with MuStARD. Essentially, sequencing depth is not crucial as our algorithm works after a ‘peak calling’ step that can be as simple as a bedtools merge command followed by a bedtools coverage filtering.

For the mouse liver small RNA-Seq dataset, we scanned 14,552 intervals (both strands derived from 7,276 peaks) with a total size of 3.9mbp, mean interval size of 268 bp and standard deviation of 50.3 bp. The MuStARD-mirSFC-U model was used with a sliding window step of 10bp and the total running time was 36 minutes (CPU usage only).

In the case of the fly embryo dataset, we scanned 1,638 intervals (819 peaks) with 526 kb total size, 321 bp mean size with standard deviation of 172 bp. We used MuStARD-mirSF model with a sliding window step of 10 bp and total running time of 3.8 minutes.

The average running time of MuStARD per peak, based on the mir-SFC-U model, was 0.17 seconds, while based on the mir-SF model was 0.13 seconds. Running times were normalized on 321 bp average interval size.

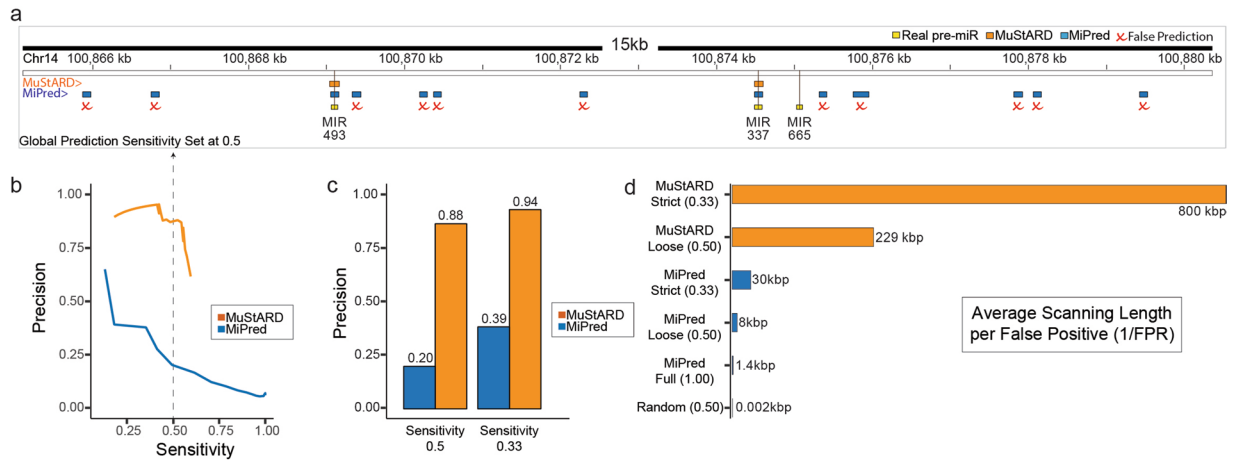
## Results

**Training of convolutional neural network model.** We compared the performance of MuStARD on all combinations of input data for the pre-miRNA prediction dataset (Supplementary Table 1). As expected, scanning test sequences with various models shows that models including a higher number of meaningful input data branches perform better in retrieval of pre-miRNAs. The model trained on secondary structure and conservation was the best performing two-input model. This result aligns with the identification of pre-miRNA hairpins by the Microprocessor complex during miRNA biogenesis primarily by characteristics of their secondary structure rather than sequence<sup>17</sup> and the fact that pre-miRNAs have highly conserved regions corresponding to the mature miRNA sequences. Surprisingly, the non-balanced model (MuStARD-mirSFC-U) performs best out of all model combinations including the balanced three input model. Since MuStARD-mirSFC-U outperforms all other models, we will only report results for this model in the following evaluations. For snoRNAs, the equivalent best performing model is MuStARD-snoSFC-U. Detailed explanation of the training scheme can be found in Supplementary Methods.

**Identification of *homo sapiens* pre-miRNA genomic loci.** While training MuStARD models, we left-out the entirety of randomly selected chromosome 14 as a final evaluation set that could be fairly used to benchmark MuStARD’s performance against the current state of the art in pre-miRNA prediction. The question of accurate pre-miRNA prediction has been thoroughly researched since there are currently over 30 published pre-miRNA prediction algorithms indexed in the OMICtools<sup>18</sup> repository. The majority of these studies could not be coerced to run on our benchmarking dataset (see Supplementary Methods for details). We managed to run and evaluate five state of the art programs: HuntMi<sup>19</sup>, microPred<sup>20</sup>, MiPred<sup>21</sup>, miRBoost<sup>22</sup> and triplet-SVM<sup>23</sup>. A list of algorithms we attempted, but failed, to evaluate can be found along with our code repository. Of these five, only triplet-SVM, MiPred and miRBoost provide probabilities as output scores allowing assessment of their performance on multiple score thresholds. HuntMi and microPred provide fixed output score/labels limiting their performance comparison on a fixed threshold (Supplementary Figure 1, Supplementary Tables 2 and 3). After evaluating all five algorithms on the chromosome 14 evaluation set, we identified MiPred as the overall optimally performing state-of-the-art algorithm, thus for the sake of brevity we will only report direct in depth comparison to MiPred. Table 1 summarizes the performance results from the pre-labelled and scanning chromosome 14 benchmarks.

Both MuStARD and MiPred report predictions with probability scores, and both programs would as default be used at a score threshold of 0.5. However, at that threshold, MiPred produces an inordinate amount of false positives (Supplementary Tables 2 and 3). For fairness of comparison of program precision, we have set a threshold on prediction sensitivity at the point where each program predicts 50% of real pre-miRNAs (Fig. 2a). MuStARD exhibits consistently high precision for any level of sensitivity (Fig. 2b,c) and at a strict threshold where 33% of real pre-miRNAs can be annotated it produces on average one false positive prediction per 800,000





**Figure 2.** Evaluation of MuStARD human predictions against MiPred, the best performing of state-of-the-art pre-miRNA prediction algorithms. **(a)** Genome browser visualization of each algorithm's performance on the scanning windows in a 15 kb locus hosting three pre-miRNAs on the left-out chromosome 14. Both evaluated programs have been benchmarked at scores that give sensitivity of 0.5 over the left-out chromosome. MuStARD correctly predicts 2/3 of the annotated pre-miRNAs (in this particular locus), same as MiPred. MuStARD produces no false positive predictions, compared to 11 for MiPred (marked with red x). **(b)** precision-sensitivity curve of MuStARD and MiPred over scanned areas of the left-out chromosome 14. **(c)** Precision of MuStARD and MiPred at loose (sensitivity 0.5) and strict (sensitivity 0.33) thresholds. **(d)** Average length in thousands of base pairs for finding each false positive prediction on the left-out chromosome. Showing MuStARD at strict and loose thresholds, and MiPred at strict, loose, and full (score 0.5 - sensitivity ~1) thresholds, and random prediction (threshold sensitivity 0.5) denoting the worst performing levels an algorithm could achieve.

scanned nucleotides (Fig. 2d) outperforming MiPred by an order of magnitude. Detailed information related to the scanning and static types of evaluation can be found in Supplementary Information.

**Identification of pre-miRNAs from small RNA-Seq data.** Our method can scan large genomic regions with unprecedented precision, but would still produce a large number of false positives in a full genome scan of several billion bases. A more realistic experimental and computational approach would be the identification of molecules from small RNA-Seq data. In this type of commonly performed experiment, RNA is isolated and filtered for sizes below a certain threshold, thus removing most mRNAs and long non-coding RNAs. However, there still remain fragments and other artifacts, and the bona fide small RNAs still need to be classified into different classes.

We used the human pre-miRNA trained model of our method to retrieve pre-miRNA predictions from three small RNA-Seq datasets in varying degrees of evolutionary distance from humans. The first dataset consists of human H1 cells, in which we only evaluated 502 small RNA-Seq enriched regions from the left-out chromosome 14. The second dataset comes from mouse liver and we evaluated on 7,276 enriched regions genome-wide. The last datasets is of higher difficulty as it was derived from drosophila melanogaster, an evolutionary distant organism for which conservation information was not readily available. We evaluated our method without the conservation branch (MuStARD-mirSF) on drosophila using the top 819 small RNA-Seq enriched regions (Table 2).

We have evaluated MuStARD and MiPred using precision/recall curves (Supplementary Figure 2) as well as the F1 measure at multiple score thresholds to gain a spherical view of the algorithms' performance (Fig. 3). For the precision/recall curves specifically, we have added the 'naive' strategy of picking multiple top percentiles of small RNA-Seq enriched loci ranked by decreasing expression level. The 'naive' strategy serves as the baseline performance that any Machine Learning algorithm should outperform.

MuStARD outperforms MiPred at every benchmark dataset while keeping a relatively balanced ratio between precision and sensitivity across multiple score thresholds. For example, in order for MiPred to reach high levels of precision (85.7%) in human, it needs to increase the score threshold at a level that reduces the sensitivity below 16% (Supplementary Table 4). MuStARD, on the other hand, is able to maintain sensitivity above 40%, even with a score threshold as high as 0.89 that translates to 82% of precision.

As expected, we notice a decreasing level of MuStARD's prediction performance with increasing evolutionary distance from our training organism with human (F1 = 0.66, Fig. 3a), mouse (F1 = 0.57, Fig. 3b) and drosophila (F1 = 0.39, Fig. 3c) at 0.5 score threshold (Supplementary Table 4). Our method can narrow down the peaks identified from small RNA-Seq and better prioritize ones that could harbor small RNAs of a specific class. Detailed information related to the small RNA-Seq based strategy can be found in Supplementary Methods.

**Cross-species identification of pre-miRNA genomic loci.** Having established a substantial increase in precision for intra-species pre-miRNA prediction we evaluated our model on an inter-species prediction. Briefly, we used the best performing pre-miRNA identification model trained on human data, to scan swathes of the mouse genome (in total ~9.8Mbps) containing 1,227 annotated mouse pre-miRNAs. The inter-species prediction

		MuStARD-mirSFC-U	MiPred	Expression
Homo Sapiens - small RNA-Seq in H1 cells	Precision	0.750	0.857	1
	Sensitivity	0.500	0.157	0.027
	F1	0.600	0.266	0.054
Mus Musculus - small RNA-Seq in Liver	Precision	0.747	0.512	0.964
	Sensitivity	0.581	0.097	0.065
	F1	0.653	0.163	0.122
Drosophila Melanogaster - small RNA-Seq in Embryo	Precision	0.526	1	0.500
	Sensitivity	0.500	0.023	0.052
	F1	0.512	0.046	0.095

**Table 2.** Performance summary based on the small RNA-Seq datasets from Homo Sapiens, Mus Musculus and Drosophila Melanogaster, at 0.84 score threshold.

correctly identified pre-miRNAs with a small number of false positives, at a rate of 1/260kbp. Figure 4a shows a browser snapshot of a mouse pre-miRNA cluster locus. As expected, the precision of the inter-species prediction was lower than the intra-species evaluation set (Fig. 4b), and even lower for pre-miRNAs that do not have a human homologue as they have lower levels of conservation which is one of our model's input branches (Fig. 4c). MuStARD exhibits exceptional levels of generalisation capacity (Supplementary Table 5) identifying correctly a large majority (94/129) of homologous pre-miRNAs and more than double (212) non-homologous pre-miRNAs. Detailed information can be found in Supplementary Methods.

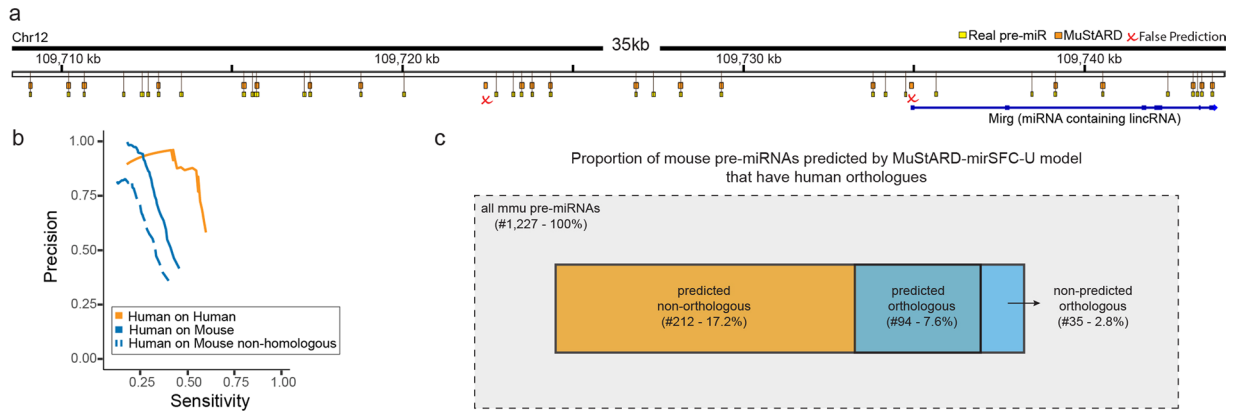
**Evaluation of miRBase retracted pre-miRNAs.** In order to evaluate our method on another difficult and realistic task, we mined all pre-miRNAs that were annotated in previous versions of miRBase (version 14 to 22.1) but have since been retracted from it. Extensive details can be found in Supplementary Methods. These are loci that show a close similarity to bona fide microRNAs, enough to be suggested by some experimental method. We evaluated 57 human and 64 mouse pre-miRNAs with our pre-miRNA prediction model trained on the latest human miRBase. Since these retracted pre-miRNAs do not exist in this latest version, our training model has never seen them before. At the 0.5 score threshold (loose) used for the evaluation, MuStARD correctly identified as negative 54/57 (95%) which increased to 56/57 (98%) at 0.85 threshold (strict) for human pre-miRNAs. For mouse pre-miRNAs, still using the human trained model, we correctly retrieved as negative 64/64 (100%) of the targets even at the most loose threshold (Supplementary Table 6).

**Identification of *homo sapiens* sno-RNA loci.** Despite its high accuracy on pre-miRNA classification, MuStARD was not specifically developed for pre-miRNA detection. To demonstrate its flexibility we trained models on a completely different class of small non-coding RNAs, small nucleolar RNAs (snoRNAs). SnoRNAs are a class of small RNAs with widely varying structure, sequence, and conservation patterns. We experimentally trained a model on all snoRNAs as well as two additional models for the most populous snoRNAs sub-families, the H/ACA and C/D box. H/ACA box snoRNAs have a secondary structure consisting of hairpins and single stranded regions. In contrast, C/D box snoRNAs have a stem-box structure that is much more variable than H/ACA box. In addition, our 'all snoRNA' dataset includes snoRNAs beyond these two sub-families. For the two sub-families we were also able to benchmark against snoReport<sup>24</sup> a state-of-the-art snoRNA prediction software developed specifically to identify each of these two categories against background (Table 3, Supplementary Table 7). We observe that MuStARD matches snoReport on the Homo Sapiens C/D box training set (Score 0.8, F1: 0.759 vs 0.769), but completely outperforms snoReport in Mus Musculus prediction for both C/D box (Score 0.8, F1: 0.704 vs 0.570) and H/ACA box (Score 0.8, F1: 0.810 vs 0.033). MuStARD also outperforms snoReport at the Homo Sapiens H/ACA box model (Score 0.8, F1: 0.755 vs 0.094). Furthermore, we tested the inter-species capabilities of the MuStARD model, by applying the human-trained snoRNA model to the mouse genome (Supplementary Table 7). These results demonstrate that the MuStARD method is capable of producing well trained models beyond the state of the art without domain knowledge, and even with relatively heterogeneous positive samples ("all snoRNAs").

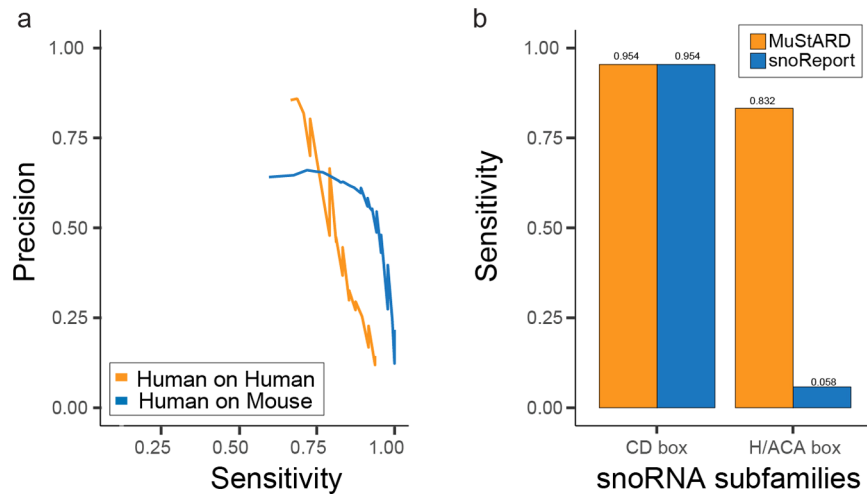
## Discussion

We present here a flexible Deep Learning framework that can be used to identify small RNA genomic loci based on the sequence, conservation, and secondary structure characteristics of the class. A model can be easily trained, without any changes on the code, to identify any class of small RNA loci provided enough examples of the class exist. Training of the model does not require expertly curated features specific to the RNA class. In contrast to highly specific methods that rely on extraction of hundreds of features, our method operates directly on raw sequence, conservation scores, and a simple linear folding representation. Despite the simplicity of the inputs, and the generality of the method, it manages to convincingly outperform all state of the art methods that have been each developed and trained on one single class of RNAs specifically.

An important aspect of our method is the ability, for the first time, to scan large genomic regions, or even several thousand sequencing peaks, at an acceptably low false discovery rate. Machine learning methods can only learn variation that is presented to them. When looking for extremely rare events, such as small RNA genomic



**Figure 3.** Evaluation of filtering for small RNA-Seq datasets for pre-miRNAs. F1 score per score threshold of the prediction method. MiPred default score threshold is 0.5. We evaluated three datasets: (a) human H1 cells, left out chromosome 14. (b) mouse liver, whole genome. (c) drosophila melanogaster embryo, whole genome. For the drosophila evaluation, vertebrate evolutionary conservation track was not available so the MuStARD-mirSF (sequence, folding) model was used instead.



**Figure 4.** Prediction of mouse pre-miRNAs by the model trained on human. (a) Genome browser visualization of MuStARD performance on the scanning of a 35 kb locus hosting 36 pre-miRNAs. MuStARD correctly identifies 20/36 pre-miRNAs with 2 false positives, out of which one falls on the first “exon” of a long non coding RNA *Mirg* annotated as “miRNA containing lincRNA”. (b) Precision-Sensitivity curve of human trained MuStARD predictions on mouse pre-miRNAs. Orange line shows the model prediction on human for reference. Solid blue line shows the prediction on all mouse pre-miRNAs, and dashed blue line shows the prediction on mouse pre-miRNAs without a direct human homologue. (c) A visualization of the mouse pre-miRNA evaluation set denoting the number of predicted and non-predicted, orthologous and non-orthologous pre-miRNAs.

loci, it becomes evident that large genomic regions will need to be scanned, and the background variation will be enormous. However, most negative loci have extremely low potential of being confused for small RNA loci. A static classification between real loci and randomly selected background is prone to overestimate the predictive power of evaluated methods. Some methods have attempted to create ‘harder’ negative sets by including sequences that have characteristics similar to the predicted class. This approach implies that the researcher already knows the major characteristics of the predicted class, and that these characteristics will remain stable for each training model. In our case neither of these prerequisites were true.

We initially prototyped our method with a small set of negatives, four for each real training example, randomly selected from regions within the coding and intronic regions of mRNAs, and long non-coding RNAs. We quickly realized that while our method could separate between these categories easily, it still produced a large amount of false positives in the more realistic large region scanning evaluation. Training several models using different, but equal, background sets showed us variability in the number and range of identified false positives. However, we noticed that a number of false positives appeared consistently in several of the trained models. These ‘hard cases’ of background variation are the ones that have sequence, conservation, and folding characteristics closest

			All snoRNAs	C/D box		H/ACA box	
			MuStARD	MuStARD	snoReport	MuStARD	snoReport
Score Threshold 0.5	Homo Sapiens	Precision	0.545	0.476	0.512	0.705	0.100
		Sensitivity	0.791	0.954	0.954	0.764	0.058
		F1	0.645	0.635	0.666	0.733	0.073
	Mus Musculus	Precision	0.549	0.494	0.332	0.730	0.103
		Sensitivity	0.928	0.969	0.897	0.951	0.146
		F1	0.689	0.654	0.484	0.825	0.120
Score Threshold 0.8	Homo Sapiens	Precision	0.820	0.645	0.666	0.909	0.250
		Sensitivity	0.708	0.954	0.954	0.647	0.058
		F1	0.759	0.769	0.784	0.755	0.094
	Mus Musculus	Precision	0.656	0.580	0.420	0.772	0.055
		Sensitivity	0.769	0.897	0.887	0.853	0.024
		F1	0.708	0.704	0.570	0.810	0.033

**Table 3.** Evaluation of prediction for all snoRNA, and CD-box or H/ACA-box subfamilies separately.

to the real training examples and are thus harder to differentiate. We decided to attempt an iterative enrichment technique for the training background in which ‘hard cases’ that confuse our models consistently are added into the training set for a second round of training. This method achieved a great leap in performance when evaluated in completely independent data. Importantly, this automatic iterative method does not rely on an expert user to select the characteristics of importance. The ‘hard cases’ are identified by the training model itself and will fit to whatever positive set it is training on. The enriched background for pre-miRNA and sno-RNA models is radically different, representing the differences of these classes between them, and allowing the models to be easily and accurately trained on any positive set of small RNA loci.

Using a number of pre-miRNA prediction algorithms for region scanning was time consuming and arduous labor. To calculate hundreds of features on regions spanning less than one percent of the human genome, all other algorithms (with miRBoost being the sole exception) required to group the scanning region into smaller batches of 2000 sequences in order to parallelize the analysis into a computer cluster (MetaCentrum-CERIT). Even so, the computing time for each single batch was approximately 4 days. In contrast, our algorithm was able to scan the mouse benchmark dataset that includes several million base pairs in a few hours on a single CPU.

We have demonstrated that our method can be used for cross-species prediction of small RNAs. As a proof of concept we trained models on human pre-miRNAs and snoRNAs and then identified their counterparts in mouse, a pair of well annotated species that have considerable evolutionary distance. The pre-miRNAs we correctly identified on the mouse genome were enriched in evolutionary conserved pre-miRNAs in human (approximately 30% of our true positive predictions vs 10% of all mouse miRNAs). That said, the majority (70%) of our predicted pre-miRNAs are not homologous to human pre-miRNAs and would not be easily identified by a simple homology search.

The method presented here can be generalized for any class of small RNAs on any species. We chose to highlight two examples (pre-miRNAs and sno-RNAs) that differ radically. Where pre-miRNAs have high levels of conservation and fold into characteristic hairpin structures, sno-RNAs show a much wider size distribution (118.8 mean / 59.1 sd vs 81.9 mean / 16.9 sd) and have a variety of subclasses with variable secondary structure and evolutionary patterns, making their identification harder. Thousands of known pre-miRNA sequences against a few hundred sno-RNAs reduce the size of the training set, adding a level of difficulty to the task. It follows that several methods for pre-miRNA identification have been developed to date, while sno-RNA identification methods have not been developed in the past decade. The need for modern, easy to use, easy to train, methods becomes self-evident, especially for RNA classes with fewer members, for which no new development is performed. It is beyond the scope of this paper to develop models for each class of small RNAs, but using our openly available method researchers can easily produce such models for their own RNAs of interest. MuStARD has been specifically designed to automate this process and facilitate ease-of-use by simplifying the input requirements. Regions of the targeted small RNA class can be loaded as a bed file, and MuStARD handles all pre-processing steps such as sequence and evolutionary conservation extraction as well as secondary structure calculation. Additionally, the iterative training module provides an interface for the automatic selection of background genomic loci that optimally represent the negative set, specifically tailored for the specific small RNA class. MuStARD can be easily applied on any small RNA identification problem that would not be easily identifiable by using older methods. Extensive documentation and tutorials on using MuStARD for novel RNA class predictions are available along with the MuStARD code repository at [gitlab.com/RBP\\_Bioinformatics/mustard](https://gitlab.com/RBP_Bioinformatics/mustard).

## Conclusion

To conclude, we have developed a method that is easy to train and deploy for any class of small RNA genomic loci. Using the novel iterative background selection our method can choose the background ‘hard cases’ specific for each training, boosting performance. We show that our method outperforms class specific methods, both in accuracy, and computational performance. We achieved cross species identification of small RNAs beyond homology, and also highlighted a realistic use case in the identification of pre-miRNAs out of small RNA-Seq peaks.

## References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
4. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
5. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nat. Genet.* **41**, 572–578 (2009).
6. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
7. Wang, Q.-H. *et al.* Systematic analysis of human microRNA divergence based on evolutionary emergence. *FEBS Letters* **585**, 240–248 (2011).
8. Saçar Demirci, M. D., Baumbach, J. & Allmer, J. On the performance of pre-microRNA detection algorithms. *Nat. Commun.* **8**, 330 (2017).
9. Lestrade, L. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research* **34**, D158–D162 (2006).
10. Xie, J. *et al.* Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. *Nucleic Acids Res.* **35**, D183–7 (2007).
11. Makarova, J. A. & Kramerov, D. A. SNOntology: Myriads of novel snornas or just a mirage? *BMC Genomics* vol. 12 (2011).
12. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
13. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
14. Thurmond, J. *et al.* FlyBase 2.0: the next generation. *Nucleic Acids Res.* **47**, D759–D765 (2019).
15. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gky1141> (2018).
16. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–6 (2004).
17. Roden, C. *et al.* Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation. *Genome Res.* **27**, 374–384 (2017).
18. Henry, V. J., Bandrowski, A. E., Pepin, A.-S., Gonzalez, B. J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database* **2014** (2014).
19. Gudyś, A., Szcześniak, M. W., Sikora, M. & Makałowska, I. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics* **14**, 83 (2013).
20. Batuwita, R. & Palade, V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* **25**, 989–995 (2009).
21. Jiang, P. *et al.* MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* **35**, W339–44 (2007).
22. Tran, V. D. T., Tempel, S., Zerath, B., Zehraoui, F. & Tahri, F. miRBoost: boosting support vector machines for microRNA precursor classification. *RNA* **21**, 775–785 (2015).
23. Xue, C. *et al.* Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* **6**, 310 (2005).
24. Hertel, J., Hofacker, I. L. & Stadler, P. F. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* **24**, 158–164 (2008).

## Acknowledgements

This research was funded by Postdoc@MUNI with project registration number CZ.02.2.69/0.0/0.0/16\_027/00083 60 grant to G.K.G.; the fellowship program Brno Ph.D. talent 2017 and the Associazione Italiana Ricerca Cancro's fellowship 2018 no. 22620 to A.G. Grantová Agentura České Republiky, 19-10976Y Grant to P.A. Funding for open access charge: Postdoc@MUNI with project registration number CZ.02.2.69/0.0/0.0/16\_027/0008360.

## Author contributions

P.A. and G.K.G. conceived the study. P.A. oversaw the whole study. G.K.G. developed the code, implemented all analyses and comparisons. A.G. applied HuntMi and microPred on the benchmark datasets and performed the mouse/human pre-miRNA homology analysis. G.K.G. and E.C. applied miPred on the benchmark datasets. K.G.L. and F.C.P. applied miRBoost and triplet-SVM on the benchmark datasets. P.A. and G.K.G. wrote the manuscript and prepared the figures.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-66454-3>.

**Correspondence** and requests for materials should be addressed to P.A.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

## Chapter 4

Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study.

Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V, Fazio G, Froňková E, Giraud M, **Griani A**, Hancock J, Herrmann D, Jiménez C, Krejci A, Moppett J, Reigl T, Salson M, Scheijen B, Schwarz M, Songia S, Svaton M, van Dongen JJM, Villarese P, Wakeman S, Wright G, Cazzaniga G, Davi F, García-Sanz R, Gonzalez D, Groenen PJTA, Hummel M, Macintyre EA, Stamatopoulos K, Pott C, Trka J, Darzentas N, Langerak AW; EuroClonality-NGS working group. Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. *Leukemia*. 2019 Sep;33(9):2241-2253. doi: 10.1038/s41375-019-0496-7. Epub 2019 Jun 26. PMID: 31243313; PMCID: PMC6756028.



Minimal residual disease

# Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study

Monika Brüggemann<sup>1</sup> · Michaela Kotrová<sup>1,2</sup> · Henrik Knecht<sup>1</sup> · Jack Bartram<sup>3</sup> · Myriam Boudjogrha<sup>4</sup> · Vojtech Bystry<sup>5</sup> · Grazia Fazio<sup>6</sup> · Eva Froňková<sup>2</sup> · Mathieu Giraud<sup>7</sup> · Andrea Grioni<sup>6</sup> · Jeremy Hancock<sup>8</sup> · Dietrich Herrmann<sup>1</sup> · Cristina Jiménez<sup>9</sup> · Adam Krejci<sup>5</sup> · John Moppett<sup>10</sup> · Tomas Reigl<sup>5</sup> · Mikael Salson<sup>7</sup> · Blanca Scheijen<sup>11</sup> · Martin Schwarz<sup>1</sup> · Simona Songia<sup>6</sup> · Michael Svaton<sup>2</sup> · Jacques J. M. van Dongen<sup>12</sup> · Patrick Villarese<sup>13</sup> · Stephanie Wakeman<sup>8</sup> · Gary Wright<sup>3</sup> · Giovanni Cazzaniga<sup>6</sup> · Frédéric Davi<sup>4</sup> · Ramón García-Sanz<sup>9</sup> · David Gonzalez<sup>14</sup> · Patricia J. T. A. Groenen<sup>11</sup> · Michael Hummel<sup>15</sup> · Elizabeth A. Macintyre<sup>13</sup> · Kostas Stamatopoulos<sup>16</sup> · Christiane Pott<sup>1</sup> · Jan Trka<sup>2</sup> · Nikos Darzentas<sup>1,5</sup> · Anton W. Langerak<sup>17</sup> · on behalf of the EuroClonality-NGS working group

Received: 15 January 2019 / Accepted: 20 February 2019 / Published online: 26 June 2019  
© The Author(s) 2019. This article is published with open access

## Abstract

Amplicon-based next-generation sequencing (NGS) of immunoglobulin (IG) and T-cell receptor (TR) gene rearrangements for clonality assessment, marker identification and quantification of minimal residual disease (MRD) in lymphoid neoplasms has been the focus of intense research, development and application. However, standardization and validation in a scientifically controlled multicentre setting is still lacking. Therefore, IG/TR assay development and design, including bioinformatics, was performed within the EuroClonality-NGS working group and validated for MRD marker identification in acute lymphoblastic leukaemia (ALL). Five EuroMRD ALL reference laboratories performed IG/TR NGS in 50 diagnostic ALL samples, and compared results with those generated through routine IG/TR Sanger sequencing. A central polytarget quality control (cPT-QC) was used to monitor primer performance, and a central in-tube quality control (cIT-QC) was spiked into each sample as a library-specific quality control and calibrator. NGS identified 259 (average 5.2/sample, range 0–14) clonal sequences vs. Sanger-sequencing 248 (average 5.0/sample, range 0–14). NGS primers covered possible IG/TR rearrangement types more completely compared with local multiplex PCR sets and enabled sequencing of bi-allelic rearrangements and weak PCR products. The cPT-QC showed high reproducibility across all laboratories. These validated and reproducible quality-controlled EuroClonality-NGS assays can be used for standardized NGS-based identification of IG/TR markers in lymphoid malignancies.

These authors contributed equally: Monika Brüggemann, Michaela Kotrová

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41375-019-0496-7>) contains supplementary material, which is available to authorized users.

✉ Anton W. Langerak  
a.langerak@erasmusmc.nl

Extended author information available on the last page of the article.

## Introduction

Specific antigen recognition by cells of the adaptive immune system (B cells, T cells) is mediated through receptors (immunoglobulin, IG, and T-cell receptor, TR) that are uniquely formed during immune development in bone marrow and thymus, respectively. Through recombination of IG/TR loci a diverse (polyclonal) repertoire of unique IG/TR receptors is created. In certain autoimmune diseases this repertoire is skewed (oligoclonal), whereas in lymphoid malignancies receptors are largely identical (monoclonal) [1–7]. IG/TR rearrangements thus form



unique genetic biomarkers (molecular signatures) for studying immune cells for clinical, diagnostic and research applications [8–11]. Classically, methods for immunogenetic analysis mostly concern fragment analysis and Sanger-based sequencing. The introduction of NGS makes deeper analysis of IG/TR rearrangements possible, with impact on the main immunogenetic applications: clonality assessment, MRD detection, repertoire analysis [12–29].

The EuroClonality-NGS working group (euro-clonalityngs.org; Supplementary Figure 1) has ample expertise in development, standardization and validation of IG/TR assays, to address the challenges in the translational research towards clinical application.

Here we report on the development and standardization (see also accompanying manuscript by Knecht et al. [30]) of novel amplicon-based IG/TR NGS assays between September 2012 and October 2017, via a total of 14 international coordination and evaluation meetings (Supplementary Table 1). This study focuses on IG/TR marker identification in lymphoid malignancies for subsequent MRD analysis, and their multicentre validation in acute lymphoblastic leukaemia (ALL). Assay optimizations and modifications for other applications of IG/TR NGS are partly still ongoing and will be reported in separate publications.

## Materials and methods

### General concept of assay design

With the objective of developing a universal amplicon-based NGS approach for IG/TR sequence analysis at the DNA level, applicable in all lymphoid malignancies, assays for multiple IG/TR loci were designed: IG heavy (IGH), IG kappa (IGK), TR beta (TRB), TR gamma (TRG) and TR delta (TRD), including complete and incomplete rearrangements whenever applicable. IG lambda (IGL) was excluded due to its limited complementarity to other IG loci and its reduced diversity. TR alpha (TRA) was excluded due to its high complexity, severely hampering a reasonable multiplex PCR approach at the DNA level.

The IGH locus is rearranged in two steps. After initial coupling of a single IGHD gene to an IGHJ gene, an IGHV gene is joined to the incomplete IGHD–IGHJ rearrangement, resulting in a complete IGHV–IGHJ rearrangement. For amplification of complete IGH rearrangements, primers located in the FR1, FR2 and FR3 regions were designed, but here we only discuss the FR1 assay for marker identification in ALL (for application of IGH-VJ-FR3 assay in clonality testing see accompanying manuscript by Scheijen et al. [31]). IGHD–IGHJ rearrangements were amplified in a separate multiplex PCR reaction. The IGK light chain locus is composed of functional IGKV and IGKJ genes, as well as

the so-called kappa deleting element (Kde) that can rearrange to IGKV genes, or to a recombination signal sequence (RSS) in the IGKJ–IGKC intron, leading to functional inactivation of the IGK allele. The IGKV forward primers were designed to be used in combination with IGKJ and Kde reverse primers in one multiplex reaction, whereas a second PCR was developed for the forward intron RSS and reverse Kde primers.

The TRB locus also features a two-step process with initial formation of incomplete TRBD–TRBJ rearrangements followed by complete TRBV–TRBJ rearrangements. Incomplete and complete TRB rearrangements are detected in two separate multiplex PCR reactions. As TRG locus rearrangements are one-step VJ recombinations involving a limited number of TRGV and TRGJ genes, a single multiplex assay could be developed. Finally, in the TRD locus, complete VJ rearrangements are preceded by DD, VD and DJ rearrangements. In addition, certain TRAV genes can rearrange to both TRDJ and TRAJ, whereas TRDV–TRAJ rearrangements, usually involving TRAJ29, can also occur. All of these rearrangements were designed to be amplified in one multiplex PCR assay.

Both the design and further testing were coordinated by the respective ‘Target’ network leaders: IGH-VJ by C. Pott, Kiel and R. Garcia Sanz, Salamanca; IGH-DJ by F. Davi, Paris and K. Stamatopoulos, Thessaloniki; IGK-V/intron-IGKJ/Kde by P.J.T.A. Groenen, Nijmegen and A.W. Langerak, Rotterdam; TRB by M. Brüggemann, Kiel and M. Hummel, Berlin; TRG by G. Cazzaniga, Monza and J.J.M. van Dongen, Leiden; and TRD by E. Macintyre, Paris. Initial testing of each assay was performed by 2–3 experienced laboratories per target and final assays were validated for IG/TR marker identification in ALL in a multicentre setting. In addition, central quality control procedures were developed to monitor assay performance.

The bioinformatic platform ARResT/Interrogate [32], developed from the ground-up within the EuroClonality-NGS to assist with its multi-faceted activities, was further adapted for this study as described below.

### Primer design and technical validation of primer performance

Primers were designed to be gene-specific, but in case of allelic variants, degenerate primers were designed to avoid differential annealing in individuals with different allelic variants. For the same reason, single mismatches in the middle or at the 5′-end of the primer were accepted.

Primer3 [33], Primer Digital (PrimerDigital Ltd, Helsinki, Finland) MFEprimer-2.0 [34] and Oligo (Molecular Biology Insights, Inc., Colorado, USA) were used for checking primer specificity and multiplexing. Common primer design criteria were followed for all loci:

primer melting temperature 57–63 °C; comparable size of final amplicon; primer length 20–24 nt; avoidance of primer dimers; minimal distance of 3' primer end to the junctional region of, preferably, >10–15 bp to avoid false-negativity for rearrangements with larger nucleotide deletions from the germline sequence; avoidance of regions with known single nucleotide polymorphisms to allow identical primer annealing for all alleles of the respective V, D or J genes; targeting of, preferably, all V, D and J genes known to be rearranged plus the intronRSS and Kde regions for IGK.

Following *in silico* design, primers were first tested in monoplex and multiplex reactions using primary patient samples or cell lines with defined rearrangements. In occasional cases where no such samples were available, healthy tonsil or mononuclear DNA samples were employed. Oligoclonal template pools were then created from mixtures of rearranged cell lines and diagnostic samples with defined rearrangements covering many different V, D and/or J genes. Alternatively, for some loci, plasmid pools were produced, covering as many different rearrangements as possible. These multi-target pools allowed fine-tuning of reaction conditions and/or primer concentrations to assess comparable amplification efficiencies. This iterative process of testing also led to a reduction of primers if these appeared redundant. Further multicentre testing was performed with a limited number of monoclonal and poly/oligoclonal samples on different sequencing platforms, which allowed assessment of robustness of the primer mixes and protocols.

As assays were designed with the aim to be platform-independent, a two-step PCR was employed, that enabled switching of sequencing adaptors and to reduce the total number of primers even if many barcodes are necessary. Also, maximal amplicon lengths were defined with respect to the possible maximal sequencing read lengths of current sequencers. PCR conditions were optimized with the aim to find optimal conditions common for all reactions, thus allowing for parallel library preparation. Various numbers of PCR cycles in 1st and 2nd PCR, different polymerases and several library purification methods were tested and compared.

### **Multicentre validation of assays for MRD marker identification in ALL**

Five experienced laboratories tested the robustness and applicability of the optimized assays for NGS-based IG/TR marker identification in ALL in comparison to standard techniques. All laboratories (Bristol/London, Paris, Monza, Prague and Kiel) are members of the EuroMRD consortium and reference laboratories for ALL MRD analysis. Each of them performed NGS-based IG/TR MRD marker identification in 10 patients with B- or T-lineage ALL. A central

standard operating procedure was strictly followed. The study was executed using the Illumina MiSeq (2 × 250 bp v2 kit). NGS analyses were performed fully in parallel to conventional PCR plus Sanger sequencing of clonal products following standard guidelines [11]. For a part of the cases with unexplained discrepant results between the two methods, allele-specific PCR assays (either for digital droplet PCR or real-time quantitative PCR) were designed to clarify if the respective clonal rearrangement represented the leukaemic bulk. EuroMRD guidelines were used to design and interpret allele-specific PCR assays [35, 36].

## **Results**

### **Primer design and technical validation of primer performance**

Based on the results of the testing and validation phases (Supplementary Table 2), the final IG/TR primer mixes consisted of eight tubes with 92 forward and 30 reverse primers, 15 of the latter being used in pairs of different tubes (Supplementary Table 3). Primer positions and sequences are presented in Fig. 1.

### **Implementation of quality control procedures**

Quality control of robust amplification, library preparation and sequencing are of utmost importance for these complex assays. Different primers need to work under the same reaction conditions, while additional variability can be introduced by sample characteristics and sequencing. Primer performance must be monitored longitudinally, and for the exact estimation of clonal abundance it is important to correct for the number of sequencing reads per input molecule.

To address these issues, we established and validated two types of quality control procedures: (i) a 'central in-tube quality control' (cIT-QC) spiked to each tube as library control and calibrator, and (ii) a 'central poly-target quality control' (cPT-QC), or run control, to monitor general primer performance and sequencing.

To compose the cIT-QC, IG/TR rearrangements of many human lymphoid cell lines were comprehensively characterized by amplicon- and capture-based NGS and Sanger sequencing. Nine cell lines were selected to form the cIT-QC with at least three different clonal rearrangements for each of the eight PCR tubes, totalling 24 rearrangements. The current design requires an equal number of cell line DNA copies to be spiked into each tube, as described below.

For the cPT-QC a mixture of different lymphoid specimens was considered to cover the whole IG/TR repertoire



**Fig. 1** Schematic diagrams of rearrangements and primer sets. **a** Schematic diagrams of IGHV-IGHJ and IGHD-IGHJ rearrangements. The relative position of the VH family primers, DH family primers and consensus JH primers is given according to their most 5' nucleotide upstream (–) or downstream (+) of the involved RSS. **b** Schematic diagrams of IGKV-IGKJ rearrangement and the two types of Kde rearrangements (V-Kde and intronRSS–Kde). The relative position of the IGKV, IGKJ, Kde, and intronRSS (INTR) primers is given according to their most 5' nucleotide upstream (–) or downstream (+) of the involved RSS. **c** Schematic diagrams of TRBV-TRBJ rearrangement and TRBD-TRBJ rearrangement. The relative position of the TRBV family primers, TRBD primers and the TRBJ primers is given

according to their most 5' nucleotide upstream (–) or downstream (+) of the involved RSS. **d** Schematic diagrams of TRGV-TRGJ rearrangement and the relative position of the TRGV and the TRGJ primers. The relative position of the TRGV primers and the TRGJ primers is given according to their most 5' nucleotide upstream (–) or downstream (+) of the involved RSS. **e** Schematic diagram of TRDV-TRDD-TRDJ, TRDD-TRDD, and TRDV-TRDD, TRDV-TRAJ29 rearrangements, showing the positioning of TRDV, TRDJ, TRDD, and TRAJ29 primers, all combined in a single tube. The relative position of the TRDV, TRDD, and TRDJ primers is indicated according to their most 50 nucleotides upstream (–) or downstream (+) of the involved RSS

more comprehensively. To this end we produced material consisting of equal ratios of DNA from peripheral blood mononuclear cells (MNCs), thymus and tonsil. For more details see accompanying manuscript by Knecht et al. [30].

**Laboratory protocol**

Primers were tailed with universal and T7-linker sequences, and divided over eight tubes (IGH-VJ, IGH-DJ, IGK-VJ-Kde, intron-Kde, TRB-VJ, TRB-DJ, TRG, TRD). The PCR

protocol is summarized in Table 1. Sequencing libraries were prepared via a two-step PCR, each using a final reaction volume of 50 µl with 100 ng diagnostic DNA and 10 ng of polyclonal DNA. For the cIT-QC, 40 cell equivalents of the nine different cell lines were spiked into all samples (see accompanying manuscript by Knecht et al. [30]). MgCl<sub>2</sub> was intended to be used at a final concentration of 1.5 mM, but needed optimization for some tubes. Therefore, master-mixes for the 1st PCR were tube-specific, but the temperature profile was uniform for all tubes.

**Table 1** Standardized PCR protocol

(a) Reaction conditions of 1st and 2nd PCR

	IGH V-J		IGH D-J		IGK-VJ-Kde, intron-Kde		TRB V-J, D-J		TRG		TRD	
	Final concentration	μl/library	Final concentration	μl/library	Final concentration	μl/library	Final concentration	μl/library	Final concentration	μl/library	Final concentration	μl/library
PCR Buffer II	1x	5	1x	5	1x	5	1x	5	1x	5	1x	5
MgCl <sub>2</sub>	2.5 mM	5	3 mM	6	1.5 mM	3	4 mM	8	4 mM	8	2 mM	4
dNTP-Mix	0.2 mM	1	0.4 mM	2.0	0.2 mM	1	0.2 mM	1	0.2 mM	1	0.2 mM	1
EagleTaq/ AmpliTaq Gold	1 U/rxn	0.2	1.5 U/rxn	0.3	1 U/rxn	0.2	1 U/rxn	0.2	1 U/rxn	0.2	1 U/rxn	0.2
2nd PCR	Stock concentration											
	all tubes											
	Final concentration											
PCR buffer with MgCl <sub>2</sub>	10x											
	1.8 mM											
dNTP-Mix	0.2 mM											
Fast Start High Fidelity polymerase	2.5 U/rxn											
(b) Cycling conditions												
1st PCR	2nd PCR											
1 cycle	Initial denaturation	94 °C	10 min	1 cycle	Initial denaturation	95 °C	2 min					
35 cycles	Denaturation	94 °C	1 min	20 cycles	Denaturation	94 °C	30 s					
	Annealing	63 °C	1 min		Annealing	63 °C	30 s					
	Extension	72 °C	30 s		Extension	72 °C	30 s					
1 cycle	Final extension	72 °C	30 min	1 cycle	Final extension	72 °C	5 min					
		12 °C	∞			12 °C	∞					

Reaction volume: 50 μl

**Table 2** Mean size of PCR products after the 2nd PCR (containing the Illumina sequencing adaptors and barcodes)

Gene	Amplicon length (bp)
TRB-VJ	309–407
TRB-DJ	300–408
TRG	256–360
TRD	309–450
IGH-VJ	484–681
IGH-DJ	266–358
IGK-VJ-Kde	296–384
intron-Kde	309–382

Concentrations of all primers are shown in Supplementary Table 3. After 1st PCR, gel electrophoresis was performed to check for successful amplification of all targets. For TRB, gel extraction of the specific PCR products was performed prior to the 2nd PCR.

All 1st round PCR products, except TRB PCR products, were diluted 1:50 unless amplicons were very weak. TRB PCR products and PCR products with weak amplicons were used undiluted. Master-mixes for the 2nd PCR and the temperature profiles were identical for all tubes (Table 1). Primers for the 2nd PCR contained sequencing adaptors and sequencing indexes (barcodes). Unique combination of forward and reverse indexes was used for each library. Three microlitres of undiluted TRB PCR products and 1  $\mu$ l of 1:50-diluted IGH, IGK, TRG and TRD PCR products were amplified in the 2nd PCR.

Following 2nd PCR, products from all samples of a run were pooled in equimolar ratios into eight tube-wise subpools and purified by gel extraction (see Table 2 for the amplicon lengths). Finally, the subpools were pooled equimolarly into one final pool. Sequencing was performed on Illumina MiSeq sequencers, using 2  $\times$  250 bp v2 chemistry with a final concentration of 7 pM for the amplicon library and 10% PhiX control added to avoid low-complexity library issues. The detailed standard operating procedure is provided as supplementary information.

### Bioinformatic protocol

ARResT/Interrogate [32] was the main bioinformatics platform used in this study. Both Vidjil [37] and IMGT [38] resources are available through ARResT/Interrogate as built-in tools and were employed for specific aspects of this work, mainly analysis of rearrangements with unclear annotation. Data are deposited at EMBL/EBI European Nucleotide Archive (ENA), accession code PRJEB32668.

Demultiplexing was performed accepting no mismatches. Reads were annotated with EuroClonality-NGS primer sequences (to trim non-amplicon sequences, and for

the cIT-QC-based quality control), paired-end joined, dereplicated, immunogenetically annotated [39], and eventually classified into rearrangement types (complete and incomplete, and other special types like intron-Kde rearrangements), or ‘junction classes’. Reads without rearrangement were excluded from the total read count used for relative abundances.

cIT-QC sequences described above and elsewhere (see accompanying manuscript by Knecht et al. [30]), were identified in the data through their immunogenetic annotation. Their counts served both as ‘in-tube’ control and for normalization per primer set: total cIT-QC cells are divided by cIT-QC total reads, the resulting factor used to convert rearrangement reads to cells, and those cells then further divided by total input cells (15,000 in this study). Identified IG/TR sequences were defined as index sequences if their abundance after cIT-QC normalisation exceeded 5%.

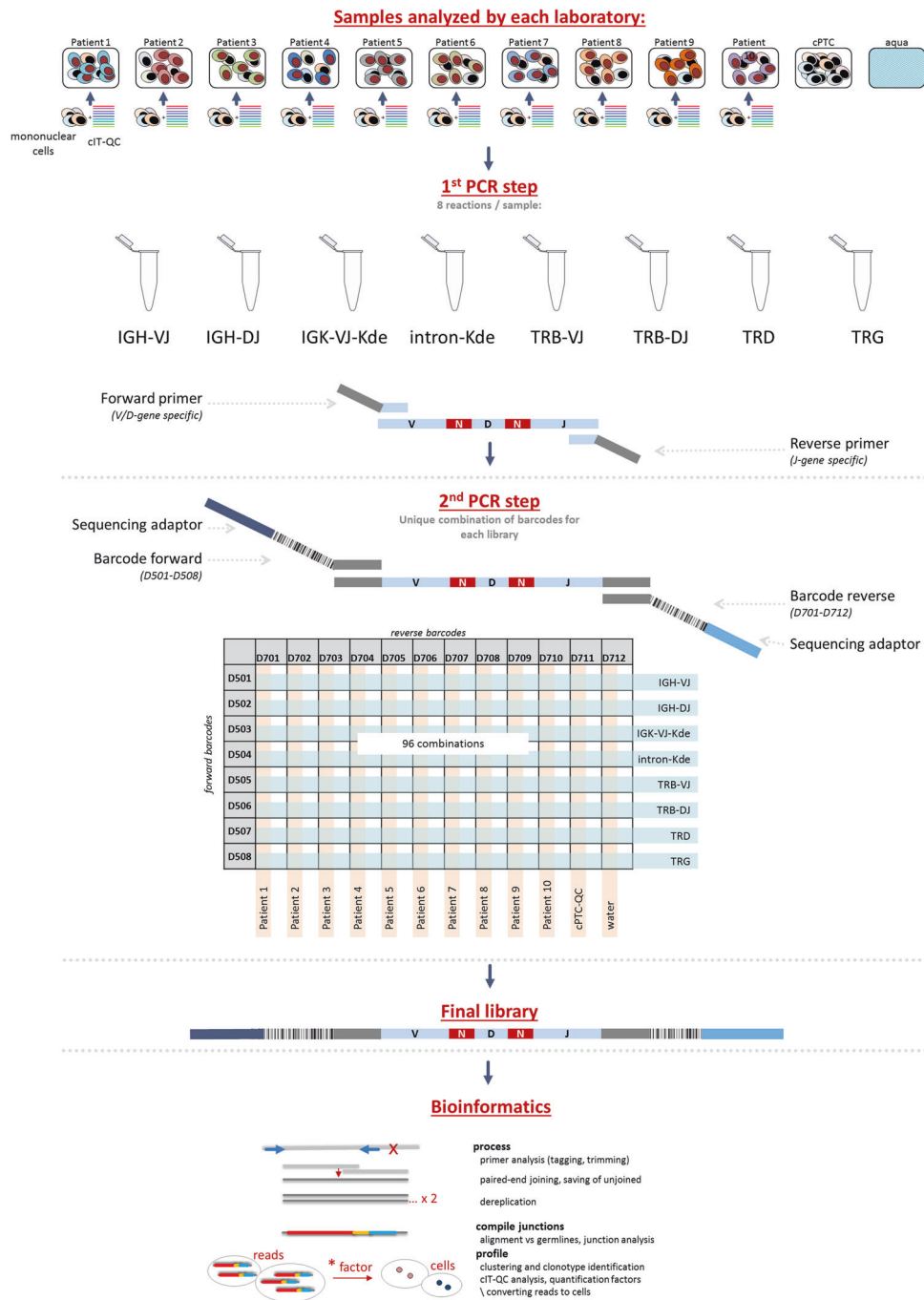
ARResT/Interrogate can track the DNJ 3’ stem of a junction, the sequence remaining stable during IGH or TRB clonal evolution in case of V replacement or ongoing V to DJ rearrangements. The stem consists of the last  $\leq$  3nt of D (or of the NDN if no D is identifiable), any and all of N2 nucleotides, and the J nucleotides of the junction. This stem is available as a separate immunogenetic feature across all samples and thus can be linked to other features, e.g. clonotypes.

### Multicentre validation of assays for MRD marker identification in ALL

Next, 50 ALL diagnostic samples (29 BCP-ALL and 21 T-ALL; Supplementary Table 4) were analysed for the multicentre validation study. Each of the five participating laboratories received preconfigured 96-well plates containing the different multiplexed NGS primer combinations per target (Fig. 2).

In total, 96 libraries were generated per lab (total of 480 libraries), and sequenced with a collective output of 47M reads ( $\approx$  9.2 M/lab). Centralised analysis was performed with ARResT/Interrogate [32] using IMGT germline sequences [39]—further analyses and verifications were performed with Vidjil [37] and IMGT/V-QUEST [38].

Overall, 311 clonal IG/TR rearrangements (clonotypes) were identified, with a mean of 5.2 (0–14)/sample by NGS (a 5% threshold was applied for NGS after cIT-QC-based normalization) vs. 5.0 (0–14)/sample by Sanger, while 217 (45%) libraries demonstrated no clonotypes above threshold by either method. A total of 196/311 (63%) clonotypes were fully concordant between NGS and Sanger (Fig. 3). NGS exclusively identified 63/311 (20%) index sequences, whereas 52/311 (17%) IG/TR Sanger sequences were not assigned as NGS index sequence by ARResT/Interrogate. 26/63 NGS positive/Sanger negative cases showed a clonal PCR product



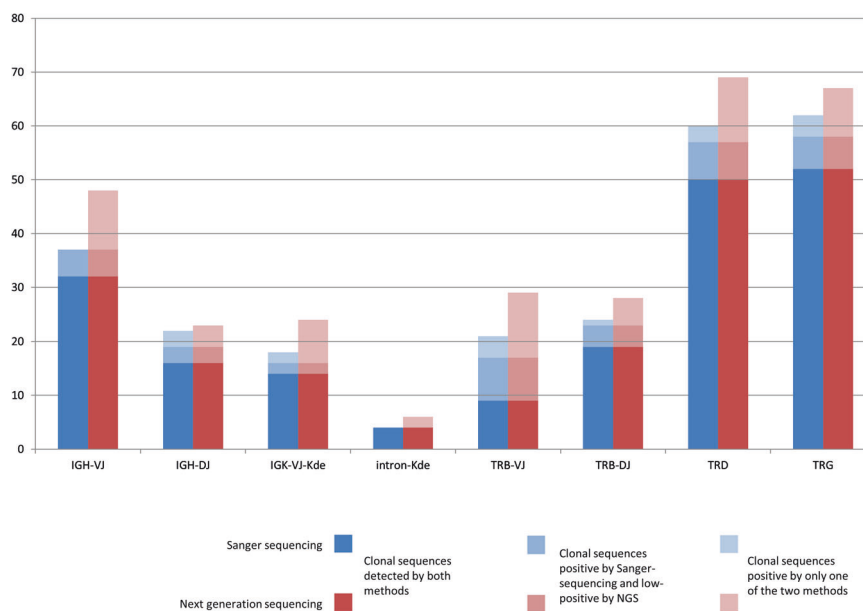
**Fig. 2** Schematic overview of the workflow for multicentre validation of IG/TR NGS assays for MRD marker identification in ALL. The IG and TR gene rearrangements are amplified in a two-step approach using multiplex PCR assays. Each of the participating laboratories performed NGS-based IG/TR MRD marker identification in 10

patients with ALL. A central polytarget control (cPT-QC) was used to monitor primer performance, and central in-tube controls (cIT-QC) were spiked to each sample as library-specific quality control and calibrator. Pipetting was performed in a 96-well format. The data analysis was performed using ARRES/Interrogate

also in the respective low-throughput approach but subsequent Sanger sequencing failed due to polyclonal background, mixed sequences or weak PCR products. In an additional 6/63 NGS positive/Sanger negative cases the respective primer was missing in the low-throughput approach. For the remaining 31/63 discrepancies no

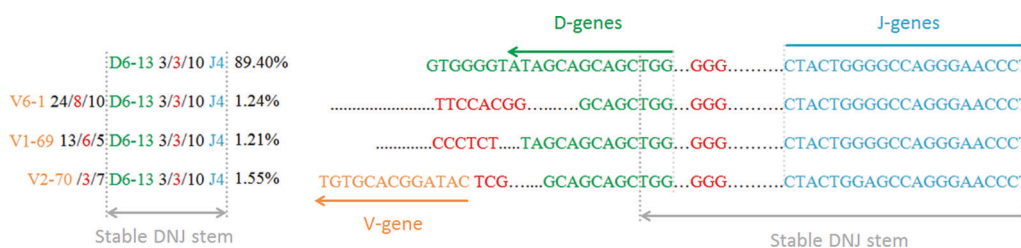
technical explanation for Sanger failure could be found. In 16/19 q/ddPCR evaluated cases the rearrangement was confirmed by ASO-PCR, in three of these on a subclonal level.

Conversely, 52/311 clonal IG/TR rearrangements were detected by Sanger sequencing only, when applying the 5% NGS threshold: for 5/52 sequences (1 TRG, 2 TRB-VJ and



**Fig. 3** Results of multicentre validation of assays for MRD marker identification in ALL. Blue: Index sequences identified by Sanger sequencing. Red: Index sequences identified by NGS. Darkest blue/red are clonal sequences identified by both methods; lightest blue/red are

sequences identified only by the respective method. Median blue/red are clonal sequences identified by both methods, but by NGS with an abundance of <5% after normalization



**Fig. 4** Clonal evolution in a BCP-ALL patient. The dominant incomplete IGH rearrangement (IGHD6-13 - IGHJ4) was identified with an abundance of 89.4% together with three additional complete

IGH rearrangements with lower abundance (1.21–1.55%) and the same DNJ sequence. Only the CDR3 region is shown for each sequence

2 IGH-DJ) the relevant primer was not present in the NGS primer set, in 12/52 cases no explanation was found for the discrepancy. However, in most discordant cases (35/52) the Sanger identified sequences (7 TRD, 8 TRB-VJ, 6 TRG, 4 TRB-DJ, 2 IGK-VJ-Kde, 5 IGH-VJ and 3 IGH-DJ) that were also detectable by NGS, but with an abundance below 5%. In 36/39 q/ddPCR evaluated cases the rearrangement was confirmed by ASO-PCR (including all low NGS positive sequences), in 14 of these on a subclonal level. The overall concordance between Sanger and NGS, including negative libraries, was 78%.

Interestingly, in 12/29 B-lineage ALL samples the evolution of the dominant clonal IGH sequence was identified employing a special tool in ARResT/Interrogate. The evolved clonotypes shared the DNJ stem with the dominant one, but the VND part of the rearrangement differed (example in Fig. 4).

Assay performance was also analysed by standardized evaluation of QC samples (cIT-QC and cPT-QC, see

accompanying manuscript by Knecht et al. [30]) and showed high intra- and inter-lab consistency without statistically significant differences between the five labs.

### Modifications of the central SOP

During the process of multicentre validation, modifications of the SOP were tested in particular laboratories as parallel projects.

### One-step versus two-step PCR

It was decided to use two-step PCR to enable switching of sequencing adaptors and to limit the total number of required primer batches even if a large number of barcodes is required. As first round PCR products are not barcoded, identification of contamination phenomena is hampered in this approach. Therefore, a one-step PCR was tested in a single centre (Paris). The one-step approach reduces the risk of

contamination and thus favours use of NGS not only for marker identification, but also for MRD assessment. The standard operating procedure is shown in Supplementary information.

### Use of Ion Torrent platform

Ion Torrent platform was tested in a single-centre setting (Prague) and showed a very good concordance ( $R^2 = 0.89$ ) with the standard approach. The standard operating procedure is shown in Supplementary information.

### Removal of polyclonal DNA from reaction mix

Polyclonal DNA was added to each reaction in order to prevent excessive primer dimer formation in samples lacking particular rearrangements. The addition of polyclonal DNA, however, alters the composition of polyclonal background of the samples and hampers the analysis of the immune repertoire. We therefore performed testing on four samples with B- and four samples with T-cell aplasia and showed that addition of cIT-QC is sufficient to prevent the excessive formation of unspecific PCR products (see Supplementary information).

### Bead extraction

During the single target evaluation and validation phase, gel extraction of the specific TRB amplicons turned out to lead to more specific libraries compared with bead extraction. However, gel extraction is not used in all laboratories, therefore, in a later phase of the study bead purification of all libraries was also tested. Optimization of the purification processes led to comparable ratios of specific reads irrespective of the type of library purification (Supplementary Table 5).

## Discussion

Amplicon-based IG/TR NGS provides an elegant method to detect clonality, identify MRD markers and monitor MRD in lymphoid malignancies. However, comprehensive SOPs for all relevant IG/TR targets, applicable QC procedures, suitable bioinformatic tools, and validation of the technology in a scientifically controlled, multicentre setting are still lacking [19].

Here we describe an *in vitro* and *in silico* protocol for the diagnostic identification of IG/TR MRD markers in ALL, and demonstrate its robustness and applicability across five European laboratories. EuroClonality-NGS primer sets were successfully used with high reproducibility and good concordance to Sanger sequencing, identifying on average 4% more markers per patient than classical low-throughput

methods. NGS was particularly successful in correctly identifying bi-allelic rearrangements, which are technically challenging for Sanger sequencing because this requires prior separation of the respective clonal PCR products. NGS also performs better in the presence of a background of polyclonal rearrangements. Besides, it allows a more comprehensive coverage of rearrangement types. The EuroClonality-NGS TRD assay for example not only detects all types of complete and incomplete TRD gene rearrangements but also VD-JA29 recombinations [40], present in about 20% of all B-cell precursor (BCP) ALLs. In our current series, these TRDV2-JA29 rearrangements were detected in 7/29 BCP-ALL patients (24%), providing an attractive target for MRD monitoring. Notably, rearrangement coverage is not complete. The IGH-DJ tube lacks an IGHD7 primer because that would predominantly amplify the germline-configured IGH-IGHD7-IGHJ1.

Low-throughput sequencing of clonal IG/TR gene rearrangements is often cumbersome. This particularly holds true for TRB, where Sanger sequencing of clonal TRB BIOMED-2 amplicons requires a multistep approach: first with the complete set of primers to identify the rearranged genes, and second, a repetition of the sequencing reaction with gene-specific primers. In contrast, the EuroClonality-NGS assays do not require specific workflows for particular targets, thus enormously streamlining the process of MRD marker identification. This becomes increasingly important in times of MRD-based treatment requiring early patient assignment to the respective MRD risk group.

Critically, our assays provide ways to evaluate primer performance and overall quality of the whole NGS run (primers in the cPT-QC) and of each tube (spike-ins in the cIT-QC, see accompanying manuscript by Knecht et al. [30]). Such functionalities are embedded in the ARResT/Interrogate pipeline, further standardizing the whole workflow. A challenge for correct MRD marker identification in NGS data is the phenomenon of accompanying lymphoid clones that might be mixed up with the leukaemia-specific ones. Therefore, information regarding blast infiltration of the analysed sample must be related to the combined abundance information of the clonal rearrangement, the polyclonal background and the cIT-QC sequences. The integration of all this information allows for a more specific assignment compared with published approaches that define an index sequence simply as sequence with an abundance of >5% [16]. This is particularly necessary for tubes that exclusively cover rearrangements being present only in a minority of lymphoid cells (especially the TRD and intron-Kde tubes). TRD genes are not rearranged in normal B cells and are deleted in most TR $\gamma$  $\delta$  cells [41]. Therefore, oligoclonal TCR $\gamma$  $\delta$  T cells might give rise to dominant clonotypes in TRD NGS assay, in particular as the normal TCR $\gamma$  $\delta$  T-cell repertoire is strikingly skewed during childhood.



Here the cIT-QC-based abundance correction is of utmost importance to avoid miss-assignment of (minor) clonal TRD rearrangements from minor TCR $\gamma\delta$  cell populations as leukaemic rearrangements. Also, knowledge on rearrangement patterns in ALL is important. BCP-ALL features neither complete TRD, nor TRBJ1 gene rearrangements, T-ALL in contrast generally does not harbour complete IGH or IGK gene rearrangements [42]. Hence, identification of such rearrangements would actually reflect more the presence of accompanying T- and B-cell clones, respectively. This immunogenetic knowledge is of particular importance if marker identification is performed, e.g. at relapse after stem cell transplantation, when patients often harbour a restricted B- and T-cell repertoire. The EuroClonality-NGS approach allows for the bioinformatic identification and correction of this phenomenon, whereas conventional low-throughput approaches do not harbour correction mechanisms. Nevertheless, we urge caution in assignment of minor clones to the ALL. Although smaller subclones might be missed based on an abundance threshold (which largely explains discrepancies between Sanger sequencing and NGS in our study), decreasing the threshold would be at the expense of specificity.

Oligoclonality is a well-known phenomenon in ALL that hampers conventional IG/TR MRD [43] assessment, but this can be better identified by NGS. Multiple IG/TR gene rearrangements in ALL result from both continuing rearrangement processes (e.g. continuing IGHV to DJ joining) and from secondary rearrangements (e.g. IGH-DJ replacements, V replacement in a complete IGH rearrangement) [23, 44–49]. In 12 of 29 (41.4%) patients with B-lineage ALL, a dominant clonal IGH rearrangement was subjected to clonal evolution, resulting in the presence of smaller subclones with the same D-J stem, but different V-genes. D-J replacements are also an evolutionary possibility but cannot be unambiguously discriminated from unrelated lymphoid clones even with sophisticated bioinformatic tools.

Modifications to the here described EuroClonality-NGS assays would be possible, and have actually been tested and approved to be suitable within the working group. In particular, a one-step instead of the two-step PCR presented here might be a reasonable alternative for sites that would apply NGS not only for marker identification but also for MRD assessment. Finally, the Ion Torrent platform was successfully tested as a replacement for the Illumina MiSeq used in this study, and has subsequently also been applied more extensively for clonality assessment in formalin-fixed paraffin-embedded tissue (see accompanying manuscript by Scheijen et al. [31]).

In summary, the EuroClonality-NGS developed an IG/TR marker identification protocol, which was validated across many expert European laboratories. It covers *in vitro*

and *in silico* requirements and allows for quality-controlled, streamlined, comprehensive detection of clonal IG/TR rearrangements in ALL. Compared with low-throughput methods, more MRD markers are identified, sensitivity is increased, processing time is reduced and labour-intensive conventional methods to resolve mixed sequences in case of bi-allelic rearrangements or background are avoided. In parallel, the ARResT/Interrogate bioinformatic platform has been developed with specific functionalities addressing potential pitfalls of IG/TR marker identification in ALL, thus enabling a standardized workflow. In addition, the presented approach forms the basis for future applications in clonality assessment, repertoire analysis and MRD quantification in a quality-controlled and accreditable assay with the potential to meet the upcoming European criteria (EU Regulation 2017/746) for *in vitro* diagnostics.

**Acknowledgements** Analyses in Brno were supported by Ministry of Health of the Czech Republic, grant no. 16-34272A. Computational resources by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, under the programme “Projects of Large Research, Development, and Innovations Infrastructures”. Analyses in the Monza (Centro Ricerca Tettamanti, SS, AG, GF and GC) laboratory were supported by the Italian Association for Cancer Research (AIRC) and Comitato Maria Letizia Verga. Analyses in the Paris (Necker, AP-HP) laboratory were supported by the Ile de France CancéroPôle. Analyses in Prague were supported by AZV 16-32568A, and PRIMUS/17/MED/11. Design of IGK assays and analyses in Nijmegen and Rotterdam laboratories were supported by an Innovation project granted by the Zorgverzekeraars Nederland (number 2017-3442). This publication presents independent research commissioned by the Health Innovation Challenge Fund R9-486, a parallel funding partnership between the Department of Health & Social Care and Wellcome Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health & Social Care or Wellcome Trust.

## Compliance with ethical standards

**Conflict of interest** The EuroClonality-NGS Working Group is an independent scientific subdivision of EuroClonality that aims at innovation, standardization and education in the field of diagnostic clonality analysis. The revenues of the previously obtained patent (PCT/NL2003/000690), which is collectively owned by the EuroClonality Foundation and licensed to InVivoScribe, are exclusively used for EuroClonality activities, such as for covering costs of the Working Group meetings, collective WorkPackages and the EuroClonality Educational Workshops. The EuroClonality consortium operates under an umbrella of ESLHO, which is an official EHA Scientific Working Group. MB: contract research for Affimed, Amgen, Regeneron, advisory board of Amgen, Incyte, Speaker bureau of Janssen, Pfizer, Roche. AWL: contract research for Roche-Genentech, research support from Gilead, advisory board for AbbVie, speaker for Gilead, Janssen. RG-S: research grants from Gilead, Takeda, Amgen, and the Spanish government; and reports consulting fees from Janssen, Takeda, Incyte, and BMS. KS: research support from Janssen, Abbvie, Gilead; speaker for Janssen, Abbvie, Gilead; advisory board for Janssen, Abbvie, Gilead. PG: speaker for Gilead.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.




**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Tonegawa S. Somatic generation of antibody diversity. *Nature*. 1983;302:575–81.
2. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature*. 1988;334:395–402.
3. Schlissel MS. Regulating antigen-receptor gene assembly. *Nat Rev Immunol*. 2003;3:890–9.
4. Lefranc M-P, Lefranc G. The T cell receptor factsbook. Academic Press; 2001. <https://www.sciencedirect.com/science/book/9780124413528>. Accessed 22 Mar 2018.
5. Lefranc M-P, Lefranc G. The immunoglobulin factsbook. Academic Press; 2001.
6. Monroe JG, Dorshkind K. Fate decisions regulating bone marrow and peripheral B lymphocyte development. *Adv Immunol*. 2007;95:1–50.
7. von Boehmer H, Melchers F. Checkpoints in lymphocyte development and autoimmune disease. *Nat Immunol*. 2010;11:14–20.
8. Evans PAS, Pott C, Groenen PJTA, Salles G, Davi F, Berger F, et al. Significantly improved PCR-based clonality testing in B-cell malignancies by use of multiple immunoglobulin gene targets. Report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia*. 2007;21:207–14.
9. Brüggemann M, White H, Gaulard P, Garcia-Sanz R, Gameiro P, Oeschger S, et al. Powerful strategy for polymerase chain reaction-based clonality assessment in T-cell malignancies Report of the BIOMED-2 Concerted Action BHM4 CT98-3936. *Leukemia*. 2007;21:215–21.
10. Langerak AW, Groenen PJTA, Brüggemann M, Beldjord K, Bellan C, Bonello L, et al. EuroClonality/BIOMED-2 guidelines for interpretation and reporting of Ig/TCR clonality testing in suspected lymphoproliferations. *Leukemia*. 2012;26:2159–71.
11. van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*. 2003;17:2257–317.
12. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med*. 2009;1:12ra23.
13. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol*. 2013;31:166–9.
14. Bartram J, Goulden N, Wright G, Adams S, Brooks T, Edwards D, et al. High throughput sequencing in acute lymphoblastic leukemia reveals clonal architecture of central nervous system and bone marrow compartments. *Haematologica*. 2018;103:e110–e114.
15. Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, et al. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood*. 2012;120:5173–80.
16. Ladetto M, Brüggemann M, Monitillo L, Ferrero S, Pepin F, Drandi D, et al. Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders. *Leukemia*. 2014;28:1299–307.
17. Pulsipher MA, Carlson C, Langholz B, Wall DA, Schultz KR, Bunin N, et al. IgH-V(D)J NGS-MRD measurement pre- and early post- allo-transplant defines very low and very high risk ALL patients. *Blood*. 2015;125:3501–8.
18. Kotrova M, Muzikova K, Mejstrikova E, Novakova M, Bakardjieva-Mihaylova V, Fiser K, et al. The predictive strength of next-generation sequencing MRD detection for relapse compared with current methods in childhood ALL. *Blood*. 2015;126:1045–7.
19. Langerak AW, Brüggemann M, Davi F, Darzentas N, Gonzalez D, Cazzaniga G, et al. High throughput immunogenetics for clinical and research applications in immunohematology: potential and challenges. *J Immunol*. 2017;198:3765–74.
20. Kotrova M, van der Velden VHJ, van Dongen JJM, Formankova R, Sedlacek P, Brüggemann M, et al. Next-generation sequencing indicates false-positive MRD results and better predicts prognosis after SCT in patients with childhood ALL. *Bone Marrow Transpl*. 2017;52:962–8.
21. Kotrova M, Trka J, Kneba M, Brüggemann M. Is next-generation sequencing the way to go for residual disease monitoring in acute lymphoblastic leukemia? *Mol Diagn Ther*. 2017. <https://doi.org/10.1007/s40291-017-0277-9>.
22. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res*. 2009;19:1817–24.
23. Gawad C, Pepin F, Carlton VEH, Klinger M, Logan AC, Miklos DB, et al. Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood*. 2012;120:4407–17.
24. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, et al. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc Natl Acad Sci USA*. 2011;108:21194–9.
25. Logan AC, Zhang B, Narasimhan B, Carlton V, Zheng J, Moorhead M, et al. Minimal residual disease quantification using consensus primers and high-throughput IGH sequencing predicts post-transplant relapse in chronic lymphocytic leukemia. *Leukemia*. 2013;27:1659–65.
26. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med*. 2010;2:47ra64–47ra64.
27. Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci*. 2010;107:1518–23.
28. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci Transl Med*. 2012;4:134ra63–134ra63.
29. Wu Y-C, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood*. 2010;116:1070–8.
30. Knecht H, Reigl T, Kotrová M, Appelt F, Stewart P, Bystry V, et al. Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic

- functionalities by EuroClonality-NGS. *Leukemia*; revision. [Epub ahead of print]
31. Scheijen B, Meijers R, Rijntjes J, van der Klift M, Möbs M, Steinhilber J, et al. Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. *Leukemia*; revision. [Epub ahead of print]
  32. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A, et al. ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. *Bioinformatics*. 2016;33:btw634.
  33. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*. 2000;132:365–86.
  34. Qu W, Zhou Y, Zhang Y, Lu Y, Wang X, Zhao D, et al. MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Res*. 2012;40:W205–8.
  35. van der Velden VHJ, Cazzaniga G, Schrauder A, Hancock J, Bader P, Panzer-Grumayer ER, et al. Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data. *Leukemia*. 2007;21:604–11.
  36. Pongers-Willems MJ, Seriu T, Stolz F, D'Aniello E, Gameiro P, Pisa P, et al. Primers and protocols for standardized detection of minimal residual disease in acute lymphoblastic leukemia using immunoglobulin and T cell receptor gene rearrangements and TAL1 deletions as PCR targets: report of the BIOMED-1 CONCERTED ACTION. *Leukemia*. 1999;13:110–8.
  37. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F, et al. Vidjil: a web platform for analysis of high-throughput repertoire sequencing. *PLoS ONE*. 2016;11:e0166126.
  38. Giudicelli V, Brochet X, Lefranc M-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell Recept (TR) Nucleotide Seq Cold Spring Harb Protoc. 2011;2011:695–715.
  39. Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res*. 2005;33:D256–61.
  40. Szczepanski T, Van Der Velden VHJ, Hoogeveen PG, De Bie M, Jacobs CH, Van Wering ER, et al. Vdelta2-Jalpha rearrangements are frequent in precursor-B-acute lymphoblastic leukemia but rare in normal lymphoid cells. *Blood*. 2004;103:3798–804.
  41. Lefranc MP, Rabbitts TH. Genetic organization of the human T-cell receptor gamma and delta loci. *Res Immunol*. 1990;141:565–77.
  42. Dongen J van, Szczepanski T, Adriaansen H. *Immunobiology of leukemia*. 7th edn. WB Saunders Company: Philadelphia; 2002.
  43. van Dongen JJ, Seriu T, Panzer-Grumayer ER, Biondi A, Pongers-Willems MJ, Corral L, et al. Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood. *Lancet*. 1998;352:1731–8.
  44. Kitchingman GR. Immunoglobulin heavy chain gene VH-D junctional diversity at diagnosis in patients with acute lymphoblastic leukemia. *Blood*. 1993;81:775–82.
  45. Steenbergen EJ, Verhagen OJ, van Leeuwen EF, von dem Borne AE, van der Schoot CE. Distinct ongoing Ig heavy chain rearrangement processes in childhood B-precursor acute lymphoblastic leukemia. *Blood*. 1993;82:581–9.
  46. Szczepanski T, Willems MJ, Brinkhof B, van Wering ER, van der Burg M, van Dongen JJM. Comparative analysis of Ig and TCR gene rearrangements at diagnosis and at relapse of childhood precursor-B-ALL provides improved strategies for selection of stable PCR targets for monitoring of minimal residual disease. *Blood*. 2002;99:2315–23.
  47. de Haas V, Verhagen OJ, von dem Borne AE, Kroes W, van den Berg H, van der Schoot CE. Quantification of minimal residual disease in children with oligoclonal B-precursor acute lymphoblastic leukemia indicates that the clones that grow out during relapse already have the slowest rate of reduction during induction therapy. *Leukemia*. 2001;15:134–40.
  48. Germano G, del Giudice L, Palatron S, Giarin E, Cazzaniga G, Biondi A, et al. Clonality profile in relapsed precursor-B-ALL children by GeneScan and sequencing analyses. Consequences on minimal residual disease monitoring. *Leukemia*. 2003;17:1573–82.
  49. Theunissen PMJ, van Zessen D, Stubbs AP, Faham M, Zwaan CM, van Dongen JJM, et al. Antigen receptor sequencing of paired bone marrow samples shows homogeneous distribution of acute lymphoblastic leukemia subclones. *Haematologica*. 2017;102:1869–77.

## Affiliations

Monika Brüggemann<sup>1</sup> · Michaela Kotrová<sup>1,2</sup> · Henrik Knecht<sup>1</sup> · Jack Bartram<sup>3</sup> · Myriam Boudjoghra<sup>4</sup> · Vojtech Bystry<sup>5</sup> · Grazia Fazio <sup>6</sup> · Eva Froňková<sup>2</sup> · Mathieu Giraud <sup>7</sup> · Andrea Grioni<sup>6</sup> · Jeremy Hancock<sup>8</sup> · Dietrich Herrmann<sup>1</sup> · Cristina Jiménez<sup>9</sup> · Adam Krejci<sup>5</sup> · John Moppett <sup>10</sup> · Tomas Reigl<sup>5</sup> · Mikael Salson<sup>7</sup> · Blanca Scheijen<sup>11</sup> · Martin Schwarz<sup>1</sup> · Simona Songia<sup>6</sup> · Michael Svaton<sup>2</sup> · Jacques J. M. van Dongen<sup>12</sup> · Patrick Villarese<sup>13</sup> · Stephanie Wakeman<sup>8</sup> · Gary Wright<sup>3</sup> · Giovanni Cazzaniga<sup>6</sup> · Frédéric Davi<sup>4</sup> · Ramón García-Sanz<sup>9</sup> · David Gonzalez<sup>14</sup> · Patricia J. T. A. Groenen<sup>11</sup> · Michael Hummel<sup>15</sup> · Elizabeth A. Macintyre<sup>13</sup> · Kostas Stamatopoulos<sup>16</sup> · Christiane Pott<sup>1</sup> · Jan Trka<sup>2</sup> · Nikos Darzentas<sup>1,5</sup> · Anton W. Langerak<sup>17</sup> · on behalf of the EuroClonality-NGS working group

<sup>1</sup> Department of Hematology, University Hospital Schleswig-Holstein, Kiel, Germany

<sup>2</sup> CLIP - Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University, University Hospital Motol, Prague, Czech Republic

<sup>3</sup> Department of Paediatric Haematology, Great Ormond Street Hospital, London, UK

<sup>4</sup> Department of Hematology, Hopital Pitié-Salpêtrière,

Paris, France

<sup>5</sup> Central European Institute of Technology, Masaryk University, Brno, Czech Republic

<sup>6</sup> Centro Ricerca Tettamanti, University of Milano Bicocca, Monza, Italy

<sup>7</sup> CNRS, CRISTAL, Université Lille, Inria Lille, France

<sup>8</sup> Bristol Genetics Laboratory, Southmead Hospital, Bristol, UK

<sup>9</sup> Hospital Universitario de Salamanca-IBSAL, Salamanca, Spain

- 
- <sup>10</sup> Department of Pediatric Haematology, Bristol Royal Hospital for Children, Bristol, UK
- <sup>11</sup> Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands
- <sup>12</sup> Department of Immunohematology and Blood Transfusion (IHB), Leiden University Medical Center, Leiden, The Netherlands
- <sup>13</sup> Department of Hematology, APHP Necker-Enfants Malades and Paris Descartes University, Paris, France
- <sup>14</sup> Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, UK
- <sup>15</sup> Institute of Pathology, Charité – Universitätsmedizin Berlin, Berlin, Germany
- <sup>16</sup> Institute of Applied Biosciences, Thessaloniki, Greece
- <sup>17</sup> Department of Immunology, Laboratory Medical Immunology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands

## Chapter 5

Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS.

Knecht H, Reigl T, Kotrová M, Appelt F, Stewart P, Bystry V, Krejci A, **Grióni A**, Pal K, Stranska K, Plevova K, Rijntjes J, Songia S, Svatoň M, Froňková E, Bartram J, Scheijen B, Herrmann D, García-Sanz R, Hancock J, Moppett J, van Dongen JJM, Cazzaniga G, Davi F, Groenen PJTA, Hummel M, Macintyre EA, Stamatopoulos K, Trka J, Langerak AW, Gonzalez D, Pott C, Brüggemann M, Darzentas N; EuroClonality-NGS Working Group. Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. *Leukemia*. 2019 Sep;33(9):2254-2265. doi: 10.1038/s41375-019-0499-4. Epub 2019 Jun 21. PMID: 31227779; PMCID: PMC6756032.



Minimal residual disease

# Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS

Henrik Knecht<sup>1</sup> · Tomas Reigl<sup>2</sup> · Michaela Kotrová <sup>1</sup> · Franziska Appelt<sup>1</sup> · Peter Stewart<sup>3</sup> · Vojtech Bystry<sup>2</sup> · Adam Krejci<sup>2</sup> · Andrea Grioni<sup>4</sup> · Karol Pal<sup>2</sup> · Kamila Stranska<sup>2,5</sup> · Karla Plevova<sup>2,5</sup> · Jos Rijntjes<sup>6</sup> · Simona Songia<sup>4</sup> · Michael Svatoň<sup>7</sup> · Eva Froňková<sup>7</sup> · Jack Bartram<sup>8</sup> · Blanca Scheijen<sup>6</sup> · Dietrich Herrmann<sup>1</sup> · Ramón García-Sanz <sup>9</sup> · Jeremy Hancock<sup>10</sup> · John Moppett <sup>11</sup> · Jacques J. M. van Dongen<sup>12</sup> · Giovanni Cazzaniga <sup>4</sup> · Frédéric Davi<sup>13</sup> · Patricia J. T. A. Groenen<sup>6</sup> · Michael Hummel<sup>14</sup> · Elizabeth A. Macintyre<sup>15</sup> · Kostas Stamatopoulos<sup>16</sup> · Jan Trka<sup>7</sup> · Anton W. Langerak<sup>17</sup> · David Gonzalez<sup>3</sup> · Christiane Pott<sup>1</sup> · Monika Brüggemann<sup>1</sup> · Nikos Darzentas<sup>1,2</sup> · on behalf of the EuroClonality-NGS Working Group

Received: 15 January 2019 / Revised: 23 March 2019 / Accepted: 23 April 2019 / Published online: 21 June 2019  
© The Author(s) 2019. This article is published with open access

## Abstract

Assessment of clonality, marker identification and measurement of minimal residual disease (MRD) of immunoglobulin (IG) and T cell receptor (TR) gene rearrangements in lymphoid neoplasms using next-generation sequencing (NGS) is currently under intensive development for use in clinical diagnostics. So far, however, there is a lack of suitable quality control (QC) options with regard to standardisation and quality metrics to ensure robust clinical application of such approaches. The EuroClonality-NGS Working Group has therefore established two types of QCs to accompany the NGS-based IG/TR assays. First, a central polytarget QC (cPT-QC) is used to monitor the primer performance of each of the EuroClonality multiplex NGS assays; second, a standardised human cell line-based DNA control is spiked into each patient DNA sample to work as a central in-tube QC and calibrator for MRD quantification (cIT-QC). Having integrated those two reference standards in the ARResT/Interrogate bioinformatic platform, EuroClonality-NGS provides a complete protocol for standardised IG/TR gene rearrangement analysis by NGS with high reproducibility, accuracy and precision for valid marker identification and quantification in diagnostics of lymphoid malignancies.

## Introduction

Identification and assessment of clonal immunoglobulin (IG) and T cell receptor (TR) gene rearrangements is a widely used tool for the diagnosis of lymphoid malignancies, and is also essential for monitoring minimal residual disease (MRD) [1–6].

Next-generation sequencing (NGS) of IG/TR gene rearrangements is gaining popularity in clinical laboratories, as it avoids laborious design of patient-specific real-time

quantitative (RQ)-PCR assays and provides the capability to sequence multiple rearrangements and rearrangement types within a single sequencing run. It also allows detection of MRD with a more specific readout than RQ-PCR [7]. Hence, several methods have already been described for high-throughput profiling of IG/TR rearrangements at diagnosis and follow-up in acute lymphoblastic leukaemia (ALL), chronic lymphocytic leukaemia (CLL) and other lymphoid malignancies [8–13].

NGS assays, especially those based on amplicons, pose major challenges, as multiple primers need to anneal under the same reaction conditions, while many technical variables may be introduced by library preparation, sequencing and bioinformatics, potentially leading to inaccurate results [14]. Particularly in a clinical context, strategies for standardisation of laboratory protocols and quality control (QC) of each component of an NGS assay are highly desirable, if not required.

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41375-019-0499-4>) contains supplementary material, which is available to authorized users.

✉ Anton W. Langerak  
a.langerak@erasmusmc.nl

Extended author information available on the last page of the article.

Reference standards are essential for the evaluation of wet-lab and *in silico* NGS processes to ensure the analytical validity of test results prior to implementation of an NGS technology into clinical practice [15–17]. Reference DNA materials should be stable sources of rearrangements that can be sequenced and used for measuring qualitative and quantitative properties. However, previously published standards have a limited scope and utility, since they (1) do not cover all relevant IG/TR loci, (2) do not report on the quality of the sequencing run or the performance of samples and primers and/or (3) are synthetic constructs that may not reflect the complexity of native genomic DNA [9, 18, 19].

The EuroClonality-NGS Working Group was initiated to develop, standardise and validate protocols for IG/TR NGS applications, as introduced in Langerak et al. [20] and described in the accompanying manuscripts by Brüggemann et al. [21] and Scheijen et al. [22]. Innovatively, the EuroClonality-NGS assays include two types of QCs, both based on basic assay components, and both fully integrated in ARResT/Interrogate [23], the interactive bioinformatics platform developed within the Working Group:

1. A central polytarget QC (cPT-QC) consisting of a standardised mixture of lymphoid specimens, representing a full repertoire of IG/TR genes. It serves to assess performance biases or unusual amplification shifts in a sequencing run by tracking primer usage and comparison with stored reference profiles.
2. A central in-tube quality/quantification control (cIT-QC) consisting of human B and T cell lines with well-defined IG/TR rearrangements. The cIT-QC is directly added to a sample to undergo concurrent library preparation and sequencing, acting as in-tube qualitative and quantitative standard that is subjected to the same technical downstream variables.

Here we describe, evaluate and showcase these concepts and functionalities. We tested the developed protocol on a dataset of polyclonal samples, B-ALL and T-ALL diagnostic materials and follow-ups of patients with substantial treatment-induced shifts in IG/TR repertoires. We show its successful application and robustness for clinical laboratories that want to implement the EuroClonality-NGS assays for marker identification and quantification. Figure 1 provides an overview of the study.

## Materials and methods

### EuroClonality-NGS assay

The EuroClonality-NGS assay for marker identification used herein is the two-step PCR protocol with eight primer

sets (IGH-VJ, IGH-DJ, IGK-VJ-Kde, intron-Kde, TRB-VJ, TRB-DJ, TRG, TRD)—hereafter termed ‘tubes’—per sample, as described in the accompanying manuscript by Brüggemann et al. [21].

### ARResT/Interrogate

ARResT/Interrogate uses a web browser-based interface to (1) run an analytical pipeline to identify different types of rearrangements—‘junction classes’—across all IG/TR loci (Supplementary Table S1), (2) store, retrieve and report on runs, (3) allow highly varied analyses and visualisations and (4) enable purpose-built meta-analyses and applications. Bioinformatic analyses were performed with ARResT/Interrogate and purpose-built tools unless otherwise stated. Further implementation details are provided below and as Supplementary Information. The platform is currently freely available at [arrest.tools/interrogate](http://arrest.tools/interrogate), hosted at the Meta-Centrum and CERIT-SC centres in the Czech Republic.

### Implementation of the cPT-QC

#### Sources and methods

The cPT-QC consists of genomic DNA isolated from healthy human thymus, tonsil and peripheral blood mononuclear cells (MNCs) in a 1:1:1 ratio (see Supplementary Information). The cPT-QC undergoes library preparation alongside the investigated samples (Figs. 1 and 2).

#### Implementation

Primers are bioinformatically identified in the reads of each of the eight cPT-QC tubes of the run and their abundances compared to stored cPT-QC reference results using the test of proportions.

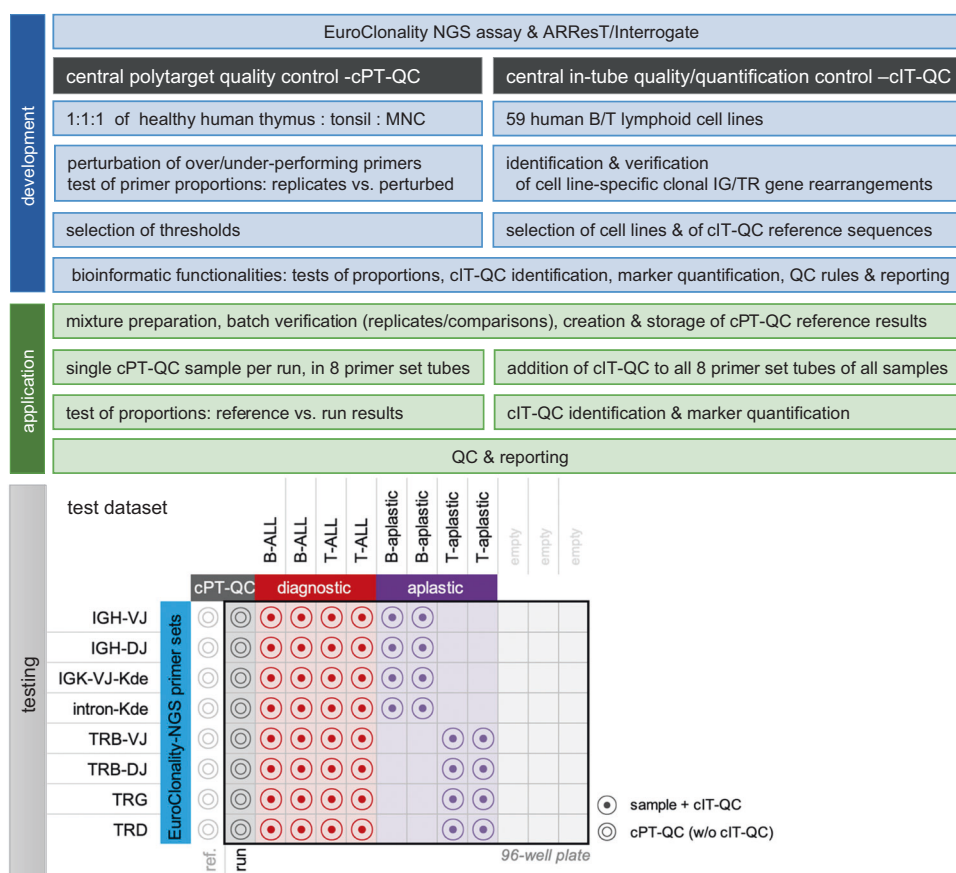
Stored reference results are the output of ARResT/Interrogate from the analysis of a cPT-QC sample. These results should be confirmed through replicate runs over time in each lab to accommodate for technical variability (see Discussion). The results (and not the raw NGS data) are stored to ensure that the bioinformatic analysis is not compromised inadvertently by the user; this means that the results are updated with every major release of ARResT/Interrogate to ensure compatibility with new runs.

Issues with abundances of primers of a specific primer set are used to tag the corresponding cPT-QC samples and all user samples of the same primer set as ‘QC-failed’.

#### Replicates

As reproducibility is important for a QC of this type, we performed replicate runs of cPT-QC and also of MNC (four

**Fig. 1** Study design: components and steps of development (in blue), application (in green) and testing for the central polytarget quality control (cPT-QC) and central in-tube quality/quantification control (cIT-QC), including a schematic overview of the test dataset based on a 96-well plate. Text boxes are either shared across cPT-QC and cIT-QC or describing equivalent steps if on same row. MNC = mononuclear cells, QC = quality control, ref. = reference, w/o = without



libraries in total); MNCs are regularly used and could serve as an alternative. Relative abundances of 5' primers were compared employing the test of proportions.

### Primer perturbations

To investigate whether and how the cPT-QC can be used to detect issues with primer performance, artificial perturbations of primer concentrations were created to simulate missing pipetting a primer or pipetting the wrong primer concentration.

First, 5' primer usage was analysed in a cPT-QC sample. Two primers of differing abundances were selected from each primer set, skipping intron-Kde that only has two primers: IGH-VJ-FR1-M-1, IGHV-FR1-O-1; IGHD-B-1, IGHD-E-1; IGK-V-G-1, IGK-V-I-1; TRB-V-AD-1, TRB-V-G-1; TRB-D-A-1, TRB-D-B-1; TRG-V-F-1, TRG-V-E-1; TRD-D-A-1, TRD-V-B-1. Second, these primers were perturbed by fully excluding them from the primer pool (0%) and by changing their concentration by reduction to 10% and by increase to 200%. Replicate runs of these three primer-perturbed cPT-QC libraries (six in total) were performed; however, since the replicates were consistent (data not shown), only the first

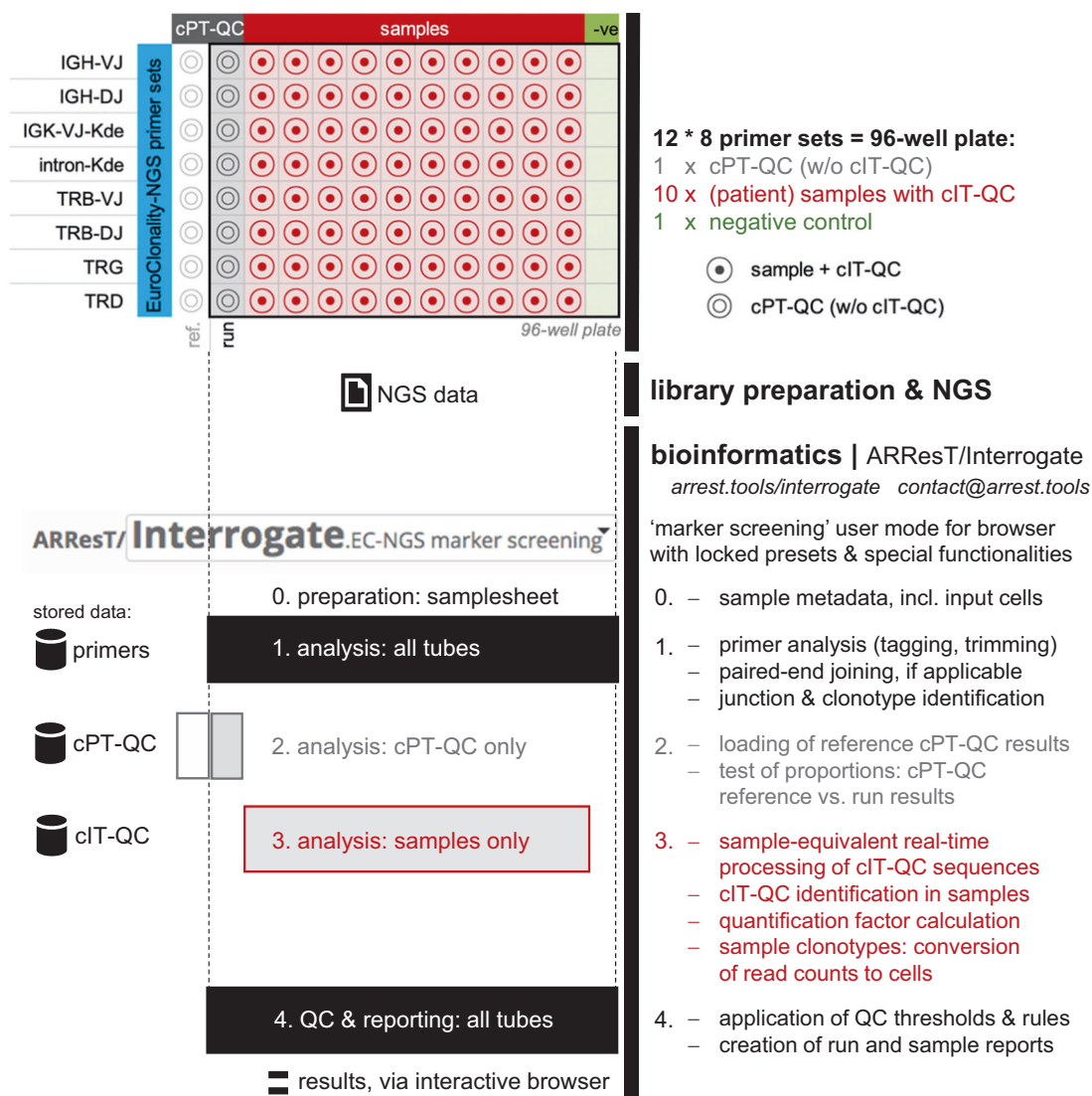
replicate of each is shown in Results. Finally, relative abundances of 5' primers were compared between normal replicates and between normal replicates and the perturbed libraries using the test of proportions.

### Design and validation of the cIT-QC

#### Sources and methods

In total, 59 human B ( $n = 30$ ) and T ( $n = 29$ ) lymphoid cell lines were obtained from the American Type Culture Collection (ATCC, Manassas, VA, USA; [www.lgcpromochem-atcc.com](http://www.lgcpromochem-atcc.com)) and the German Collection of Microorganisms and Cell Cultures GmbH (DSMZ, Braunschweig, Germany; [www.dsmz.de](http://www.dsmz.de)), or were derived from internal cell line banks. Supplementary Table S2 gives an overview of the cell lines. DNA from cultured cell lines was isolated using a phenol–chloroform extraction protocol, followed by ethanol precipitation and elution in Tris ethylenediaminetetra-acetic acid buffer. Alternatively, DNA was isolated with the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich, St. Louis, MO, USA) according to the manufacturer's protocol.





**Fig. 2** EuroClonality-NGS (next-generation sequencing) protocol for quality control and quantification in marker identification: 96-well plate set-up, including central polytarget quality control (cPT-QC) and central in-tube quality/quantification control (cIT-QC), library preparation and NGS, bioinformatics with ARResT/Interrogate. The

bioinformatics are additionally organised per sample type to showcase distinct steps and functionalities listed on the right: all tubes (1 and 4, in black), cPT-QC (2, in grey), (patient) samples (3, in red)—these colours are shared with the well plate. ref. = reference, QC = quality control, w/o = without

### Identification of cell line-specific clonal IG/TR gene rearrangements

Each of the 59 cell lines was screened for clonal IG/TR gene rearrangements using the EuroClonality-NGS assay with 100 ng of DNA (quantified with Qubit 3.0, Thermo Fisher Scientific) from each cell line, without the addition of MNC. Paired-end sequencing (2 × 250 bp) was performed on Illumina MiSeq (Illumina, San Diego, CA, USA) with a final concentration of 7 pM per library aiming for at least 2000 reads per sample. To avoid low-complexity issues, 10% PhiX control was added to each sequencing run.

### Verification of cell line-specific clonal IG/TR gene rearrangements

Additional methods were used to verify the NGS amplicon-identified cell line rearrangements:

1. A capture-based protocol, established within EuroClonality-NGS Working Group and covering the coding V, D and J genes of IG/TR loci [13]: in short, cell line DNA was fragmented and processed with the KAPA Hyperplus Kit with Library Amplification (Roche Sequencing Solutions, Pleasanton, CA, USA); hybridisation of libraries was performed

with customised SeqCap EZ Choice Probes (Roche Sequencing Solutions, Pleasanton, CA, USA), developed based on Wren et al. [13] 2 × 150 bp paired-end sequencing was performed on Illumina NextSeq.

2. Multiplex amplification and Sanger sequencing according to the BIOMED-2 protocol: PCR products were checked for fragment sizes and clonality in the QIAXCEL Advanced System [24, 25]. Clonal PCR products were subjected to heteroduplex analysis and sequenced on either an ABI 3130 or ABI 3500 platform (Applied Biosystems, Foster City, CA, USA).

IG/TR rearrangement profiles of all cell lines were compared between the different methods.

For cases with discrepant results between the three methods, IG/TR allele-specific PCR assays were designed for digital droplet PCR (ddPCR) (QX200™ Droplet Digital™ PCR System, Bio-Rad) to verify the respective rearrangement. Absolute quantification of IG/TR gene rearrangements by ddPCR was performed using two different genomic DNA amounts (50 ng, 100 ng) (Supplementary Information). Each experiment included a polyclonal MNC control and a no-template control.

### Cell line selection criteria

For establishment of the cIT-QC from the spectrum of IG/TR gene rearrangements of the 59 cell lines, the following selection criteria were defined:

1. The final set should consist of as few cell lines as possible, while covering each primer set by at least three different rearrangements, hence aiming for ALL cell lines harbouring not only lineage characteristic but also cross-lineage rearrangements.
2. The rearrangements should be unambiguously detectable with Sanger sequencing and amplicon-based NGS.
3. The variable region of IGHV-(IGHD)-IGHJ gene rearrangements should preferably be unmutated in order to avoid issues with primer annealing.

### Implementation

For cIT-QC mixture preparation see Supplementary Information.

Bioinformatically, cIT-QC reads are identified using an immunogenetic annotation-based approach that is extremely fast while allowing for variations in sequence, avoiding compute-intensive and potentially inaccurate alignment.

For QC, we expect identification of at least one read per cIT-QC rearrangement and of at least as many total cIT-QC reads as total cIT-QC cells, otherwise the tube is tagged as ‘QC-failed’ (see below for how this is used in ARResT/Interrogate).

Quantification applies the quantification factor—calculated per primer set by dividing total cIT-QC cells by total cIT-QC reads—to convert read counts of a clonotype to cell counts, and then calculate its relative abundance against the total sample input cells.

### Creation of a test dataset

To evaluate and showcase the aforementioned concepts and functionalities, we compiled a test dataset with:

1. Four diagnostic bone marrow B-/T-ALL samples with high leukaemic infiltration (assessed by routine cytology to be 60–80%).
2. Four samples of patients with B/T cell aplasia after antibody treatment. The two samples with B cell aplasia were CLL samples after Rituximab (anti-CD20) treatment and the two samples with T cell aplasia were T cell prolymphocytic leukaemia samples after Alemtuzumab (anti-CD52) treatment. In all these samples lineage-specific aplasia was confirmed by flow cytometry.
3. cPT-QC for all primer sets, but with the TRB-VJ primer set results swapped with perturbed results from experiments outlined above. To showcase generic QC functionalities, one diagnostic sample was sub-sampled to <1000 random reads.

The diagnostic samples and the cPT-QC were run with all primer sets as described in the accompanying manuscript by Brüggemann et al. [21], while the aplastic follow-up samples only with the corresponding primer sets, that is, the IG sets for samples with B cell aplasia, and the TR sets for samples with T cell aplasia. Figure 1 includes a schematic of the test dataset. Finally, the follow-up samples were run without the addition of MNC to test that the addition of cIT-QC is sufficient to stabilise the samples for sequencing without compromising their immunogenetic profile.

### Results

The resulting protocol and functionalities for QC and quantification in IG/TR NGS marker identification are depicted in Fig. 2. We present and further discuss the underlying results below.

**Table 1** cPT-QC: replicates and primer perturbations. Relative abundances (%) of selected 5' primers across all primer sets. Top group of primers were perturbed as described in Materials and methods; bottom group is a selection of primers that were left un-perturbed: one per primer set selected alphabetically, plus two examples where the primer behaviour is of interest to the discussion (see text). Results are shown from two cPT-QC replicates (blue column) and from replicate 1 of the blue column ("rep1") vs. cPT-QC libraries where primers were excluded (0%, orange column), reduced to 10% (yellow column) and increased to 200% (green column). Changes in abundance compared to cPT-QC rep1 are shown separately (column "% or rep1", in italics) and coloured from red (0%) to white (100%, i.e. no change) to green (200%). Actual primer abundances are coloured based on the *p* value from the test of proportions, with grey indicating a noticeable change according to our threshold of  $1e-200$  (*p* value  $<1e-199$  highlighted in dark grey, and  $<1e-99$  in light grey, otherwise in white)

all numbers are percentages (%) ; rep:replicate ; test of proportions vs cPT-QC rep1, dark grey: $<1e-199$ , light grey: $<1e-99$

primer set	primers primer name	cPT-QC		vs. 0%		vs. 10%		vs. 200%		
		rep1	% of rep1	rep2	% of rep1	% of rep1	% of rep1	% of rep1		
IGH-VJ-FR1	IGH-V-FR1-M-1	27.44	81.05	22.24	2.66	0.73	7.35	2.02	128.13	35.16
IGH-VJ-FR1	IGH-V-FR1-O-1	1.18	92.48	1.10	5.33	0.06	5.74	0.07	241.98	2.87
IGH-DJ	IGH-D-B-1:#1:14C	7.32	101.64	7.44	0.00	0.00	0.65	0.05	197.73	14.47
IGH-DJ	IGH-D-B-1:#2:14T	11.74	104.09	12.22	0.01	0.00	0.74	0.09	197.79	23.22
IGH-DJ	IGH-D-E-1:#4:14G22G	1.86	94.69	1.77	0.29	0.01	0.59	0.01	89.27	1.66
IGK-VJ-Kde	IGK-V-G-1	6.08	102.78	6.25	2.07	0.13	2.78	0.17	223.52	13.59
IGK-VJ-Kde	IGK-V-I-1	8.85	100.64	8.91	0.66	0.06	3.99	0.35	234.06	20.71
TRB-VJ	TRB-V-AD-1	31.76	105.92	33.64	1.11	0.35	15.44	4.91	112.37	35.69
TRB-VJ	TRB-V-G-1	10.09	94.90	9.58	0.27	0.03	1.99	0.20	117.44	11.85
TRB-DJ	TRB-D-A-1	63.20	101.50	64.15	0.02	0.01	22.64	14.31	110.33	69.73
TRB-DJ	TRB-D-B-1	36.14	96.24	34.78	0.22	0.08	8.08	2.92	135.17	48.85
TRD	TRD-V-B-1	12.55	118.57	14.88	0.49	0.06	3.27	0.41	344.94	43.29
TRD	TRD-D-A-1	64.60	109.85	70.96	0.14	0.09	3.35	2.16	88.53	57.19
TRG	TRG-V-E-1	3.52	96.79	3.40	0.09	0.00	1.70	0.06	257.81	9.06
TRG	TRG-V-F-1	14.48	99.45	14.40	0.75	0.11	0.20	0.03	162.50	23.53
IGH-VJ-FR1	IGH-V-FR1-A-1	15.34	111.08	17.04	94.20	14.45	76.21	11.69	148.31	22.75
IGH-VJ-FR1	IGH-V-FR1-D-1	16.41	90.13	14.79	259.54	42.59	237.96	39.05	39.07	6.41
IGH-DJ	IGH-D-A-1:#1:6C	8.29	118.24	9.80	121.46	10.07	115.17	9.55	93.87	7.78
IGK-VJ-Kde	IGK-V-A-1	9.79	100.82	9.87	139.47	13.65	134.77	13.19	101.50	9.93
TRB-VJ	TRB-V-AB-1	1.42	103.79	1.48	204.01	2.90	136.33	1.94	95.15	1.35
TRD	TRD-V-A-1	14.37	50.49	7.26	165.69	23.81	156.51	22.49	68.63	9.86
TRG	TRG-V-A-1	18.71	109.09	20.41	116.35	21.77	110.15	20.61	85.94	16.08

### cPT-QC allows to assess primer performance

We compared normal cPT-QC and MNC replicate libraries and primer-perturbed cPT-QC replicate libraries (10 libraries in total) to investigate the use of cPT-QC in assessing primer performance. We applied the test of proportions on 5' primer relative abundances in those libraries, which showed that there is a clear difference in *p* values between un-perturbed (high *p* values indicating insignificant changes) and perturbed (low *p* values) primers. In other words, *p* values of the differences in abundance of the perturbed primers are noticeably lower, an observation we can use to highlight such cases.

Table 1 presents a simplified view of the results, focusing on perturbed primers plus at least one other un-perturbed primer per primer set, either to show their normal behaviour or discuss their abnormal behaviour. At a *p* value threshold of  $1e^{-200}$  none of the primers are flagged in the cPT-QC (white cells), which highlights the reproducibility of the assay, while all the perturbed primers are flagged in the perturbed libraries (light/dark grey cells). Significant changes in abundance are also visible in other cells, with the most likely explanation that those primers were indirectly

affected by perturbations of other primers. That is, a primer 'taking over' when an initially abundant primer was excluded, such as IGHV-FR1-D-1 when IGH-VJ-FR1-M-1 is perturbed either way, especially since these primers amplify partially overlapping lists of genes. Supplementary Table S3 presents the full set of results, including the actual *p* values and results from the replicate MNC libraries.

### Composing the cIT-QC sample from human B and T cell lines

Following the criteria outlined above, we selected six B cell lines: ALL/MIK (ALL), Raji (Burkitt lymphoma), REH (B cell precursor ALL), TMM (CML-BC/EBV + B-LCL), TOM-1 (ALL) and WSU-NHL (B cell lymphoma, histiocytic lymphoma); and three T cell lines: JB6 (ALCL), Karpas299 (ALCL) and MOLT-13 (ALL). The nine cell lines featured a total of 46 rearrangements, all of which are used as part of the cIT-QC. All but two rearrangements that were not detected by capture NGS were detected by all three sequencing methods. Also, another two were of very low abundance and/or trimmed in the capture NGS data, but since the junction segmentation was clearly the same, they

were still tagged as confirmed. Table 2 presents the full list of the 46 rearrangements, with the NGS amplicon-based reference nucleotide sequences in Supplementary Table S4.

### QC aspects can be evaluated in ARResT/Interrogate

Information on the *in silico* QC based on both the cPT-QC and cIT-QC is available in ARResT/Interrogate (Supplementary Figure S1). Generic QC is also performed on samples, specifically to check for low number of raw reads and low percentage of reads with an identified junction. Such samples are tagged as ‘QC-failed’ and excluded by default to prevent the user from their unintended use. However, the user is notified and has the option to include them back in the analysis.

### Marker identification and quantification

Abundances of lymphocyte subpopulations are frequently not available for samples of patients with lymphoid malignancies. Furthermore, as IG/TR NGS only reflects relative representation of the rearrangements, it was important to establish a calibrator that would allow us to normalise sequencing reads to input DNA cells.

Analysis of our test dataset showed the utility of the cIT-QC in marker identification and quantification. Excluding cIT-QC reads, both diagnostic and aplastic samples seem to harbour few highly abundant clones if simply based on the number of reads (Fig. 3, Supplementary Table S5). However, the very high number of reads from only a very limited number of cIT-QC cells (120–440, dependent on the number of cIT-QC rearrangements per primer set), in all aplastic and a few of the diagnostic samples, are an indirect yet clear indication of the restricted numbers of patient cells harbouring rearrangements in those samples. From another perspective, the total percentage of reads of cIT-QC is much greater than that of patient rearrangements in those samples, suggesting that also cIT-QC cells are more numerous than patient cells with rearrangements. Consequently, after quantification with the cIT-QC, marker abundances fall well below the threshold indicating clonality. On the other hand, and as expected, in most diagnostic samples cIT-QC reads constitute a minority, indicating the true abundant presence of patient cells with clonal rearrangements. Hence, using the cIT-QC, a marker can be more accurately quantified and identified.

### ARResT/Interrogate user mode for marker identification

A critical aspect of bioinformatic-based protocols is their standardisation and usability, as evident from our experiences within EuroClonality-NGS and EuroMRD. We have

thus designed ARResT/Interrogate to be flexible but also ‘lockable’. Flexibility comes from a deep parameterisation of many aspects of the pipeline and the browser. At the same time, we can lock down important parameters so that users cannot inadvertently compromise the analysis. This concept is called ‘user mode’ in ARResT/Interrogate, and as a result of this study we have created a marker identification user mode.

In this user mode, EuroClonality-NGS primer sets and cIT-QC sequences are pre-selected and locked, as are other pipeline options. A special samplesheet is available to annotate samples with metadata, including providing numbers of sample input cells for quantification. The user interface is simplified, with many non-essential functionalities (including many of the visualisations normally available) hidden from view, and with less user actions required to load results. The minimum read-based percentage abundance for a clonotype is pre-set to 5% for marker identification.

## Discussion

In this study, we introduce protocols developed within the EuroClonality-NGS Working Group for QC and quantification in NGS-based IG/TR marker identification. Both laboratory and bioinformatic protocols are presented and showcased on clinically relevant data.

The cPT-QC is used to monitor the primer performance of each of the EuroClonality multiplex NGS assays; the cIT-QC is spiked into each patient DNA sample for QC and quantification. The use of ‘central’ highlights that these controls should be as stable as possible and thus centrally available at an applicable level (minimum at an intra-laboratory level)—this is further discussed below in the context of the cPT-QC.

Our experiments show that the cPT-QC is a valuable tool to monitor reproducibility of results and to identify primer perturbations and other deviations in the wet-lab protocol, as they introduce detectable changes to the sequencing profile. The addition of cPT-QC to each analysis allows to check the primer and assay performance after sequencing. Accidental deviations in the concentrations of single primers within the multiplexed IG/TR primer sets can be detected, performance failures of single primers can be traced and consequences for the IG/TR analysis can be estimated by analysis of cPT-QC data.

In our study, replicates of cPT-QC demonstrated high reproducibility. Nevertheless, we are aware that reproducibility across labs may be affected by a large number of other variables, from consumables and equipment to users. Only centralised access to consumables, for example, in the form of a kit, and a comprehensive protocol, including the

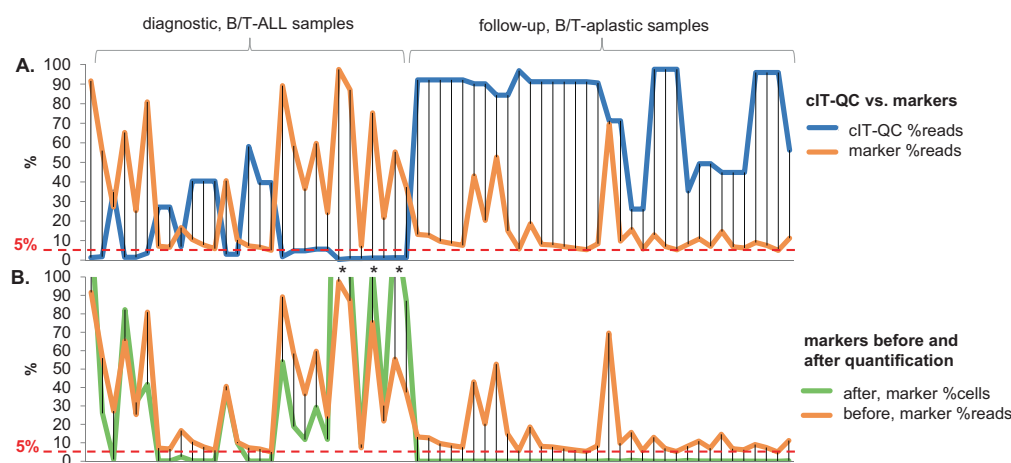
**Table 2** cIT-QC: full list of gene rearrangements per primer set and human B/T cell line, with notes on their verification and clonotype annotation

Primer set	Cell line	Notes	Clonotype (see Supplementary Information—Materials and methods)
TRB-VJ	JB6		VJ:Vb-(Db)-Jb V12-3 = V12-4 6/14/4 J2-3 CASRLAGGPDQTQYF pro
TRB-DJ	JB6		DJ:Db-Jb D1 7/6/4 J2-2 VGTETITGELFF pop
TRG	JB6		VJ:Vg-Jg V10 7/12/12 J1 = J2 CAAWS*GW#KLF unp
TRG	JB6		VJ:Vg-Jg V2 5/13/ J1 = J2 CATWGSIVNYYKLF unp
TRB-VJ	Karpas299		VJ:Vb-(Db)-Jb V20-1 1/22/6 J2-7 CSARAQIGSSPLEQYF pro
TRB-DJ	Karpas299		DJ:Db-Jb D1 /2/6 J1-6 VGTGGLNSPLHF pop
TRG	Karpas299		VJ:Vg-Jg V2 /13/4 JP2 CATWDGG*VP#SDWIKTF unp
TRG	Karpas299		VJ:Vg-Jg V8 /2/5 J1 = J2 CATWDR##YKLF unp
IGH-VJ-FR1	ALL/MIK		VJ:Vh-(Dh)-Jh V3-72 16/24/ J4 SPCPPRKN#YFDYW unp
IGH-VJ-FR1	ALL/MIK		VJ:Vh-(Dh)-Jh V7-4-1 11/40/27 J4 TPYYDSSGY*VP unp
IGK-VJ-Kde	ALL/MIK		Vk-Kde V2-24 = V2D-24 26/6/20 Kde LGGR unk
IGK-VJ-Kde	ALL/MIK		VJ:Vk-Jk V1-39 = VID-39 6/7/5 J3 CQQSYSTGA#F unp
intron-Kde	ALL/MIK		Intron-Kde intron 4/2/ Kde PCVCPIDAAVASFP##SPSGSPGR unk
Intron-Kde	ALL/MIK	Capture: low%	Intron-Kde intron 4/6/1 Kde PCVCPIDAAVASFPSL#SPSGSPGR unk
TRD	ALL/MIK		VJ:Vd-(Dd)-Ja V2 5/21/4 J29 CACAQGGPRS#SGNTPLVF unp
TRG	ALL/MIK		VJ:Vg-Jg V2 /5/8 JP1 CATWDGP#GWFKIF unp
TRG	ALL/MIK		VJ:Vg-Jg V5 2/3/ JP1 CATWDTYTTGWFKIF pro
TRB-VJ	MOLT-13		VJ:Vb-(Db)-Jb V10-1 6/18/1 J1-1 CASRRVRRDRNTEAFF unp
TRB-DJ	MOLT-13		DJ:Db-Jb D1 //6 J1-5 VGTGG#QPQHF pop
TRB-DJ	MOLT-13		DJ:Db-Jb D2 /4/3 J2-3 VGTSGRA#TDTQYF pop
TRD	MOLT-13		VJ:Vd-(Dd)-Jd V1 1/9/ J1 CALGEPGGYTDKLIF pro
TRG	MOLT-13		VJ:Vg-Jg V3 /8/9 J1 = J2 CATWDRPRLKLF pro
TRG	MOLT-13		VJ:Vg-Jg V8 3//3 JP1 CATWD#TGWFKIF unp
IGH-VJ-FR1	Raji	Capture: low%	VJ:Vh-(Dh)-Jh V3-11 = V3-21 = V3-48 2/40/3 J4 CARQRNDFSDNNSYYSNFDWF pro
IGH-DJ	Raji		DJ:Dh-Jh D6-13 8/12/6 J1 VGYSSIPPP#YFQHW pop
IGK-VJ-Kde	Raji		Vk-Kde V1-8 2/2/4 Kde CQQYYSYSVPSGSPGR unk
IGH-VJ-FR1	REH		VJ:Vh-(Dh)-Jh V3-15 1/21/5 J6 CTTGMVRGVI#YYYYGMDVW unp
IGK-VJ-Kde	REH		VJ:Vk-Jk V2-29 5/4/ J4 *MQGIHLS#LTF unp
IGK-VJ-Kde	REH		Vk-Kde V3-20 = V3D-20 4/1/ Kde CQQYGSS##SPSGSPGR unk
Intron-Kde	REH		Intron-Kde intron 5// Kde PCVCPINAAVASF##SPSGSPGR unk
TRB-VJ	REH		VJ:Vb-(Db)-Jb V20-1 1/2/26 J2-7 CSARG unp
TRD	REH		VD:Vd-Dd3 V2 7/3/ D3 CACLLGDTH unk
TRD	REH		VJ:Vd-(Dd)-Ja V2 3/22/5 J29 CACDPYGGGSP#SGNTPLVF unp
TRG	REH		VJ:Vg-Jg V9 1/2/3 J1 = J2 CALWEV#YYKLF unp
TRG	REH		VJ:Vg-Jg V4 10/14/3 J1 = J2 CATLF*R#YYKLF unp
IGH-VJ-FR1	TMM		VJ:Vh-(Dh)-Jh V1-24 /28/8 J5 CATDQAISGVVKSFDPW pro

**Table 2** (continued)

Primer set	Cell line	Notes	Clonotype (see Supplementary Information—Materials and methods)
IGH-DJ	TMM		DJ:Dh-Jh D2-2 3/13/ J3 VRIL**YQLLLNSANDAFDIW pop
IGK-VJ-Kde	TMM		Vk-Kde V2-30 = V2D-30 /7/3 Kde CMQGTHWRPGR#PSGSPGR unk
IGH-VJ-FR1	TOM-1		VJ:Vh-(Dh)-Jh V4-55 1/17/10 J6 CARWAGTTG#YYGMDVW unp
TRD	TOM-1		VD:Vd-Dd3 V2 3/3/2 D3 CACDL#GDTH unk
TRD	TOM-1		VD:Vd-Dd3 V2 8/4/ D3 CAFLLGDTH unk
TRG	TOM-1		VJ:Vg-Jg V5 8//18 J1 = J2 CAT#F unp
IGH-VJ-FR1	WSU-NHL		VJ:Vh-(Dh)-Jh V6-1 1/22/19 J6 CARGTYAAKASMDVW pro
IGH-DJ	WSU-NHL		DJ:Dh-Jh D2-2 1/1/8 J4 VRIL**YQLLY#DYW pop
IGK-VJ-Kde	WSU-NHL	Not in capture	VJ:Vk-Jk V1-17 = V1D-17 1//4 J4 CLQHNSYP#TF unp
Intron-Kde	WSU-NHL	Not in capture	Intron-Kde intron 2//3 Kde PCVCPIDAAVASFP##PSGSPGR unk

See Supplementary Table S4 for NGS amplicon-based full nucleotide reference sequences. *cIT-QC* central in-tube quality/quantification control



**Fig. 3** Abundances of central in-tube quality/quantification control (cIT-QC) and of markers before and after quantification, in the test dataset. The line of marker abundances before quantification (in orange) is shared in both plots for reference. The 5% threshold used for marker identification is shown in both plots. **a** Abundance in percentage of reads (“%reads”) of cIT-QC (in blue) and of markers before quantification (in orange), in diagnostic (left half) and follow-up aplastic (right half) samples. As expected because of the nature of the samples, the cIT-QC is generally most abundant where patient cells with clonal rearrangements are not, and vice versa. Note: For cIT-QC (in blue), the denominator is all reads with junction; for markers (in

orange), it is what we term ‘usable’ reads with junction, which excludes cIT-QC reads; this may lead to sums of those two numbers that exceed 100% per sample. **b** Abundance of markers before (in orange) and after (in green) cIT-QC-based quantification to percentage of patient input cells (“%cells”). Quantification of markers in the aplastic samples places their abundances below the 5% threshold routinely used in marker identification and in the EuroClonality-NGS protocols. Note: When cIT-QC read counts are very low, indicating clonality, quantification factors may lead %cells to exceed 100%; three such cases in the test dataset are indicated by an asterisk (“\*”)

equipment used, will further improve inter-laboratory comparability of results. Besides, activities such as the QC rounds organised bi-annually by ESLHO (eslho.org) are an opportunity to gather data and experience, compare assay performance and identify relevant factors introducing variations. Until full inter-laboratory standardisation is

guaranteed, the implementation of the cPT-QC will require that the reference samples are analysed in each laboratory separately, and updated with every new batch of reagents, while keeping track of equipment and users. These reference data can then be stored in ARResT/Interrogate, which has the ability to store as many different such sets of

reference data as needed, for example, linking a specific set to a specific user if necessary.

In this study we also highlighted a number of unique and advantageous properties of the cIT-QC. In contrast to plasmids or synthetic reference templates, cIT-QC cell lines are particularly well suited to be used as control because they are sources of large quantities of genomic DNA. Second, the nine cell lines with a total of 46 rearrangements represent as few cell lines as possible while covering each primer set by at least three different rearrangements, taking advantage of ALL cell lines harbouring not only lineage-associated but also cross-lineage rearrangements. Third, the rearrangements are unambiguously detectable with amplicon-based NGS. Fourth, the variable region of IGHV-(IGHD)-IGHJ gene rearrangements are not/lowly mutated and therefore minimise issues with primer annealing. Fifth, cIT-QC rearrangements represent 2/3 of the amplifiable junction classes (in italics in Supplementary Table S1) over all eight primer sets, and thus offer an opportunity to highlight a number of issues, most obviously over-/under-amplification, but also bioinformatic misidentification. Additionally, cIT-QC rearrangements can replace MNC for PCR stability without influencing the patient immune repertoire (since cIT-QC rearrangements are identified and by default excluded from the results).

Our cIT-QC enables the conversion from reads to cells, which is of utmost importance for clinical use. Diagnostic material being analysed for MRD marker identification can show abundances of particular clonotypes that do not reflect the clonal composition of the sample. For example, if the diagnostic sample is highly infiltrated by a lymphoid malignancy that does not harbour a targetable rearrangement, the (few) residual lymphoid cells would generate the whole spectrum of detectable rearrangements; in such situations minor accompanying physiological B or T cell clones could be misassigned as clones with leukaemic markers. In the accompanying study by Brüggemann et al. [21], where 134 clonal signals with abundance >5% were detected by NGS but not by Sanger sequencing, cIT-QC quantification reduced the abundances of 71 (53%) of them below the 5% threshold.

In addition to its use in marker identification, and as exemplarily shown for B and T cell depletion in aplastic follow-up samples, the cIT-QC is of utmost relevance for MRD quantification in samples on or after treatment, in particular if B or T cell-directed therapy was applied, which minimises the background of polyclonal gene rearrangements. If the relative tumour burden is calculated by the ratio of leukaemia-specific reads to all annotated reads without any quantification, the quotient reflects the marker frequency only among cells carrying a particular type of rearrangement (e.g. IG rearrangements in B cells) and might thus heavily overestimate the tumour load [26].

Quantification values over 100% (examples in Fig. 3b and Supplementary Table S5) show that using the cIT-QC is still a semi-quantitative approach, potentially affected by amplification biases. However, there is to date no other scientific or commercial solution available that exceeds our methodology in its broad applicability (universal IG/TR approach) and/or allows precise absolute quantification [12, 27–29].

Finally, the QC protocols are embedded in ARResT/Interrogate, which informs users with reports and messages and allows them, for example, to include the QC-failed samples back into the analysis. The logic behind this is that the ‘fail’ flag simply indicates that our pre-defined QC criteria were not met, and not that the data are corrupt beyond use. Nevertheless, flagged data should always be used with caution, and dependent on the application or question.

In summary, our study showcases the applicability of two reference standards, developed by the EuroClonality-NGS Working Group, which allow standardised analysis of IG/TR NGS data (using the EuroClonality-NGS primer sets) with high reproducibility, accuracy and precision in marker identification. With ARResT/Interrogate, a complete *in silico* solution accompanying the *in vitro* assays was built, enabling an analysis of IG/TR sequences including all quality criteria and quantification concepts necessary for valid marker identification in lymphoid malignancies.

**Acknowledgements** This work was supported by Ministry of Health of the Czech Republic, grant no. 16-34272A; computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures”. Analyses in Prague (JT, EF and MS) were supported by Ministry of Health, Czech Republic, grant no. 00064203, and by PRIMUS/17/MED/11. Analyses in the Monza (Centro Ricerca Tettamanti, SS, AG and GC) laboratory were supported by the Italian Association for Cancer Research (AIRC) and Comitato Maria Letizia Verga.

## Compliance with ethical standards

**Conflict of interest** The EuroClonality-NGS Working Group is an independent scientific subdivision of EuroClonality that aims at innovation, standardisation and education in the field of diagnostic clonality analysis. The revenues of the previously obtained patent (PCT/NL2003/000690), which is collectively owned by the EuroClonality Foundation and licensed to InVivoScribe, are exclusively used for EuroClonality activities, such as for covering costs of the Working Group meetings, collective WorkPackages and the EuroClonality Educational Workshops. MB: contract research for Affimed, Amgen, Regeneron, advisory board of Amgen, Incyte, Speaker bureau of Janssen, Pfizer, Roche; AWL: contract research for Roche-Genentech, research support from Gilead, advisory board for AbbVie, speaker for Gilead, Janssen; RG-S: research grants from Gilead, Takeda, Amgen and the Spanish government; and reports consulting fees from Janssen, Takeda, Incyte and BMS; KS: research support from Janssen, Abbvie, Gilead; speaker for Janssen, Abbvie, Gilead;

advisory board for Janssen, Abbvie, Gilead; PG: speaker for Gilead. The other authors declare that they have no conflict of interest.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.





## References

- Pott C. Minimal residual disease detection in mantle cell lymphoma: technical aspects and clinical relevance. *Semin Hematol.* 2011;48:172–84.
- Ferrero S, Drandi D, Mantoan B, Ghione P, Omedè P, Ladetto M. Minimal residual disease detection in lymphoma and multiple myeloma: Impact on therapeutic paradigms. *Hematol Oncol.* 2011;29:167–76.
- Brüggemann M, Gökbuget N, Kneba M. Acute lymphoblastic leukemia: monitoring minimal residual disease as a therapeutic principle. *Semin Oncol.* 2012;39:47–57.
- Brüggemann M, Raff T, Kneba M. Has MRD monitoring superseded other prognostic factors in adult ALL? *Blood.* 2012;120:4470–81.
- Van Dongen JJM, Seriu T, Panzer-Grümayer ER, Biondi A, Pongers-Willems MJ, Corral L, et al. Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood. *Lancet.* 1998;352:1731–8.
- Brüggemann M, Kotrova M. Minimal residual disease in adult ALL: technical aspects and implications for correct clinical interpretation. *Hematol Am Soc Hematol Educ Progr.* 2017;2017(1):13–21.
- Kotrova M, Van Der Velden VHJ, Van Dongen JJM, Formankova R, Sedlacek P, Brüggemann M, et al. Next-generation sequencing indicates false-positive MRD results and better predicts prognosis after SCT in patients with childhood ALL. *Bone Marrow Transplant.* 2017;52:962–8.
- Logan AC, Vashi N, Faham M, Carlton V, Kong K, Buño I, et al. Immunoglobulin and t cell receptor gene high-throughput sequencing quantifies minimal residual disease in acute lymphoblastic leukemia and predicts post-transplantation relapse and survival. *Biol Blood Marrow Transplant.* 2014;20:1307–13.
- Faham M, Zheng J, Moorhead M, Carlton VEH, Stow P, Coustan-Smith E, et al. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood.* 2012;120:5173–80.
- Logan AC, Zhang B, Narasimhan B, Carlton V, Zheng J, Moorhead M, et al. Minimal residual disease quantification using consensus primers and high-throughput IGH sequencing predicts post-transplant relapse in chronic lymphocytic leukemia. *Leukemia.* 2013;27:1659–65.
- Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, et al. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc Natl Acad Sci USA.* 2011;108:21194–9.
- Ladetto M, Brüggemann M, Monitillo L, Ferrero S, Pepin F, Drandi D, et al. Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders. *Leukemia.* 2014;28:1299–307.
- Wren D, Walker BA, Brüggemann M, Catherwood MA, Pott C, Stamatopoulos K, et al. Comprehensive translocation and clonality detection in lymphoproliferative disorders by next-generation sequencing. *Haematologica.* 2017;102:e57–e60.
- Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. *Nat Rev Genet.* 2017;18:473–84.
- Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J Clin Microbiol.* 2016;54:2857–65.
- Endrullat C, Glökler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. *Appl Transl Genom.* 2016;10:2–9.
- Kotrova M, Trka J, Kneba M, Brüggemann M. Is next-generation sequencing the way to go for residual disease monitoring in acute lymphoblastic leukemia? *Mol Diagn Ther.* 2017;21:481–92.
- Kurtz DM, Green MR, Bratman SV, Scherer F, Liu CL, Kunder CA, et al. Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood.* 2015;125:3679–87.
- Pulsipher Ma, Carlson C, Langholz B, Wall Da, Schultz KR, Bunin N, et al. IgH-V (D) J NGS-MRD measurement pre- and early post-allotransplant de fi nes very low- and very high-risk ALL patients. *Blood.* 2015;125:3501–9.
- Langerak AW, Brüggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D, et al. High-throughput immunogenetics for clinical and research applications in immunohematology: potential and challenges. *J Immunol.* 2017;198:3765–4.
- Brüggemann M, Kotrova M, Knecht H, Bartram J, Boudjoghra M, Bystry, V et al. Next-generation sequencing of immunoglobulin and T-cell receptor gene rearrangements for MRD marker identification in acute lymphoblastic leukemia: a validation study by EuroClonality-NGS. *Leukemia.* 2019. In press.
- Scheijen B, Meijers RW, Rijntjes J, van der Klift MY, Möbs M, Steinhilber J et al. Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. *Leukemia.* 2019. In press.
- Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A, et al. ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. *Bioinformatics.* 2017;33:435–7.
- van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 concerted action BMH4-CT98-3936. *Leukemia.* 2003;17:2257–317.
- Langerak A W, Szczepański T, Van Der Burg M, ILM Wolvers-Tettero, JJM VanDongen. Heteroduplex PCR analysis of rearranged T cell receptor genes for clonality assessment in suspect T cell proliferations. *Leukemia.* 1997;11:2192–9.
- Grupp SA, Kalos M, Barrett D, Aplenc R, Porter DL, Rheingold SR, et al. Chimeric antigen receptor-modified T cells for acute lymphoid leukemia. *N Engl J Med.* 2013;368:1509–18.
- Salson M, Giraud M, Caillaud A, Grardel N, Duployez N, Ferret Y, et al. High-throughput sequencing in acute lymphoblastic leukemia: follow-up of minimal residual disease and emergence of new clones. *Leuk Res.* 2017;53:1–7.



28. Takamatsu H, Wee RK, Zaimoku Y, Murata R, Zheng J, Moorhead M, et al. A comparison of minimal residual disease detection in autografts among ASO-qPCR, droplet digital PCR, and next-generation sequencing in patients with multiple myeloma who underwent autologous stem cell transplantation. *Br J Haematol*. 2017. <https://doi.org/10.1111/bjh.15002>.
29. Wood B, Wu D, Crossley B, Dai Y, Williamson D, Gawad C, et al. Measurable residual disease detection by high-throughput sequencing improves risk stratification for pediatric B-ALL. *Blood*. 2018;131:1350–9.

## Affiliations

Henrik Knecht<sup>1</sup> · Tomas Reigl<sup>2</sup> · Michaela Kotrová<sup>1</sup>  · Franziska Appelt<sup>1</sup> · Peter Stewart<sup>3</sup> · Vojtech Bystry<sup>2</sup> · Adam Krejci<sup>2</sup> · Andrea Grioni<sup>4</sup> · Karol Pal<sup>2</sup> · Kamila Stranska<sup>2,5</sup> · Karla Plevova<sup>2,5</sup> · Jos Rijntjes<sup>6</sup> · Simona Songia<sup>4</sup> · Michael Svatoň<sup>7</sup> · Eva Froňková<sup>7</sup> · Jack Bartram<sup>8</sup> · Blanca Scheijen<sup>6</sup> · Dietrich Herrmann<sup>1</sup> · Ramón García-Sanz<sup>9</sup>  · Jeremy Hancock<sup>10</sup> · John Moppett<sup>11</sup>  · Jacques J. M. van Dongen<sup>12</sup> · Giovanni Cazzaniga<sup>13</sup>  · Frédéric Davi<sup>13</sup> · Patricia J. T. A. Groenen<sup>6</sup> · Michael Hummel<sup>14</sup> · Elizabeth A. Macintyre<sup>15</sup> · Kostas Stamatopoulos<sup>16</sup> · Jan Trka<sup>7</sup> · Anton W. Langerak<sup>17</sup> · David Gonzalez<sup>3</sup> · Christiane Pott<sup>1</sup> · Monika Brüggemann<sup>1</sup> · Nikos Darzentas<sup>1,2</sup> on behalf of the EuroClonality-NGS Working Group

<sup>1</sup> Department of Hematology, University Hospital Schleswig-Holstein, Kiel, Germany

<sup>2</sup> Central European Institute of Technology, Masaryk University, Brno, Czech Republic

<sup>3</sup> Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, UK

<sup>4</sup> Centro Ricerca Tettamanti, University of Milano Bicocca, Monza, Italy

<sup>5</sup> Department of Internal Medicine – Hematology and Oncology, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic

<sup>6</sup> Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands

<sup>7</sup> CLIP – Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University, University Hospital Motol, Prague, Czech Republic

<sup>8</sup> Department of Paediatric Haematology, Great Ormond Street Hospital, London, UK

<sup>9</sup> IBMCC-CSIC, Hospital Universitario de Salamanca-IBSAL, Salamanca, Spain

<sup>10</sup> Bristol Genetics Laboratory, Southmead Hospital, Bristol, UK

<sup>11</sup> Department of Pediatric Haematology, Bristol Royal Hospital for Children, Bristol, UK

<sup>12</sup> Department of Immunohematology and Blood Transfusion (IHB), Leiden University Medical Center, Leiden, The Netherlands

<sup>13</sup> Department of Hematology, Hopital Pitié-Salpêtrière, Paris, France

<sup>14</sup> Institute of Pathology, Charité – Universitätsmedizin Berlin, Berlin, Germany

<sup>15</sup> Department of Hematology, APHP Necker-Enfants Malades and Paris Descartes University, Paris, France

<sup>16</sup> Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece

<sup>17</sup> Department of Immunology, Laboratory Medical Immunology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands

## Chapter 6

A novel EP300 mutation associated with Rubinstein-Taybi syndrome type 2 presenting as combined immunodeficiency.

Saettini F, Moratto D, **Grioni A**, Maitz S, Iascone M, Rizzari C, Pavan F, Spinelli M, Bettini LR, Biondi A, Badolato R. A novel EP300 mutation associated with Rubinstein-Taybi syndrome type 2 presenting as combined immunodeficiency. *Pediatr Allergy Immunol.* 2018 Nov;29(7):776-781. doi: 10.1111/pai.12968. Epub 2018 Sep 28. PMID: 30076641.



Article type : Letter to the Editor

## **A novel *EP300* mutation associated with Rubinstein-Taybi syndrome type 2 presenting as combined immunodeficiency**

Francesco Saettini, MD<sup>a</sup>, Daniele Moratto, PhD<sup>b</sup>, Andrea Grioni<sup>a,c,d</sup>, Silvia Maitz, MD<sup>e</sup>, Maria Iascone, MD<sup>f</sup>, Carmelo Rizzari, MD<sup>a</sup>, Fabio Pavan, MD<sup>a</sup>, Marco Spinelli, MD<sup>a</sup>, Laura Rachele Bettini, MD<sup>a</sup>, Andrea Biondi, MD<sup>a,d</sup>, Raffaele Badolato, MD, PhD<sup>b,g</sup>

<sup>a</sup>Department of Pediatrics, Fondazione MBBM, University of Milan-Bicocca, Monza, Italy.

<sup>b</sup>Institute for Molecular Medicine A. Nocivelli, and Department of Pathology, Laboratory of Genetic Disorders of Childhood, Department of Molecular and Translational Medicine, University of Brescia, Spedali Civili, Brescia, Italy.

<sup>c</sup>National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic

<sup>d</sup>Centro Ricerca Tettamanti, Fondazione MBBM, Monza, Italy.

<sup>e</sup>Clinical Pediatric Genetics Unit, Pediatrics Clinics, Fondazione MBBM, San Gerardo Hospital, Monza, Italy.

<sup>f</sup>USSD Laboratorio di Genetica Medica, Azienda Socio Sanitaria Territoriale Papa Giovanni XXIII, Bergamo, Italy.

<sup>g</sup>Department of Clinical and Experimental Sciences, Pediatrics Clinic and Institute for Molecular Medicine A. Nocivelli, University of Brescia, Spedali Civili di Brescia, Italy

Corresponding author:

Francesco Saettini. Department of Pediatrics, University of Milan-Bicocca, Via Cadore, 20900 Monza, Italy.  
Mail: francescosaettini@yahoo.it

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/pai.12968

This article is protected by copyright. All rights reserved.

**KEY WORDS:** Rubinstein-Taybi syndrome, common variable immunodeficiency, autoimmune lymphoproliferative syndrome, combined immunodeficiency, syndromic immunodeficiency.

#### **CONFLICT OF INTEREST**

Authors declare no conflict of interest

#### **AUTHOR CONTRIBUTION**

FS, DM, AG, and RB wrote the manuscript. DM, MI, and AG performed the experiments. FS, SM, CR, FP, MS, LRB, AB, and RB critically revised the manuscript.

*To the editor:*

The Rubinstein-Taybi syndrome (RSTS; OMIM #180849, #613684) is a rare developmental disorder characterized by craniofacial dysmorphisms, broad thumbs and toes, and mental and growth deficiency. It affects equally males and females, with prevalence of 1:100.000 to 1:125.000 liveborn infants. Mutations in the cyclic adenosine monophosphate response element-binding protein (CREB)-binding protein (*CREBBP*) or in the E1A-associated protein p300 (*EP300*) have been demonstrated in 55% (RSTS1) and up to 8% of the patients (RSTS2), respectively. Hypogammaglobulinemia has been described in RSTS1 patients, while no immunological defect was reported in patients with RSTS2 (1-4).

Herein, we describe a 15-year-old male patient with novel heterozygous mutation of *EP300* gene associated with lymphopenia and hypogammaglobulinemia. His manifestations were initially characterized by elevated CD3+TCR $\alpha\beta$ +CD4 $-$ CD8 $-$  double negative T-cells (DNT), Fisher-Evans' syndrome, and hypogammaglobulinemia suggesting Autoimmune Lymphoproliferative Syndrome (ALPS) evolving into Common Variable Immunodeficiency (CVID). But, later on, he developed lymphopenia suggesting a combined immunodeficiency (CID).

The patient was born to unrelated healthy Italian parents at 34 weeks' gestation with adequate weight for gestational age. Shortly after birth, he underwent several surgical procedures due to interventricular defect, aortic coarctation, double outlet right ventricle, open Botalli's duct, and gastroesophageal reflux. At the age of four, he came to our attention due to stomatitis. Clinical examination revealed dysmorphism (microcephaly, wide forehead, sparse eyebrows, high nasal root,

low-hanging columella, thick lips, micrognathia), splenomegaly (spleen diameter 11.6 cm at abdominal ultrasound), and severe developmental delay. In the course of the infectious episode, blood tests showed leukopenia associated with neutropenia (white blood cells  $2.210/\text{mm}^3$ ; neutrophils  $20/\text{mm}^3$ ) and thrombocytopenia (platelets  $1.000/\text{mm}^3$ ). Analysis of bone marrow aspirate revealed normal differentiation of both myeloid and erythroid lineages and no significant abnormalities in megakaryocytes numbers and morphology were detected. Treatment with high doses of intravenous immunoglobulin (IVIG) resulted in increase of platelet counts (up to  $44000/\text{mm}^3$  after 1 month), while neutrophil counts spontaneously recovered after the infectious episode. Thrombocytopenia relapsed after 2 months ( $2000/\text{mm}^3$ ) but intravenous high-doses of corticosteroids were not effective to restore normal platelet count. When the child was six, oral corticosteroid treatment was started, but this therapy could not prevent autoimmune hemolytic anemia episodes. From six to twelve years of age, low dose steroids have been administered and the patient presented several infections (stomatitis, upper respiratory tract infections, and skin abscesses), none requiring hospitalization but one episode of hypovolemic shock due to severe diarrhea. Because of the history of infections, the patient was started to IVIG replacement at the age of fourteen. Despite this treatment, he had relapsing bilateral pneumonia requiring assisted ventilation and/or admission to pediatric intensive care unit due to acute respiratory failure with evidence of *Mycoplasma pneumoniae* and Rhinovirus infections.

Immunological evaluation during his follow-up (Figure 1A and 1B, Table 1, Table S1, and Table S2) showed severely progressive lymphopenia (lymphocyte counts ranging from  $350/\text{mm}^3$  to  $2100/\text{mm}^3$ ), hypogammaglobulinemia, intermittent thrombocytopenia, undetectable anti-diphtheria and anti-tetanus toxoid antibodies, and splenomegaly. Interestingly, analysis of isohemoagglutinins while he was off corticosteroid treatment revealed low titers of anti-A (1:8) at 4 years of age, but normal immunoglobulins. Immunological re-evaluation when the child was seven, showed reduced lymphocyte proliferation to mitogens, hypogammaglobulinemia, increased DNT cells and impaired FAS-mediated apoptosis in two separate occasions (Table S3). Evaluation of B-cells when he was fourteen showed other abnormalities of B lymphocyte subsets, including reduction of switched memory B-cells and increased  $\text{CD}21^{\text{lo}}\text{CD}38^{\text{lo}}$  B-cells (Figure 1C). Analysis of T-cell compartment unveiled a decreased proportion of  $\text{CD}4^+\text{CD}31^+\text{CCR}7^+\text{CD}45\text{RA}^+$  recent thymic emigrants (RTE) cells and of naïve T-cells ( $\text{CD}4^+\text{CCR}7^+\text{CD}45\text{RA}^+$  and  $\text{CD}8^+\text{CCR}7^+\text{CD}45\text{RA}^+$ ), with prevalence of effector memory T-cells ( $\text{CD}8^+\text{CCR}7^-\text{CD}45\text{RA}^-$ ) (Table S1). DNT cells were persistently elevated.

Proband and parents whole-exome sequencing (trio-WES) revealed a novel *de novo* heterozygous missense mutation (NM\_001429.3:c.4763T>C, p.Met1588Thr) in the exon 29 of the gene *EP300* encoding the Histone Acetyltransferase (HAT) protein p300. This protein is a transcription factor that, like CREBBP, is recruited with NF- $\kappa$ B to bind promoter sites and is preloaded in most of the promoters and enhancers of NF- $\kappa$ B regulated genes(5). We herein report on a novel *EP300* missense mutation causing the substitution of the non-polar amino acid Met1588 with the polar amino acid threonine (Thr). This substitution most probably disrupts the C/H3 domain folding causing the loss of protein function (Figure 1D and 1E) and is considered probably damaging by software analysis, suggesting as probably causative of the disease. P300 HAT domain is a highly-conserved zinc finger domain affecting acetyltransferase activity, promoting histone acetylation, and DNA access for gene transcription. The HAT region spanning amino acid position 1587-1817 contains the C/H3 domain necessary for the interaction with different proteins (e.g. GATA4)(6).

Immunological features have been analyzed in a limited number of RSTS1 patients [1-4], but never reported in patients with RSTS2 (Table S4). However, there are striking similarities between the immunological features of the patient with RTSS2 we describe in this report and what observed in previous studies performed in RSTS1 patients (1,3,4). Our patient presented with autoimmune cytopenia, splenomegaly, and defective lymphocyte apoptosis with increased DNT cell count, leading to diagnosis of ALPS. But, the appearance of hypogammaglobulinemia, poor antibody response but progressive B- and T-cell lymphopenia could also suggest CVID. In fact, when the patient was 7, flow cytometry revealed expansion of CD19<sup>hi</sup>CD21<sup>lo</sup>CD38<sup>lo</sup> B-cells, that is frequently associated with splenomegaly in CVID patients(7), and reduced number of switched memory B-cells, similarly to what previously reported in a RSTS patient with *CREBBP* mutation (4). These changes of B cell subsets are in keeping with the expansion of CD21<sup>lo</sup>CD38<sup>lo</sup> B-cells that was reported in a patient with NF- $\kappa$ B1 haploinsufficiency(8), suggesting that alterations in the NF- $\kappa$ B pathway due to *EP300* mutations might also affect B-cell differentiation.

Immunological studies when the patient was 14 years old showed low IgG (321 mg/dl), markedly elevated IgM levels and decrease of T-cells including CD3<sup>+</sup>, CD4<sup>+</sup>, CD8<sup>+</sup>, and naïve CD4<sup>+</sup> cells (Figure 1 and Table S2). According to analysis of B-cell subsets, high IgM levels are probably related to high proportion of terminal differentiated IgM<sup>+</sup> B-cells. These immunological features, associated with the history of invasive infections and immune dysregulation (lymphoproliferation and Fisher-Evans' syndrome), suggest that RSTS2 can manifest as CID(10,11).

This case of RSTS2 underlines the value of WES in patients with complex clinical phenotype. Because no RSTS typical trait (i.e. broad halluces and thumbs) was identified in this patient (Table S5), several syndromes and primary immunodeficiencies were considered before performing WES (Table S6). In addition, this report expands the clinical spectrum of RSTS2, suggesting that *EP300* mutations should be suspected in patients with clinical and immunological features resembling distinct immunological defects which share common manifestations such as CVID, ALPS, and CID. At disease onset, the immunological and clinical features of this case were reminiscent of ALPS, and because of the appearance of hypogammaglobulinemia of CVID, but in the following years the clinical picture evolved as CID. Furthermore, this study suggests that immunological work-up should be taken into consideration in RSTS patients to identify those immunological abnormalities that may lead to development of severe immune-hematological complications.

**Figure 1.** A. Immunoglobulin serum levels (black arrow indicates starting of Ig replacement) and E. lymphocyte subsets during follow-up. B. Flow cytometric analysis of the B cell compartment. Identification and quantification of total B cells (first column) and B cell subsets according to the gating strategy used in fig.1C of Lougaris et al. [4] for a RSTS1 patient: second column transitional and terminally differentiated cells (CD38hiCD21lo/dim) (green gate), mature cells (CD38lo/dimCD21hi) (orange gate) and CD21loCD38lo cells (yellow gate); third column naïve (IgD+CD27-), switched memory (IgD-CD27+) and unswitched memory cells (IgD+CD27+); fourth column transitional (CD20+CD27-) and terminally differentiated cells (CD20-CD27hi). C. 3D structure representation of region spanning amino acid 1587 to 1817 of EP300 and containing the C/H3 region. Red arrow points to non-polar Met1588. D. Red arrow indicating the point mutation with the polar amino acid Treonine.

Francesco Saettini, MD<sup>a</sup>, Daniele Moratto, PhD<sup>b</sup>, Andrea Grioni<sup>a,c,d</sup>, Silvia Maitz, MD<sup>e</sup>, Maria Iascone, MD<sup>f</sup>, Carmelo Rizzari, MD<sup>a</sup>, Fabio Pavan, MD<sup>a</sup>, Marco Spinelli, MD<sup>a</sup>, Andrea Biondi, MD<sup>a,d</sup>, Raffaele Badolato, MD, PhD<sup>b,g</sup>

<sup>a</sup>Department of Pediatrics, Fondazione MBBM, University of Milan-Bicocca, Monza, Italy.

<sup>b</sup>Institute for Molecular Medicine A. Nocivelli, and Department of Pathology, Laboratory of Genetic Disorders of Childhood, Department of Molecular and Translational Medicine, University of Brescia, Spedali Civili, Brescia, Italy.

<sup>c</sup>National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic

<sup>d</sup>Centro Ricerca Tettamanti, Fondazione MBBM, Monza, Italy.

<sup>e</sup>Clinical Pediatric Genetics Unit, Pediatrics Clinics, Fondazione MBBM, San Gerardo Hospital, Monza, Italy.

<sup>f</sup>USSD Laboratorio di Genetica Medica, Azienda Socio Sanitaria Territoriale Papa Giovanni XXIII, Bergamo, Italy.

<sup>g</sup>Department of Clinical and Experimental Sciences, Pediatrics Clinic and Institute for Molecular Medicine A. Nocivelli, University of Brescia, Spedali Civili di Brescia, Italy

Corresponding author:

Francesco Saettini. Department of Pediatrics, University of Milan-Bicocca, Via Cadore, 20900 Monza, Italy.  
Mail: francescosaettini@yahoo.it

## REFERENCES

(1)Villella A, Bialostocky D, Lori E, et al. Rubinstein-Taybi syndrome with humoral and cellular defects: a case report. *Arch Dis Child* 2000;83(4):360–361.

(2)Torres LC, Sugayama SMM, Arslanian C, et al. Evaluation of the immune humoral response of Brazilian patients with Rubinstein-Taybi syndrome. *Braz J Med Biol Res* 2010;43(12):1215-1224. doi: 10.1590/S0100-879X2010007500119

(3)Pasic S. Rubinstein-Taybi syndrome associated with humoral immunodeficiency. *J Investig Allergol Clin Immunol* 2015;25(2):137–138.

(4)Lougaris V, Facchini E, Baronio M, et al. Progressive severe B cell deficiency in pediatric Rubinstein-Taybi syndrome. *Clin Imm* 2016;173:181-183.



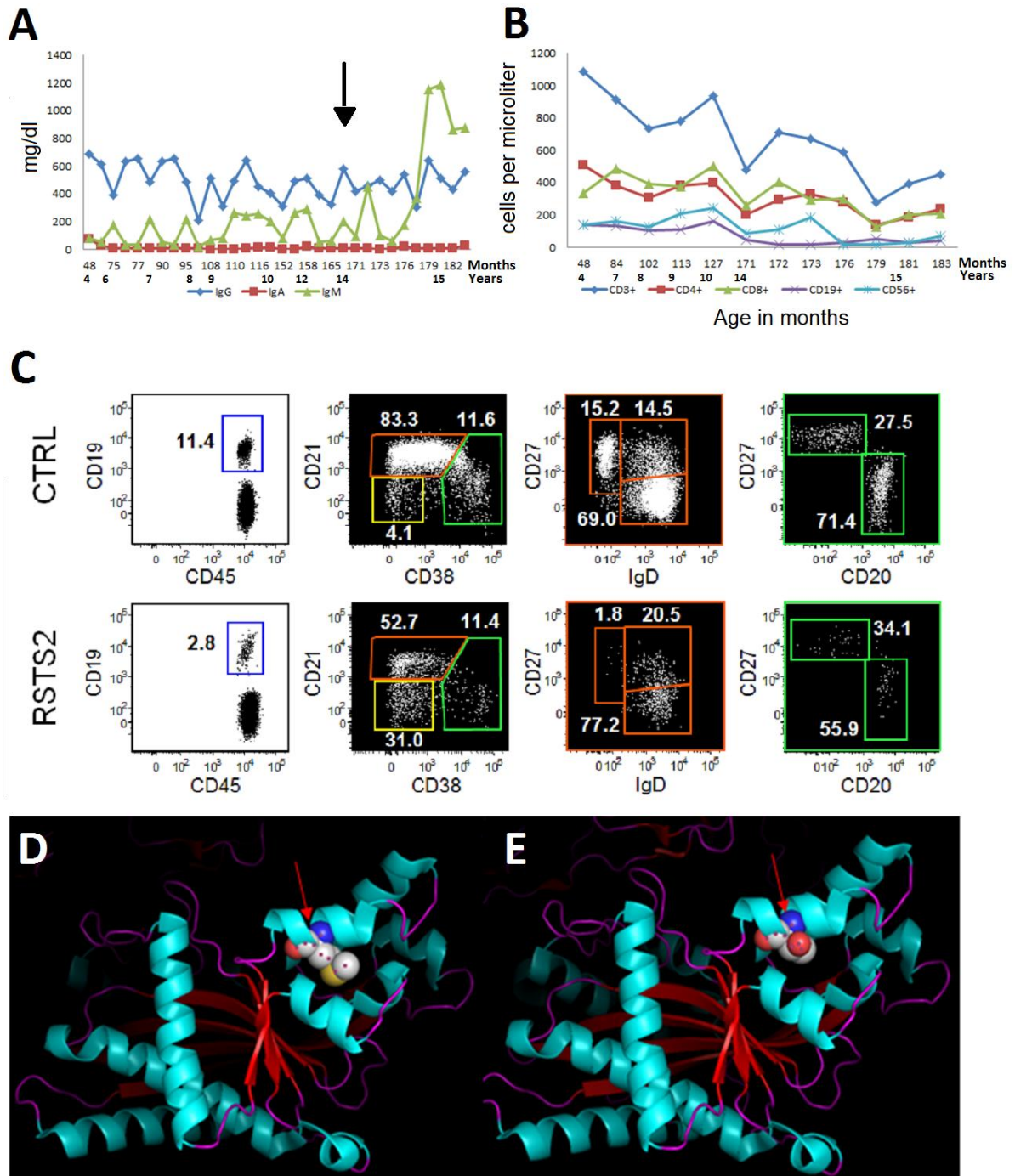
- (5) Mukherjee SP, Behar M, Birnbaum HA, et al. Analysis of the RelA:CBP/p300 Interaction Reveals Its Involvement in NF- $\kappa$ B-Driven Transcription. *PLoS Biol* 2013;11(9): e1001647. <https://doi.org/10.1371/journal.pbio.1001647>
- (6) Dai YS and Markham BE. p300 functions as a coactivator of transcription factor GATA-4. *J Biol Chem*. 2001;276:37178–37185
- (7) Wehr C, Kivioja T, Schmitt C, et al. The EUROclass trial: defining subgroups in common variable immunodeficiency. *Blood* 2008;111(1):76-85.
- (8) Lougaris V, Moratto D, Baronio M, et al. Early and late B-cell developmental impairment in nuclear factor kappa B, subunit 1-mutated common variable immunodeficiency disease. *J Allergy Clin Immunol* 2016;139(1):349-352. <http://dx.doi.org/10.1016/j.jaci.2016.05.045>
- (9) Park J, Munagala I, Xu H, et al. Interferon Signature in the Blood in Inflammatory Common Variable Immune Deficiency. *PLoS One* 2013;8(9):e74893.10.
- (10) ESID. ESID Registry—working definitions for clinical diagnosis of PID. [http://esid.org/content/download/13053/372959/file/ESIDRegistry\\_ClinicalCriteria.pdf](http://esid.org/content/download/13053/372959/file/ESIDRegistry_ClinicalCriteria.pdf) 2015.
- (11) Speckmann C, Doerken S, Aiuti A, et al. P-CID study of the Inborn Errors Working Party of the EBMT. A prospective study on the natural history of patients with profound combined immunodeficiency: An interim analysis. *J Allergy Clin Immunol*. 2017;139(4):1302-1310.e4. doi: 10.1016/j.jaci.2016.07.040.

**Table 1. Immunological and haematological features of the index patient.**

	<b>4 years of age</b>	<b>7 years of age</b>	<b>14 years of age</b>
<b>Steroid treatment</b>	-	0.35 mg/kg/die	0.16 mg/kg/die
<b>WBC</b>	1970/mmc (5200-11000)	5040/mmc (4400-9500)	4820/mmc (4400-8100)
<b>Lymphocytes</b>	1370/mmc (2300-5400)	1220/mmc (1900-3700)	820/mmc (1400-3300)
<b>Neutrophils</b>	200/mmc	3330/mmc	3480/mmc
<b>Platelets</b>	33000/mmc	139000/mmc	70000/mmc
<b>CD3+</b>	79.4% (56-78.9)	74.6% (59.1-80.9)	85.2% (58.1-80.1)
	1088/mmc (1300-4500)	910/mmc (900-3200)	699/mmc (750-2700)
<b>CD4+</b>	37.0% (29.4-55.7)	31.5% (24.9-51.1)	32.5% (27.9-53.4)
	507/mmc (600-2760)	384/mmc (500-2100)	267/mmc (380-1730)
<b>HLA-DR+</b>	6.5% (1.3-12.1)		19.2% (1.4-11.5)
	33/mmc (17-225)		51/mmc (17-244)
<b>Naïve CD45RA+CCR7+</b>	50.3% (49.2-85.8)		10.7% (35.1-82.2)
	255/mmc (440-2050)		29/mmc (205-1140)
<b>RTE CD45RA+CCR7+CD31+</b>			1.4% (26.2-67.1)
			4/mmc (180-750)
<b>Central memory CD45RA-CCR7+</b>	40.5% (9.6-31.9)		63.1% (10.7-44.3)
	205/mmc (210-540)		168/mmc (130-490)
<b>Effector Memory CD45RA-CCR7-</b>	7.3% (2.8-16.9)		24.7% (5.4-25.3)
	37/mmc (40-240)		66/mmc (60-275)
<b>Terminal differentiated CD45RA+CCR7-</b>	1.8% (0.7-4.8)		1.7% (0.6-6.5)
	9/mmc (8-110)		5/mmc (3-31)
<b>CD8+</b>	24.6% (11.6-32.4)	40% (13.8-31.2)	46.9% (12.3-31.9)
	337/mmc (410-1360)	488/mmc (400-1150)	385/mmc (270-800)
<b>HLA-DR+</b>	11.0% (0.9-33.2)		44.8% (1.8-31.4)
	37/mmc (20-500)		172/mmc (12-320)
<b>Naïve CD45RA+CCR7+</b>	47.2% (22.8-79.9)		9.5% (15.1-76.7)
	159/mmc (160-760)		37/mmc (60-530)
<b>Central memory CD45RA-CCR7+</b>	7.1% (0.9-11.3)		4.4% (1.3-15.4)
	24/mmc (7-115)		17/mmc (15-80)

<b>Effector Memory</b>	33.4% (4.7-31.3)		62.7% (6.1-40.1)
<b>CD45RA-CCR7-</b>			
	113/mmc (30-380)		241/mmc (60-280)
<b>Terminal differentiated CD45RA+CCR7-</b>	12.3% (6.8-52.7)		23.5% (6.8-46.7)
	41/mmc (80-620)		90/mmc (30-300)
<b>CD19+</b>	10.0% (10.7-34.9)	10.7% (8.6-26.3)	2.8% (6.3-25.4)
	137/mmc (260-1750)	131/mmc (215-1000)	23/mmc (106-1200)
<b>RBE CD38hiCD10+</b>	13.3% (10.6-42.6)		7.7% (4.2-28.5)
	18/mmc (40-540)		2/mmc (20-175)
<b>Naïve IgD+CD21hiCD10-CD27-</b>	46.3% (34.2-65.5)		40.3% (39.9-74.1)
	63/mmc (125-800)		9/mmc (60-360)
<b>CD19hiCD21lo</b>	25.6% (1.5-9.8)		31% (1.0-15.8)
	35/mmc (8-70)		7/mmc (4-87)
<b>Switched memory IgD-CD27+CD21hi</b>	1.3% (1.5-14.2)		1.3% (2.7-16.5)
	2/mmc (18-140)		0/mmc (6-90)
<b>IgM Memory IgD+CD27+CD21hi</b>	9.9% (2.9-15.3)		11.1% (3.4-18.0)
	14/mmc (22-135)		3/mmc (5-108)
<b>Terminal differentiated CD38hiCD27hiCD20-</b>	2.9% (0.4-15.3)		4.7% (0.2-7.1)
	4/mmc (4-130)		1/mmc (1-32)
<b>PC CD38hiCD27hiCD20-CD138+</b>	n.a.		0.7% (0.1-2.4)
			0/mmc (0-8)
<b>CD56+CD16+</b>	9.9% (3.0-21.3)	13% (3.3-22.8)	10.7% (3.8-24.6)
	136/mmc (90-850)	159/mmc (120-900)	87/mmc (80-830)
<b>CD3+CD4+CD8- a/b (total/CD3+ lymphocytes)</b>	n.a.	2.66%/3.57% (1.5-2.5)	4.6%/5.7% (1.5-2.5)
<b>TCR <math>\gamma/\delta</math></b>	13.6%		1.3%
<b>Isohemoagglutinin</b>	Anti-B absent, Anti-A 1:8		
<b>Mitogen proliferation</b>	Normal	Reduced to PHA and ConA	
<b>TRECs</b>	33147 (3521+-17922)		
<b>IgG / A / M</b>	685 / 73 / 83 mg/dl	274 / 17 / 185 mg/dl	321 / 7 / 59 mg/dl
<b>Anti-tetanus / Anti-diphtheria Ab</b>	Absent / Absent	Absent / Absent	Absent / Absent
<b>Spleen diameter</b>	11.6 cm	14 cm	17 cm

Between brackets reference values for absolute and percentages are shown according to age. ConA = concavalin A. Ig = immunoglobulin. PC = plasmacells. PHA = phytohemagglutinin. RTE = recent thymic emigrants. RBE = recent B emigrants. TCR = T cell receptor. TRECs = T-cell receptor excision circles. WBC = white blood cells.



## Chapter 7

First evidence of a paediatric patient with Cornelia de Lange syndrome with acute lymphoblastic leukaemia.

Fazio G, Massa V, **Grioni A**, Bystry V, Rigamonti S, Saitta C, Galbiati M, Rizzari C, Consarino C, Biondi A, Selicorni A, Cazzaniga G. First evidence of a paediatric patient with Cornelia de Lange syndrome with acute lymphoblastic leukaemia. *J Clin Pathol*. 2019 Aug;72(8):558-561. doi:10.1136/jclinpath-2019-205707. Epub 2019 Apr 4. PMID: 30948435.

# First evidence of a paediatric patient with Cornelia de Lange syndrome with acute lymphoblastic leukaemia

Grazia Fazio,<sup>1</sup> Valentina Massa,<sup>2</sup> Andrea Grioni,<sup>1,3</sup> Vojtech Bystry,<sup>3</sup> Silvia Rigamonti,<sup>1,2</sup> Claudia Saitta,<sup>1</sup> Marta Galbiati,<sup>1</sup> Carmelo Rizzari,<sup>4</sup> Caterina Consarino,<sup>5</sup> Andrea Biondi,<sup>1,4</sup> Angelo Selicorni,<sup>6</sup> Giovanni Cazzaniga<sup>1,7</sup>

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/jclinpath-2019-205707>).

<sup>1</sup>Centro di Ricerca Tettamanti, Clinica Pediatrica, Università di Milano, Bicocca, Monza, Italy

<sup>2</sup>Dipartimento di Scienze della Salute, Università degli Studi di Milano, Milano, Italy

<sup>3</sup>Central European Institute of Technology, Masarykova Univerzita, Brno, Czech Republic

<sup>4</sup>Pediatric Department, Monza Brianza per il Bambino e la sua Mamma (MBBM) Foundation, Monza, Italy

<sup>5</sup>Ematologia ed Oncologia Pediatrica, Presidio Ospedaliero Ciaccio-De Lellis, Catanzaro, Italy

<sup>6</sup>Department of Pediatrics, Presidio S. Fermo, ASST Lariana, Como, Italy

<sup>7</sup>Dipartimento di Medicina e Chirurgia, Università degli Studi di Milano-Bicocca, Monza, Italy

## Correspondence to

Dr Giovanni Cazzaniga, Centro di Ricerca Tettamanti, Clinica Pediatrica, Università di Milano-Bicocca, Monza 20900, Italy; [gianni.cazzaniga@hsgerardo.org](mailto:gianni.cazzaniga@hsgerardo.org) and Dr Angelo Selicorni, Department of Pediatrics, Presidio S. Fermo, ASST Lariana, Como, Italy; [angelo.selicorni61@gmail.com](mailto:angelo.selicorni61@gmail.com)

GF and VM contributed equally.

Received 7 January 2019

Revised 15 March 2019

Accepted 16 March 2019



© Author(s) (or their employer(s)) 2019. No commercial re-use. See rights and permissions. Published by BMJ.

**To cite:** Fazio G, Massa V, Grioni A, et al. *J Clin Pathol* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jclinpath-2019-205707

## ABSTRACT

Cornelia de Lange syndrome (CdLS) is a rare autosomal-dominant genetic disorder characterised by prenatal and postnatal growth and mental retardation, facial dysmorphism and upper limb abnormalities. Germline mutations of cohesin complex genes *SMC1A*, *SMC3*, *RAD21* or their regulators *NIPBL* and *HDAC8* have been identified in CdLS as well as somatic mutations in myeloid disorders. We describe the first case of a paediatric patient with CdLS with B-cell precursor Acute Lymphoblastic Leukaemia (ALL). The patient did not show any unusual cytogenetic abnormality, and he was enrolled into the high risk arm of AIEOP-BFM ALL2009 protocol because of slow early response, but 3 years after discontinuation, he experienced an ALL relapse. We identified a heterozygous mutation in exon 46 of *NIPBL*, causing frameshift and a premature stop codon (RNA-Targeted Next generation Sequencing Analysis). The analysis of the family indicated a de novo origin of this previously not reported deleterious variant. As for somatic cohesin mutations in acute myeloid leukaemia, also this ALL case was not affected by aneuploidy, thus suggesting a major impact of the non-canonical role of *NIPBL* in gene regulation. A potential biological role of *NIPBL* in leukaemia has still to be dissected.

## INTRODUCTION

The multimeric cohesin complex is highly evolutionary conserved and consists of core proteins (*SMC1A*, *SMC3*, *RAD21* and either *SA1* or *SA2* in vertebrate) and associated proteins (*PDS5A*, *WAPL* and *Sororin*). The cohesin complex has been shown to play a pivotal role in sister chromatid cohesion to spindle apparatus preceding chromosome segregation and in fundamental cellular events such as postreplicative DNA repair, checkpoint activation and transcription regulation.<sup>1</sup> Indeed, transcription factors clusters are formed around cohesins that warrant both DNA accessibility and binding. Germline mutations in cohesin genes cause autosomal dominant genetic disorders termed *cohesinopathies*, among which Cornelia de Lange syndrome (CdLS, OMIM 122470, 300590, 610759, 614701, 300882) and Roberts syndrome (RBS, OMIM 268300) are the best described.<sup>2</sup> Patients with CdLS present developmental delay, specific facial features, behavioural abnormalities and major malformations. CdLS is caused by heterozygous variant in *NIPBL*, *RAD21* or *SMC3* or hemizygous variant in *HDAC8* or *SMC1A*.<sup>3</sup>

Somatic mutations in cohesins have been rarely reported in cancer.<sup>4</sup> However, in 6%–13% of acute myeloid leukaemia (AML) and other myeloid neoplasms, their mutations involving multiple components of the cohesin-complex, including *STAG1/SA1*, *STAG2*, *RAD21*, *SMC1A* and *SMC3* have been reported.<sup>5–7</sup> Interestingly, recently, a single CdLS/AMKL (acute megakaryoblastic leukaemia) case has been reported with a constitutional *NIPBL* mutation.<sup>8</sup>

Herein, we describe the first patient with CdLS with acute lymphoblastic leukaemia (ALL) carrying a novel *NIPBL* pathogenetic variant.

## METHODS

The index patient was enrolled in the AIEOP-BFM ALL2009 study and written informed consent was obtained from parents within the protocol. DNA samples from buccal smears have been collected under Informed Consent for research purposes, in accordance with the Declaration of Helsinki. Details of molecular biology, bioinformatics and statistics analyses are described in online supplementary file 1. Briefly, conventional cytogenetic, RT-PCR, PCR-based MRD molecular approaches have been applied, in addition to RNA targeted Next Generation Sequencing (NGS) panel (TruSight Pan-Cancer, Illumina, San Diego, California, USA; FASTQ files available on ENA database, accession PRJEB29923). Diagnostic and remission DNA samples were genotyped by CytoScan HD Array (Affymetrix, Santa Clara, California, USA; submitted to GEO GSE122859), in addition to buccal DNA, used for PCR variants analyses.

## RESULTS AND DISCUSSION

The index patient (PT1) was a male diagnosed at the age of 3 years as CdLS based on clinical features, such as short stature, typical facial dysmorphism, small hands, bilateral clinodactyly of the fifth finger, gastro-oesophageal reflux, back hirsutism, microcephaly and intellectual disabilities. Moreover, CdLS diagnosis was supported by the use of the recently published scoring system which turned to be positive (12 points with three cardinal features).<sup>9</sup> At the age of 18 months, he had a surgical intervention of anti-inflammatory plastic to cardiac oedema and bilateral biinguinal hernia. When 8 years old, he was diagnosed with B-cell precursor ALL. He had no central nervous system involvement and he was allocated in the high risk treatment group arm of the AIEOP-BFM ALL 2009 study protocol

because of slow early response pattern as assessed by Minimal Residual Disease monitoring. The patient did not experience major complications during treatment, but 3 years after discontinuation, he experienced an ALL relapse. The investigations on bone marrow (BM) at diagnosis did not reveal any prognostically relevant fusion transcript (*ETV6/RUNX1*, *BCR/ABL1*, *MLL/AF4*, *TCF3/PBX1*). Further, we applied RNA targeted TruSight Pan-Cancer panel, which contains probes for 1385 genes involved in cancer, B-cell leukaemia included (such as *ABL1*, *JAK2*, *EBF1*, *IKZF1*), to assess the presence of other alternative fusion transcript and to identify causative CdLS mutation. Bioinformatics analysis excluded the presence of fusions, while two known germline variants of *JAK3* and one of *TP53* have been detected (online supplementary table S1), confirmed by PCR (online supplementary table S2) on diagnostic/remission/buccal samples in PT1 and family (online supplementary table S3). The *JAK3* rs7254346 variant (shared between PT1 and father) is frequent (MAF=0.2815 in 1000Genomes, dbSNP database) and annotated as benign in ClinVar repository. However, *JAK3* rs3213409 has been assessed as heterozygous in all family members, although rare (MAF=0.0036 in 1000Genomes, dbSNP). Importantly, this variant has been described as somatic but not germline mutation both in AML and ALL.<sup>10</sup>

Also, the *TP53* rs1042522, shared between PT1 and mother is a common variant as well as the identified homozygous genotype G/G, whose biological significance is reported as benign, with a potential involvement in drug response.

Moreover, FISH analysis identified *IGH/CRLF2* translocation (online supplementary figure S1), while whole genome SNParray analysis identified recurrent somatic copy number variations, including *IKZF1* deletion (online supplementary table S4).

Among cohesin genes included in PanCancer panel, variant analysis identified four variants in PT1 (table 1, online supplementary data), which includes also the cohesin complex gene, such as *SMC1A*, *SMC3*, *NIPBL*, *STAG2*, *RAD21*.

Importantly, a novel c.7977dupT:p.P2659fs insertion variant was identified in exon 46 of *NIPBL* (table 1 and figure 1A). We further validated this mutation on BM diagnosis and remission DNA samples as well as on DNA from buccal smear collected after stop therapy (Sanger analyses are shown in figure 1B and PCR in online supplementary figure S2A). The analyses of this variant in buccal swabs from parents and the unaffected brother were negative as expected (figure 1B, online supplementary figure S2B), demonstrating its germline de novo origin in PT1. The predicted analysis of the Open Reading Frame indicated that the frameshift variant caused a premature stop codon and a truncated protein of 2659 AAs, instead of the expected 2804 AAs full length wild type *NIPBL* (online supplementary figure S3 and table S5). Additionally, this variant occurs within a highly

conserved region among species (online supplementary figure S4).

Furthermore, the intronic rs7075340 in *SMC3* and the exonic synonymous rs2419565 were annotated as non-pathogenic (dbSNP).

The fourth identified variant was the rs1264011 (dbSNP) hemizygous mutation (table 1), located in 5'UTR region of *SMC1A* gene (chromosome X) and annotated as benign (ClinVar, last Update: 8 September 2018).<sup>11</sup> It has been validated in PT1 diagnosis, remission and buccal swab DNA (online supplementary figure S5). We also confirmed its obvious maternal origin and the presence in the not affected brother (online supplementary figure S6).

Comprehensively, figure 1C shows the pedigree of family, with indicated the distribution of relevant variants.

According to the multiple hit hypothesis on cancer origin, there is growing evidence that lesions associated to genetic syndromes can predispose to cancer.<sup>12</sup> Some cancer predisposing syndromes have been associated either with ALL, that is, Down, Noonan, Li-Fraumeni and Williams syndromes,<sup>12</sup> or correlated to AML/MDS,<sup>13</sup> for example, Baraitser–Winter Cerebrofrontofacial Syndrome.<sup>14</sup>

Although mutations in cohesins have been found in many tumours myeloid neoplasms,<sup>5–7</sup> only few cases have been reported in patients with CdLS, also as accidental autoptoc findings.<sup>15</sup> Very recently, a patients with CdLS was reported with AMKL,<sup>8</sup> harbouring somatic aberrations characteristics of Down syndrome associated AMKL (ie, transient myeloid disorder, somatic trisomy 21 and *GATA1* mutation), a condition in which cohesin mutations have been found in more than 50% of cases,<sup>16</sup> thus suggesting their biological relevant role.

In the present study, we report the first case of a paediatric patient with CdLS with concomitant ALL, carrying a germline novel frameshift mutation in *NIPBL*, which is most likely responsible for the CdL autosomal dominant syndrome.

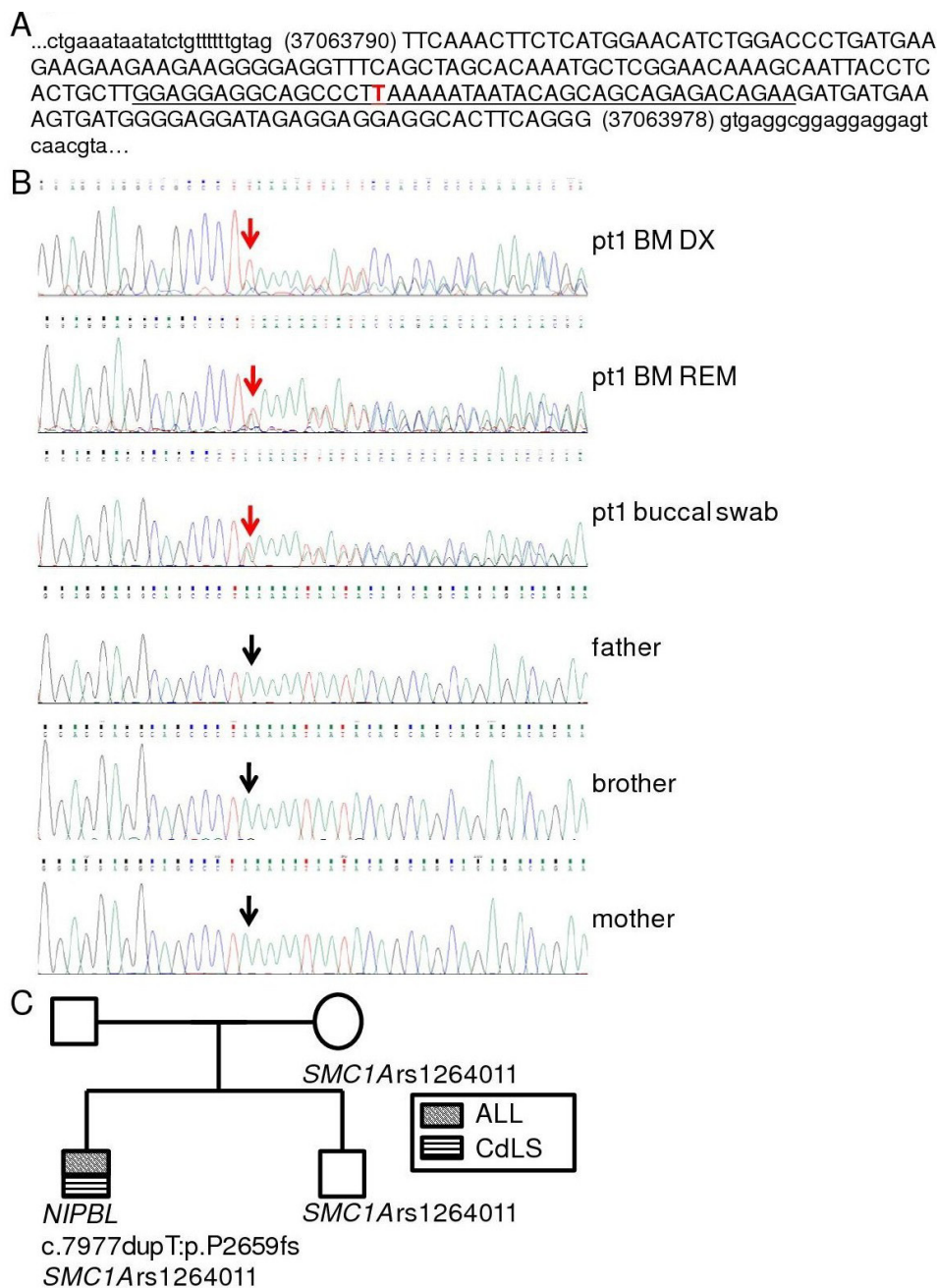
Those observations, although based on single cases, lead to hypothesis that unique mutational combinations of cohesins (either constitutive or somatic) with somatic leukemic variants can be related to specific myeloid or lymphoid leukaemias.

The expected result of the involvement of germline and somatic variants of cohesin genes in leukaemia would consist in aneuploidy, as a consequence of their canonical role in the assembly of the mitotic spindle.<sup>1</sup> Alternatively, the non-canonical role of cohesins in gene regulation might be determining a higher susceptibility to DNA damage or can potentially affect B-cell differentiation, because of immunoglobulin class switch recombination impairment, as previously shown in CdLS B-lymphoblastoid cell lines.<sup>17</sup> As hypothesis and schematically represented in online supplementary figure S7, these mechanisms still

**Table 1** Cohesins variants by targeted RNA NGS analysis

Gene	Chr	Start	End	Ref	Alt	VF	Func.refGene	Exonic Func. refGene	AAChange.refGene	dbSNP	clinvar
<i>NIPBL</i>	5	37 063 905	37 063 905	-	T	0.60	Exonic	Frameshift insertion	<i>NIPBL</i> :NM_015384:exon46:c.7977dupT:p.P2659fs, <i>NIPBL</i> :NM_133433:exon46:c.7977dupT:p.P2659fs	NOVEL	n.a.
<i>SMC3</i>	10	110 596 573	110 596 573	G	A	1	Intronic	None	n.a.	rs7075340	CLINSIG=benign
		110 602 112	110 602 112	A	G	1	Exonic	Synonymous SNV/ silent	<i>SMC3</i> :NM_005445:exon25:c.A3039G;p.S1013S	rs2419565	CLINSIG=non-pathogenic
<i>SMC1A</i>	X	53 422 619	53 422 619	G	A	1	5'UTR	5_prime_UTR_variant 109	NM_006306:c.-19C>T, NM_001281463:c.-682C>T	rs1264011	CLINSIG=non-pathogenic

VF=variant frequency = N°Alt Reads/ (N°Alt Reads+N°Ref Reads).  
NGS, Next Generation Sequencing; n.a., not applicable.



**Figure 1** *NIPBL* mutation in PT1 affected by CdLS and BCP-ALL. (A) Genomic region sequence on Chromosome 5, with details of *NIPBL* exon 46 and 25 bp of flanking introns. The underlined sequence showed the novel mutation, highlighted in red, *NIPBL*: NM\_015384:exon46:c.7977dupT:p.P2659fs. (B) Chromatograms showed the correspondent underlined sequence of exon 46, obtained by PCR and Sanger analyses. From top to the bottom, the three upper panels represented PT1 diagnostic bone marrow (DX BM), remission bone marrow and germinal DNA isolated from buccal swab, respectively. The three lower panels showed the same *NIPBL* gene region in family buccal swab DNA. (C) Cohesin genes variants pattern in patient 1 and his family. BCP-ALL, B-cell precursor acute lymphoblastic leukaemia; CdLS, Cornelia de Lange syndrome.

need to be established and further explored in larger populations to comprehend the functional link between cohesin and leukaemia.

**Handling editor** Mary Frances McMullin.

**Acknowledgements** The authors thank Dr Cristina Gervasini for comments and they are grateful to the Italian National Association of Volunteers Cornelia de Lange for support and inspiration.

**Contributors** Conception and design of the article by GF, VM, AS and GC. GF, VM, AG, SR, CS and MG performed experiments. GF, VM, AG and VB analysed data. GF and VM analysed data and assembled figures. GF, VM, CR, CC, AB, AS and GC

actively analysed and interpreted findings. GF, VM, AS and GC wrote the manuscript. All authors read, edited and approved the final manuscript.

**Funding** This work has been supported by Fondazione Cariplo (grant no. 2015-0783 to VM), by the Associazione Italiana per la Ricerca sul Cancro (AIRC IG2015 grant no. 17593 and IG2018 grant no. 21999 to GC; IG2017 grant no. 20564 to AB; fellowship 2018 no. 22620 to AG), by the fellowship program 'PhD talent 2017' to AG and by DIMET PhD program University of Milano-Bicocca to CS.

**Competing interests** None declared.

**Provenance and peer review** Not commissioned; externally peer reviewed.



## REFERENCES

- 1 Jessberger R. Cohesin's dual role in the DNA damage response: repair and checkpoint activation. *Embo J* 2009;28:2491–3.
- 2 Banerji R, Skibbens RV, Iovine MK. How many roads lead to cohesinopathies? *Dev Dyn* 2017;246:881–8.
- 3 Deardorff MA, Noon SE, Krantz ID, et al. Cornelia de Lange syndrome. In: Adam MP, Ardinger HH, Pagon RA, et al, eds. *GeneReviews*(®). Seattle (WA), 1993.
- 4 Losada A. Cohesin in cancer: chromosome segregation and beyond. *Nat Rev Cancer* 2014;14:389–93.
- 5 Kon A, Shih L-Y, Minamino M, et al. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat Genet* 2013;45:1232–7.
- 6 Thol F, Bollin R, Gehlhaar M, et al. Mutations in the cohesin complex in acute myeloid leukemia: clinical and prognostic implications. *Blood* 2014;123:914–20.
- 7 Shiba N, Yoshida K, Shiraishi Y, et al. Whole-exome sequencing reveals the spectrum of gene mutations and the clonal evolution patterns in paediatric acute myeloid leukaemia. *Br J Haematol* 2016;175:476–89.
- 8 Vial Y, Lachenaud J, Verloes A, et al. Down syndrome-like acute megakaryoblastic leukemia in a patient with Cornelia de Lange syndrome. *Haematologica* 2018;103.
- 9 Kline AD, Moss JF, Selicorni A, et al. Diagnosis and management of Cornelia de Lange syndrome: first international consensus statement. *Nat Rev Genet* 2018;19:649–66.
- 10 Walters DK, Mercher T, Gu T-L, et al. Activating alleles of Jak3 in acute megakaryoblastic leukemia. *Cancer Cell* 2006;10:65–75.
- 11 Deardorff MA, Kaur M, Yeager D, et al. Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of Cornelia de Lange syndrome with predominant mental retardation. *Am J Hum Genet* 2007;80:485–94.
- 12 Ripperger T, Bielack SS, Borkhardt A, et al. Childhood cancer predisposition syndromes-A Concise review and recommendations by the cancer predisposition Working Group of the Society for pediatric oncology and hematology. *Am J Med Genet A* 2017;173:1017–37.
- 13 Tawana K, Drazer MW, Churpek JE. Universal genetic testing for inherited susceptibility in children and adults with myelodysplastic syndrome and acute myeloid leukemia: are we there yet? *Leukemia* 2018;32:1482–92.
- 14 Cianci P, Fazio G, Casagrande S, et al. Acute myeloid leukemia in Baraitser-Winter cerebrofrontofacial syndrome. *Am J Med Genet A* 2017;173:546–9.
- 15 Santoro C, Apicella A, Casale F, et al. Unusual association of non-anaplastic Wilms tumor and Cornelia de Lange syndrome: case report. *BMC Cancer* 2016;16.
- 16 Yoshida K, Toki T, Okuno Y, et al. The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nat Genet* 2013;45:1293–9.
- 17 Enverald E, Du L, Visnes T, et al. A regulatory role for the cohesin loader NIPBL in nonhomologous end joining during immunoglobulin class switch recombination. *J Exp Med* 2013;210:2503–13.

## Chapter 8

A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia.

**Grioni A**, Fazio G, Rigamonti S, Bystry V, Daniele G, Dostalova Z, Quadri M, Saitta C, Silvestri D, Songia S, Storlazzi CT, Biondi A, Darzentas N, Cazzaniga G. A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia. *Hemasphere*. 2019Jun 4;3(3):e250. doi: 10.1097/HS9.0000000000000250. PMID: 31723839; PMCID:PMC6746019.

# A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia

Andrea Grioni<sup>1,2</sup>, Grazia Fazio<sup>1</sup>, Silvia Rigamonti<sup>1</sup>, Vojtech Bystry<sup>2</sup>, Giulia Daniele<sup>3</sup>, Zuzana Dostalova<sup>2</sup>, Manuel Quadri<sup>1</sup>, Claudia Saitta<sup>1,6</sup>, Daniela Silvestri<sup>7,8</sup>, Simona Songia<sup>1</sup>, Clelia T. Storlazzi<sup>3</sup>, Andrea Biondi<sup>1,5</sup>, Nikos Darzentas<sup>2,4</sup>, Giovanni Cazzaniga<sup>1</sup>

**Correspondence:** Giovanni Cazzaniga (e-mail: gianni.cazzaniga@hsgerardo.org).

## Abstract

Acute lymphoblastic leukemia (ALL) is the most frequent pediatric cancer. Fusion genes are hallmarks of ALL, and they are used as biomarkers for risk stratification as well as targets for precision medicine. Hence, clinical diagnostics pursues broad and comprehensive strategies for accurate discovery of fusion genes. Currently, the gold standard methodologies for fusion gene detection are fluorescence in situ hybridization and polymerase chain reaction; these, however, lack sensitivity for the identification of new fusion genes and breakpoints. In this study, we implemented a simple operating procedure (OP) for detecting fusion genes. The OP employs RNA CaptureSeq, a versatile and effortless next-generation sequencing assay, and an in-house as well as a purpose-built bioinformatics pipeline for the subsequent data analysis. The OP was evaluated on a cohort of 89 B-cell precursor ALL (BCP-ALL) pediatric samples annotated as negative for fusion genes by the standard techniques. The OP confirmed 51 samples as negative for fusion genes, and, more importantly, it identified known (*KMT2A* rearrangements) as well as new fusion events (*JAK2* rearrangements) in the remaining 38 investigated samples, of which 16 fusion genes had prognostic significance. Herein, we describe the OP and its deployment into routine ALL diagnostics, which will allow substantial improvements in both patient risk stratification and precision medicine.

## Introduction

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer.<sup>1</sup> The 5-year survival rate exceeds 85% in children, but the survival following relapse is poor.<sup>2</sup> Analysis of paired diagnosis/relapse ALL samples shows clonal diversity that

arises from the accumulation of new deletions and mutations over time. Despite that, the founding fusion genes are usually conserved from diagnosis to relapse, indicating that the predominant clones observed at diagnosis and relapse are clones derived from a common ‘preleukemic’ clone.<sup>3</sup> Fusion genes arise from chromosomal translocations and intrachromosomal

*This work was partially supported by the program ‘Passaporto genetico’ of the Parents Committee ‘Comitato Maria Letizia Verga ONLUS’, as well as by the IG17593 and IG20564 Grants of the Associazione Italiana per la Ricerca sul Cancro (AIRC) to GC and AB, respectively; the IG2014 grant no. 15413 to CTS; the fellowship program Brno PhD talent 2017 as well as fellowship AIRC 2018 no. 22620 to AG; by the DIMET PhD program - University of Milano-Bicocca to GS. Bioinformatics of CEITEC Masaryk University is gratefully acknowledged for the obtaining of the scientific data presented in this paper. AG, ND and VB contributed new analytic tools, and participated in the data analysis and writing of the paper. AB, AG, GF and GC participated in research design and writing of the paper. GF, SR, GD and CTS participated in the performance of the research. CS and MQ performed PCR validations. DS provided samples information. ZD participated in the writing and correction of the paper.*

*The authors declare no conflicts of interest.*

<sup>1</sup>Centro Ricerca Tettamanti, Clinica Pediatrica, Università degli Studi di Milano-Bicocca, Fondazione MBBM/Ospedale S. Gerardo, Monza, Italy

<sup>2</sup>Central European Institute of Technology, Masaryk University, Brno, Czech Republic

<sup>3</sup>Department of Biology, University of Bari “Aldo Moro”, Bari, Italy

<sup>4</sup>Department of Hematology, University Hospital Schleswig-Holstein, Kiel, Germany

<sup>5</sup>Clinica Pediatrica, Università degli Studi di Milano-Bicocca, Fondazione MBBM/Ospedale S. Gerardo, Monza, Italy

<sup>6</sup>Cancer Center, Humanitas Research Hospital, Humanitas University, Rozzano, Milan, Italy

<sup>7</sup>Center of Biostatistics for Clinical Epidemiology, Department of Health Science, University of Milano-Bicocca, Milan, Italy

<sup>8</sup>Pediatric Hematology-Oncology Unit, Department of Pediatrics, University of Milano-Bicocca, MBBM Foundation/ASST Monza, Monza, Italy.

Copyright © 2019 the Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the European Hematology Association. This is an open access article distributed under the Creative Commons Attribution-ShareAlike License 4.0, which allows others to remix, tweak, and build upon the work, even for commercial purposes, as long as the author is credited and the new creations are licensed under the identical terms.

HemaSphere (2019) 3:3(e250)

Received: 27 December 2018 / Received in final form: 4 April 2019 / Accepted: 4 April 2019

**Citation:** Grioni A, Fazio G, Rigamonti S, Bystry V, Daniele G, Dostalova Z, Quadri M, Saitta C, Silvestri D, Songia S, Storlazzi CT, Biondi A, Darzentas N, Cazzaniga G. A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia. *HemaSphere*, 2019;3:3. <http://dx.doi.org/10.1097/HS9.0000000000000250>

rearrangements that mainly disrupt genetic regulators of normal hematopoiesis as well as lymphoid development (e.g., those involving *RUNX1* and *ETV6*) and constitutively activate tyrosine kinases<sup>4</sup> (e.g., *ABL1* chimeras). Thus, fusion genes are hallmarks of ALL that play a pivotal role in leukemogenesis, and their identification is crucial for patient risk stratification.<sup>5</sup>

Common fusion genes in B-lineage ALL are: t(12;21)(p13;q22), encoding *ETV6-RUNX1* (TEL-AML); t(1;19)(q23;p13), encoding *TCF3-PBX1* (E2A-PBX1)<sup>6</sup>; t(9;22)(q34;q11.2), resulting in formation of the “Philadelphia” chromosome, encoding *BCR-ABL1*; rearrangements of *KMT2A* (*MLL*) at 11q23 to a range of fusion partners<sup>7</sup>; and rearrangements of the cytokine receptor gene *CRLF2* at the pseudo autosomal region 1 (PAR1) at Xp22.3/Yp11.3.<sup>8,9</sup> Fusion genes correlate with the clinical outcome, and they are used as biomarkers for patient risk stratification<sup>10</sup>: for example, patients positive for t(12;21)/*ETV6-RUNX1* have the most favorable prognosis, whereas t(9;22)/*BCR-ABL1*, t(1;19)/*TCF3-PBX1*, and *KMT2A-AFF1* correlate with a brief disease latency and have a poor prognosis.<sup>10,11</sup> Moreover, specific drug inhibitors antagonizing the fusion proteins provide a more efficient and less toxic tool for disease eradication (precision medicine): for example, the imatinib tyrosine kinase inhibitor inhibits the oncogenic deregulation caused by the (9;22)/*BCR-ABL1* fusion protein.<sup>12</sup>

Before the next generation sequencing (NGS) era, elaborate and extensive cytogenetic studies lead to the description of few recurrent and highly expressed fusion genes,<sup>13</sup> such as *BCR-ABL1* and *ETV6-RUNX1*. The characterization of their breakpoint coordinates enabled the design of diagnostic screening by both quantitative multiplex polymerase chain reaction (qPCR) and fluorescence in situ hybridization (FISH).<sup>14</sup> The recent introduction of NGS allowed a fast and accurate screening of the patient’s genome at the nucleotide level, which led to the discovery of a broad array of previously unknown fusion genes.<sup>15</sup> This reflects the increased capability of NGS to recognize subtle chromosomal rearrangements. On the contrary, FISH may only detect exchanges of considerably larger chromosome segments, without nucleotide precision, while qPCR screenings can identify already known fusion gene breakpoints only.<sup>16</sup>

Whole transcriptome sequencing (RNAseq), together with open-source bioinformatics tools, has already been applied to identifying fusion genes.<sup>17</sup> Whole RNAseq performs well in the detection and quantification of highly and medium abundant transcripts, but it may fail in cases of low abundance transcripts.<sup>18</sup> The RNA capture sequencing (RNA CaptureSeq) is a probe-based assay for capturing, amplifying, and sequencing genomic regions of interest only (targets). The RNA CaptureSeq generates libraries of small fragments (250–300 bp) in a short time (2.5 days) compared to whole RNAseq, and it is compatible with the well-known MiSeq and NextSeq Illumina NGS platforms. RNA CaptureSeq is sensitive to low abundance transcript variants of targeted genes<sup>19</sup>; however, the detection of fusion transcripts may be compromised when the fusion partner gene is not part of the capture procedure (unknown partner). This scenario reduces discoverability of fusion transcripts to only those fragments that span the target gene breakpoint.

We have developed and herein present a simple, efficient, and ready-to-use operating procedure (OP) for the clinical identification of fusion genes in B-cell ALL. The OP is based on RNA CaptureSeq, and it is supported by an in-house bioinformatics pipeline that is purpose-built to detect and extend fragments spanning the fusion gene breakpoint. We applied the OP to a cohort of 89 B-cell ALL pediatric patients enrolled in the AIEOP-BFM ALL clinical protocol<sup>20</sup> that were annotated as negative to fusion genes by the standard screening methods. This paper

summarizes the results of the OP applied to clinical diagnostics and discusses its implications for patient risk stratification.

## Results

### Comparison of available bioinformatics pipelines

We developed a bioinformatic method for fusion gene assessment from RNA CaptureSeq datasets and evaluated it on a training dataset composed of 23 samples evaluated as positive to 6 different fusion genes, namely t(9;22)/*BCR-ABL1*, t(12;21)/*ETV6-RUNX1*, t(4;11)/*KMT2A-AFF1*, del(X)/P2RY8-CRLF2, t(1;19)/*TCF3-PBX1*, and t(9;11)/*KMT2A-MLLT3*, by standard methods. Our method distinguished all 6 sample-specific fusion genes within the dataset. In addition, we analyzed the same training dataset through Illumina BaseSpace, STAR-Fusion,<sup>21</sup> and the customized pipeline described by Jennifer L. Winters et al.<sup>22</sup> The STAR-Fusion tool did not detect 1 out of 6 fusion genes (del(X)/P2RY8-CRLF2), while the Illumina BaseSpace did not detect 2 out of 6 fusion genes (t(9;11)/*KMT2A-MLLT3* and t(4;11)/*KMT2A-AFF1*). The method described by Jennifer L. Winters et al. did not detect 3 out of 6 fusion genes (t(1;19)/*TCF3-PBX1*, t(9;11)/*KMT2A-MLLT3*, and del(X)/P2RY8-CRLF2) (Table 1).

The ability of our procedure to detect all fusion transcripts derives from the fine-tuning of the bioinformatics pipeline to cover the specific RNA target–capture scenario, where both genes involved in the fusion are not always captured (see Material and Methods and Fig. 1). For these reasons, we applied only our method in the subsequent analyses.

### Evaluation of the OP in clinical diagnosis

RNA material obtained from patient bone marrow mononuclear cells at the onset or relapse of the disease was sequenced using the RNA PanCancer (Illumina, San Diego, CA). Raw FASTQ files underwent quality control and were afterwards analyzed through our system. A detailed description of the OP strategy is available in the Materials and Methods section. The time required for the procedure from library preparation to obtaining results was 2.5 days.

We screened a cohort of 89 samples of B-cell ALL leukemia (test set) for positivity to fusion genes. All samples were negative for the fusion genes t(12;21)/*ETV6-RUNX1*, t(9;22)/*BCR-ABL1*, t(4;11)/*KMT2A-AFF1*, and t(1;19)/*TCF3-PBX1* by the standard screening methods. The test set was divided into 3 groups: frontline high-risk (HR), relapse (RL), and patients with a high value of minimal residual disease (MRD) at day 33 of chemotherapy induction (TP1+). Overall, the OP identified 26 different fusion genes in 38 out of the 89 investigated samples, with the transcripts of 16 of them being of prognostic value (Table 2 and Suppl. Table 1, Supplemental Digital Content, <http://links.lww.com/HS/A34>). New fusion genes in B-cell ALL and not recorded in public databases were validated through reverse transcription PCR (RT-PCR) or FISH to discern between false and true positives (Supplementary Table 2, Supplemental Digital Content, <http://links.lww.com/HS/A34>).

### OP applied to the frontline HR group

Seven out of 16 samples (43%) resulted as positive for fusion genes (Fig. 2a). Four samples carried fusion genes recurrently associated to B-cell ALL: t(5;5)/*EBF1-PDGFRB* (n=2), t(9;9)/*PAX5-JAK2* (n=1), and t(12;19)/*ZNF384-TCF3* (n=1) and 3 samples were positive for t(19;19)/*TCF3-OAZ1* (n=1), t(7;7)/*IKZF1-DDC* (n=1), t(2;9)/*ZEB2-JAK2* (n=1), and t(9;17)/*MPRIIP-JAK2* (n=1)

**Table 1****Comparison of available bioinformatics pipelines.**

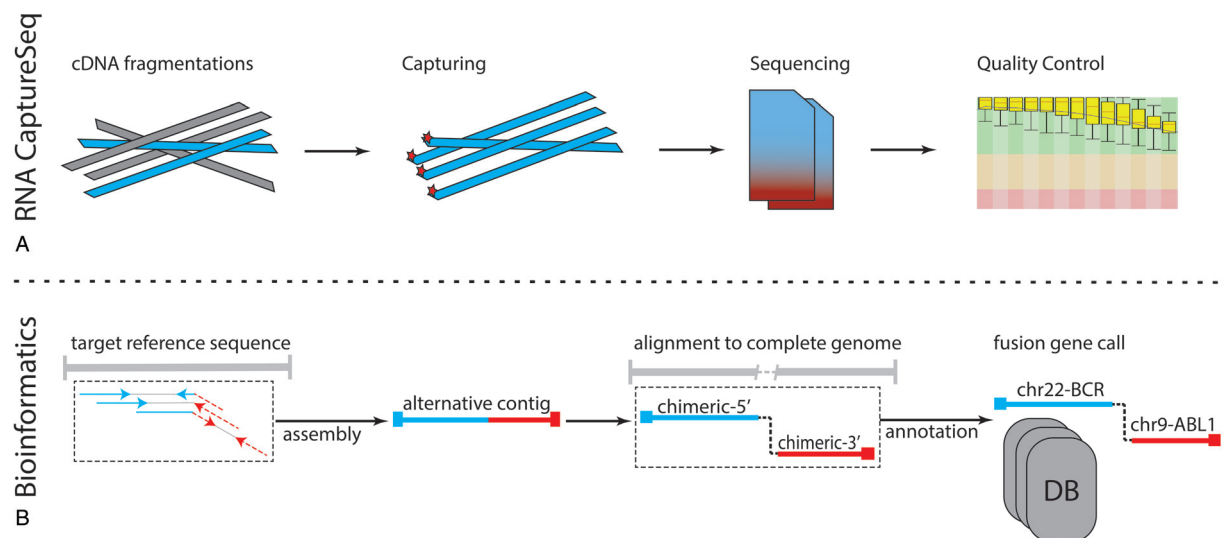
Sample	Blast%	Metadata			Bioinformatics Pipeline				
		Fusion gene	Raw-reads	FASTQC	Probes	Internal	BaseSpace	TopHat	Star-Fusion
KN1	90	t(9;22) BCR-ABL1	3.22E+06	+	t/p	+	+	+	+
KN2	90	t(12;21) ETV6-RUNX1	5.19E+06	+	t/p	+	+	ND	+
KN3	92	t(4;11) KMT2A-AFF1	5.60E+06	+	t/p	+	+	ND	+
KN4	90	t(9;22) BCR-ABL1	4.76E+06	+	t/p	+	+	ND	+
KN5	93	t(9;22) BCR-ABL1	6.06E+06	+	t/p	+	+	ND	+
KN5	93	t(12;21) ETV6-RUNX1	6.06E+06	+	t/p	+	+	ND	+
KN6	98	del(X) P2RY8-CRLF2	3.81E+06	+	t/p	+	+	ND	ND
KN7	NA	t(9;22) BCR-ABL1	2.41E+06	+	t/p	+	+	ND	+
KN8	NA	t(4;11) KMT2A-AFF1	2.57E+06	+	t/p	+	+	ND	+
KN9	91	t(1;19) TCF3-PBX	2.46E+06	+	t/p	+	+	ND	+
KN10	64	t(12;21) ETV6-RUNX1	2.30E+06	+	t/p	+	+	ND	+
KN11	NA	t(9;11) KMT2A-MLLT3	2.50E+06	+	t/p	+	+	ND	+
KN12	NA	t(9;22) BCR-ABL1	1.62E+06	+	t/p	+	+	+	+
KN13	NA	t(4;11) KMT2A-AFF1	2.40E+06	+	t/p	+	+	+	+
KN14	91	t(1;19) TCF3-PBX	6.53E+05	+	t/p	+	+	ND	+
KN15	64	t(12;21) ETV6-RUNX1	2.47E+06	+	t/p	+	+	ND	+
KN16	NA	t(9;11) KMT2A-MLLT3	6.21E+06	+	t/p	+	ND	ND	+
KN17	NA	t(9;22) BCR-ABL1	5.29E+06	+	t/p	+	+	ND	+
KN18	93	t(4;11) KMT2A-AFF1	3.17E+06	+	t/p	+	ND	ND	+
KN19	90	t(4;11) KMT2A-AFF1	6.56E+06	+	t/p	+	+	+	+
KN20	93	t(1;19) TCF3-PBX	6.73E+06	+	t/p	+	+	ND	+
KN21	94	t(4;11) KMT2A-AFF1	4.50E+06	+	t/p	+	+	ND	+
KN22	70	t(12;21) ETV6-RUNX1	4.66E+06	+	t/p	+	+	+	+
KN23	97	t(9;22) BCR-ABL1	5.44E+06	+	t/p	+	+	+	+

fusion genes. All fusion transcripts were confirmed by RT-PCR, while the novel fusion genes t(2;9)/ZEB2-JAK2 (n = 1) and t(9;17)/MPRIIP-JAK2 were validated through FISH (Suppl. Fig. 1, Supplemental Digital Content, <http://links.lww.com/HS/A34>).

### OP applied to the TP1+ group

The OP identified fusion genes in 19 out of 49 samples (38.8%) (Fig. 2b). Nine samples were evaluated as positive for fusion

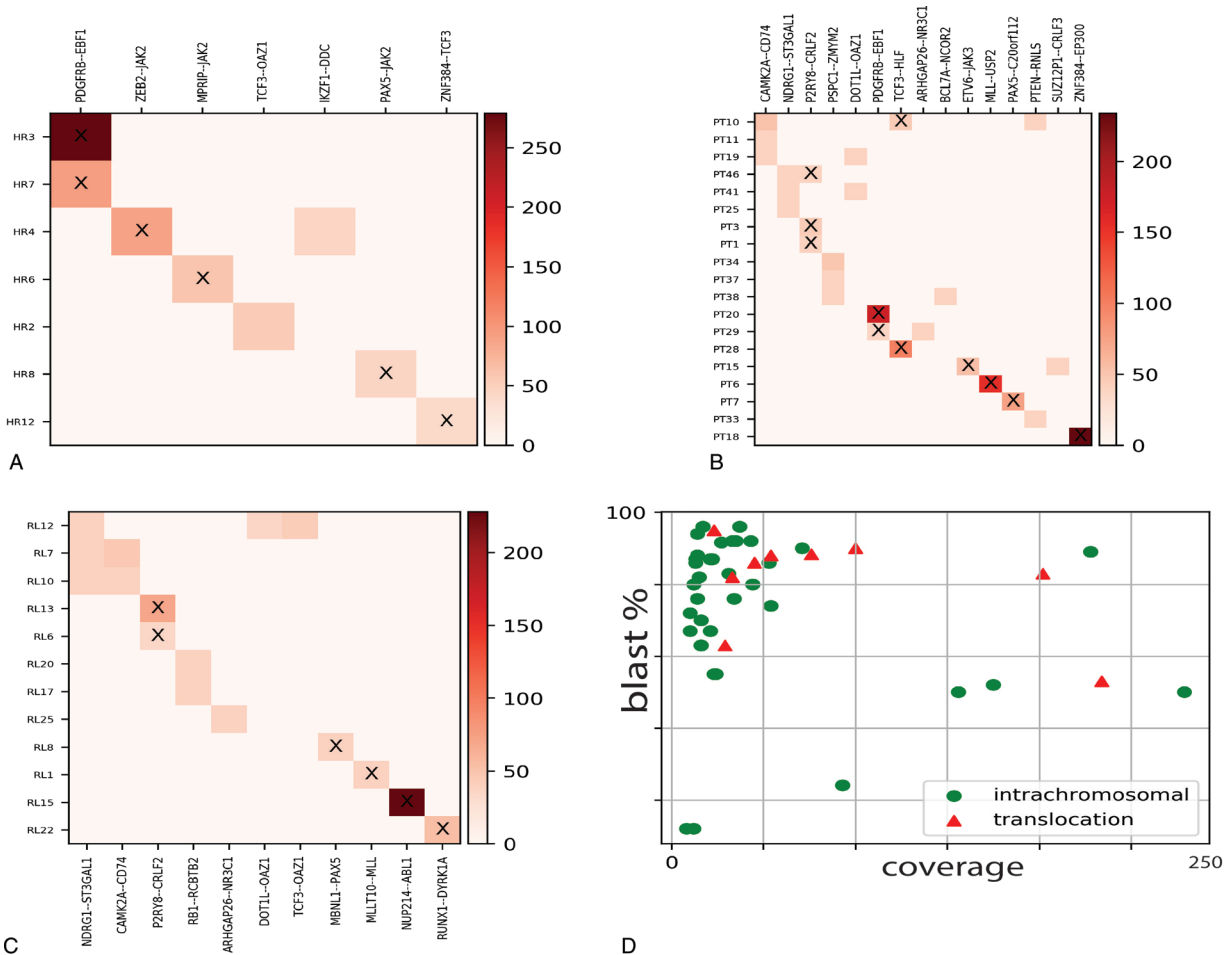
genes that are frequent in B-cell ALL: t(17;19)/TCF3-HLF (n = 2), del(X)/P2RY8-CRLF2 (n = 3), t(5;5)/EBF1-PDGFRB (n = 2), t(12;19)/ETV6-JAK3 (n = 1), t(12;22)/ZNF384-EP300 (n = 1). We also identified a novel inter-chromosomal rearrangement, t(9;20)/PAX5-C20orf112 (n = 1), and a variety of intra-chromosomal fusion genes (n = 9) that were already annotated in public databases, and we validated them by RT-PCR (Suppl. Table 1, Supplemental Digital Content, <http://links.lww.com/HS/A34>).



**FIGURE 1.** The standard operating procedure: (A) RNA CaptureSeq protocol allows the isolation of specific genomic regions (targets) through complementary probes; then, the captured fragments are sequenced, and the FASTQ file quality is evaluated. (B) The bioinformatics pipeline includes four sequential steps, which allows the identification of fusion genes through the identification of putative break-points on the genomic sequences of targeted genes.

**Table 2**  
**RNAseq Fusion transcripts identified by our OP.**

tz	Fusion gene	Probes	Progn.	PCR	FusionHub
6	t(8;8) NDRG1-ST3GAL1	t	-	+	['CHIMERSEQ', 'Tumor_Fusion_GDP', 'HPA', 'Banned_dataset', 'Known_Fusions']
5	t(5;5) CAMK2A-CD74	t/p	-	+	['Known_Fusions']
5	del(X) P2RY8-CRLF2	t/p	+	+	['CHIMERPUB', 'FARE-CAFE', 'TICDB']
4	t(5;5) PDGFRB-EBF1	t/p	+	+	['CHIMERSEQ', 'CHITARS', 'Known_Fusions']
3	t(13;13) PSPC1-ZMYM2	p	-	+	['Banned_Dataset', 'GTex']
3	t(19;19) DOT1L-OAZ1	t	-	+	['HPA', 'Banned_Dataset']
2	t(10;10) PTEN-RNLS	t	-	+	['Tumor_Fusion_GDP']
2	5(13;13) RB1-RCBTB2	t	-	+	['GTex']
2	t(17;19) TCF3-HLF	t/p	+	+	['CHIMERKB', 'CHIMERPUB', 'FARE-CAFE', 'TICDB']
2	t(19;19) TCF3-OAZ1	t	-	+	NOVEL
2	t(5;5) ARHGAP26-NR3C1	t/p	-	+	['HPA', 'Banned_Dataset', 'GTex']
1	t(10;11) MLLT10-KMT2A	t/p	+	+	['CHIMERKB', 'CHIMERPUB']
1	5(11;11) KMT2A-USP2	t/p	+	+	['Known_Fusions']
1	t(12;12) BCL7A-NCOR2	t/p	-	+	['Known_Fusions']
1	t(12;19) ETV6-JAK3	t/p	+	+	NOVEL
1	t(12;19) ZNF384-TCF3	t/p	+	+	['CHIMERSEQ', 'CHITARS', 'FARE-CAFE', 'TICDB', 'Known_Fusions']
1	t(12;22) ZNF384-EP300	t/p	+	+	['CHIMERPUB']
1	t(17;17) SUZ12P1-CRLF3	t	-	+	['18_Cancers']
1	t(9;17) MPRIP-JAK2	p	+	+	NOVEL
1	t(21;21) RUNX1-DYRK1A	t	+	+	['GTex']
1	t(2;9) ZEB2-JAK2	p	+	+	NOVEL
1	t(3;9) MBNL1-PAX5	t/p	+	+	['Known_Fusions']
1	t(7;7) IKZF1-DDC	t	-	+	NOVEL
1	t(9;20) PAX5-C20orf112	t	+	+	['CHIMERSEQ', 'CHITARS', 'FARE-CAFE', 'TICDB']
1	t(9;9) NUP214-ABL1	t/p	+	+	['COSMIC', 'CHIMERAKB', 'CHIMERPUB', 'CHIMERSEQ', 'FARE-CAFE', 'TICDB', 'TUMOR_Fusion_GDP', 'Oesophagus_Dataset']
1	t(9;9) PAX5-JAK2	t/p	+	+	['COSMIC', 'CHIMERKB', 'FARE-CAFE', 'TICDB']



**FIGURE 2.** (A), (B), and (C) Heatmaps of detected fusion genes among different risk groups. The axes correspond to the detected fusion genes (X) and sample names (Y). The color code represents the coverage on the fusion gene breakpoint as reported by the scale on the right. The 'X' tag highlights fusion genes of prognostics relevance. (D) Fusion genes distribution in terms of intrachromosomal (green dots) or interchromosomal translocations (red triangles) in relations to the breakpoint read coverage and percentage of blast cells.

## OP applied to the RL group

The OP identified fusion genes in 12 out of 24 samples of the RL group (~50%) (Fig. 2c): t(9;9)/NUP214-ABL1 (n=1), del(X)/P2RY8-CRLF2 (n=2), t(10;11)/MLLT10-KMT2A (n=1), t(21;21)/RUNX1-DYRK1A (n=1), and t(3;9)/PAX5-MBLN1 (n=1) fusion genes were associated with ALL and of clinical relevance for the patients and were hence immediately validated by RT-PCR. On the other hand, the OP identified additional fusion genes derived from intra-chromosomal rearrangements, such as t(8;8)/NDRG1-ST3GAL1 (n=3), t(13;13)/RB1-RCBTB2 (n=2), t(19;19)/DOT1L-OAZ1 (n=1), t(19;19)/TCF3-OAZ1 (n=1), t(5;5)/ARHGAP26-NR3C1 (n=1), and t(5;5)/CAMK2A-CD74 (n=2), which were already annotated in public databases.

## Enrichment of intra-chromosomal fusion genes

The OP identified 26 fusion genes in 38 investigated patients (HR, RL, and TP1+ groups). Among them, 17 (65%) fusion genes derived from intra-chromosomal rearrangements and were

supported by a low read coverage (~20× to ~50×) in coexistence with high levels of blast cells in the BM (~70% to ~96%) (Fig. 2d). We did not observe a correlation between intra-chromosomal fusion genes associated with recurrent chromosomal translocations in B-cell ALL (Table 3). RT-PCR confirmed frequent B-cell ALL intra-chromosomal fusion genes, such as PDGFRB-EBF1, NUP214-ABL1, and PAX5-JAK2 (Suppl. Table 2, Supplemental Digital Content, <http://links.lww.com/HS/A34>). P2RY8-CRLF2 fusions were not confirmed by RT-PCR since those samples correlated with del(X)(p22p22) detected by multiplex ligation-dependent probe amplification and highly expressed CRLF2 detected by gene expression profile (data not presented). We further investigated gene expression levels in healthy whole-blood samples for genes involved in intra-chromosomal fusions as well as those not known in B-cell ALL (n=21, gene set) through the GTEx portal.<sup>2,3</sup> Sixteen genes had transcript per million (TPM) expression levels from medium to high (TPM greater than 5.4), while 5 of them had low levels (TPM between 1 and 5.4) (Fig. 3). Also, some intra-chromosome fusion transcripts involved genes spatially close, within a range of

**Table 3**  
Sample-specific fusion transcripts.

Sample	Fusion gene	Chromosome	% Leukemic cell in BM	Sex	Karyotype
HR2	TCF3-OAZ1	t(19;19)	98	F	
HR3	PDGFRB-EBF1	t(5;5)	60	M	
HR4	ZEB2-JAK2 IKZF1-DDC	t(2;9) t(7;7)	NA	M	
HR6	MPRIIP-JAK2	t(9;17)	NA	M	
HR7	PDGFRB-EBF1	t(5;5)	53	M	46,XY,der(1)inv(1)(q21q31)dup(1)(q31q32)[8]/46,XY[14]
HR8	PAX5-JAK2	t(9;9)	NA	M	
HR12	ZNF384-TCF3	t(12;19)	90	F	
PT1	P2RY8-CRLF2	del(X)	91	M	46,XY, der(9)T(9;?)p13;?, -13, add(13)(q34), +21 [10]/47,XY,+21[4]
PT3	P2RY8-CRLF2	del(X)	95	M	
PT6	KMT2A-USP2	t(11;11)	NA	M	
PT7	PAX5-C20orf112	t(9;20)	NA	M	
PT10	TCF3-HLF CAMK2A-CD74 PTEN-RNLS	t(17;19) t(5;5) t(10;10)	NA	F	
PT11	CAMK2A-CD74	t(5;5)	90	M	
PT15	ETV6-JAK3 SUZ12P1-CRLF3	t(12;19) t(17;17)	90	F	
PT18	ZNF384-EP300	chr12-chr22	85	M	
PT19	CAMK2A-CD74 DOT1L-OAZ1	t(5;5) t(19;19)	NA	M	
PT20	PDGFRB-EBF1	t(5;5)	80	F	
PT25	NDRG1-ST3GAL1	t(8;8)	NA	F	
PT28	TCF3-HLF	t(17;19)	95	F	
PT29	PDGFRB-EBF1 ARHGAP26-NR3C1	t(5;5) t(5;5)	98	F	
PT33	PTEN-RNLS	t(10;10)	NA	M	
PT34	PSPC1-ZMYM2	t(13;13)	NA	F	
PT37	PSPC1-ZMYM2	t(13;13)	80	M	
PT38	BCL7A-NCOR2 PSPC1-ZMYM2	t(12;12) t(13;13)	NA	F	
PT41	DOT1L-OAZ1 NDRG1-ST3GAL1	t(19;19) t(8;8)	NA	M	
PT46	P2RY8-CRLF2 NDRG1-ST3GAL1	del(X) t(8;8)	NA	F	
RL1	MLLT10-KMT2A	t(10;11)	90	M	
RL6	P2RY8-CRLF2	del(X)	76	M	
RL7	CAMK2A-CD74 NDRG1-ST3GAL1	t(5;5) t(8;8)	70	M	
RL8	MBNL1-PAX5	t(3;9)	NA	M	
RL10	CAMK2A-CD74 NDRG1-ST3GAL1	t(5;5) t(8;8)	NA	M	
RL12	NDRG1-ST3GAL1 TCF3-OAZ1 DOT1L-OAZ1	t(8;8) t(19;19) t(19;19)	97	M	
RL13	P2RY8-CRLF2	del(X)	98	F	47,XX,+21c[14]
RL15	NUP214-ABL1	t(9;9)	92	F	
RL17	RB1-RCBTB2	t(13;13)	40	M	
RL20	RB1-RCBTB2	t(13;13)	NA	M	
RL22	RUNX1-DYRK1A	t(21;21) 117	NA	M	
RL25	ARHGAP26-NR3C1	t(5;5)	99	F	



**FIGURE 3.** Gene expression profile of genes involved in intra-chromosomal fusion genes but not associated to ALL.

150 to 250kb, and annotated as conjoined genes. Indeed, we validated those fusion gene events by RT-PCR and confirmed their nucleotide sequences by Sanger sequencing (Suppl. Table 2, Supplemental Digital Content, <http://links.lww.com/HS/A34>).

## Discussion

Fusion genes are hallmarks of ALL both in pediatric and adult patients; their identification is crucial to design a risk-reducing-driven chemotherapy treatment (precision medicine). Precision medicine allows either very low-risk patients to proceed with standard therapy or very high-risk patients to be candidates for experimental and/or targeted therapies. For this purpose, sensitive, specific, and comprehensive screening of selected genomic regions prone to chromosomal breaks are needed in routine diagnostics to identify the increasing variety of fusion genes.

We built a versatile and straightforward OP to recognize fusion genes at nucleotide resolution without any a priori knowledge, which overcomes the limitations of qPCR and FISH. The OP employs an RNA CaptureSeq panel that allows targeted transcriptome sequencing through a simple library preparation protocol. For the subsequent data analysis, we fine-tuned a bioinformatics pipeline that deploys robust and stable tools, which can be easily set up on any operative system through the Anaconda Platform. Our bioinformatics pipeline recognized all fusion genes harbored by samples within the training dataset, while the Star-Fusion, Illumina BaseSpace, and the strategy proposed by Winter et al reached 83%, 66%, and 50% success in fusion transcripts identification, respectively. Prognostically significant and frequent B-cell precursor ALL fusion genes such

as *KMT2A* rearrangements and P2RY8-CRLF2 were not fully detected by the external tools. Patients harboring *KMT2A* rearrangements have a particularly unfavorable prognosis.<sup>10,24,25</sup> *KMT2A* is prone to breaks in various genomic location with several partners, thus making the detection of its resulting fusion genes challenging. On the other hand, the repetitive nature of the chromosome X may compromise read alignment and the identification of the P2RY8-CRLF2 fusion gene. Our results indicated that our purpose-built, disease- and NGS-strategy specific bioinformatics pipeline is required for covering many possible scenarios causing fusion genes. The evaluation of the OP through the analysis of 89 pediatric B-cell precursor ALL samples identified 26 different fusion genes among 38 samples that were undetectable by the standard routine diagnostics. Sixteen of those fusion transcripts have prognostic value since they involved rearrangements in genes driving leukemogenesis (*KMT2A*, *JAK2*, and *PAX5*). Moreover, the newly identified fusion genes t(2;9)/ZEB2-JAK2 and t(9;17)/MPRIP-JAK2, which are possibly targetable by JAK/STAT inhibitors, highlight the potential of our OP for precision medicine and biomarker discovery. Additionally, we detected a case of NUP214/ABL1 fusion genes in B-cell ALL, which only 2 cases were previously reported.<sup>26</sup> We confirmed the increased capability provided by RNA CaptureSeq to detect small local structural variants through the identification of a variety of intra-chromosomal fusion genes (n = 17). Multiple intra-chromosomal fusion genes were the only detected in the sample within our set of genes (n = 1385); hence, it is not possible to state any functional correlation between those rearrangements and the recurrent fusion genes (such as BCR-ABL1, ETV6-RUNX1, and *KMT2A* rearrangements). Some intra-chromosomal fusion transcripts, namely PSPC1-ZMYM2, DOT1L-OAZ1, RB1-RCBTB2, ARHGAP26-NR3C1, were also observed in NGS studies<sup>27,28,29</sup> of healthy populations (e.g., GTEx, Banned\_dataset, and HPA), or annotated as conjoined genes.<sup>30,31</sup> We also detected intra-chromosomal fusion transcripts involving recurrent leukemogenic genes (IKZF1-DDC, P2RY8-CRLF2, *KMT2A*-UPS2, MLLT10-*KMT2A*) that are prone to deletions and with a prognostic value (such as IKZF1,<sup>32</sup> and *KMT2A*<sup>33</sup>). Despite RNA CaptureSeq cannot discerns between inter- and intra- chromosome fusion genes when the same chromosomes are involved, these previous studies suggested an intra-chromosome origin.

In conclusion, herein we have described an NGS-based approach suitable for the detection of fusion genes, regardless of their expression levels, that may be incorporated into routine ALL diagnostics, with the advantage of a substantial improvement of precision medicine. Despite the OP lacks ISO certification, our finding highlights its potential and the need to develop bioinformatics tools addressing fusion genes detections from the RNA CaptureSeq scenario with precision. For this purpose, our OP may offer an idea for their implementation. Nonetheless, further studies are required to understand the biological significance and the potential therapeutic implication of the additional discoveries allowed by this tool.

## Materials and methods

### Patient cohort

A cohort of 89 B-cell precursor (BCP) ALL patients enrolled in the AIEOP-BFM ALL2009 protocol in Italy was sequenced by Illumina RNA CaptureSeq PanCancer to discern prognostic fusion genes. The cohort was composed of: 16 patients from the



frontline HR group, with a level of MRD above  $5 \times 10^{-4}$  at day +78 (TP2), who were shown as fusion gene-negative during the screening; 49 patients TP1+, that is, with a high level of PCR-MRD ( $>5 \times 10^{-4}$  compared to diagnostic value) at day +33 from the start of the induction therapy; and 24 patients from the RL (defined as having at least  $5 \times 10^{-2}$  blast cells after complete remission, CR). See Suppl. Table 3 (Supplemental Digital Content, <http://links.lww.com/HSA34>).

## Training dataset

A subgroup of 23 pediatric ALL patients enrolled in the AIEOP-BFM ALL2009 protocol, who were positive for fusion genes by standard clinical diagnosis, were selected. We used this subgroup as a training dataset for the development and evaluation of our bioinformatics pipeline of analysis for the assessment of fusion genes.

## FISH analysis for validating the identified fusion genes

The experiments were performed on BM metaphases from archival methanol:acetic acid-fixed chromosome suspensions, as previously described.<sup>17</sup> Bacterial Artificial Chromosome (BAC) clones were opportunely selected according to the NGS data from the University of California Santa Cruz (UCSC) database (release of December 2013, GRCh38/hg38) and previously tested on normal human metaphases. Briefly, chromosome preparations from BM cells were hybridized in situ with 1  $\mu$ g of each BAC probe labeled by nick translation. Hybridization was performed at 37°C in  $2 \times$  saline-sodium citrate (SSC), 50% (vol/vol) formamide, 10% (w/vol) dextran sulfate, 5  $\mu$ g Cot-1 DNA (Bethesda Research Laboratories, Gaithersburg, MD, USA), and 3  $\mu$ g sonicated salmon sperm DNA in a volume of 10  $\mu$ L. Post-hybridization washings were performed at 60°C in  $0.1 \times$  SSC (3 times). In co-hybridization experiments, the probes were directly labeled with fluorescein, Cy3, and Cy5 or indirectly with biotin-dUTP and subsequently detected by 7-(diethylamino)coumarin-3-carboxylic acid *N*-succinimidyl ester-conjugated streptavidin. Chromosomes were identified by DAPI staining. Digital images were obtained using a Leica DMRXA epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments, Boston, MA). All fluorescence signals that were detected using specific filters were recorded separately as gray-scale images. Pseudo-coloring and merging of images were performed with Adobe Photoshop software.

## Enrichment analysis

Ensembl gene IDs were extracted through the BioMart API (<https://www.ensembl.org/biomart>). Gene expression profile data from non-diseased samples were obtained from the GTEx portal through submission of the corresponding ENSEMBL gene ID (<https://gtexportal.org/home/>).

## External tools for fusion gene assessment

The Illumina BaseSpace pipeline for the identification of fusion genes first aligns filtered FASTQ files to the reference human genome through the TopHat<sup>34</sup> (v. 2.1.0) or STAR<sup>35</sup> aligner (v. 2.5.0a). Then, the STAR aligner supports Manta-fusion and the TopHat aligner supports the TopHat-fusion<sup>36</sup> to identify

candidate fusion genes. For the purpose of our analysis, we required the Illumina BaseSpace to recognize the sample-specific fusion gene by at least one application. The STAR-Fusion tool, v. 1.5.0, was utilized with standard parameters on the GRCh38.p12 genome reference and the corresponding Gencode<sup>37</sup> annotation set. We simulated the customized pipeline described by Jennifer L. Winters et al by deploying TopHat v. 2.1.1, which included TopHat-Fusion, and running the TopHat-Fusion pipeline with the Bowtie1<sup>38</sup> flag activated.

## Operating procedure

The OP consists of a laboratory and a bioinformatics module that has been built to both maximize the efficiency and minimize the time of ALL clinical diagnostics. Each element of the laboratory module is fully customizable and commercially available, whereas each tool deployed for the bioinformatics module is freely available through the Anaconda Platform (<https://www.anaconda.com/>).

### Laboratory module

RNA extraction protocol. Total RNA was extracted during diagnosis from bone marrow mononuclear cells by the guanidinium thiocyanate-phenol-chloroform method. Guanidine methods were used for total RNA preparation, as described by Sacchi et al.<sup>39</sup>

RNA CaptureSeq and sample sequencing. The RNA CaptureSeq 'TruSight RNA PanCancer' (Illumina), which includes 57,010 probes complementary to 21,043 coding regions for a total of 1385 cancer-related RNA transcripts, was applied (Fig. 1a). The protocol required 2.5 days, from library preparation to NGS sequencing. The sample libraries were prepared per the manufacturer's protocol using 10 ng of total RNA. Batches of 8 samples per run were sequenced through cartridge V3 on the Illumina MiSeq platform in a 75 bp paired-end setting for a total of 25 million paired-end reads (PE reads). The cost per sample was about 250 USD. A detailed list of targeted regions can be obtained from Illumina ([https://support.illumina.com/sequencing/sequencing\\_kits/trusight-rna-pan-cancer-panel/downloads.html](https://support.illumina.com/sequencing/sequencing_kits/trusight-rna-pan-cancer-panel/downloads.html)).

### Bioinformatics module

FASTQ file quality control. The raw FASTQ quality control was performed using the FASTQC tool (<https://www.bioinformatics.babraham.ac.uk/>), which provided information on reads in terms of sequence duplication levels, per base and per sequence average quality score, sequence length distribution, and adapter content.

Fusion gene assessment. A purpose-built bioinformatics pipeline was developed to detect fusion genes from RNA CaptureSeq datasets. The pipeline deploys stable and open-source bioinformatics tools in a sequential mode (Fig. 1b):

- *Alignment to targets.* BWA-MEM<sup>40</sup> v. 0.7.15-r1140 aligned PE reads to the genomic sequences of the targeted genes. The PE reads that did not map entirely on the reference genome through SAMTOOLS<sup>41</sup> v. 1.8 were isolated; these PE reads (informative) may derive from fragments of the fusion gene breakpoint.
- *Assembly.* The informative reads are assembled into longer sequences (contigs) through the SPAdes<sup>42</sup> v. 3.12.0 tool.

SPAdes was run with 3 different settings of k-mer size (25, 31, and 51) to cover any possible contig scenarios, thus maximizing the sensitivity of our strategy. This step is critical since more extended sequences have a higher chance of correctly aligning on the fusion gene partner at the genomic level.

- *Alignment to the complete genome.* BWA-MEM aligned contig sequences to the complete human genome (GRCh38.p12). SAMTOOLS then retrieved contig sequences that showed chimeric features, thus mapping the 5′- and 3′-sides of different genomic locations.
- *Gene annotation and fusion gene assessment.* The chimeric sequences were annotated with BEDTOOLS<sup>43</sup> v. 2.27.0 and GENCODE<sup>37</sup> release 29 (GRCh38.p12) annotation. Any chimeric sequence with different gene annotation between the 5′- and 3′-side were termed fusion genes. These were queried to the web-application FusionHub<sup>44</sup> to highlight fusion genes already described in other studies.
- Description of public databases is provided by the FusionHub's authors (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5929557/table/pone.0196588.t001/?report=objectonly>).

## REFERENCES

1. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. *Lancet*. 2013;381:1943–1955. doi:10.1016/S0140-6736(12)62187-4.
2. Nguyen K, Devidas M, Cheng S-C, et al. Factors influencing survival after relapse from acute lymphoblastic leukemia: a Children's Oncology Group study. *Leukemia*. 2008;22:2142–2150. doi:10.1038/leu.2008.251.
3. Hunger SP, Mullighan CG. Redefining ALL classification: toward detecting high-risk ALL and implementing precision medicine. *Blood*. 2015;125:3977–3987. doi:10.1182/blood-2015-02-580043.
4. Iacobucci I, Mullighan CG. Genetic basis of acute lymphoblastic leukemia. *J Clin Oncol*. 2017;35:975–983. doi:10.1200/JCO.2016.70.7836.
5. Harrison CJ. Cytogenetics of paediatric and adolescent acute lymphoblastic leukaemia. *Br J Haematol*. 2009;144:147–156. doi:10.1111/j.1365-2141.2008.07417.x.
6. Felice MS, Gallego MS, Alonso CN, et al. Prognostic impact of t(1;19)/TCF3-PBX1 in childhood acute lymphoblastic leukemia in the context of Berlin-Frankfurt-Münster-based protocols. *Leuk Lymphoma*. 2011;52:1215–1221. doi:10.3109/10428194.2011.565436.
7. Winters AC, Bernt KM. MLL-rearranged leukemias—an update on science and clinical approaches. *Front Pediatr*. 2017;5:4doi:10.3389/fped.2017.00004.
8. Harvey RC, Mullighan CG, Chen I-M, et al. Rearrangement of CRLF2 is associated with mutation of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome in pediatric B-progenitor acute lymphoblastic leukemia. *Blood*. 2010;115:5312–5321. doi:10.1182/blood-2009-09-245944.
9. Russell LJ, Capasso M, Vater I, et al. Deregulated expression of cytokine receptor gene, CRLF2, is involved in lymphoid transformation in B-cell precursor acute lymphoblastic leukemia. *Blood*. 2009;114:2688–2698. doi:10.1182/blood-2009-03-208397.
10. Pui C-H, Robison LL, Look AT. Acute lymphoblastic leukaemia. *Lancet*. 2008;371:1030–1043. doi:10.1016/S0140-6736(08)60457-2.
11. Stam RW. MLL-AF4 driven leukemogenesis: what are we missing? *Cell Res*. 2012;22:948–949. doi:10.1038/cr.2012.16.
12. Iqbal N, Iqbal N. Imatinib: a breakthrough of targeted therapy in cancer. *Chemother Res Pract*. 2014;2014:357027doi:10.1155/2014/357027.
13. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst*. 1960;25:85–109. <http://www.ncbi.nlm.nih.gov/pubmed/14427847>. Accessed February 19, 2019.
14. Iijima-Yamashita Y, Matsuo H, Yamada M, et al. Multiplex fusion gene testing in pediatric acute myeloid leukemia. *Pediatr Int*. 2018;60:47–51. doi:10.1111/ped.13451.
15. Mertens F, Johansson B, Fioretos T, et al. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer*. 2015;15:371–381. doi:10.1038/nrc3947.
16. Bacher U, Shumilov E, Flach J, et al. Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. *Blood Cancer J*. 2018;8:113doi:10.1038/s41408-018-0148-6.
17. Kumar S, Vo AD, Qin F, et al. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci Rep*. 2016;6:21597doi:10.1038/srep21597.
18. Mercer TR, Clark MB, Crawford J, et al. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat Protoc*. 2014;9:989–1009. doi:10.1038/nprot.2014.058.
19. Clark MB, Mercer TR, Bussotti G, et al. Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat Methods*. 2015;12:339–342. doi:10.1038/nmeth.3321.
20. Conter V, Bartram CR, Valsecchi MG, et al. Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. *Blood*. 2010;115:3206–3214. doi:10.1182/blood-2009-10-248146.
21. Haas B, Dobin A, Stransky N, et al. STAR-fusion: fast and accurate fusion transcript detection from RNA-seq. *bioRxiv*. 2017;120295. doi: <https://doi.org/10.1101/120295>.
22. Winters JL, Davila JI, McDonald AM, et al. Development and verification of an RNA sequencing (RNA-Seq) assay for the detection of gene fusions in tumors. *J Mol Diagn*. 2018;20:495–511. doi:10.1016/j.jmoldx.2018.03.007.
23. GTEx Consortium TGTThe genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;45:580–585. doi:10.1038/ng.2653.
24. van der Linden MH, Valsecchi MG, De Lorenzo P, et al. Outcome of congenital acute lymphoblastic leukemia treated on the Interfant-99 protocol. *Blood*. 2009;114:3764–3768. doi:10.1182/blood-2009-02-204214.
25. Pieters R, Schrappe M, De Lorenzo P, et al. A treatment protocol for infants younger than 1 year with acute lymphoblastic leukaemia (Interfant-99): an observational study and a multicentre randomised trial. *Lancet*. 2007;370:240–250. doi:10.1016/S0140-6736(07)61126-X.
26. Roberts KG, Morin RD, Zhang J, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell*. 2012;22:153–166. doi:10.1016/j.ccr.2012.06.005.
27. Puig-Oliveras A, Revilla M, Castelló A, et al. Expression-based GWAS identifies variants, gene interactions and key regulators affecting intramuscular fatty acid content and composition in porcine meat. *Sci Rep*. 2016;6:31803doi:10.1038/srep31803.
28. Babiceanu M, Qin F, Xie Z, et al. Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res*. 2016;44:2859–2872. doi:10.1093/nar/gkw032.
29. Nicorici D, Şatalan M, Edgren H, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. 2014;011650. doi:10.1101/011650.
30. Kim RN, Kim A, Choi S-H, et al. Novel mechanism of conjoined gene formation in the human genome. *Funct Integr Genomics*. 2012;12:45–61. doi:10.1007/s10142-011-0260-1.
31. Prakash T, Sharma VK, Adati N, et al. Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. *PLoS One*. 2010;5:e13284doi:10.1371/journal.pone.0013284.
32. Mullighan CG, Su X, Zhang J, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N Engl J Med*. 2009;360:470–480. doi:10.1056/NEJMoa0808253.
33. Sevov M, Bunikis I, Häggqvist S, et al. Targeted RNA sequencing assay efficiently identifies cryptic KMT2A (MLL)-fusions in acute leukemia patients. *Blood*. 2014;124: <http://www.bloodjournal.org/content/124/21/2406?sso-checked=true>. Accessed March 2, 2019.
34. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25:1105–1111. doi:10.1093/bioinformatics/btp120.
35. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21. doi:10.1093/bioinformatics/bts635.
36. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol*. 2011;12:R72doi:10.1186/gb-2011-12-8-r72.
37. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22:1760–1774. doi:10.1101/gr.135350.111.
38. Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25doi:10.1186/gb-2009-10-3-r25.

39. Chomzynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Anal Biochem.* 1987;162:156–159. doi:10.1006/abio.1987.9999.
40. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2010;26:589–595. doi:10.1093/bioinformatics/btp698.
41. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25:2078–2079. doi:10.1093/bioinformatics/btp352.
42. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19:455–477. doi:10.1089/cmb.2012.0021.
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–842. doi:10.1093/bioinformatics/btq033.
44. Panigrahi P, Jere A, Anamika K. FusionHub: a unified web platform for annotation and visualization of gene fusion events in human cancer. Kumar-Sinha C, ed. *PLoS One.* 2018;13:e0196588doi:10.1371/journal.pone.0196588.

## Chapter 9

ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. *Bioinformatics*.

Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, **Griani A**, Knecht H, Schlitt M, Dreger P, Sellner L, Herrmann D, Pingeon M, Boudjoghra M, Rijntjes J, Pott C, Langerak AW, Groenen PJTA, Davi F, Brüggemann M, Darzentas N; EuroClonality-NGS. ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. *Bioinformatics*. 2017 Feb 1;33(3):435-437. doi: 10.1093/bioinformatics/btw634. PMID: 28172348.

## Sequence analysis

# ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data

Vojtech Bystry<sup>1,†</sup>, Tomas Reigl<sup>1,†</sup>, Adam Krejci<sup>1,2,†</sup>, Martin Demko<sup>1</sup>, Barbora Hanakova<sup>1</sup>, Andrea Grioni<sup>1,3</sup>, Henrik Knecht<sup>4</sup>, Max Schlitt<sup>4</sup>, Peter Dreger<sup>5</sup>, Leopold Sellner<sup>5</sup>, Dietrich Herrmann<sup>4</sup>, Marine Pingeon<sup>6</sup>, Myriam Boudjoghra<sup>6</sup>, Jos Rijntjes<sup>7</sup>, Christiane Pott<sup>4</sup>, Anton W. Langerak<sup>8</sup>, Patricia J. T.A. Groenen<sup>7</sup>, Frederic Davi<sup>6</sup>, Monika Brüggemann<sup>4</sup> and Nikos Darzentas<sup>1,\*</sup> also on Behalf of EuroClonality-NGS

<sup>1</sup>CEITEC – Central European Institute of Technology, Masaryk University, Brno, Czech Republic, <sup>2</sup>RECAMO, Masaryk Memorial Cancer Institute, Brno, Czech Republic, <sup>3</sup>Centro Ricerca Tettamanti, Clinica Pediatrica, Università di Milano-Bicocca, Ospedale San Gerardo/Fondazione MBBM, Monza, Italy, <sup>4</sup>Department of Hematology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany, <sup>5</sup>Department of Medicine V, University Hospital Heidelberg, Heidelberg, Germany, <sup>6</sup>Department of Hematology, Hopital Pitié-Salpêtrière and Pierre et Marie Curie University, Paris, France, <sup>7</sup>Department of Pathology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands and <sup>8</sup>Department of Immunology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Inanc Birol

Received on January 28, 2016; revised on September 9, 2016; accepted on September 29, 2016

## Abstract

**Motivation:** The study of immunoglobulins and T cell receptors using next-generation sequencing has finally allowed exploring immune repertoires and responses in their immense variability and complexity. Unsurprisingly, their analysis and interpretation is a highly convoluted task.

**Results:** We thus implemented ARResT/Interrogate, a web-based, interactive application. It can organize and filter large amounts of immunogenetic data by numerous criteria, calculate several relevant statistics, and present results in the form of multiple interconnected visualizations.

**Availability and Implementation:** ARResT/Interrogate is implemented primarily in R, and is freely available at <http://bat.infospire.org/arrest/interrogate/>

**Contact:** nikos.darzentas@gmail.com

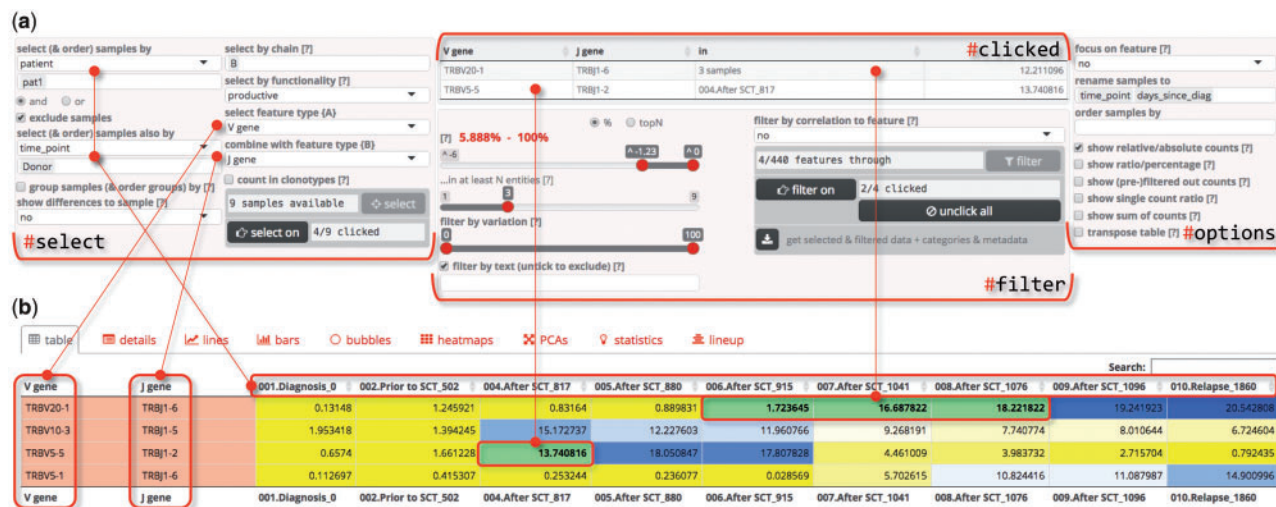
**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Immunoglobulins (IG) and T cell receptors (TR) are highly adaptive molecular receptors responsible for antigen recognition in immunological responses. Fundamental to their adaptiveness is their enormous inherent variability, achieved through stochastic processes during B and T cell maturation. The advent of high-throughput profiling of IG and TR repertoires (Benichou *et al.*, 2012) has been instrumental for understanding normal and pathologic immune responses, which include a

wide range of diseases with an underlying immune cause. This unprecedented capability has also brought along novel and unique challenges.

The first task of immunoprofiling is sequence annotation, such as which variable (V), diversity (D) and joining (J) genes have been rearranged, or what is the sequence of the hypervariable complementarity-determining region 3 (CDR3). IMGT® (Lefranc *et al.*, 2015) is the global reference in the field of antigen receptor sequence analysis and immunogenetic annotation.



**Fig. 1.** The controls panel (a) with the table view (b). (a) The controls panel is divided into 3 parts: select, filter and options. The former two are common for all the visualizations, while options change depending on which visualization view is currently selected. The table ('#clicked') above the filters shows features and samples currently highlighted in the visualizations and it is updated on the fly as the user clicks in the visualizations. (b) In the table, for each feature in a row, abundance values are shown in columns of samples. Cells with features are colored in groups (in this case by receptor and chain i.e. 'TRB'), cells with abundance values are colored in a heatmap-like fashion (Color version of this figure is available at *Bioinformatics* online.)

Mining these inherently complex immunogenetic annotations of usually millions of reads and tens to hundreds of samples for biologically relevant information is a non-trivial task. There is an increasing number of published software applications to tackle this challenge, all with their unique features and advantages, but also limitations like limited interactivity (Alamyar *et al.*, 2012; Shugay, 2015) or scope restricted to repertoire studies (Moorhouse *et al.*, 2014) or minimal residual disease (MRD) monitoring (Giraud *et al.*, 2014).

In this work, we put together in one application features and functionalities we believe are needed for wide-ranging *in silico* immunoprofiling. These insights are a result of collaborative efforts within the EuroClonality-NGS consortium, which strives to develop, standardize and validate *in vitro* assays and bioinformatics for IG/ TR NGS analysis.

## 2 Methods

ARResT/Interrogate is primarily based on R and Shiny, a framework for user interactivity and web-based accessibility. The analytical core relies on the 'data.table' R package for efficient data handling based on advanced indexing techniques. Therefore, ARResT/Interrogate is able to maintain sufficient responsiveness even with datasets of tens of thousands of clonotypes from millions of reads and dozens of samples.

ARResT/Interrogate has four step-wise functions: input processing, data selection and filtering, comparative calculations and visualization.

### 2.1 Input processing

An integrated parser processes multiple IMGT/HighV-QUEST runs and their major immunogenetic annotations. Of these, the V, D and J genes and alleles are combined with the amino acid sequence of the junction (which encompasses the CDR3) to construct IMGT-like clonotypes (Li *et al.*, 2013). These annotations are referred to as 'feature types' and their corresponding individual values as 'features'; for example, feature type 'V gene' contains feature 'TRBV20-1' (Fig. 1).

### 2.2 Data selection and filtering

Users can annotate samples with arbitrary metadata (e.g. patient data, sampling dates) and use these to select and group samples of interest. The next necessary step is to select feature types to focus on. This creates a table of abundance per feature per sample, with abundance expressed as relative or absolute count of reads or clonotypes. Individual features can be filtered in or out using a combination of four filters: abundance, variation across samples, correlation of abundance profiles across samples, and text regular expression (see [Supplementary Section S2.1](#)).

### 2.3 Comparative calculations

ARResT/Interrogate can calculate and visualize differences between samples and features. Samples are compared on the basis of the abundance of a single feature (e.g. TRBV20-1), or an entire feature type (e.g. V gene). Features are compared on the basis of their abundance distributions across samples. Groups of samples can also be statistically compared, for example, to assess immunogenetic differences before and after therapy (S2.3). ARResT/Interrogate can also perform principal component analyses (PCA) of samples and features.

### 2.4 Visualization

Interactive views include tables; line charts, suitable for time-series analyses of clonal kinetics including MRD monitoring; bar charts, popular in clonality testing for lymphoma diagnostics; bubble charts; heatmaps, for sample-sample distance and sample-feature distributions; PCA scatterplots; statistical plots; and multiple sequence alignments. Customizing the visualizations (Fig. 1a, #options) includes changing axis properties like values, labels, scales, orientation; and using extra virtual features such as sums of abundances. Interactivity includes zooming, feature highlighting or hiding, and tooltips with detailed information on any data point. Finally, visualizations are interconnected, with features selected in one automatically highlighted in others.

### 3 Results

Results from the validation and the expert evaluation of ARResT/Interrogate based on actual research data, as well as a running example, are available in the [Supplementary Material](#).

### 4 Conclusions

We presented ARResT/Interrogate, an interactive data manipulation and visualization application for NGS-based immunoprofiling. It offers a wide variety of options and aims to serve as a user-friendly platform with flexible and powerful analytical capabilities.

### Acknowledgements

Computational resources in CEITEC MU were provided by MetaCentrum (LM2010005), and CERIT-SC (CERIT Scientific Cloud, Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144).

### Funding

Authors from CEITEC MU were supported by research grant AZV-MZ-CR 16-34272A-4/2016, project CEITEC 2020 (LQ1601) and ESLHO::EuroClonality, and an OVC Pet Trust Research Grant with the University of Guelph (051699); A.K. was additionally supported by project MEYS-NPS I-LO1413.

*Conflict of Interest:* none declared.

### References

- Alamyar, E. *et al.* (2012) IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res.*, 8, 26.
- Benichou, J. *et al.* (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135, 183–191.
- Giraud, M. *et al.* (2014) Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics*, 15, 409.
- Lefranc, M.P. *et al.* (2015) IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.*, 43, D413–D422.
- Li, S. *et al.* (2013) IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat. Commun.*, 4, 2333.
- Moorhouse, M. *et al.* (2014) ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *BMC Immunology*, 15.
- Shugay, M. 2015. VDJtools: a framework for post-analysis of repertoire sequencing data. Release 1.0.4

## Chapter 10

High resolution IgH repertoire analysis reveals fetal liver as the likely origin of life-long, innate B lymphopoiesis in humans.

Roy A, Bystry V, Bohn G, Goudevenou K, Reigl T, Papaioannou M, Krejci A, O'Byrne S, Chaidos A, **Griani A**, Darzentas N, Roberts IAG, Karadimitris A. High resolution IgH repertoire analysis reveals fetal liver as the likely origin of life-long, innate B lymphopoiesis in humans. *Clin Immunol.* 2017 Oct;183:8-16. doi: 10.1016/j.clim.2017.06.005. Epub 2017 Jun 20. PMID: 28645875; PMCID:PMC5678457.





# High resolution IgH repertoire analysis reveals fetal liver as the likely origin of life-long, innate B lymphopoiesis in humans



Anindita Roy<sup>a,1</sup>, Vojtech Bystry<sup>b,1</sup>, Georg Bohn<sup>c</sup>, Katerina Goudevenou<sup>c</sup>, Tomas Reigl<sup>b</sup>, Maria Papaioannou<sup>c</sup>, Adam Krejci<sup>c,d</sup>, Sorcha O'Byrne<sup>a</sup>, Aristeidis Chaidos<sup>c</sup>, Andrea Griioni<sup>a,e</sup>, Nikos Darzentas<sup>b</sup>, Irene A.G. Roberts<sup>a,f,\*\*,1</sup>, Anastasios Karadimitris<sup>c,\*,1</sup>

<sup>a</sup> Department of Paediatrics, University of Oxford, Brno, Czech Republic

<sup>b</sup> CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic

<sup>c</sup> Centre for Haematology, Department of Medicine, Imperial College London, Imperial College Healthcare NHS Trust, Hammersmith Hospital, London, UK

<sup>d</sup> RECAMO, Masaryk Memorial Cancer Institute, Brno, Czech Republic

<sup>e</sup> Centro Ricerca Tettamanti, Clinica Pediatrica, Università di Milano-Bicocca, Ospedale San Gerardo/Fondazione MBBM, Monza, Italy

<sup>f</sup> MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford and BRC Blood Theme, NIHR Oxford Biomedical Centre, Oxford, UK

## ARTICLE INFO

### Article history:

Received 15 May 2017

Received in revised form 16 June 2017

Accepted with revision 16 June 2017

Available online 20 June 2017

### Keywords:

Human

Fetal

IgH repertoire

## ABSTRACT

The ontogeny of the natural, public IgM repertoire remains incompletely explored. Here, high-resolution immunogenetic analysis of B cells from (unrelated) fetal, child, and adult samples, shows that although fetal liver (FL) and bone marrow (FBM) IgM repertoires are equally diversified, FL is the main source of IgM natural immunity during the 2nd trimester. Strikingly, 0.25% of all prenatal clonotypes, comprising 18.7% of the expressed repertoire, are shared with the postnatal samples, consistent with persisting fetal IgM + B cells being a source of natural IgM repertoire in adult life. Further, the origins of specific stereotypic IgM + B cell receptors associated with chronic lymphocytic leukemia, can be traced back to fetal B cell lymphopoiesis, suggesting that persisting fetal B cells can be subject to malignant transformation late in life. Overall, these novel data provide unique insights into the ontogeny of physiological and malignant B lymphopoiesis that spans the human lifetime.

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Mature B-cell development in humans starts in the fetal liver (FL) in early fetal life, and becomes well established at this site by the start of the second trimester [1,2]. Subsequently, during the second trimester, bone marrow (BM) becomes the main site of B lymphopoiesis [3] and remains so throughout post-natal life.

Development of mature B-cells depends upon, and proceeds commensurately with expression of a functional B-cell receptor (BCR) and of its constituent immunoglobulin (Ig) heavy(H) and light(L) chains. The molecular hallmark of B-cell development, somatic recombination of the genes that encode the IGH(V, D and J) and IGL(V and J) chains, takes place in early B-cell progenitors in primary B lymphopoiesis sites (i.e. FL, FBM and adult BM). This ensures the first wave of Ig

repertoire diversification, with antigen specificity primarily encoded by the complementarity determining region 3 (CDR3). This process is a pre-requisite for efficient humoral immunity, even early in fetal life [4]. The first mature B-cells that emerge from FL and FBM are transitional B-cells that co-express IgM, IgD and CD10 [5,6]. Transitional B-cells mature into CD10neg naïve B-cells that express less IgM. In postnatal life, but not fetal life, naïve B-cells enter a germinal centre reaction in secondary lymphoid organs, undergoing isotype class switch to IgG/IgA and somatic hypermutation, a process that ensures the second wave of Ig repertoire diversification and the production of high affinity soluble antibodies. By contrast, the majority of the fetal life IgM repertoire comprises antibodies that are self- and poly-reactive [7]. This so called 'natural' IgM antibody repertoire is public, i.e., shared by different individuals at birth and is present in adult life as part of the normal, non-pathogenic innate Ig repertoire, albeit at lower frequencies than in the newborn [8,9]. Self-reactive and poly-reactive IgM antibodies, and in particular those using the IGHV6-1 gene, are dominant in FL B-cells [10]. In adult life, self-reactive IgM antibodies may play a role in protection from pathogens and autoimmunity [11]. In mice, the natural IgM repertoire is largely linked to B-1a cells which once developed and selected in FL, persist for the animal's lifespan through their ability for self-renewal rather than iterative development and selection [12].

\* Correspondence to: Anastasios Karadimitris, Centre for Haematology, Imperial College London, 4<sup>th</sup> Floor, Commonwealth Building, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK.

\*\* Correspondence to: Irene AG Roberts, Department of Paediatrics, University of Oxford, Brno, Czech Republic.

E-mail addresses: [irene.roberts@paediatrics.ox.ac.uk](mailto:irene.roberts@paediatrics.ox.ac.uk) (I.A.G. Roberts), [a.karadimitris@imperial.ac.uk](mailto:a.karadimitris@imperial.ac.uk) (A. Karadimitris).

<sup>1</sup> These authors contributed equally.

Recent evidence suggests that B-1a-like cells also exist in humans and may contribute to the development of the natural IgM repertoire [13].

Profiling of the expressed IgH gene repertoire at mRNA level has helped to understand the dynamics of humoral immunity development. However, the relationship of the fetal B-cell IgM repertoire to post-natal child and adult B-cells is incompletely understood and has mostly been approached by low-throughput analyses [14, 15]. A recent high-throughput study of the IgH repertoire of circulating fetal blood B-cells provided some insights into Ig repertoire ontogeny [16]. However, the spatiotemporal relationship between the IgH repertoire in FL with that in FBM, and the impact of the fetal Ig repertoire on the long-term repertoire present in post-natal life, as well as the link between this and the development of disease, are unknown.

Here, to address these issues and to gain insights into the ontogeny of the human innate B-cell repertoire, we take advantage of a high-resolution analysis of the IgH-Cmu repertoire of normal human FL, FBM and post-natal B-cells from healthy infants, young children and adults.

## 2. Materials and methods

### 2.1. Samples

Human FL and BM cells (Table S1) were provided by the Human Developmental Biology Resource ([www.hdbr.org](http://www.hdbr.org)). Surplus blood from samples collected from healthy children was obtained under national ethics committee approval (MREC12/LO/0425). For each sample, CD34-CD19 + mature B-cells (Table S1) were FACS sorted on BD FACSAriaII (Becton Dickinson, Oxford, UK) for BCR repertoire analysis by 454 sequencing.

### 2.2. Bioinformatics

To reduce repertoire sampling biases, we included in the analysis only samples with a comparable number of B-cells when possible (Table S1). The raw NGS data were processed, annotated with germline sequences from IMGT® and/or using IMGT/V-QUEST and IMGT/HighV-QUEST (<http://www.imgt.org>), and analysed through ARResT/Interrogate [17]. As part of ARResT/Interrogate, and with the use of the R language for statistical computing [[www.R-project.org](http://www.R-project.org)]; the Jensen-Shannon divergence was used to compute repertoire similarity between pairs of samples; the inverse Simpson concentration [18], which favors abundant clonotypes over rare ones, was used on vectors of clonotype abundances to calculate clonotypic diversity. Sequences were assigned to the 19 major subsets of stereotyped B-cell receptors in chronic lymphocytic leukemia (CLL) using ARResT/AssignSubsets [19].

Further methodological details are provided in Supplementary methods.

## 3. Results

### 3.1. High-resolution analysis of fetal and postnatal IgHmu repertoires

For initial assessment of the IgM repertoire ontogeny in FL and FBM B-cells, we analysed flow-sorted CD34-CD19 + B-cells. These express cytoplasmic IgM( $\mu$ ) and/or surface (s)IgM and comprise pre-B-cells, immature, transitional and naïve B-cells [5,20]. Spectratyping of IGHV-Cmu mRNA IGHV1-IGHV6 amplicons from a 2nd trimester FL sample (gestational age [GA], 15<sup>+3</sup> weeks), a 2nd trimester FBM sample of the same GA(15<sup>+3</sup> weeks) and B-cells from healthy children and adults revealed a polyclonal repertoire in both FL and FBM that was comparable to the postnatal samples (Fig. 1a)

To gain further insights into the ontogeny of IgH diversification, we sequenced the IGHV-Cmu mRNA IGHV1-IGHV7 family amplicons from

FL, FBM and postnatal samples using next-generation sequencing (NGS) and the 454 technology. In total, 20 libraries generated from 17 individual, flow-sorted CD34-CD19 + B-cell samples were sequenced: 5 FL (4 performed in independent duplicate libraries; GA 14–18 weeks), 3 FBM (GA 13–17 weeks; different fetuses from the FL samples), 3 child peripheral blood (cPB) and 5 adult PB (aPB) B-cell samples. We obtained 117,757 unique clonotypes of which 76%(90,238) were productive, with the remainder representing unproductive rearrangements (Table S1).

Reproducibility was tested by comparing the duplicate libraries from the 4 FL B-cell samples generated and sequenced in 2 independent experiments. Principal component analysis of different combinations of immunogenetic features demonstrated clear demarcation and tight clustering of duplicate pairs (Fig. S1), showing the high degree of accuracy and reproducibility of the assay.

### 3.2. Diversification of the fetal IGHV, IGHD and IGHJ repertoires

Further dissection of the complexity of IgM repertoire development showed that all 52 member genes of the IGHV1-IGHV7 families were used at varying and often significantly different frequencies in all 4 developmental stages (Fig. 1b and Table S2). In line with previous reports [14–16], the most notable difference in IGHD genes usage frequency was the >10 fold higher IGHD7-27 frequency in fetal compared to postnatal samples (Fig. 1c). The pattern of IGHJ repertoire usage was nearly identical between FL and FBM, and between cPB and aPB B-cells. IGHJ4 was the most frequently used J gene in all developmental stages (Fig. 1d) and, consistent with previous reports [15,16], there was reciprocal presence of 4 IGHJ genes: IGHJ6 and IGHJ2 were significantly over-represented in postnatal B-cells ( $p < 0.001$ ), while IGHJ3 and IGHJ5 were significantly over-represented in fetal B-cells ( $p < 0.001$ ; Fig. 1d). Finally, as previously described [15,16], average CDR3 length was significantly shorter in fetal than postnatal B-cells, 14.8 amino acids (aa) vs. 17.3aa ( $p = 0.001$ ; Fig. 1e)

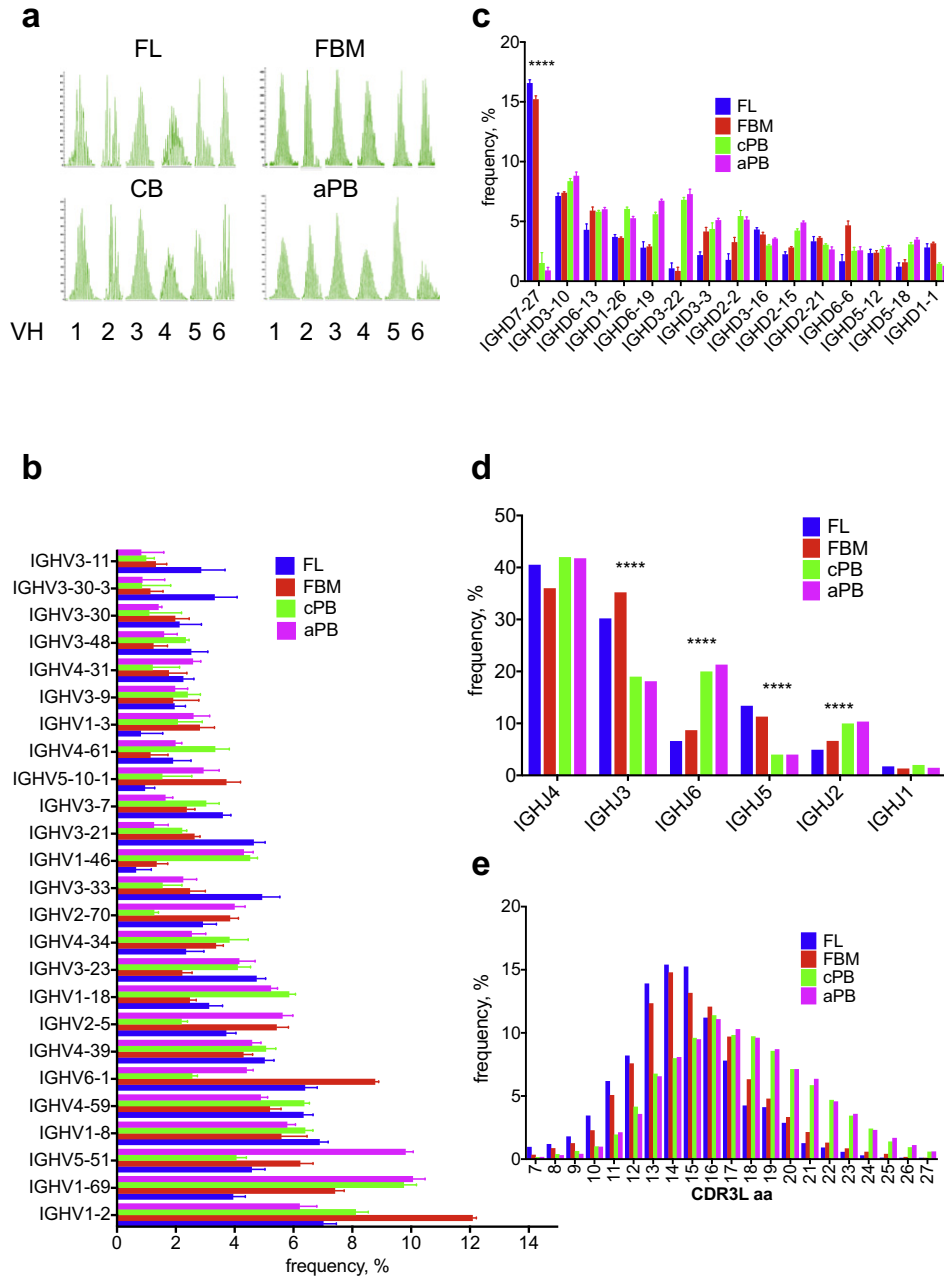
These data show the molecular mechanisms responsible for VDJ recombination-dependent repertoire diversification are active and efficient early in B-cell development in both FL and FBM and that on a qualitative level, comparably diversified B-cell lymphopoiesis exists contemporaneously in FL and FBM.

### 3.3. Evidence of antigen-driven clonotypic expansions in FL B-cells

Antibodies produced by the fetus are mostly IgM and are self- and poly-reactive; however, the source of fetal IgM in FL or FBM B-cells is not known. Hypothesising that B-cells producing IgM would have undergone clonotypic expansion in response to self-antigenic stimulus, we sought to identify such expansions by studying the 100 most abundant clonotypes in each stage. Mean clonotype abundance of the top 100 most abundant clonotypes in each of the 4 stages, was 10-fold lower in FBM B-cells (0.12%) than in FL B-cells (1.2%,  $p < 0.0001$ ), while corresponding abundances in postnatal PB B-cells were intermediate (cPB: 0.54%; aPB: 0.41%; Fig. 2a). Reflecting the paucity of expanded clonotypes amongst FBM B-cells, analysis of the 100 most abundant clonotypes from across all 4 stages (i.e., 100 of 90,238 clonotypes, Table S1) showed that none were present in FBM, compared to 65 in FL, 23 in cPB, and 12 in aPB B-cells (Fig. 2b)

To assess clonotypic expansion and diversity in individual stages, we estimated the inverse Simpson concentration of the clonotypic repertoires [18]. We found that FL clonotypes are the least diverse, followed by cPB and aPB, while FBM showed significantly higher clonotypic diversity compared to the other groups ( $p < 0.001$ ; Fig. 2c).

Together these results are consistent with robust IgM B-cell clonotypic expansions being prominent in FL and nearly absent in FBM of the same GA, and support the notion that the FL B-cells are the main source of fetal IgM production during the 2nd trimester.

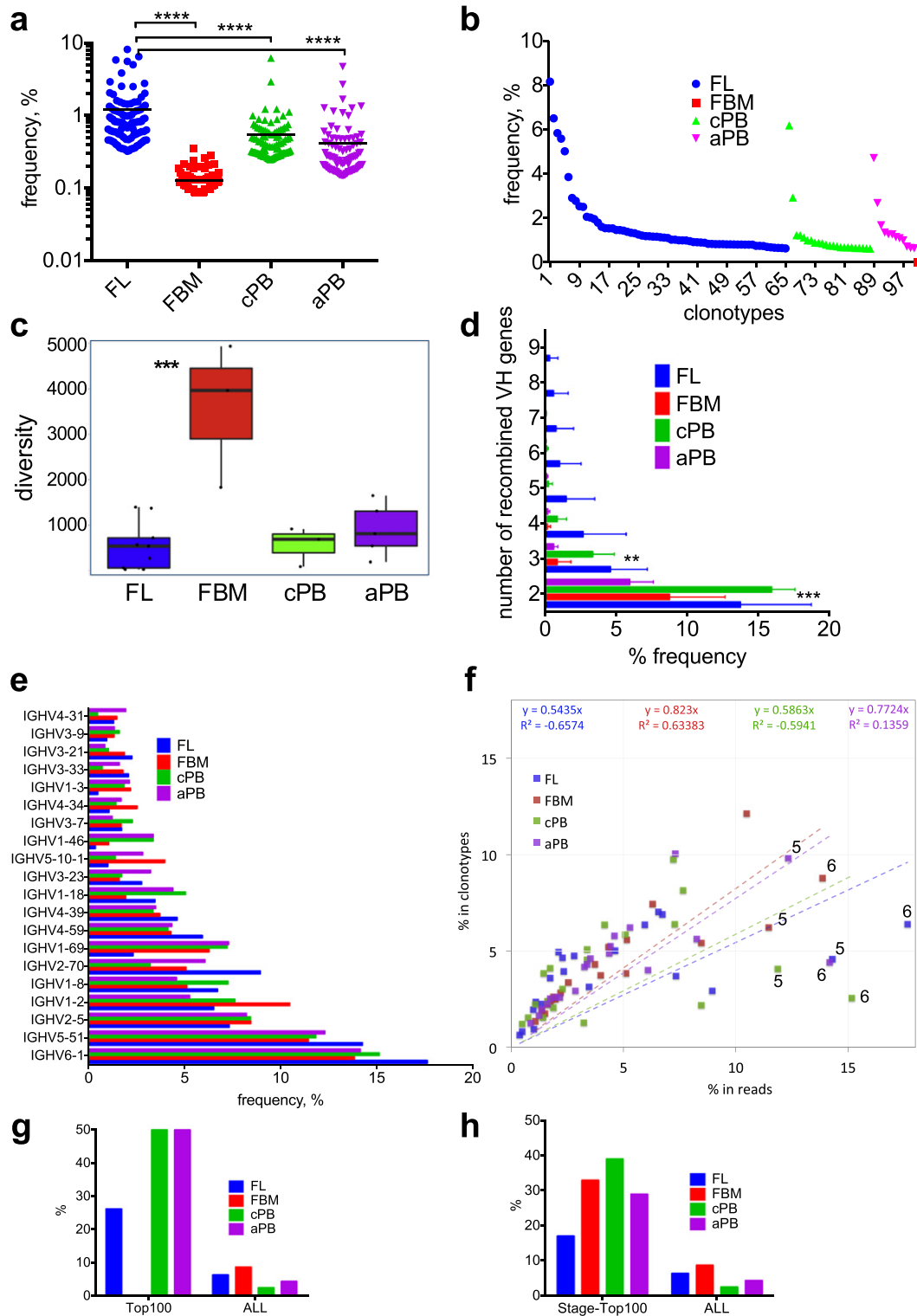


**Fig. 1.** Diversification features of fetal B cell repertoire. **a.** Spectratyping analysis of IGHV1-6 families in fetal liver (FL) GA 15<sup>+3</sup> weeks, fetal bone marrow (FBM) GA 15<sup>+3</sup> weeks, cord blood (CB) and adult peripheral blood (aPB) B-cells. **b.** IGHV gene (the top 25 out of 52 VH genes are shown) **c.** IGHD gene (the top 15 out of total 25 JH genes are shown) **d.** IGHD gene and **e.** CDR3 aa length repertoires counted in unique clonotypes in the 4 different fetal and postnatal developmental stages (cPB: child peripheral blood). **b–e:** mean values with SD are shown, except in **d** & **e** where error bars are omitted for simplicity. (\*\*\**p* < 0.001, \*\*\*\**p* < 0.0001).

### 3.4. Convergent recombination in fetal B-cells

We sought further evidence of antigen-driven responses amongst fetal B-cells by studying their clonotypes in detail, focusing first on clonotypes shared between the duplicate FL libraries to mitigate against possible PCR/sequencing artefacts. All 4 duplicate FL libraries showed evidence of distinct VDJ rearrangements encoding identical CDR3 peptide regions, involving 95–185 clonotypes (1.9–9.3% of all unique clonotypes per library; Table S3). Notably, these CDR3 regions were identical both at the aa and nucleotide level using either different IGHV or IGHD genes, but always the same IGHD gene (Tables S4 & S5). This strikingly precise selection of CDR3 regions, previously termed convergent recombination, has

been described for T-cell receptor repertoire [21,22] and recently in murine B-1a cells [23]. We investigated this further by systematically searching for multiple IGHV genes recombined to an identical CDR3 aa sequence in all samples. Across all stages we found evidence of hundreds of CDR3 sequences recombined with 2–4 different IGHV genes, with up to 9 different IGHV genes identified (Fig. 2d); in nearly all cases this involved genes of the same IGHV family. Detailed sequence analysis (Table S5) highlights the unambiguous assignment of respective germline sequences with no signs of PCR hybrids. Importantly, CDR3 sequences involved in convergent recombinations were most abundant in FL and cPB (Fig. 2d) in line with their increased incidence of prominent clonotypic expansions (Fig. 2a–c)



**Fig. 2.** Clonotypic abundance and diversity in fetal and post-natal B-cells. **a.** Frequency (abundance counted in reads) of the 100 most abundant clonotypes in each developmental stage (horizontal lines indicate mean values, \*\*\*\* $p < 0.0001$ ). **b.** Distribution of the 100 most abundant clonotypes across the 4 developmental stages. **c.** Clonotypic diversity in each developmental stage as assessed by the inverse Simpson concentration (see Methods). \*\*\*\* $p < 0.001$  for FBM as the most diverse. **d.** Frequency of CDR3 peptides generated by convergent recombination of 2–9 different IGHV genes in each developmental stage. \*\* and \*\*\* for  $p < 0.01$  and  $p < 0.001$  respectively for FL and cPB vs. FBM and aPB. **e.** IGHV gene repertoire counted in reads in fetal and postnatal B-cells. Mean values are shown (error bars omitted for simplicity). **f.** Correlation of the relative clonotype abundances of the 20 most popular IGHV genes across developmental stages with their corresponding relative read counts. Lower slopes (as indicated by 'y' values in respective colors) of the regression lines for FL and cPB indicate the predominance of high abundance clonotypes in these developmental stages. '5' and '6': outlying and highly expressed IGHV5-51 and IGHV6-1 respectively. **g** & **h.** IGHV6-1 usage in the 100 most abundant clonotypes across developmental stages (**g**), and in the 100 most abundant clonotypes in each developmental stage (**h**); these "Top100" frequencies are compared to those of IGHV6-1 usage in all clonotypes ("ALL").

Therefore, convergent recombination, a process that ensures generation of a high abundance public immune repertoire in T-cells [21,24], also appears to shape the early fetal B-cell repertoire.

### 3.5. Abundant IGHV6-1 repertoire across developmental stages

To investigate whether repertoire complexity is influenced by biases in specific IGHV family member usage, we compared the rankings of IGHV genes by their frequency of unique clonotypes (i.e. counting in unique clonotypes; Fig. 1b) and abundance (i.e. counting in sequence reads; Fig. 2e). IGHV6-1 and IGHV5-51 were the 1st and 2nd most expressed genes across all 4 developmental stages; however, both genes ranked lower when counted in unique clonotypes, especially in postnatal samples. Fig. 2f correlates relative clonotype abundances of the 20 most popular IGHV genes across developmental stages with their corresponding relative read counts. Expecting these two measures to be linearly correlated, outliers should highlight IGHV genes with highly/lowly-expressed clonotypes. We found that in all 4 developmental stages, IGHV6-1 and IGHV5-51 are placed the furthest from their projected linear distribution and strongly biased towards high expression. Therefore, although IGHV6-1 and IGHV5-51 genes did not have the most associated clonotypes, they are the most likely to participate in high abundance, expanded clonotypes. At a global level, the lower slopes of the regression lines for FL and cPB are also consistent with the higher frequency of high abundance clonotypes in those samples

We then focused on IGHV6-1 as this has previously been shown to be over-represented in fetal B-cells beyond its expected frequency of ~1.9% (i.e., 1/52) [25–27]. Fetal IGHV6-1 IgM BCRs have been reported to react against ssDNA and cardiolipin autoantigens, and are thus important sources of natural IgM [10]. We confirmed that although the relative frequency of IGHV6-1 in unique clonotypes was not higher than the expected 1.9% in FL and FBM (Fig. 1b), its abundance was indeed significantly higher (18% and 14% respectively;  $p = 0.002$ ; Fig. 2e) with similar trends in cPB and aPB. Supporting this, IGHV6-1 was identified in 17/65 (26.2%) FL, 11/23 (50%) cPB and 6/12 (50%) aPB of the 100 most expanded clonotypes across all 4 stages (Fig. 2g). Similarly, within each developmental stage, IGHV6-1 comprised 17, 33, 39 and 29 of the 100 most abundant clonotypes respectively (Fig. 2h), significantly higher frequencies ( $p < 0.01$ ) than their respective average unique clonotype frequencies (6.4, 8.7, 2.5 and 4.4%). Thus, IGHV6-1 clonotypic expansions are dominant in all developmental stages, highlighting an important role of IGHV6-1 IgM in innate humoral immunity throughout life.

### 3.6. Presence of antigen response-competent mature B-cells in FL but not FBM

The high frequency of expanded clonotypes in FL but not FBM suggests that in fetal life it is the FL rather than FBM B-cells that mount (auto-)antigen-driven responses, despite being equally diversified by VDJ recombination. To investigate this further, we compared the frequencies of B-cell sub-populations within the CD34-CD19+ compartment in FL and FBM using previously described markers especially those defining fetal B cell subsets where available [5,6,20,28,29] (see Supplementary methods). While pre-B-cells lack (s)IgM expression, immature B-cells, transitional and naïve B-cells express (s)IgM (Fig. 3a, b). Compared to FL, the FBM CD34-CD19+ compartment had a higher frequency of pre-B-cells (FL:  $52.7 \pm 5.4\%$  vs. FBM:  $69.2 \pm 1.5\%$ ,  $p < 0.01$ ) but a similar frequency of immature B-cells ( $30.7 \pm 4.6\%$  vs.  $21.0 \pm 1.6\%$ ), while transitional and naïve B-cells were significantly decreased in FBM (FL:  $4.2 \pm 0.8\%$  vs. FBM:  $1.5 \pm 0.4\%$ ,  $p < 0.01$ ; and FL:  $2.8 \pm 0.9\%$  vs. FBM:  $0.7 \pm 0.2\%$ ,  $p < 0.05$ ; Fig. 3c). This lack of developed mature B-cells explains, at least in part, the paucity of clonotypic expansions in 2nd trimester FBM. Of the three sIgM+ B-cell populations (immature, transitional and naïve) we used for IgHmu repertoire profiling, only

the transitional B-cell subset was previously shown to expand in response to antigen in a T-cell-independent fashion. Indeed, transitional B-cells are enriched in autoreactive B-cells in normal individuals and more so in patients with systemic lupus erythematosus [30]. Thus, we speculate that transitional B-cells are likely to be the main source of the FL IgM clonotypic expansions and, in contrast to previous reports [4,31], we found a very low frequency of CD34-CD19+ CD27+ B-cells in 2nd trimester FL and FBM (range 0–1.9% of total CD34-CD19+ B-cells, median 0.06%). (Fig. 3b & c).

### 3.7. High abundance FL clonotypes shared across developmental stages

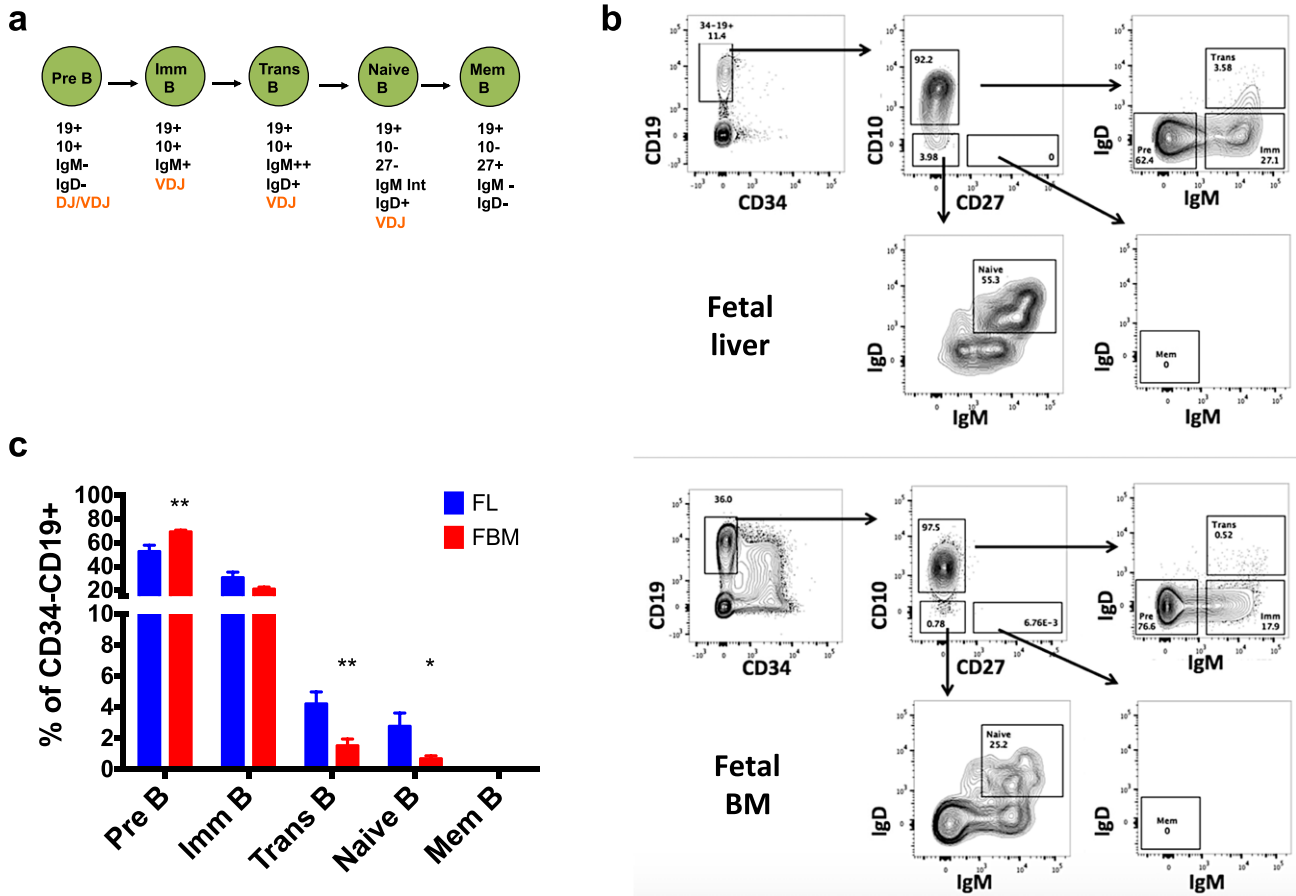
To explore continuity in IgM B-cell immunity between fetal and adult life, we searched for clonotypes shared within and between developmental stages. Overall 0.13% (122/90,238) of productive clonotypes were shared, with none shared by >2 developmental stages (Fig. 4a); 15 were shared between FL and FBM (expressed as 0.37% and 0.22% of reads respectively) (Fig. 4b and Table S6), suggesting selection of B-cells by the same antigen can occur independently in FL or FBM, or possibly migration of B-cells between sites; 2 clonotypes were shared between cPB and aPB B-cells (not shown); 22 of the total FL IgH expressed repertoire (Table S7) were shared between FL (16.3% of reads) and postnatal B-cells (0.92% of reads; Fig. 4c); and 83 were shared between FBM (2.3% of reads) and either cPB (59; 0.85% of reads) or aPB (24; 1.13% of reads; Fig. 4d, Table S8).

Reflecting the high abundance clonotypes in FL, the mean abundance of clonotypes shared between FL and postnatal B-cells was 38-fold higher than FBM (0.77% vs. 0.02%,  $p = 0.001$ ; Fig. 4e), highlighting sharing of only high abundance clonotypes between FL and postnatal B-cells (Fig. 4e). Indeed, 10/22 clonotypes shared between FL and postnatal B-cells were also amongst the 100 most abundant clonotypes across all developmental stages (Fig. 2b, Table S7) and 16/22 shared clonotypes were 41-fold more abundant in FL B-cells than in postnatal B-cells (median 0.36% vs. 0.005%,  $p < 0.0001$ ; Fig. 4f). Notably, 5/22 FL-postnatal shared clonotypes, corresponding to 2 individual CDR3 sequences, had evidence of convergent recombination (Table S7), supporting the notion that clonotypic expansions shared between fetal and post-natal IgM B-cell repertoires are antigen-driven.

Together, these observations are consistent with a fully functional FL IgM repertoire in which B-cell clonotypic expansions are robust and likely to be antigen-driven. The presence of identical clonotypes in fetal and postnatal B-cells might be the result of independent selection at different developmental stages in different individuals or, more likely, selection during fetal life and subsequent persistence in postnatal life. Further, the higher abundance of some shared clonotypes in postnatal compared to FL B-cells indicates that IgM-producing B-cells of FL origin remain functional in postnatal life and retain their ability to expand in response to recurrent antigenic stimulation. Finally, the unexpected degree of clonotype sharing (0.25% of the entire fetal IgM repertoire) between fetal and postnatal B-cells derived from samples that are HLA-disparate suggests that the selection of these shared (public) clonotypes occurs in an HLA- and thus T-cell-independent manner consistent with IgM innate humoral immune responses.

### 3.8. Fetal BCR repertoire and malignancy-associated stereotypic receptors

Our results so far suggest that fetal IgM-producing B-cells may persist into adult life and remain under antigenic stimulation throughout life, potentially increasing their risk of neoplastic transformation. Stereotypic (or quasi-identical) IgM BCR are known to be part of the normal adult B-cell repertoire (enriched in IgM+ CD5+ B-cells in particular [32]) and, importantly, they have also been demonstrated in ~30% of patients with chronic lymphocytic leukemia (CLL), one of the most common IgM+ mature B cell malignancies in humans [33–36]. Nevertheless, their developmental origins and ontogeny have not



**Fig. 3.** B cell development in FL and FBM. a. Schematic representation of proposed fetal B cell maturation according to immunophenotypic markers and stages of VDJ recombination (using human fetal B cell development data where available) [5,6,29] that was studied in 2nd trimester FL and FBM. b. Representative flow-cytometric analysis of FL and FBM of the same fetus (GA 17 weeks) showing the gating strategy used to identify the various stages of B cell maturation as described in (a). Data are from viable CD34 negative cells for the FL sample and viable mononuclear cells for the FBM sample. c. Frequencies of the B cell stages, expressed as % of CD34-CD19+ cells, are shown in the bar graph with data represented as mean ± SEM from FL (n = 13) and FBM (n = 12) samples. (Imm: immature, Trans: transitional, Mem: memory B-cells; \*p < 0.05, \*\*p < 0.01).

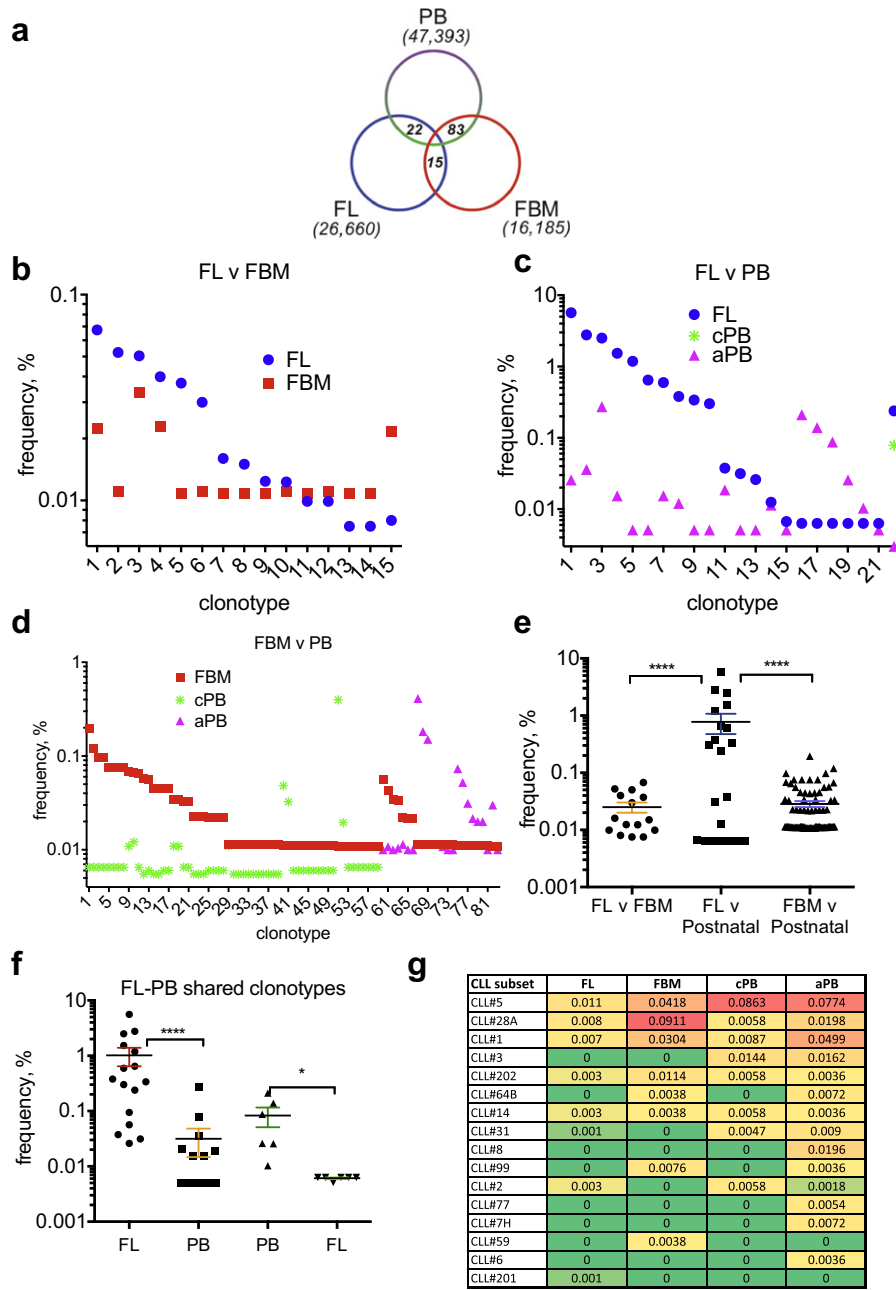
been defined. We therefore searched for evidence in our fetal and postnatal IgM-producing B-cell samples, for the 19 major stereotypic CLL IgH receptors, or major CLL subsets, previously reported in a large study of > 7500 CLL patients [33]. Overall, we found evidence of stereotypic IgH receptors corresponding to one or more of the 16 major CLL subsets in 3/5 FL B-cell samples and in all FBM and postnatal samples (Fig. 4g and not shown): 14/16 subsets were found in postnatal samples, with CLL#1, CLL#5, and CLL#28A the most prevalent but strikingly, 11/16 stereotypical subsets were also present in fetal B-cells, with 2 subsets in fetal B-cells only (Fig. 4g, Table S9). CLL#1, CLL#5 and CLL#28A subsets were again the most prevalent in FL and FBM B-cells (Fig. 4g). Importantly CLL#1, CLL#5 and CLL#28A are amongst the 10 most common CLL subsets, and CLL#1 and #5 are associated with aggressive disease [36]. These findings may provide clues into the ontogenesis of CLL and indicate that for a substantial proportion of stereotypy-associated CLL, the IgM+ B-cell that undergoes malignant transformation in adult life may originally be selected during fetal life and persist throughout adulthood.

#### 4. Discussion

Here we present a comparative, high-resolution dissection of the human IgHmu repertoire from early fetal to adult life. This is the first such analysis to include FBM and FL, the primary sites of fetal B-cell development thus allowing ontogenic and anatomical mapping of the human natural IgM repertoire.

The IgHmu repertoire in prenatal life is responsible for development of the so-called natural antibody immunity. Work in mice has shown that the development of B-cells secreting natural IgM is instructed by non-protein, lipid, phospholipid and glycan antigens often from cells undergoing apoptosis [37–39]. In this respect natural IgM are low affinity auto-reactive antibodies perhaps triggered by inadequately cleared apoptotic cells during fetal development [40]. In postnatal life, the natural IgM repertoire is further enriched with specificities against commensal flora or pathogen-derived non-protein antigens [40]. In mice, the main cellular source of natural IgM are B-1a cells that develop in FL but not FBM. After their selection and clonal expansion by auto-antigens, they persist throughout life by self-renewal.

Our analysis of the ontogeny of the corresponding human IgHmu repertoire, not previously characterised, reveals many features analogous to mice. We find that while comparably diversified B-cell lymphopoiesis exists contemporaneously in FL and FBM, the robust IgM-producing B-cell clonotypic expansions prominent in FL are virtually absent in FBM of the same GA, thus identifying human FL as the likely main source of the natural IgM repertoire in fetal life. The lack of clonotypic expansions in FBM reflects the paucity of late mature B-cells; these probably develop in late 3rd trimester to become the main source of adaptive B-cells in postnatal life. Given that the cord blood and early neonatal IgM repertoires are functionally autoreactive, these clonotypic expansions are likely to be auto-antigen-driven. Indeed, IGHV6-1 clonotypic expansions were dominant in FL, and human IGHV6-1+ fetal B-cells have previously been shown to be reactive against self-phospholipids such as cardiolipin [10]. The



**Fig. 4.** Sharing of clonotypes across developmental stages. a. Venn diagram showing the distribution of the 120 shared clonotypes between FL, FBM and post-natal PB cells. b, c & d. Sharing of clonotypes between FL and FBM, FL and post-natal, and FBM and post-natal B-cells, respectively. None of the clonotypes was shared by >2 developmental stages. Details of the clonotypes shown in b, c & d are shown in Tables S6, S7 & S8 respectively. e. Abundance of clonotypes shared between FL and FBM, FL and post-natal, and FBM and postnatal B-cells. f. Abundance of FL-PB shared clonotypes. Two left columns show shared clonotypes whose abundance is higher in FL rather than in PB, the two right-most columns show shared clonotypes with abundance higher in PB rather than in FL. g. Abundance (%) and sharing of stereotypic IgH receptors associated with CLL in fetal and postnatal B-cells showing 11 subsets shared between FL and FBM, 13 subsets shared between child and adult PB B-cells and 3 subsets shared across all 4 stages (see Table S9; \**p* < 0.05, \*\*\*\**p* < 0.0001).

corresponding orthologous VH7183.1 is also dominant in murine FL [41] revealing remarkable and refined evolutionary conservation.

As in mice, the human natural IgM repertoire is also public, comprising identical/near-identical clonotypes shared by different individuals. Our data provide the first evidence of considerable sharing of IgM clonotypes that originate in human FL amongst different fetuses. For some FL clonotypes, we documented stringent (even at the nucleotide level) convergent recombination underpinning public IgM repertoire generation. While convergent recombination occurring at the aa level has been described in the adult Ig repertoire [42–44], nucleotide-level convergent recombination was recently described in murine B1 cells [23], providing further parallels between human and murine natural IgM ontogeny.

Another distinct contribution of our work to the delineation of the ontogeny of the postnatal ‘public’ IgM repertoire is the finding that FL expanded clonotypes, including those with IGHV6-1, are also found clonally expanded in postnatal life. This most likely reflects auto-reactive IgM-producing B-cells, clonotypically expanded in FL, persisting throughout life perhaps bypassing FBM. Whether their postnatal persistence and expansion is the result of continuous antigenic stimulation (e.g., by apoptotic cells) or of their ability to self-new (analogous to murine B-1a cells) remains to be determined.

Recent IgH repertoire analysis of fetal B cell progenitors at a single cell level, demonstrated that the distinct immunogenetic features of fetal IgH repertoire are determined, at least in part, by a fetal-specific

pattern of VDJ recombination process which may be driven by differences in Tdt expression in fetal life [29].

While analysis of the human fetal IgM repertoire has revealed high concordance with the corresponding murine repertoire, our analysis of the cellular correlate of the pre-immune repertoire, i.e., of B-1a cells, has not. The existence of human counterparts of murine B-1a cells has been contentious. Recent work identified a rare IgM + CD20 + CD27 + CD43 + B-cell population in human cord blood and PB with several functional features akin to murine B-1a cells [13,31]. However, despite strict gating and use of two different anti-CD27 mAb clones (data not shown), the frequency of CD19 + CD27 + cells in the FL and FBM samples we analysed was consistently <1% with most samples having no CD19 + CD27 + cells (Fig. 3). Instead, we found FL but not FBM enriched in immunophenotypically-defined transitional B-cells, a population that in humans has also been linked with production of autoreactive IgM, autoimmune disease and a CD27-CD5 + phenotype [30,45]. This raises the possibility that in humans the B-cell subset responsible for FL IgM clonotypic expansions has features that at least in part overlap with transitional B-cells. In future work, functional characterization and high resolution analysis of the IgHmu repertoire in purified FL and FBM B-cell subsets (Fig. 3) would be required to address this question.

Another novel insight from our work is the demonstration that stereotypic, autoreactive BCR with innate function that are associated with CLL, a malignancy of older adults, may be selected during fetal life. We find that the frequencies of these stereotypic receptors in both FL and FBM are overall very low (<0.01%; Fig. 4g) and nowhere near the frequencies of expanded clonotypes in FL (up to 8%; Fig. 2b) implying that auto-antigens driving their expansion in late adult life may not present in fetal life.

Notwithstanding the very low frequency of CD27 + mature B-cells we observed in FL and FBM, this would support the notion that CD5 + B-cell CLL with unmutated BCR might have its origin in FL B-1a-like B-cells, which also express CD5 and although they are selected once during fetal life they persist long-term in postnatal life [46,47]. Alternatively, and more consistent with our immunophenotyping findings (Fig. 3), unmutated CLL has been mooted to originate from autoreactive transitional IgM + CD5 + B-cells [48]. Previous gene expression profiling of PB human B-cells identified CD27-CD5 + cells as the likely physiologic counterpart of the unmutated CLL B-cells [32]. We speculate that ongoing and life-long antigenic stimulation of these innate B-cells with stereotypic BCR originating in FL renders them susceptible to malignant transformation resulting in CLL.

In conclusion, comparative analysis of IgM repertoire development from fetal to adult IgM B-cells reveals that B-cell repertoire diversification during the 2nd trimester takes place in parallel in FL and FBM. However, since we have shown that mature B-cells capable of antigenic responses are present in FL but not in FBM, this suggests that the liver is the dominant site of likely self-antigen-driven B-cell clonotypic expansions during the 2nd trimester of fetal life. Such FL-derived expanded IgM + B-cells, including those of the IGHV6-1 gene, may persist into adult life and contribute to the auto- and poly-reactive public IgM repertoire and even become targets of malignant transformation.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.clim.2017.06.005>.

#### Authorship contributions

I.R. and A.K. designed and supervised the study. A.R., G.B., K.G., M.P., S.O.B. and A.C. performed the experiments and did the data collection. V.B., T.R., A.K., A.G. and N.D. did the data analysis and interpretation of statistical data. A.R., I.R., N.D. and A.K. wrote the paper and created the figures. All authors reviewed the drafts of the paper and gave final approval of the version to be published.

#### Conflict of interest

The authors declare no conflict of interest.

#### Acknowledgments

A.R. is supported by a Bloodwise Clinician Scientist Fellowship (grant no. 14041) and an EHA-ASH Translational Research Training in Hematology Fellowship. G.B. and K.G. were supported by a Leukaemia Lymphoma Research (Bloodwise) Lectureship. This work was supported by Oxford NIHR Biomedical Centre based at Oxford University Hospitals NHS Trust and University of Oxford and NIHR Biomedical Centre based at Imperial College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The human embryonic and fetal material was provided by the Joint MRC/Wellcome Trust (grant # 099175/Z/12/Z) Human Developmental Biology Resource ([www.hdbr.org](http://www.hdbr.org)). Authors from CEITEC MU were supported by Ministry of Health of the Czech Republic grant nr. 16-34272A, project CEITEC 2020 (LQ1601), and ESLHO::EuroClonality; A.K. was additionally supported by project MEYS - NPS I - LO1413; computational resources were provided by MetaCentrum (LM2010005) and CERIT-SC (CERIT Scientific Cloud, Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144).

#### References

- [1] C. Nunez, N. Nishimoto, G.L. Gartland, et al., B cells are generated throughout life in humans, *J. Immunol.* 156 (1996) 866–872.
- [2] A. Roy, G. Cowan, A.J. Mead, et al., Perturbation of hematopoietic stem and progenitor cell development by trisomy 21, *Proc. Natl. Acad. Sci. U. S. A.* 109 (2012) 17579–17584.
- [3] M. Tavian, K. Biasch, L. Sinka, J. Vallet, B. Peault, Embryonic origin of human hematopoiesis, *Int. J. Dev. Biol.* 54 (2010) 1061–1065.
- [4] F.A. Scheeren, M. Nagasawa, K. Weijer, et al., T cell-independent development and induction of somatic hypermutation in human IgM + IgD + CD27 + B cells, *J. Exp. Med.* 205 (2008) 2033–2042.
- [5] S. Agrawal, S.A. Smith, S.G. Tangye, W.A. Sewell, Transitional B cell subsets in human bone marrow, *Clin. Exp. Immunol.* 174 (2013) 53–59.
- [6] L. McWilliams, K.Y. Su, X. Liang, et al., The human fetal lymphocyte lineage: identification by CD27 and LIN28B expression in B cell progenitors, *J. Leukoc. Biol.* 94 (2013) 991–1001.
- [7] H. Wardemann, S. Yurasov, A. Schaefer, J.W. Young, E. Meffre, M.C. Nussenzweig, Predominant autoantibody production by early human B cell precursors, *Science* 301 (2003) 1374–1377.
- [8] E. Meffre, J.E. Salmon, Autoantibody selection and production in early human life, *J. Clin. Invest.* 117 (2007) 598–601.
- [9] Y. Merbl, M. Zucker-Toledano, F.J. Quintana, I.R. Cohen, Newborn humans manifest autoantibodies to defined self molecules detected by antigen microarray informatics, *J. Clin. Invest.* 117 (2007) 712–718.
- [10] T. Logtenberg, F.M. Young, J.H. Van Es, F.H. Gmelig-Meyling, F.W. Alt, Autoantibodies encoded by the most Jh-proximal human immunoglobulin heavy chain variable region gene, *J. Exp. Med.* 170 (1989) 1347–1355.
- [11] R. Ubelhart, H. Jumaa, Autoreactivity and the positive selection of B cells, *Eur. J. Immunol.* 45 (2015) 2971–2977.
- [12] N. Baumgarth, The double life of a B-1 cell: self-reactivity selects for protective effector functions, *Nat. Rev. Immunol.* 11 (2011) 34–46.
- [13] D.O. Griffin, N.E. Holodick, T.L. Rothstein, Human B1 cells in umbilical cord and adult peripheral blood express the novel phenotype CD20 + CD27 + CD43 + CD70, *J. Exp. Med.* 208 (2011) 67–80.
- [14] V. Pascual, L. Verkruyse, M.L. Casey, J.D. Capra, Analysis of Ig H chain gene segment utilization in human fetal liver. Revisiting the "proximal utilization hypothesis", *J. Immunol.* 151 (1993) 4164–4172.
- [15] M.M. Souto-Carneiro, G.P. Sims, H. Girschik, J. Lee, P.E. Lipsky, Developmental changes in the human heavy chain CDR3, *J. Immunol.* 175 (2005) 7425–7436.
- [16] E. Rechavi, A. Lev, Y.N. Lee, et al., Timely and spatially regulated maturation of B and T cell repertoire during human fetal development, *Sci. Transl. Med.* 7 (2015) 276ra25.
- [17] V. Bystry, T. Reigl, A. Krejci, et al., ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data, *Bioinformatics* (2016).
- [18] L. Jost, Entropy and diversity, *Oikos* 113 (2006) 363–375.
- [19] V. Bystry, A. Agathangelidis, V. Bikos, et al., ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy, *Bioinformatics* 31 (2015) 3844–3846.
- [20] M. Perez-Andres, B. Paiva, W.G. Nieto, et al., Human peripheral blood B-cell compartments: a crossroad in B-cell traffic, *Cytometry B Clin. Cytom.* 78 (Suppl. 1) (2010) S47–S60.



- [21] V. Venturi, K. Kedzierska, D.A. Price, et al., Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination, *Proc. Natl. Acad. Sci. U. S. A.* 103 (2006) 18691–18696.
- [22] V. Venturi, D.A. Price, D.C. Douek, M.P. Davenport, The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* 8 (2008) 231–238.
- [23] Y. Yang, C. Wang, Q. Yang, et al., Distinct mechanisms define murine B cell lineage immunoglobulin heavy chain (IgH) repertoires, *elife* 4 (2015) e09083.
- [24] V. Venturi, M.F. Quigley, H.Y. Greenaway, et al., A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing, *J. Immunol.* 186 (2011) 4285–4294.
- [25] J.E. Berman, K.G. Nickerson, R.R. Pollock, et al., VH gene usage in humans: biased usage of the VH6 gene in immature B lymphoid cells, *Eur. J. Immunol.* 21 (1991) 1311–1314.
- [26] J.H. Van Es, F.M. Raaphorst, M.J. van Tol, F.H. Meyling, T. Logtenberg, Expression pattern of the most JH-proximal human VH gene segment (VH6) in the B cell and antibody repertoire suggests a role of VH6-encoded IgM antibodies in early ontogeny, *J. Immunol.* 150 (1993) 161–168.
- [27] H.W. Schroeder Jr., J.Y. Wang, Preferential utilization of conserved immunoglobulin heavy chain variable gene segments during human fetal life, *Proc. Natl. Acad. Sci. U. S. A.* 87 (1990) 6146–6150.
- [28] T.W. LeBien, Fates of human B-cell precursors, *Blood* 96 (2000) 9–23.
- [29] M.B. Rother, K. Jensen, M. van der Burg, et al., Decreased IL7Ralpha and TdT expression underlie the skewed immunoglobulin repertoire of human B-cell precursors from fetal origin, *Sci Rep* 6 (2016) 33924.
- [30] A. Vossenkamper, P.M. Litalo, J. Spencer, Translational mini-review series on B cell subsets in disease. Transitional B cells in systemic lupus erythematosus and Sjogren's syndrome: clinical implications and effects of B cell-targeted therapies, *Clin. Exp. Immunol.* 167 (2012) 7–14.
- [31] C. Bueno, E.H. van Roon, A. Munoz-Lopez, et al., Immunophenotypic analysis and quantification of B-1 and B-2 B cells during human fetal hematopoietic development, *Leukemia* 30 (2016) 1603–1606.
- [32] M. Seifert, L. Sellmann, J. Bloehdorn, et al., Cellular origin and pathophysiology of chronic lymphocytic leukemia, *J. Exp. Med.* 209 (2012) 2183–2198.
- [33] A. Agathangelidis, N. Darzentas, A. Hadzidimitriou, et al., Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies, *Blood* 119 (2012) 4467–4475.
- [34] P. Baliakas, A. Hadzidimitriou, L.A. Sutton, et al., Clinical effect of stereotyped B-cell receptor immunoglobulins in chronic lymphocytic leukaemia: a retrospective multicentre study, *Lancet Haematol.* 1 (2014) e74–e84.
- [35] J.A. Burger, N. Chiorazzi, B cell receptor signaling in chronic lymphocytic leukemia, *Trends Immunol.* 34 (2013) 592–601.
- [36] N. Darzentas, K. Stamatopoulos, The significance of stereotyped B-cell receptors in chronic lymphocytic leukemia, *Hematol. Oncol. Clin. North Am.* 27 (2013) 237–250.
- [37] M.Y. Chou, L. Fogelstrand, K. Hartvigsen, et al., Oxidation-specific epitopes are dominant targets of innate natural antibodies in mice and humans, *J. Clin. Invest.* 119 (2009) 1335–1349.
- [38] J. Kim, Identification of a human monoclonal natural IgM antibody that recognizes early apoptotic cells and promotes phagocytosis, *Hybridoma (Larchmt)* 29 (2010) 275–281.
- [39] Y. Chen, Y.B. Park, E. Patel, G.J. Silverman, IgM antibodies to apoptosis-associated determinants recruit C1q and enhance dendritic cell phagocytosis of apoptotic cells, *J. Immunol.* 182 (2009) 6031–6043.
- [40] P.I. Lobo, Role of natural autoantibodies and natural IgM anti-leucocyte autoantibodies in Health and disease, *Front. Immunol.* 7 (2016) 198.
- [41] R.L. Schelonka, E. Szymanska, A.M. Vale, Y. Zhuang, G.L. Gartland, H.W. Schroeder Jr., DH and JH usage in murine fetal liver mirrors that of human fetal liver, *Immunogenetics* 62 (2010) 653–666.
- [42] K.J. Jackson, Y. Liu, K.M. Roskin, et al., Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements, *Cell Host Microbe* 16 (2014) 105–114.
- [43] J. Wrammert, D. Koutsouanos, G.M. Li, et al., Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection, *J. Exp. Med.* 208 (2011) 181–193.
- [44] J.C. Krause, T. Tsibane, T.M. Tumpey, et al., Epitope-specific human influenza antibody repertoires diversify by B cell intracлонаl sequence divergence and interclonal convergence, *J. Immunol.* 187 (2011) 3704–3711.
- [45] J. Lee, S. Kuchen, R. Fischer, S. Chang, P.E. Lipsky, Identification and characterization of a human CD5+ pre-naïve B cell population, *J. Immunol.* 182 (2009) 4116–4126.
- [46] R.R. Hardy, B-1 B cells: development, selection, natural autoantibody and leukemia, *Curr. Opin. Immunol.* 18 (2006) 547–555.
- [47] N. Darzentas, A. Hadzidimitriou, F. Murray, et al., A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: molecular and computational evidence, *Leukemia* 24 (2010) 125–132.
- [48] R. Garcia-Munoz, L. Llorente, Chronic lymphocytic leukaemia: could immunological tolerance mechanisms be the origin of lymphoid neoplasms? *Immunology* 142 (2014) 536–550.

## Conclusion

This dissertation provided an overview on the application of next-generation sequencing and bioinformatics as a tool for clinical diagnostics. We herein presented two core strategies for the identification and assessment of genetic markers in acute lymphoblastic leukaemia: the assessment of immunoglobulin and T-cell receptors for minimal residual disease and the identification of fusion genes. The assessment of MRD is a strategy that was developed as part of the EuroClonality-NGS consortium, and I was mainly involved in the development of the bioinformatics pipeline of analysis. My personal project mainly covered the identification of fusion genes from target capture NGS. In this case, I implemented the bioinformatics method used for the sensitive identification of known and new fusion genes in the clinical diagnostics of ALL in Fondazione Tettamanti. These two first projects allowed me to enhance my expertise in bioinformatics as well as programming languages. The gained expertise was then used for the implementation and automatization of an end-to-end informatics infrastructure allowing clinical biologists to run these pipelines of analysis on-demand. Additionally, I expanded my set of bioinformatics tools to cover other areas of NGS application, such as variant calling. These resulted in the analysis of single patient's samples leading to the identification of novel mutations in genes coding for the cohesin complex as well as the identification and characterisation of a novel mutation in the EP300 gene.

The recent application of DL in several scientific fields of research captured my interest owing to its potential application in genomics. In addition, the increased application of NGS in genomics highlighted the role of non-coding elements in the regulation of molecular mechanisms. Subsequently, the second part of my Ph.D. study was dedicated to acquiring DL expertise through the study of small non-coding RNA elements. This collaborative project led to the development of MuStARD, a DL model for the identification of small non-coding RNA loci from the scanning of genomic areas. Acquiring domain knowledge in both DL and non-coding RNA allowed me to start an independent project focused on the development of a model for the identification and localisation of miRNA-target sequence binding site interactions.

In summary, the introduction of next-generation sequencing required a simplified method for the massive analysis of genomes. NGS is scalable into a clinical diagnostic setting for the development of personalised therapies for patients. However, the potential application of NGS in clinical diagnosis is inevitably bound to the analysis and interpretation of these amazing datasets. The analysis of raw NGS datasets requires domain knowledge in bioinformatics and informatic technologies. This limitation can be partially overcome by the development and implementation of end-to-end informatics infrastructure allowing on-demand and standardised bioinformatics analysis. These informatics infrastructures can be provided through interactive and simplified web-services. Nevertheless, web-services need continuous maintenance and implementation to keep-up with the continuous advances of back-end and front-end software as well as NGS assays (e.g. web-frameworks

and operative systems, or assays for single cell analysis and third generation sequencing technologies).

Before NGS, we were already able to identify, with a certain grade of confidence, relevant genomic aberrations, such as chromosomal rearrangements or small deletions. NGS technologies provided the capability of sequencing genomes in large scale and on-site. Today, NGS and bioinformatics are used to discover new recurrent genomic aberrations to help the stratification of patients into new risk groups. However, effective personalised medicine requires the understanding of causative effects and interaction between several genetic aberrations carried by each patient. In addition, the non-coding part of the human genome still needs to be characterised. Non-coding genome elements, such as small and long non-coding RNAs, act as transcriptional regulators and they have been studied for decades, but their clear role in diseases and cancers is still partially unknown.

Machine learning has been largely used in bioinformatics and genomics. However, the drawback of machine learning is the requirement of user-selected features and the design of a clean dataset for the development of the model. Deep learning overcomes this limitation by embedding the feature selection within the model itself. Deep learning has been recently used in several fields of research, from physics to genomics, leading to remarkable results. The herein thesis demonstrated the application of deep learning for the study of small-non coding RNA and interactions from a massive amount of genomic data. Deep learning models showed the capability of extracting important features that unambiguously characterise miRNA bindings rules. The genomics era is leading to an exponential growth of NGS

public datasets and clinical information. Today, DL technology is already deployed in clinical imaging (e.g. radiology) for the identification and classification of solid cancers (such as lung cancer). The next challenge of genomics is the integration of DL models as a standard tool for both biological research and precision medicine.

## Acknowledgements

First, I would like to thank both my supervisors Prof. Andrea Biondi and Dr Vojtěch Bystrý to give me this opportunity of carrying out a co-join PhD program between Masaryk University and University of Milan-Bicocca. This great and unique opportunity allowed me to work with the excellent Ph.D. consultants Prof. Giovanni Cazzaniga and Dr. Panagiotis Alexiou, PhD. Altogether, their immense knowledge, motivation and patience have given me more power and spirit to excel in the research writing. Conducting the academic study with such a great amount of experience in different fields of data analysis and bioinformatics applied to medical research would not be possible without all of you.

Apart from my Supervisors and Consultants, I will not forget to express the gratitude to rest of the teams in the Czech Republic and Italy for giving the encouragement and sharing insightful suggestions. They all have played a major role in polishing my research skills.

In the end, I am grateful to Masaryk University, University of Milan-Bicocca and CEITEC to provide such an exciting research environment as well as freedom to carry out my research interests.