Università
degli Studi
di Padova

Sede amministrativa: Università degli Studi di Padova
Dipartimento di Salute della donna e del bambino

Corso di dottorato di ricerca in Medicina dello sviluppo e
scienze della programmazione sanitaria
Curricolo: Emato-oncologia, genetica, malattie rare e medicina predittiva
Ciclo XXIX

# CircRNAs: the transcriptional landscape of haematopoiesis at higher definition

Tesi redatta con il contributo finanziario della
Fondazione “Istituto di Ricerca Pediatrica Città della Speranza”

Coordinatore: prof. Carlo Giaquinto
Supervisore: prof. Giuseppe Basso
Tutor: dott. Geertruij te Kronnie
prof. Stefania Bortoluzzi

Dottoranda: Annagiulia Bonizzato

# Abstract

Cell states in haematopoiesis are controlled by complex circuits, involving master regulators transcription factors and a growing family of RNA species, shaping cell phenotype, its maintenance and plasticity. Amongst RNA species, circular RNAs (circRNAs) are rapidly gaining the status of particularly stable transcriptome members with distinctive qualities. Regarding molecular functions, circRNAs modulate host gene expression, compete for binding of microRNAs, RNA-binding proteins and translation initiation, and participate in regulatory circuits. RNA-seq studies identified thousands of circRNAs with developmental stage- and tissue-specific expression corroborating earlier suggestions that circular isoforms are a natural feature of the cell expression program. CircRNAs are abundantly expressed and highly regulated also in the haematopoietic compartment, as described by recent and preliminary studies on circRNAs in blood cells.

In my PhD project we focused on the development of a bioinformatics pipeline to detect, quantify and characterize circRNAs from RNA-seq data, by combining both publicly available tools and custom scripts. Aiming to increase the discovery power of the pipeline as well as results robustness, we combined four programs for circRNA detection in parallel.

The pipeline was tested on a publicly available dataset of haematopoietic lineage cells such as Haematopoietic Stem Cells, Lymphoid progenitors, Myeloid Progenitors and Megakaryocyte–Erythroblast Progenitors. This pilot analysis allowed to retrieve a great number of circRNAs despite features of the data that were not optimal for circRNA detection. Major results from the pilot study were the identification of distinct sets of circRNAs specifically expressed in different cell types, and the feasibility and convenience of circRNA detection in published datasets to complement the original studies.

In parallel, we studied circRNA and linear RNA expression in differentiated cells of the haematopoietic compartment, specifically B cells, T cells and Monocytes. We produced RNA-seq data of 12 samples, obtained by cell sorting from peripheral blood of healthy donors and using ribosomal RNA depletion for the library construction. Out of the over 115 000 detected backsplices supported by at least two reads, we selected putative circRNAs found by at least two methods, gaining also indirect support that most of them are truly circular forms thanks to independent evidence. This subset consists of 26 211 circRNAs expressed by 7307 different genes, with 38.6 % of genes expressing one circRNA each, and 40.7 % of genes producing from 2 to 5 different circular isoforms and the remaining genes expressing a higher

number of circRNAs. The large majority of circRNAs are exonic, 11.5 % have backsplice ends falling into intronic regions and only a few (2.5 %) probably derive from genomic regions annotated as intergenic. Comparison with the analysis of the linear transcriptome pointed out that the expression levels of linear and circular RNAs expressed from the same gene have only a very slight tendency toward positive correlation, with most of the pairs showing scarce or even negative correlations, suggesting specific regulatory mechanisms underlying the expression of circRNAs.

The comparison between B cells, T cells and Monocytes indicated groups of circRNAs expressed in all the cell types and specific of each cell type. Unsupervised analyses of expression profiles showed for the first time specificities of circRNA expression associated to different blood cells. B cells and T cells circRNA-omes are similar from quantitative and qualitative points of view, whereas Monocytes express a lower number of circRNAs and have a more specific circRNA-ome. Indeed, differential expression tests outlined sets of circRNAs with significantly variable expression in B cells compared to Monocytes (2589), B cells compared to T cells (168) and Monocytes compared to T cells (977). Differentially expressed circRNAs are associated to genes enriched in protein products involved in key blood processes and pathways. Finally, we focused on 74 circRNAs upregulated in B cells compared to both Monocytes and T cells, 40 upregulated in T cells and 159 upregulated in Monocytes, for a total of 273 circRNAs with differential expression and cell specificity. Additional criteria for circRNA prioritization selected circRNAs associated to genes with key functions in haematopoiesis, or altered/deregulated in haematologic malignancies. Prioritized circRNAs will undergo experimental validations. The sequence analyses for *in silico* prediction of possible circRNAs functions, as presence of multiple miRNA binding sites, protein binding motifs, or open reading frames, will be the starting point for experimental studies to better elucidate the functions of more promising circRNAs.

In conclusion we performed the first study of circRNAs in normal B cells, T cells and Monocytes grounding on several biological replicates of each cell type being informative on circRNA differential expression. The integration of circular and linear RNA expression profiles with gene annotations and functions, in conjunction with differential expression data, produced new and original results. We showed that taking into account circRNA expression might add definition to the representation of transcriptome variations in normal haematopoiesis, posing the basis to better comprehend the role of circRNAs in the regulatory circuits of blood cells differentiation, which is a prerequisite for transferring this knowledge to research on haematological malignancies.

# Sommario

Il differenziamento cellulare durante l'ematopoiesi è controllato da circuiti complessi, che coinvolgono fattori di trascrizione e diverse specie di RNA che concorrono a stabilire il fenotipo delle cellule e a mantenerlo, e ne assicurano anche la plasticità. Negli ultimi anni sono emerse chiaramente la diversificazione e l'importanza di varie classi di RNA non codificanti. Tra questi, gli RNA circolari (circRNA), prodotti mediante backsplicing di trascritti primari, si stanno rapidamente affermando come membri del trascrittoma particolarmente stabili e con ruoli biologici rilevanti, prevalentemente regolativi. Per quanto riguarda le loro funzioni molecolari i circRNA sono in grado di modulare l'espressione del gene da cui derivano, possono competere per il legame di microRNA, regolando quindi l'espressione dei loro target, ma anche interagire con proteine che legano l'RNA modulandone le funzioni. Diversi circRNA la cui funzione è stata chiarita recentemente partecipano ad importanti assi o circuiti regolatori, intervenendo in processi chiave, di grande rilevanza anche in ambito oncologico, quali la regolazione del ciclo cellulare, il controllo dell'espressione di oncogeni e l'attivazione di specifiche vie di segnale. Studi di RNA-seq hanno identificato migliaia di circRNA con espressione specifica per lo stato di sviluppo o per il tipo di tessuto, corroborando precedenti indicazioni che le isoforme circolari siano una sfaccettatura del programma cellulare, tanto interessante quanto precedentemente sottovalutata. I circRNA sono molto espressi e fortemente regolati anche nel comparto ematopoietico, come mostrato da alcuni studi preliminari sulla loro presenza nelle cellule del sangue. L'identificazione dei circRNA mediante RNA-seq si basa sulla ricerca di backplice, ovvero di sequenze che non mappano linearmente sul genoma ma che sono formate dalla fusione di due sequenze in maniera non colineare, e ciò richiede specifici metodi computazionali.

In questo progetto di dottorato è stata sviluppata una pipeline bioinformatica che consente di identificare, quantificare e caratterizzare i circRNA a partire da dati di RNA-seq, mediante quattro metodi computazionali già disponibili utilizzati in parallelo, e di combinare ed elaborare i risultati grazie a una serie di programmi scritti appositamente.

La pipeline è stata testata su un dataset di cellule del lineage ematopoietico disponibile nei database pubblici, che contiene dati di sequenziamento di cellule staminali ematopoietiche, di progenitori linfoidi, di progenitori mieloidi e di progenitori di megacariociti ed eritroblasti. Questa analisi pilota ci ha consentito di identificare molti circRNA nonostante le caratteristiche dei dati non fossero ottimali per questo tipo di analisi. I principali risultati di questo studio pilota sono stati l'identificazione

di sottogruppi distinti di circRNA specificamente espressi in diversi tipi cellulari, e l'indicazione di fattibilità e convenienza dell'applicazione di questo approccio anche su dati già pubblicati per ampliare e complementare gli studi originali che non avessero preso in considerazione i circRNA.

Il progetto principale si è quindi focalizzato sull'analisi dell'espressione di circRNA e RNA lineari in cellule differenziate del comparto ematopoietico. Sono stati prodotti dati RNA-seq di linfociti B, linfociti T e monociti ottenuti tramite sorting da sangue periferico di donatori sani, per un totale di 12 campioni ad alta profondità di sequenziamento e processati mediante un protocollo di sottrazione dell'RNA ribosomale particolarmente adatto per lo studio dei circRNA. Degli oltre 115.000 backsplice identificati da almeno 2 reads di sequenziamento considerando l'insieme dei 12 campioni analizzati, sono stati selezionati 26.211 circRNA identificati da almeno due metodi computazionali. Considerato che studi precedenti basati sull'arricchimento di circRNA in seguito al trattamento con RNAsi R hanno chiarito che i backsplice identificati da almeno due metodi indipendenti sono più affidabili, questo insieme di 26.211 circRNA selezionati dovrebbe risultare robusto. Essi risultano espressi da 7.307 geni diversi, di cui il 38,6% esprime un solo circRNA per gene, il 40,7% produce da 2 a 5 isoforme circolari e i restanti geni ne esprimono 6 o più. La maggioranza dei circRNA identificati è esonica, l'11,5% ha gli estremi della giunzione che mappano su regioni annotate come introniche nel genoma, e solo il 2,5% probabilmente deriva da regioni genomiche annotate come intergeniche.

I livelli di espressione dei circRNA e degli RNA lineari espressi dallo stesso gene hanno una leggere tendenza a correlare positivamente, mentre la gran parte delle coppie mostrano scarsa o negativa correlazione, suggerendo che ci siano dei meccanismi di regolazione specifici che sottendono all'espressione di circRNA. Questo dato è in linea con studi recentissimi che presentano lo splicing alternativo delle isoforme circolari come un ulteriore meccanismo che genera complessità nello splicing dei trascritti eucariotici.

L'analisi non supervisionata dei profili d'espressione dei circRNAs in linfociti B, linfociti T e monociti ha mostrato per la prima volta la specificità dell'espressione di circRNA associata ai tipi cellulari considerati. I circRNAomi di linfociti B e T risultano simili sia dal punto di vista qualitativo che quantitativo, mentre invece i monociti esprimono un numero minore di circRNA e hanno un circRNAoma più specifico. Il confronto tra tipi cellulari ha indicato gruppi di circRNA espressi in tutti e tre i tipi cellulari, e altri specificamente espressi in un solo tipo. L'analisi statistica dell'espressione differenziale ha evidenziato dei gruppi di circRNA con espressione significativamente diversa nei linfociti B confrontati con monociti (2589), linfociti B confrontati con linfociti T (168) e monociti confrontati con linfociti T (977). CircRNA

differenzialmente espressi sono associati a geni le cui proteine sono coinvolte in processi e pathway chiave nel comparto ematopoietico. Infine sono stati evidenziati circRNA differenzialmente espressi e specificamente up-regolati in un solo tipo cellulare: 74 circRNA risultano up-regolati nei linfociti B in confronto a monociti e linfociti T, 40 nei linfociti T e 159 nei monociti, per un totale di 273 circRNA con espressione differenziale e specificità cellulare. I circRNA differenzialmente espressi, altri con altissima espressione o derivati da geni particolarmente importanti nell'ematopoiesi normale o maligna, sono stati selezionati per ulteriori analisi *in silico* e alcuni verranno validati sperimentalmente. L'analisi della sequenza dei circRNA per la predizione *in silico* di siti di legame multipli per miRNA, motivi di legame per proteine oppure open reading frames, fornirà utili predizioni funzionali e sarà anche punto di partenza per studi sperimentali focalizzati su alcuni circRNA particolarmente promettenti.

In conclusione, questa tesi costituisce il primo studio sui circRNA in linfociti B, linfociti T e monociti sani, fondato su replicati biologici di ogni tipo cellulare. L'integrazione dei profili d'espressione di circRNA e RNA lineari con l'annotazione dei geni e le funzione, congiuntamente all'espressione differenziale, ha prodotto risultati nuovi e originali. Lo studio è molto informativo sull'abbondanza e la diversificazione dei circRNA espressi e fornisce numerosi nuovi dati sui circRNA, nonché robuste indicazioni sull'espressione differenziale dei circRNA nelle cellule considerate. Abbiamo dimostrato che considerare l'espressione dei circRNA aggiunge definizione alla rappresentazione delle variazioni del trascrittoma nell'ematopoiesi normale, ponendo le basi per ampliare la comprensione del ruolo dei circRNA nei circuiti regolatori del differenziamento delle cellule del sangue, prerequisito per trasferire queste conoscenze alla ricerca nell'ambito delle patologie ematopoietiche.

# Contents

# 1 Introduction

## 1.1 Transcriptome variation and regulatory circuits in haematopoiesis

Human blood production releases each day $10^{11}$ blood elements into the blood flow, mostly red blood cells and platelets, and leukocytes. These blood cell types derive from a small population of haematopoietic stem cells (HSCs), that expands and differentiates into progenitor cells with increasingly restricted lineage choice. Studying haematological malignancies and model organisms [1] helped to identify many critical genes and mechanisms regulating haematopoietic development. Underlaying haematopoiesis a large number of interleaved molecular circuits are regulated by sets of transcription factors [2].

Together with protein regulators and effectors, the mechanisms are also regulated by non-coding RNAs (ncRNAs). The action of miRNAs for example has been characterised in diverse phases of the haematopoiesis.

For example Chen *et al.* in their 2014 study [3] described that in less differentiated stages the majority of sequenced transcripts derived from non-coding genes, while in more differentiated stages it was the opposite. The non-coding family is wide and varied, and alongside with miRNAs contains long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs), a novel class of ncRNAs with a strong regulatory potential. CircRNAs are covalently closed RNA molecules produced by a process of "backsplicing", a particular type of alternative splicing [4], and have been demonstrated to be present in many different cell types and also in blood cells [5].

Alternative splicing has an important role in disease, with 15 % of disease-causing constitutional mutations located within splice sites and more than 20 % of missense mutations lying within predicted splicing elements [6]. Studies have also revealed that somatic mutations of splicing factor genes occur frequently in haematological cancers, including myelodysplasia and chronic lymphocytic leukaemia [7–9].

## 1.2 CircRNAs

CircRNAs are covalently closed RNA molecules, in which the 3′- and 5′-ends are linked in a non-collinear way by a process called backsplicing [4]. Unlike in linear RNA splicing, a splice donor site is joined to a splice acceptor site upstream

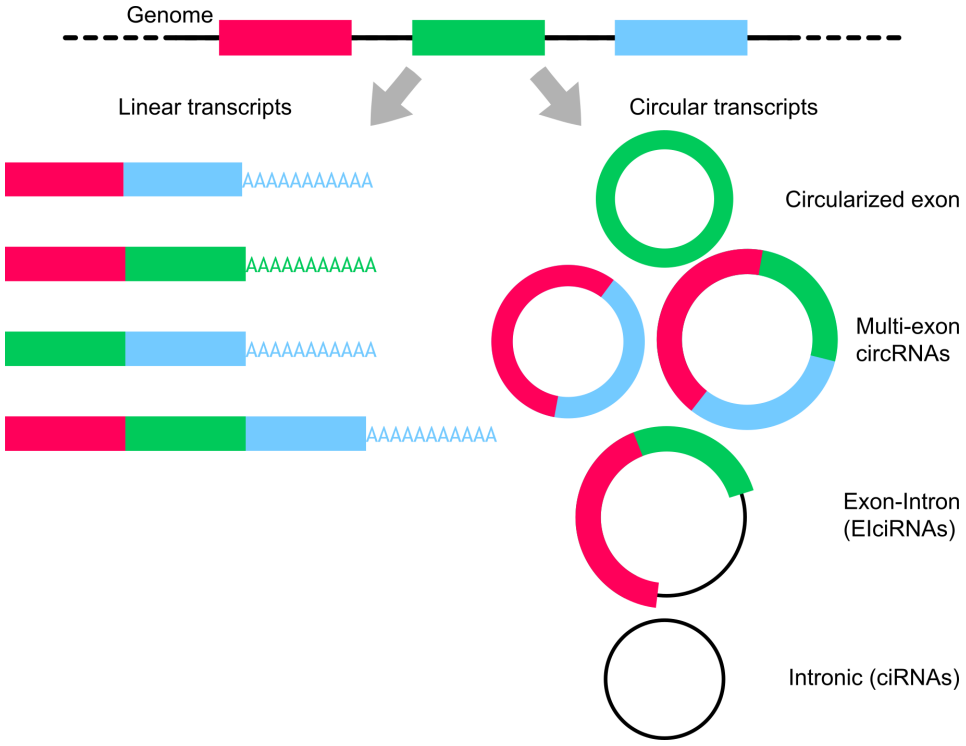



Figure 1.1: CircRNAs are produced by backsplicing and combinations of exons and introns give rise to different products, including single circularised exons, circRNAs formed by two or more exons, by exon and retained intron sequences (EI-circRNAs) and by intronic sequences only.

in the primary transcript, yielding a circRNA [4]. CircRNAs can be formed by circularisation of a single exon, two or more exons [10], both exon and intron sequences (exon–intron circRNA, EI-circRNA) [11] or intronic sequences only (circularised intron RNA, ciRNA) [12] (Figure 1.1). Several circular isoforms can be produced from a given gene and different circRNAs from the same gene may show distinct expression profiles, as reported for *circSTAU2a* and *circSTAU2b* [13].

Circularity confers specific properties to circRNAs: they are highly stable, resistant to RNAse R, have longer half-lives compared with linear RNAs [14, 15] and tend to accumulate in cells with a low proliferation rate [16]. Detection of circRNA in human body fluids such as plasma [17] and saliva [18] indicates circRNAs as potential disease biomarkers.

The first description of circRNA dates back to several decades ago. Recently, circRNAs were relaunched by RNA-seq-based studies as an RNA species with high relevance for molecular biology and molecular oncology and today over 10 000 human circRNAs have been identified [19–25].

CircRNAs are non-poly-adenylated and coincidental discovery of circRNA in the past can be attributed to RNA extraction methods that mainly used polyA selection. Naturally occurring single-stranded covalently closed RNA molecules were first

described in plant viroids [26] and were valued for their peculiar structure that allows for rolling circle replication [4]. A few studies in the nineties reported non-canonical splicing with scrambled exons of candidate tumour suppressor gene *DCC* [27], 'missplicing' of *ETS1* transcripts [28] and murine *Fmn* [29], and exon circularisation in nuclear extracts [30]. Moreover, whereas early in development the mRNA of therian *SRY* is translated into the protein that triggers the sex-determining transcriptional cascade, in adults *SRY* transcripts are found as cytoplasmic circular *SRY* (*cSRY*) not particularly bound to polysomes [31] and later proven to efficiently sponge miR-138 [32]. Other primary RNAs were found to be processed into circRNA isoforms such as *MLL* (*KMT2A*) [33], *ETS1* [34], *CYP2C18* [35], *SLC8A1* [36] and dystrophin (*DMD*) [37] transcripts. Examples of circRNAs corresponding to linear non-coding RNA, as well as antisense RNA were also detected [38, 39].

Most of the above-mentioned studies postulated relevant biological functions for circular RNAs but were only confined to certain genes. In any case, shortly after publication of circular forms of an INK4/ARF-associated non-coding RNA [38] numerous studies embarked on transcriptome-wide circRNA analysis showing developmental stage- and tissue-specific expression, and specific regulatory roles for circRNAs were suggested [10, 21]. These new data triggered the interests of the scientific community resulting in the development of molecular methods to study circRNAs and of microarray platforms to measure expression levels of circRNAs (Figure 1.2), as well as the implementation of bioinformatics software to detect and discover circRNAs from RNA-seq data (Table 1.1, p. 21) posing the basis for further experimental circRNA characterization, and for circRNA quantification and differential expression testing. Moreover, several available circRNA databases and web resources (Box 1.1) could be rather useful to explore putative circRNA interactions and functions.

Several studies on circRNAs in blood cell types and haematologic malignancies were recently conducted and will be discussed below.

### 1.2.1 CircRNAs biogenesis

**Backsplicing** As anticipated, circRNA loops are generated by backsplicing from immature RNA, where ends are joined in a non-collinear way. CircRNAs are derived from Pol II transcripts just like linear transcripts. Backsplicing requires the spliceosomal machinery [45–47] as revealed by treatment of HeLa cells with a splice inhibitor followed by nascent RNA purification.

In the majority of cases, the generation of circRNA happens at the expense of their corresponding mRNA isoforms and is characterised by the usage of canonical splice sites that precisely flank head-to-tail junctions of circular transcripts. In

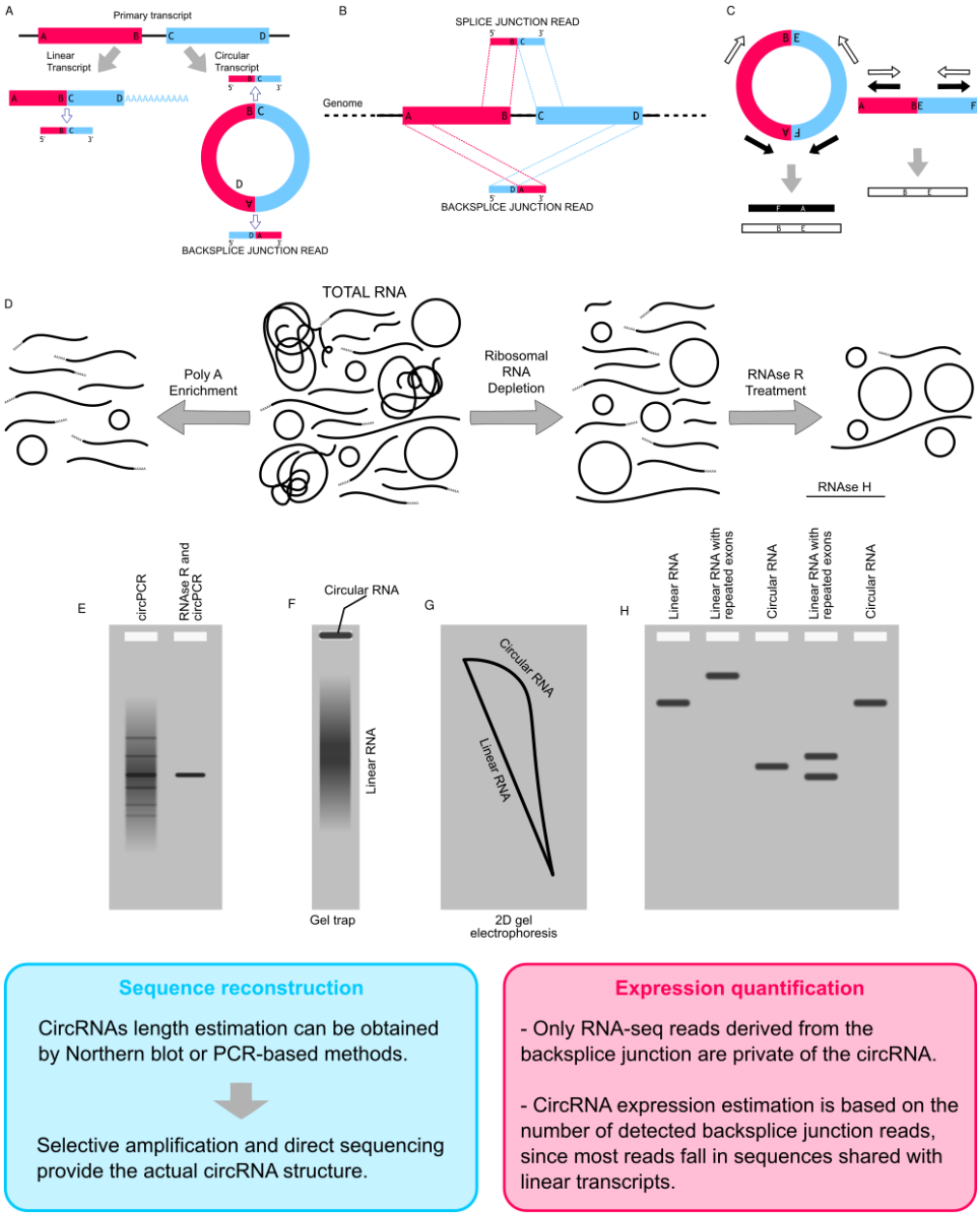



Figure 1.2: Molecular methods for circRNA detection, validation and study. (A) CircRNA detection from RNA-seq data grounds on the identification of sequence reads encompassing the backsplice junction. (B) Backsplice reads map to the genome in chiastic order (two segments of a single read align separately in reverse order) due to the backsplicing in circRNAs biogenesis. (C) Convergent primers (white arrows) designed in adjacent spliced exons amplify both linear and circular isoforms, whereas primers that are divergent in the linear transcripts (black arrows) can be used to specifically amplify the circular isoform. (D) PolyA enrichment protocols deplete circRNAs, whereas ribosome depletion and RNAse R protocols enrich circRNAs. (E) RNAse R digestion before reverse-transcription PCR lowers the amount of false-positive amplicons facilitating circRNA validation. (F) Gel trap electrophoresis allows isolate the circular and linear fractions of the input RNA, as circRNAs are hold in the well. (G) Two-dimensional acrylamide gel electrophoresis separates the circular RNA fraction in an off-diagonal curve. (H) RNA migration in agarose gel before and after a mild RNAse H treatment resulting in a single cut per molecule shows that circular molecules bearing a backsplice junction are discriminated from linear ones deriving from a duplication event, as only circRNA results in a single band after being cut once. (F)–(H) re-elaborated from Jeck & Sharpless, 2014 [44].

**circBase** Simple circRNA database that provides a searchable table of circRNAs identified by several studies (five on human data) that can be useful to sort newly identified from known circRNAs [40].
http://www.circbase.org/

**starBase v2.0** Focuses on RNA–RNA and protein–RNA interaction networks inferred from CLIP-Seq data sets. Among others, in the miRNA–lncRNA section, it includes predicted miRNA–circRNA interactions and can be searched to identify known circRNA that potentially sponge a specific miRNA [41]. http://starbase.sysu.edu.cn/

**Circ2Traits** Useful to explore potential associations of circRNAs with diseases based on predicted interactions of circRNAs with disease-associated miRNAs and on the overlap between disease-associated SNPs to circRNA *loci* [42]. http://gyanxet-beta.com/circdb/

**CircNet** Provides expression profiles of circRNAs in 464 RNA-seq samples, with circRNA sequences and annotations in term of overlapping genes and interactions [43]. http://circnet.mbc.nctu.edu.tw/

Box 1.1: Main web resources dedicated to circRNAs.

some cases, transcripts of specific genes are predominantly spliced into the circular isoform [19, 20]. Ashwal-Fluss *et al.* [45] demonstrated that circularisation and splicing of linear forms compete against each other. Kelly *et al.* [48] confirmed a direct correlation between exon skipping and circularisation. Thus, circRNA biogenesis and regulation of mRNA production are tightly linked.

**CircRNAs regulatory features** CircRNA-forming exons are often flanked by particularly long introns, possibly reducing splicing efficiency [20, 24]. Moreover, in humans these long introns are markedly enriched in ALU repeats [20] and complementary sequences in introns are involved in specific folding of primary transcripts that favour circularisation [45]. In the *Sry* gene, the activation of an upstream promoter triggers the synthesis of a primary transcript containing inverted repeats needed for circRNA production [49]. As drosophila RNA circularisation does not appear to be driven by structural complementarity of exon-bordering sequences but only determined by the length of exon-flanking introns, inverted repeats alone do not fully explain the production of circRNAs in eukaryotes. In addition, regulation of the dynamic expression of circRNAs in different cell types is likely also dependent on control by *trans*-acting factors [50].

**_Trans_-acting factors** Besides the role of flanking sequence elements, introns encasing circRNAs are highly enriched in RNA A-to-I editing events [20]. In fact, knock-down of RNA-editing enzyme ADAR1 upregulated circRNA expression, favouring a mechanism of circRNA biogenesis whereby ADAR1 antagonises circRNA expression by melting stems of RNA-RNA interactions within introns that putatively promote circularisation [20].

The Muscleblind (*MBL*) family of splicing factors was also shown to take part in the regulation of circRNA production by binding specific intronic sites flanking circularised exons [45]. Intriguingly, in the fly circRNA isoform expression of *MBL* itself is regulated by MBL protein. Decrease of circRNA expression after *MBL* knock-down supports a circRNA-promoting role for MBL proteins [45].

In addition, RNA-binding protein (RBP) Quaking (QKI) regulates the formation of circular RNAs [51]. QKI dimers bind to specific bipartite sequences termed QKI response elements that are present in many RNAs, including coding mRNAs and primary miRNAs. Conn *et al.* [51] investigated the role of QKI in promoting circRNA biogenesis in transforming growth factor-β-induced epithelial-mesenchymal transition of the epithelial HMLE cell line, demonstrating that the knock-down of the QKI-5 isoform specifically decreases the formation of circRNAs and that insertion of synthetic QKI response elements in introns mediates circRNA formation. Metabolic tagging of nascent RNAs with 4-thiouridine has been used to study the link between nascent circRNA processing and transcription [25] showing that the efficacy of circRNA processing from primary transcripts is extremely low. This study also clarified that circRNAs are largely processed post-transcriptionally and confirmed that circRNAs are stable, being thus abundant at a steady-state level and tending to accumulate particularly in cells with low proliferation rates [16].

**Backsplice as a new form of alternative splicing** Alternative RNA splicing is a complex tightly regulated phenomenon. Since the discovery of split genes robust knowledge was built on splicing prevalence, on complexity of splicing patterns and on molecular mechanisms that determine, regulate or change splicing, including RNA-protein interactions (splicing factors with *in cis* regulatory sites termed silencers or enhancers), RNA-RNA base-pairing interactions involving both *in trans* acting RNAs and *in cis* secondary structure formation, and also chromatin-based effects [52].

Disease-causing mutations occur in splice sites or in regulatory elements, as well as in genes that encode splicing factors (*U2AF1*, *SRSF2*, *SF3B1* and *ZRSR2*) and there is much interest in developing antisense oligonucleotides to control splicing patterns and using genome editing to correct disease-causing splicing defects.

Alternative splicing is highly and commonly deregulated in cancer cells [53–55] and specifically impacts prognosis and disease course of myeloid malignancies, including chronic lymphocytic leukaemia, acute lymphoblastic leukaemia (ALL), acute myeloid leukaemia and myeloproliferative neoplasms [56–59].

As circRNA biogenesis and splicing are interleaved processes, it can be hypothesised that mutations of splicing factors and/or alterations of regulatory elements have an impact on circRNA biogenesis. RBPs involved in circRNA biogenesis might drive developmental regulation of circRNA formation and show deregulation in disease. Distinct expression levels of ADAR1, MBNL1 and QKI in normal bone marrow compared with B-cell leukaemia subtypes (Figure 1.3) encourage investigations as also subtle expression variations of ADAR1 were shown to be relevant for RNA circularisation [13].

Alternative splicing is a key mechanism through which fundamental processes during haematopoiesis are regulated [3], posing the basis to interpret the consequences of genetic variation. Similarly, there is high demand to study circRNA in normal haematopoiesis, to connect biogenic mechanisms with biological functions of circRNAs, to accumulate fundamental knowledge needed to understand disease mechanisms and to inform strategies for therapeutic intervention.

**CircRNA degradation** As circRNAs are endogenous cell products one might ask which endogenous mechanisms cells have to dispose of circRNA. In general, RNA is degraded by the exosome, a multiprotein complex that reminiscently of the proteasome forms a chamber with helicase activity, which unfolds and then degrades RNA. The degradation is prevalently exoribonucleolytic from the 3′-end, but the exosome catalytic subunit RRP44 also has endonuclease activity [61].

According to available evidences circRNAs are not degraded by treatments (as tobacco acid pyrophosphatase plus terminator 5′-phosphate dependent exonuclease or highly processive 3′- to 5′-exoribonuclease RNAse R digestion) that normally degrade linear RNA with free ends [39]. Regarding degradation, in general miRNAs can regulate cleavage of circRNAs. The better characterised path toward degradation of a circRNA is that of *CDR1-as* (circular antisense transcript deriving from cerebellar degeneration-related protein 1 locus) that even presenting multiple miR-7 binding sites is completely resistant to miR-7-mediated degradation and also resistant to miR-769-mediated degradation, whereas the binding of miR-671 to *CDR1-as* directs Ago2-slicer-dependent cleavage [39].

Undoubtedly, our understanding of the regulation of circRNA turnover and endogenous degradation mechanisms is limited. We can hypothesise that not only the deregulation of circRNA synthesis but also its degradation are biologically relevant.

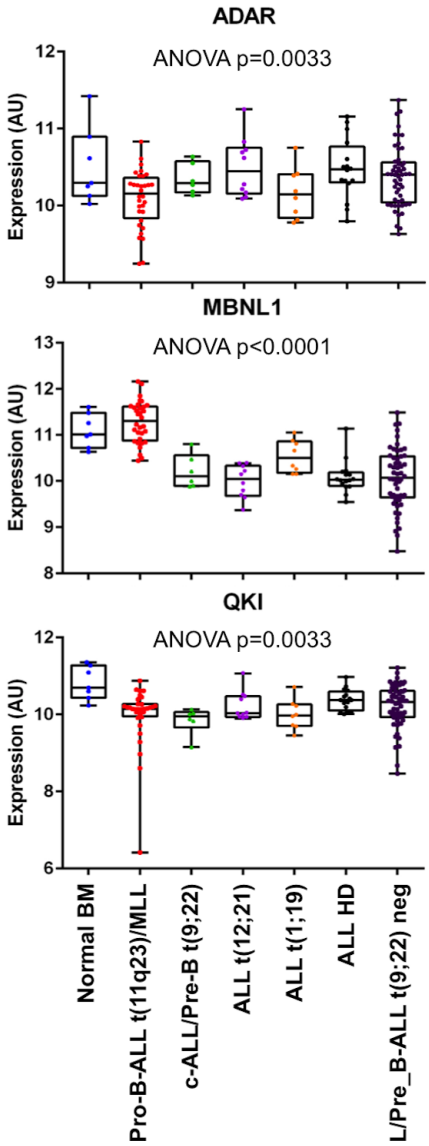



Figure 1.3: Expression variation of enzymes involved in circRNA expression. Gene expression intensities of ADAR1, MBNL1 and QKI in samples of normal bone marrow and six B-cell leukaemia subtypes carrying specific genetic aberrations (according to Haferlach *et al.* [60]); expression data obtained with HG-U133 Plus 2.0 (Affymetrix, Santa Clara, CA, USA).

For instance, a lack of cleavage might result in undesired circRNA accumulation.

### 1.2.2 Conservation and functions

**CircRNAs are evolutionarily conserved** CircRNAs were described in many eukaryotes from yeast to humans [62] and resulted very conserved at the nucleotide level: Memczak *et al.* [21] analysed sequence conservation within circRNAs and showed that 223 human circRNAs with conserved circularisation in mice were significantly more conserved in the third codon positions than exons not engaged in circular forms. CircRNAs are also depleted of polymorphisms in miRNA-binding sites [63]. Beyond apparent sequence conservation, both paralogous and orthologous gene pairs have been reported to express circular transcripts: human *HIPK2* and *HIPK3*, as well as murine *Hipk2/3* produce circRNAs [20]. Also conservation of circRNAs in terms of exonic sequences, bordering intronic sequences, precise backsplice junctions and expression patterns in mammals and to some extent in *Drosophila* has been recently reported [13]. The above indications of evolutionary preservation point to a central position of circRNAs in core biological processes.

**CircRNAs are seldom translated** It has previously been shown that eukaryotic ribosomes can initiate translation on circRNAs containing internal ribosome entry site elements (Figure 1.4) producing long-repeating polypeptides in the presence of a continuous open reading frame [64, 65]. Efficient circRNA translation can occur in HEK293 [47] and HeLa cells [66]. Intriguingly, a circRNA can be a sort of '*Möbius strip*' with translation generating proteins either recurrently, or variably depending on whether or not the sequence length is a multiple of three nucleotides. A small viroid circRNA directly translated through three completely overlapping open reading frames shifting to a new reading frame at the end of each round has been reported as a natural supercompact 'nanogenome' [67].

Even if in principle circRNAs can be translated, the majority of recently discovered and characterised circRNAs seem to have limited coding potential: seldom associated neither with messenger ribonucleoprotein particles nor with translationally active polyribosomes, suggesting that circRNAs, as a species, are unlikely to be translated into peptides [68]. The fact that in the same study mass spectrometry failed to identify peptides encoded by backsplice junctions of circRNAs could be due to low sensitivity or to the position of open reading frames outside junctions and does not rule out that part of circRNAs can be translated in some cell types and/or conditions.

A circRNA produced by murine *Fmn* [29] contains an active translation start site not leading to protein synthesis. In this way, the circular form competes

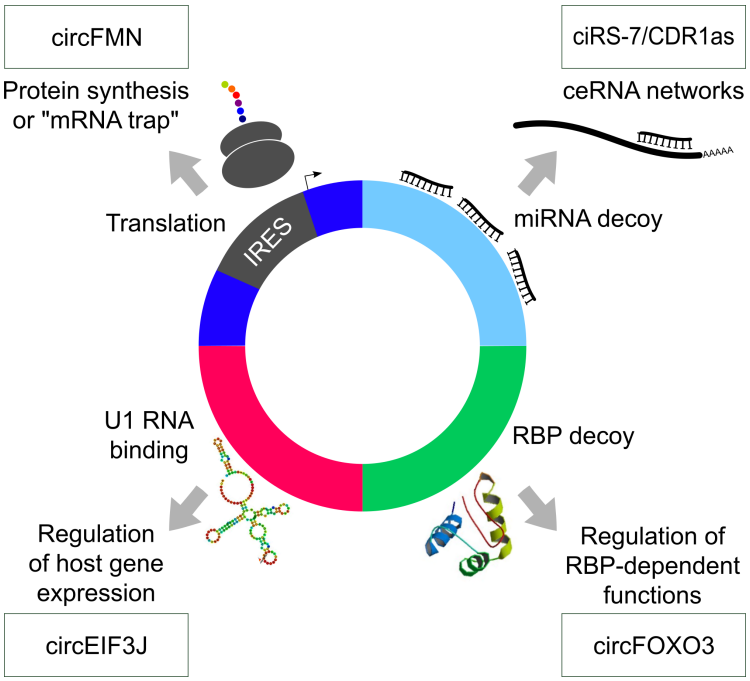



Figure 1.4: CircRNA functions. Elucidated circRNA functions include the ability to sponge miRNAs thus regulating the silencing of canonical targets (for example, *CDR1-as*/*ciRS-7* harbours 63 binding sites for miR-7) and participating to ceRNA networks; similarly circRNAs could decoy RBPs ultimately regulating the functions in which RBPs are implicated (for example, *circ-Foxo3* forms a ternary complex with p21 and CDK2 arresting cell cycle progression); circRNAs can also regulate in cis the expression if the gene from which they derive through interactions with the U1 RNA in the U1 RNP in the nucleus (for example, *circEIF3J*); moreover, circRNAs harbouring an IRES could be translated to produce peptides or compete with mRNA translation (for example, *circFMN* contains an active translation start site not leading to the protein synthesis).

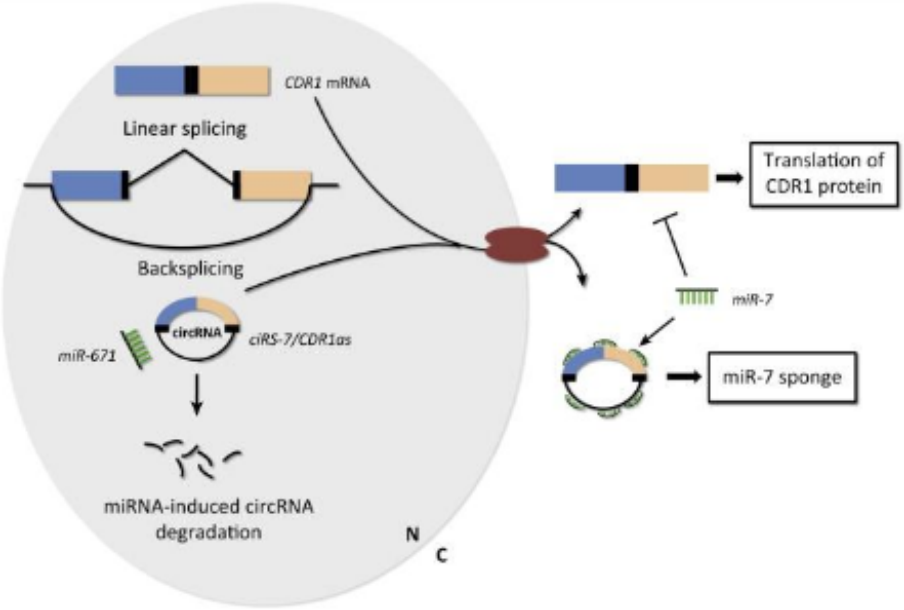



Figure 1.5: *CDR1-as* mechanism of action as a ceRNA. *CDR1-as* acts as a sponge for miR-7, resisting miR-7 mediated degradation whereas it is sensible to miR-671 mediated degradation. Figure from Guil & Esteller, 2015 [71].

with the linear mRNA both impacting on the linear transcript abundance and providing an 'mRNA trap' that can sequester proteins of the translation initiation complex. Also, Jeck & Sharpless [44] uncovered that many single-exon circRNAs contain a translation start site [69], further exemplifying this mechanism of protein expression regulation by circRNAs.

**CircRNAs are efficient miRNA sponges and participate to competing endogenous RNA networks** As demonstrated by several studies, circRNAs with multiple miRNA-binding sites are efficient miRNA sponges that participate in the regulation of specific cellular pathways (Figure 1.4) [32, 70].

*CDR1-as* harbours 63 conserved binding sites for miR-7 displaying high miRNA-binding capacity and miRNA antagonist activity in the brain (Figure 1.5) [21, 32]. Following this functional description *CDR1-as* was renamed as circular transcript *ciRS-7* (circular RNA sponge for miR-7), called also *CDR1-as*/*ciRS-7*. Notably, circRNA is completely resistant to miR-7-mediated target destabilization and strongly suppresses miR-7 activity, resulting in increased levels of miR-7 targets, including *EGFR* and *IRS2*.

The previously mentioned *cSRY* decoys miR-138, for which it displays 16 target sites [20, 32]. Circ-ITCH, a circRNA downregulated in carcinomas, was demonstrated to be a sponge for miR-7, miR-17 and miR-214, by increasing the level of ITCH and ultimately inhibiting the Wnt/β-catenin pathway [70, 72]. Another

circRNA (*hsa_circ_001569*) that acts as a sponge for miR-145 upregulates its targets enhancing cell proliferation and invasion of colorectal cancer [73].

Recently, specific circRNA–miRNA axes have been shown to regulate cancer-related processes. CircRNAs can have both cancer-promoting and -suppressing roles, depending on the molecular circuits in which they are involved and on the role of the interactors [70, 72, 74]. CircRNAs can exhibit anticancer effects: as synthetic circular sponges displayed superior anticancer activities compared with the linear sponges, RNA circles open new ways to deliver miRNA sponges with persistent effects [34].

CircRNAs like linear isoforms can act as competing endogenous RNAs (ceRNAs) that decoy miRNAs and indirectly regulate protein-coding gene expression (Figure 1.4) [75]. ceRNAs are implied in the progression of cancer and impact on cancer hallmarks [69]. Being resistant to miRNA-mediated degradation circRNA can presumably also tether RISC components depriving the cellular pool of both miRNAs and RISC effectors [70].

Following the observations of circRNAs acting as ceRNAs it has been asserted that such a mechanism may be common to all circRNAs. The latter finds support from previously mentioned shortage of polymorphisms in circRNAs' putative miRNA-binding sites. Other studies showed instead that only a minority of expressed circRNAs present multiple binding sites for specific miRNAs and according to their observations circRNAs are, in general, not bound to miRNA-loaded Argonaute proteins [44, 76]. In addition, argonaute co-immunoprecipitation experiments did not indicate an appreciable enrichment of circRNA-derived exons among argonaute family-bound transcripts, which would be expected if circRNAs were prevalently acting as ceRNAs. See Thomson & Dinger [77] for a review on evidences and open questions on endogenous miRNA sponges. According to available data, supported by recent findings [11, 72] we conclude that some circRNAs can act as ceRNAs, whereas others may be involved in a variety of other molecular mechanisms.

**Interactions with RBPs** The decoy activity of circRNAs could be important also for RBPs (Figure 1.4). CircRNAs like linear RNAs may interact with RBPs in a sequence-specific and structural motif determined way. CircRNAs could function to store, sort or localise RBPs. Recently, the interaction between *Foxo3* circular RNA and specific proteins was shown to delay cell cycle progression [78]. Foxo3 is a forkhead box O transcription factor and may behave as a tumour suppressor protein that limits cell proliferation and induces apoptosis and is frequently altered in cancer, shown to be deleted in lymphomas (diffuse large B-cell lymphoma) and translocated with *MLL* in leukaemia [79]. In healthy cells high *circ-Foxo3* expression

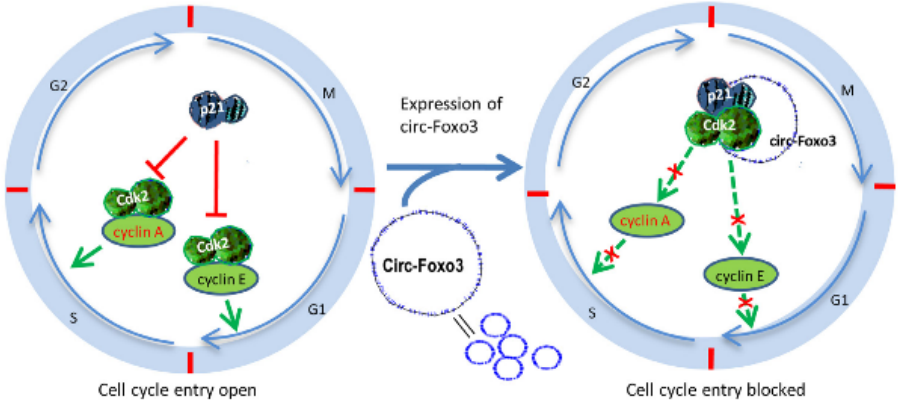



Figure 1.6: *Circ-Foxo3* dependent mechanism of cell cycle regulation. *Circ-Foxo3* plays a role in the regulation of cell cycle interacting with p21–CDK2 complex. Figure from Du *et al.*, 2016 [78].

was found to be associated with cell cycle progression. Silencing endogenous *circ-Foxo3* promoted cell proliferation, whereas ectopic expression of *circ-Foxo3* repressed cell cycle progression by binding to cell cycle proteins cyclin-dependent kinase 2 (CDK2) and cyclin-dependent kinase inhibitor 1 (or p21). Normally, CDK2 interacts with cyclin A and cyclin E to facilitate cell cycle entry, while p21 inhibits these interactions and arrests cell cycle progression. The formation of the circ-Foxo3–p21–CDK2 ternary complex arrests the function of CDK2 and blocks cell cycle progression (Figure 1.6). This study identified an oncogenic function of a circRNA and indirectly demonstrated that circRNAs can have distinct functions with respect to that of protein products encoded by the same gene [78].

***Cis* regulation of gene expression by circRNAs** Another line of evidence reported *cis*-regulatory roles for specific circRNAs. Exon–intron circRNAs derived from circularisation of RNA with intron retention were identified as a subclass of ciRNAs, enriched in the nucleus, associated with Pol II [11]. Further analyses of two exon–intron circRNAs (*circEIF3J* and *circPAIP2*) showed interactions with Pol II, U1 snRNP and parental gene promoters through sequence complementarity between the U1 snRNA and an U1-binding site, which eventually promote the transcription of the gene from which they derived (host genes), triggering a positive-feedback loop (Figure 1.7).

Zhang *et al.* [12] described circular intronic RNAs (ciRNAs) that were found to accumulate in human cells due to a failure in debranching and showed that knock-down of ciRNAs reduced expression of their host genes. One of these abundant RNAs, ci-ankrd52, largely accumulates to its sites of transcription, associates with the elongation Pol II machinery and acts as a positive regulator of Pol II transcription [12]. Apparently also non-coding intronic segments of ciRNA tran-

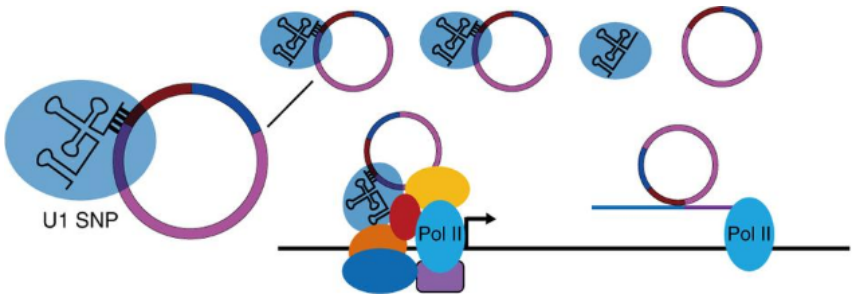



Figure 1.7: EI-circRNA mechanism of regulation of gene expression. EI-circRNAs interact with the transcription complex in the nucleus. Figure from Li *et al.*, 2015 [11].

scripts can have a *cis*-regulatory role. *CDR1-as* stabilizes CDR1 mRNA expression, probably with a sense–antisense-based feedback mechanism, where the antisense circRNA stimulates or stabilises the sense mRNA with subsequent negative impact on antisense levels [39].

The fairly well-characterized role of circRNAs to positively modulate host gene expression has foreseeable implications for stability of cell commitment choices such as in haematopoiesis.

In summary, circRNAs can regulate host gene expression but also participate to complex networks in which they compete for the binding of miRNAs (ceRNA networks) of RBPs or even for translation initiation. CircRNAs with different composition in terms of exon–intron inclusion result in a multitude of mechanisms that can affect transcriptome and proteome regulation. Through studies of specific tissues, cell types and conditions the evident versatility of circRNA is expected to reveal insight in all kinds of cellular processes.

### 1.2.3 CircRNAs in haematopoietic compartment

RNA-seq analyses showed dynamic expression of circular isoforms independent of linear transcript dynamics from the same gene [13] and cell- and differentiation stage-specific expression [10, 21] prompting lines of research that focused on specific tissues. Knowing that circRNAs interfere in key cellular processes like self-renewal, proliferation and apoptosis there is growing interest in studying circRNA in the haematopoietic compartment. In the haematopoietic tissue, pioneering studies reported circRNA isoforms of key genes such as *MLL* [33, 80] without receiving much resonance in the research community, probably due to the reported low expression of circular isoforms compared with abundant mRNAs encoding these key transcriptional regulators. In hindsight low expression of circRNAs of the above transcription factor can be understood from recent observations that in specific cell types highly expressed genes (in terms of expressed linear isoforms)

give rise to relatively less circRNA compared with moderately- or low- expressed genes [13].

**CircRNA discovery by RNA-seq in hyperdiploid B-cell precursor ALL and in sorted normal leukocyte cell populations** Aiming to discover new cancer-specific fusion transcripts in hyperdiploid B-lineage acute lymphoblastic leukaemia, Salzman *et al.* [19] exploited RNA-seq and found many transcripts with permutated exon order, which they called 'scrambled exons' and attributed to circularised RNAs. In five samples of hyperdiploid B-cell precursor acute lymphoblastic leukaemia they detected hundreds of circRNA transcripts with more than 700 circular isoforms comprising more than 10 % of all transcript isoforms produced from a comparable number of genes. These circRNAs were however not a specific feature of the leukaemic cells; PCR verified that scrambled exons were also detected in remission samples of the patients, in HeLa cells and in normal primary human cells. Also in sorted cells populations, naïve B cells (CD19+), haematopoietic stem cells (CD34+) and neutrophils, circRNA isoforms expressed by more than 800 genes were identified, with circRNA expression accounting for more than 10 % of gene expression. This study showed that a particular gene can produce circRNAs in more than one leukocyte type, but single replicate-based preliminary estimations suggest quantitative differences among cell types. This first indication that circRNAs are expressed both in normal and malignant haematopoietic cells informs the number of circRNAs in immature and lineage-specific blood cells but does not provide a more specific and useful interpretation of circRNA relevance for haematopoietic cell functions and pathology. For instance the study reported and validated a few most abundant transcripts with scrambled exons (*ESYT2*, *FBXW4*, *CAMSAP1*, *KIAA0368*, *CLNS1A*, *FAM120A*, *MAP3K1*, *ZKSCAN1*, *MANBA*, *ZBTB46*, *NUP54*, *RARS* and *MGA*) but did not pay particular attention to circRNAs from numerous genes that are important for normal haematopoiesis and present genomic aberrations or deregulated expression in leukaemia.

In accordance with published data [19] we observed that circRNAs from genes related to B-cell differentiation and acute lymphoblastic leukaemia (*JAK2*, *PAX5*, *IKZF1*, *ETV6* and *EBF1*) are prevalently present in hyperdiploid leukaemia compared with normal leukocytes samples. The latter is likely an underestimation, as only 54 genes were screened and circRNAs supported by only one read were not considered.

The same authors also investigated circRNA expression in 15 different cancer and non-cancer cell lines, detecting around 47 000 circle-specific splice junctions from 8500 genes. The validation of 8 candidates confirmed that all were true circRNAs [10]. The study further specified that highly expressed circRNA showed

cell-type-specific increase in expression that was not associated to an increase of the corresponding linear RNA. Notably, among others, the leukaemia cell line K562 presented the largest number of genes (16 559) with evidence of circular RNA expression.

A reanalysis of the same data set using the circRNAs detecting tool CIRI [80] indicated that more universally shared circRNAs tend to have higher expression levels and verified that the expression patterns of linear transcripts of circRNA-encoding exons are more similar in cancer cells compared with non-cancer cell types, whereas cancer cells appear to have more diverse circRNA expression profiles, both considering exonic and intronic circRNAs.

Subsequently, a comparison of CD34+, CD19+, neutrophils and HEK293 (human embryonic kidney cells) considering only a single biological replicate per cell type (with sequencing depth around 20 million reads per samples) was reported [21]. The study detected 1950 circRNAs of which 939 are exclusively expressed in CD19+ cells, 333 in CD34+, 194 in neutrophils and 60 in HEK293 cells. Nineteen circRNAs resulted to be shared between these cell types. The emphasis of this study was on the demonstration that circRNAs are in part cell-type-specific and are expressed in a developmental stage-related manner.

**CircRNAs in whole-blood samples from healthy individuals** RNA-seq analysis of whole-blood samples [5] showed that whole blood was very rich in circRNAs, comparable to the cerebellum, with consistent data comparing two biological replicates. Also in this study the emphasis laid on the numeric evaluation and the demonstration that circRNAs are a natural component of the transcriptome. Whole blood is composed of a gamma of different cell types. Moreover, the plasma component may also be a sink for circulating circRNA of non-haematopoietic cell origin and can be in fact explored for disease biomarkers, as previously proposed for solid tumours [17]. We cannot exclude that plasma samples could also contain exogenous RNA. In malaria infection, thousands of very short circRNAs are produced by *Plasmodium falciparum* [81], including dozens of circRNAs harbouring more than 100 binding sites for a given human miRNA, pointing to highly versatile parasite–host interactions. Similarly, as already proposed for virus–host interactions, circRNAs of viral origin might sponge host miRNAs and vice versa [42].

**CircRNAs in platelets** Alhasan *et al.* [82] reported circRNA enrichment in platelets that they ascribed mainly to differential decay of linear RNA, considering the particular circRNA resistance to degradation. In the past integration of transcriptome and proteome data of platelets had given somewhat conflicting results [83].

Extensive degradation of linear RNA isoforms leaving circRNAs intact, which results in an extensive reduction of the translatable RNAs provides now a straightforward explanation for this apparent disparity. This study demonstrated that circRNAs are highly enriched not only in platelets but also in erythrocytes relative to nucleated cells finding that more than 3000 genes show 17- to 188-fold relative enrichment of circRNAs.

### 1.2.4 Fusion-circRNAs

**Fusion-circRNA discovery** Very recently fusion-circRNAs (f-circRNAs) derived from transcribed exons of chimeric genes generated by cancer-associated chromosomal translocation were discovered and proven to be oncogenic according to *in vitro* and *in vivo* experiments [84] (Figure 1.8). Instead of a discovery-driven approach this study used informed guessing to directly detect transcript circularisation around the breakpoint/fusion region of two well-known recurrent leukaemia-related translocations. The authors hypothesized that juxtaposition of complementary sequences in introns at either side of the fusion regions could favour the formation of circRNAs and searched specifically for circRNAs expressed from fusion genes. f-circRNAs were thus detected by RT-PCR and then confirmed using RNA-seq and custom bioinformatics procedures, in promyelocytic leukaemia with a *PML/RARα* and acute myeloid leukaemia with an *MLL/AF9* fusion (Figure 1.8A). Both translocations gave rise to more than one f-circRNA characterized by different backsplice junctions, both in patient samples and in patient- derived cell lines. The discovery was also extended to solid tumours showing f-circRNAs transcription in translocated Ewing sarcoma and lung cancer.

**F-circRNAs are proto-oncogenic and a requisite for leukaemic cells viability** Guarnerio *et al.* [84] showed that f-circRNAs (*f-circPR* and *f-circM9*) expression in leukaemic cells increases cell proliferation and clonogenicity and that f-circRNA silencing reverted the phenotype, demonstrating that these f-circRNAs are biologically active and exert pro-proliferative and proto-oncogenic activities (Figure 1.8B). Moreover, shRNA-based knock-down of *f-circM9* in leukaemic THP1 cells resulted in increased apoptosis showing that f-circRNAs have an important role in maintaining the viability of leukaemic cells.

***In vivo* study of f-circRNAs** Human leukaemic cells expressing f-circRNAs *in vivo* sustain disease progression in mouse. On the other hand, f-circRNAs alone did not trigger leukaemia. Cells expressing *f-circM9* together with the *MLL/AF9* fusion protein have an increased ability to proliferate and form colonies than cells

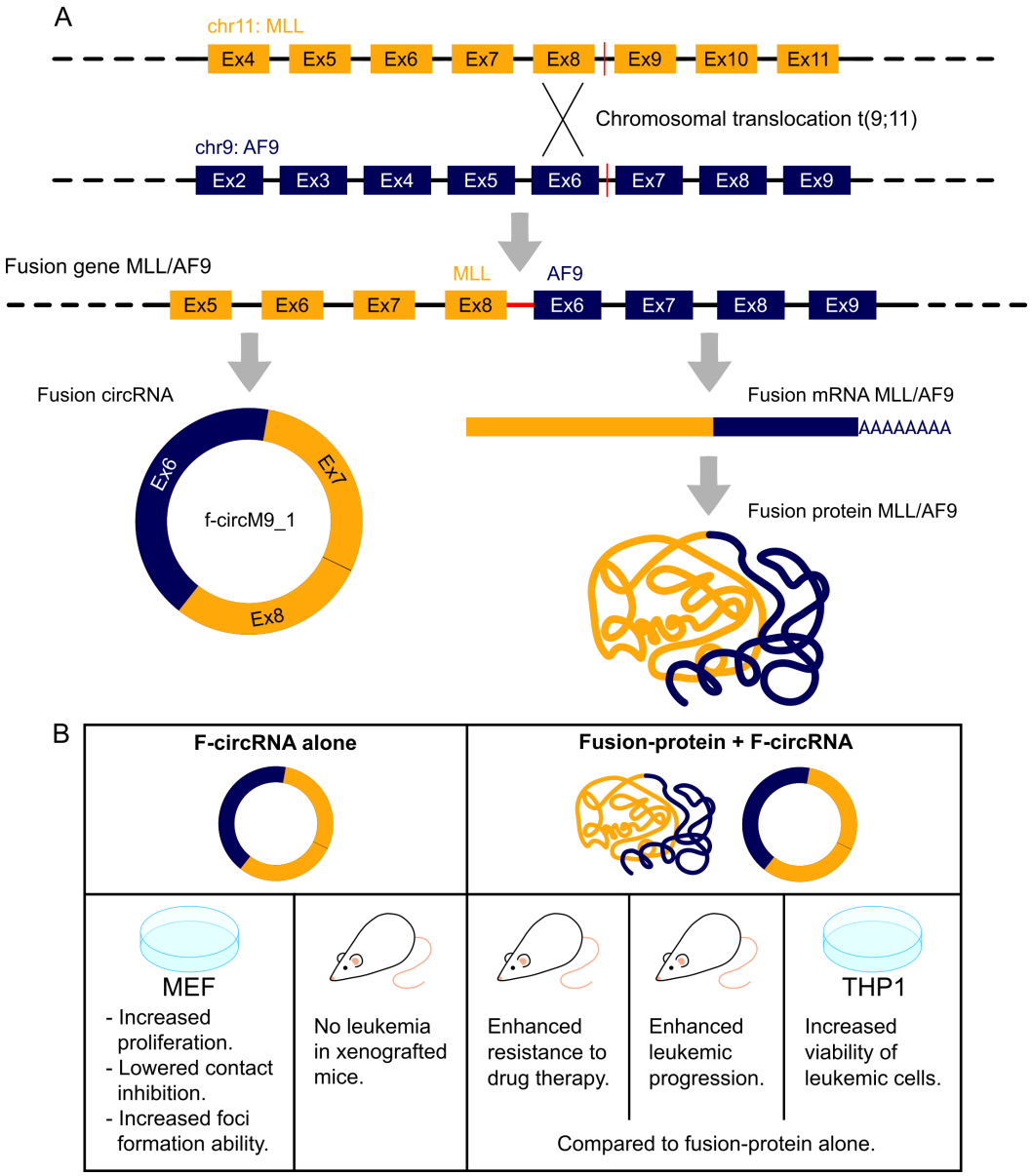



Figure 1.8: F-circRNAs derived from chromosomal translocations have oncogenic role. (A) Transcription of fusion genes generated by cancer-associated chromosomal translocation could generate both linear mRNA coding for oncogenic fusion proteins and f-circRNAs [84]. The figure depicts the example of *f-circM9_1* expressed in cells harbouring the well-known acute myeloid leukaemia *MLL/AF9* t(9;11)) translocation: *f-circM9_1* includes two sequences not present in the normal genome, the *MLL* exon 8 and *AF9* exon 6 fusion junction derived from the chromosomal translocation and the backsplice junction connecting *MLL* exon 7 with *AF9* exon 6. (B) *f-circM9* was demonstrated to be proto-oncogenic *in vitro* (increasing proliferation rate and foci forming ability in mouse embryonic fibroblasts, MEF) and required for leukaemic cell (THP1) viability. *f-circM9* alone resulted not sufficient to trigger leukaemia in vivo when expressed in HSC xenografted in mice. Concurrent expression of *f-circM9* and MLL/AF9 fusion protein contributed to leukaemia progression *in vivo* and *ex vivo* cells expressing *f-circM9* and MLL/AF9 displayed increased ability to proliferate and to form colonies. Furthermore, *f-circM9* expression in *MLL/AF9* mouse model cells increased the resistance to leukaemia treatments suggesting that *f-circM9* impacts to pre-clinical therapeutic outcome.

expressing the fusion protein alone, strengthening the hypothesis that f-circRNAs contribute to the leukaemogenic process *in vivo* (Figure 1.8B). Human leukaemic cells *in vivo* expressing f-circRNAs sustain disease progression in mouse. Moreover, f-circRNAs expression could provide tumour cells protection to standard leukaemia treatment, as arsenic trioxide, and confer survival advantage to leukaemic cells in response in addition to standard-of-care leukaemia treatment with cytarabine. Thus, according to these experiments in a pre-clinical setting *f-circM9* could impact therapeutic outcomes.

Interestingly, Guarnerio *et al.* [84] argued that the latency of leukaemia development in animal models could be due to the absence of f-circRNAs in modelled expression of intron-less fusion genes. Even if it does provide neither data nor hypotheses on the mechanisms underlying the observed pathogenetic effect of f-circRNA, the study [84] is a breakthrough for leukaemia research. In addition, as noticed by the authors, not only the expression of f-circRNA but also the reduction of circRNAs expressed from the non-translocated allele partner genes can contribute to the pathogenic effect.

## 1.3 Detection of circRNAs from RNA-seq: bioinformatics challenges and approaches

Bioinformatics is an essential part of circRNAs analysis, since the approach of detecting their presence in a transcriptome-wide manner produces a very large amount of data to be handled. The high throughput of next generation sequencing techniques has been recognized as one of the major advantages of this approach, but at the same time confronted scientists with the question of how to deal with the large size of the data output. Bioinformatics developed techniques to disentangle and polish the raw data first and then to extract reliable information from them, both concerning the identification of molecules and the quantification of their levels of expression.

In the case of circRNAs the standard analysis techniques used to handle RNA-seq data needed to be adapted to be able to identify the backsplice junction reads. CircRNAs can be detected thanks to the presence of this particular splicing junction, that would not occur if the molecule were not circular. In fact, reads that map on the exons involved in the circularisation, but not spanning the backsplicing junction, can be aligned both on the circular and on the linear form. That being the case these reads cannot be used to detect circRNAs.

In the past reads mapping to the reference genome with non-collinear exon

junctions were often considered artefacts and discarded; with the rise of interest in fusion transcripts produced from rearranged genomes, trans-splicing and circRNAs, RNA-seq aligners were adapted to consider also these 'exotic' events. In those reads, parts of the transcript sequence correspond to regions that are remote in the reference genome (fusion and trans-spliced transcripts) or that map to the genome in a chiastic way (that is, two segments of a single read align separately in reverse order due to the backsplicing in circRNA biogenesis; see also Figure 1.2B).

Table 1.1 presents ten computational methods that allow backsplice events identification using different strategies and read aligners; some approaches are specific to predict circRNAs, whereas others were developed with more general purposes, such as read alignment and/or detection of fusion events, and allow also circRNAs detection. For each method, the table indicates the read aligners and number of mapping steps implemented in the prediction strategy. It also clarifies whether the method requires genome annotations in input and if it provides an annotation of predicted backsplice junctions in terms of overlapping genes. Finally it is described whether the tools has been designed explicitly for circRNA identification or if they are more general purpose software and the last two columns report additional notes on specific software features and references. Recently, five circRNA prediction tools (circRNA_finder, Find_circ, CIRCexplorer, CIRI and MapSplice) have been compared by evaluating the levels of true- and false-positive circRNAs based on RNAse R resistance data, showing that not in all cases the most abundant circRNAs are true positives, that circRNAs identified by a single method only are in general less reliable and that the combination of at least two methods increases specificity [85].

## 1.4 Open questions and challenges regarding circRNAs in haematopoiesis

CircRNA expression further challenges a simplistic definition of 'gene', reinforcing the concept that genes are complex transcriptional units and that the sequence of a given genomic region is a sort of palimpsest[1] that contains multiple, interleaved and overlapping information parcels [90]. Transcripts from the same *locus* use a common sequence in different ways and perform distinct biological roles. In addition, circRNAs add new hints to our understanding of the alternative use and reuse of RNA sequences to produce different products and even small RNAs, as known in the case of miRtrons [91] and of tRNA- or snoRNA-derived miRNAs [90, 92] and moRNAs [93–95], that are expressed in blood cells [94, 95] and were shown to have

[1] A palimpsest is an ancient parchment on which the original text was overwritten multiple times.

| Name | Aligner | Map | Ann | Genes | Only | Notes | Ref |
|---|---|---|---|---|---|---|---|
| CIRI | BWA-Mem | 1 | No* | Yes | Yes | Accounts for uncertainty of read mapping to the junction | [80] |
| Find_circ | Bowtie | 2 | No | No | Yes | | [21] |
| CIRCexplorer | STAR, TopHat | 2 | Yes | Yes | Yes | | [23] |
| Testrealign | Segemehl | 1 | No | No | No | Parses Segemehl alignments | [86] |
| UROBORUS | TopHat, Bowtie | 2 | No | Yes | Yes | Does not underestimate expression; filters spurious alignments | [87] |
| NCLscan | BWA, Novoalign, BLAT | 3 | Yes | Yes | No | $>$ 98 % precision, test on polyA+/- libraries | [88] |
| MapSplice | MapSplice | 1 | Yes | Yes | No | Circular RNA explicit detection from MapSplice v2m (2/2013) | [85] |
| circRNA_finder | STAR | 1 | No | No | Yes | | [22] |
| KNIFE | Bowtie | 2 | Yes | Yes | Yes | Statistical approach to enrich true positives | [89] |
| PTESFinder | Bowtie | 1 | No | No | No | | [40] |

Table 1.1: Computational methods available for circRNAs discovery, characterisation and quantification from RNA-seq data. Map: mapping steps performed; Ann: whether annotation is needed or not; Genes: whether the tool gives gene names or not; Only: whether the tool gives only circRNAs or other molecular types; * = only for report.

pathogenic relevance in B-cell lymphomas [96]. Ultimately gene expression studies need to disentangle the expression of linear and circular RNAs expressed from each *locus* in order to dissect the distinct or complementary processes in which they participate.

A large body of data regarding molecular circuits that control cellular differentiation of the haematopoietic system is available and its deregulation in malignancies in association with genomic lesions is increasingly understood. In haematopoiesis, differentiated cell states are controlled by densely interconnected transcriptional circuits [97] in a seemingly hierarchical process of binary fate decisions, but the stiffness of cell fate may be more fluid [98] allowing for epigenetic regulation in response to mature blood cell demand. We envisage that circRNAs studies of haematopoietic cell stages will further elucidate how cell fate fluidity may depend on stably present circRNAs of key cell stage mRNAs.

Gene expression profiling of the protein-coding transcriptome has been very useful for the study of haematopoietic malignancies but will become more complete when integrated with circRNA expression data. Among other things it can be expected to elucidate discordant gene-protein expression often revealed for marker proteins of haematopoietic cell stages (for example, CD10, CD22 and CD38). Furthermore, induced pluripotent stem cell modelling of haematopoietic diseases may also benefit from circRNA studies, hitherto nothing is known about the behaviour of circRNAs in reprogramming procedures of induced pluripotent stem cell generation.

Recent RNA-seq data outlined transcriptional diversity in terms of (linear) alternative isoform-ratio variations among haematopoietic cells [3, 99] and of non-coding RNA's impact in haematopoietic lineage differentiation [100]. Specific miRNAs are expressed in a developmental-stage-specific manner [101, 102]. miRNAs and other small RNAs are differentially expressed in disease [95].

As in last years we abandoned the concept of the centrality of coding fraction of the transcriptome, the discovery of circRNAs made clear that the study of the linear RNAs only provides an incomplete picture of the cellular complexity. Focusing on linear transcription only we miss important elements, both for data interpretation and experimental design.

Today we have an increased appreciation of circRNA abundance, evolutionary conservation and diversity of functions and interactions. Specific data emerged of high and regulated circRNA expression in normal and malignant blood cells. The recent discovery and functional study of f-circRNAs provided important clues of the oncogenic role of this aberrant circRNA in leukaemogenesis and of their relevance in modulating therapeutic outcome. Together these data clearly indicate

that further studies of circular isoforms from different cell types and stages of the haematopoietic compartment and by rearranged or mutated genomes are warranted to better estimate the position of these new regulators in haematopoietic cell development and derived malignancies. The route toward elucidating circRNA biology is still long. Even a consistent nomenclature for circRNAs is sorely missed.

Definitely, circRNAs and their diverse molecular interactions participate to the circuitries that regulate the final cellular protein output adding to the richness and complexity of the underlying mechanisms. The particular stability of circRNAs may also make them valuable disease markers that can be identified in various body fluids and we envisage that a better understanding of circRNA biology will inform innovative therapeutic targets.

## 1.5 Aim of the research

The aim of the research is to expand the knowledge of the transcriptional landscape of normal and malignant haematopoiesis to shed new light on the mechanisms underlying molecular circuits. Since circularisation can compete with mRNA splicing favouring exon skipping, alternative splicing deregulation impacts specifically on disease course in leukaemias and myeloproliferative neoplasms, and somatic mutations of splicing factors are frequent in haematological malignancies, we hypothesize that circRNA expression might be linked to aberrant splicing patterns, with effects on the proteome and on the non-coding transcriptome as well. We aim to identify qualitative and quantitative specificities of circRNAs in normal haematopoiesis, that will be subsequently validated. The knowledge of circRNAs in normal haematopoiesis will be precious to better understand regulatory circuits and aberrant functions in blood malignancies and other diseases. Actually the simple discovery and initial characterization of circRNAs is not sufficient to obtain a significant advance on functional knowledge. We envisage that predictions of circRNA functions and interactions will facilitate the prioritization of candidate circRNAs for functional investigation.

# 2 Bioinformatics

The detection of circRNAs in RNA-seq data is performed by identifying reads containing a backsplice junction. In the classical analyses of RNA-seq data, reads that contain backsplice junctions are discarded because they can not be mapped on the reference transcriptome. Next generation sequencing has revived the interest in circRNAs, so algorithms written specifically for circRNAs detection from RNA-seq data were needed, and various tools with different approaches to the detection have been published.

In collaboration with dr. Enrico Gaffo and prof. Stefania Bortoluzzi (Department of Molecular Medicine, University of Padova), we built up circPipeline, a computational pipeline to perform the detection and analysis of circRNAs (and linear RNAs) from RNA-seq data. A simplified and refined version of the circPipeline, that performs circRNAs analysis, has been described in the paper "CirComPara: a multi-method comparative bioinformatics pipeline to detect and study circRNAs from RNA-seq data" by Gaffo E. et al. (in press Journal Non-coding RNA). This pipeline is available at https://github.com/egaffo/CirComPara.

Amongst tools and methods listed in Table 1.1, Find_circ and Testrealign, and then CIRI were available at the beginning of our project hence were initially used for our analyses. In particular Find_circ was developed by Memczak and colleagues, who were the first to study circRNAs using next generation sequencing data. Testrealign has an interesting feature that circRNAs detection is implemented inside the mapping algorithm. CIRI uses a different detection algorithm compared to Find_circ and Testrealign, and in particular pays attention to false positive filtering. Thanks to the modularity of the pipeline, we were later on able to add another program, CIRCexplorer. The latter program was analysed by Hansen *et al.* in 2015. These authors compared several available methods for circRNAs detection to analyse their performance, also in combination with each others. Their analysis in fact highlighted that CIRCexplorer, combined with any other program allowed to perform the best false positive filtering.

## 2.1 Algorithms description

The algorithms used in the different programs can be divided in two major groups, those working *de novo* and those needing an annotation of the reference genome and transcriptome. *De novo* algorithms perform backsplice detection basing on intrinsic features of the sequencing reads, thus allowing to find circRNAs deriving

also from non annotated regions. Algorithms that need an annotation can only identify circRNAs that belong to known exons and introns, and therefore are expected to be more conservative in their results. All programs need as an input the reads aligned to the reference genome by a specific reads aligner to obtain files compatible to their algorithm.

### 2.1.1 CIRI

CIRI requires as input the FASTA file of the reference genome used for the alignment and the aligned reads in SAM format generated by BWA-Mem.

The circRNAs detection by CIRI is based on the recognition of the PCC signal (Paired Chiastic Clipping) which is encoded in the CIGAR string of the alignment. The CIGAR string is an alphanumeric description of the mapping of each read to the reference genome, in which every letter has a specific meaning, such as `M` for matched, `D` for deletion and so on. When one read is split to be aligned in two separate places of the genome it appears two times in the SAM file, once for the first part of the alignment and once for the second portion aligned. The unaligned portion in these two records is marked as 'clipped' in the CIGAR string with `S` or `H`. CIRI's first step is to retrieve the reads that display the PCC signal: the same read repeated with a CIGAR string similar to `MxSy` in the first occurrence and `SxMy` in the second one.[1] When CIRI detects such a signal it then checks the strand information and the mapping position in the SAM alignment. If the two segments align on the same strand of the same chromosome and within a reasonable distance (basing on the actual distance of the mate-pair) the read is considered as a candidate junction read with positive PCC signal. CIRI then uses the additional information provided by the mate-pair layout to preliminarily filter the false positive PCC signals: a candidate PCC-harbouring read passes the filter only if its paired read aligns within the region of the putative circRNA range. This is a procedure only carried out if the reads are mate-paired. Apart from this CIRI also uses the reference genome to refine the list of putative circRNAs by checking if the `AG` and `GT` dinucleotides of the canonical splicing sites actually flank the individuated backsplicing junction. The program also accepts a list of other non-canonical splicing sites to complement the filter of false positive circRNAs. Once filtered for the false positives, CIRI scans the alignment file a second time to perform a thorough investigation of the putative junctions that passed the filtering. In this step CIRI summarizes the mapping positions of the different reads related to a certain candidate junction, compares the counts and the mapping lengths, thus determining if the reads reliably reflect

[1] For the first occurrence of the read `x` is the number of matched bases (`M`) and `y` is the number of clipped bases (`S`) and vice versa for the second occurrence.

a circRNA junction and if it should be kept in the final output. At the end of the run the output consists of a SAM-like file with information on the number of reads covering a certain junction, on the number of mate-reads that are coherent with each putative junction, and with a simple genomic annotation.

### 2.1.2 CIRCexplorer

CIRCexplorer is the circRNAs-detecting tool described in the 2014 paper by Zhang *et al.* [103]. Sequence reads from each sample are first mapped using TopHat, and unmapped reads are then extracted and mapped onto the reference genome using TopHat-Fusion. Reads that split and align on the same chromosome but in non-collinear order (indicated with special XF tags in output BAM files) are extracted as candidate backspliced junction reads. Backspliced junction reads are then further realigned against existing gene annotations to determine the precise positions of donor or acceptor splice site for each backspliced event. Junction reads with shifted alignments against canonical splice sites (with the consensus sequence flanking the donor splice site and the acceptor splice site) are adjusted to the correct positions, and reads with alignments on different genes or non-canonical splice sites (largely arising from trans-splicing or PCR errors) are discarded. Finally, the remaining backspliced junction reads are combined and scaled to RPM (Reads Per Million) mapped reads to quantify each backspliced event. The output is a BED-like file with information on the read counts for each junction and the annotation.

### 2.1.3 Find_circ

The Find_circ tool is described in the paper by Memczak *et al.*, 2013 [21]. The first step requires the sequencing reads to be aligned with Bowtie to filter out the reads that align contiguously and fully to the reference genome. Then the unmapped reads are passed to Find_circ for circRNAs detection, which is the strategy we decided to use in our pipeline too. The tool extracts 20mers from both ends of the read and aligns them independently to establish their unique anchor position on the reference genome. Only anchors that align with reversed orientation (head-to-tail) allow to detect a putative circular junction. Once established the anchor positions, the 20mers are extended so that the whole read aligns and the break point is flanked by `GU–AG` splicing sites. Before finalizing the output the putative backsplice junctions undergo a quality filtering step to increase the robustness of the results. Among the filters applied by Find_circ is the requirement of at least 2 reads covering a certain backsplice junction, that the distance between the two ends of the junction is not longer than 100 kb and an high quality alignment of

the reads (unique mapping of the anchors and not more than two mistakes in the extension procedure). Find_circ outputs the detected splicing junctions in a BED format file with additional fields containing quality statistics and reads counts.

### 2.1.4 Testrealign

CircRNAs detection by Testrealign is based on mapping seeds identified by the aligner Segemehl [104]. Because of the structure of the spliced read, a semi-global alignment is likely to fail, while the method used by Segemehl will identify several seeds matching different locations. The algorithm used, which is able to identify both splicing events (normal, trans-splicing and backsplicing) and gene fusion sites, is based on a greedy seed chaining followed by a Smith–Waterman-like alignment. The final output of Testrealign is a BED file containing both normal splicing, trans-splicing and circular splicing events.

## 2.2 circPipeline: a pipeline for detection and analysis of circRNAs from RNA-seq data

The pipeline (Figure 2.1) puts together both published tools and custom scripts. It is modular and extensible, and most importantly it is automated in many of its branches, to decrease human errors, speed up the analyses and allow re-analysis using different settings.

The pipeline is divided in two main sections: the detection module and the analysis module. The first one performs the reads alignment and annotation, starting from the raw reads produced by the sequencer, and is fully automated. Among the several tools for circRNA detection published in the last few years, we used the four described in Section 2.1. The results are then fed to the second section that performs the sequence reconstruction and expression levels quantification, and is partly automated to allow specific analyses decided by the operator.

In parallel to the circRNAs analyses the pipeline also performs the classical RNA-seq analyses for the comparison of the expression levels of linear and circular transcripts.

The raw sequencing reads are preliminarily aligned by HISAT2 that produces two separate files, one with the reads mapped on the reference genome and one with the unmapped reads. The first one is passed to the linear pipeline, while the second is used by the circular pipeline (Figure 2.2). This choice was made to decrease the computational burden of the alignment step of the pipeline and to increase the retrieval power of the circRNA-detecting programs. In this way the set

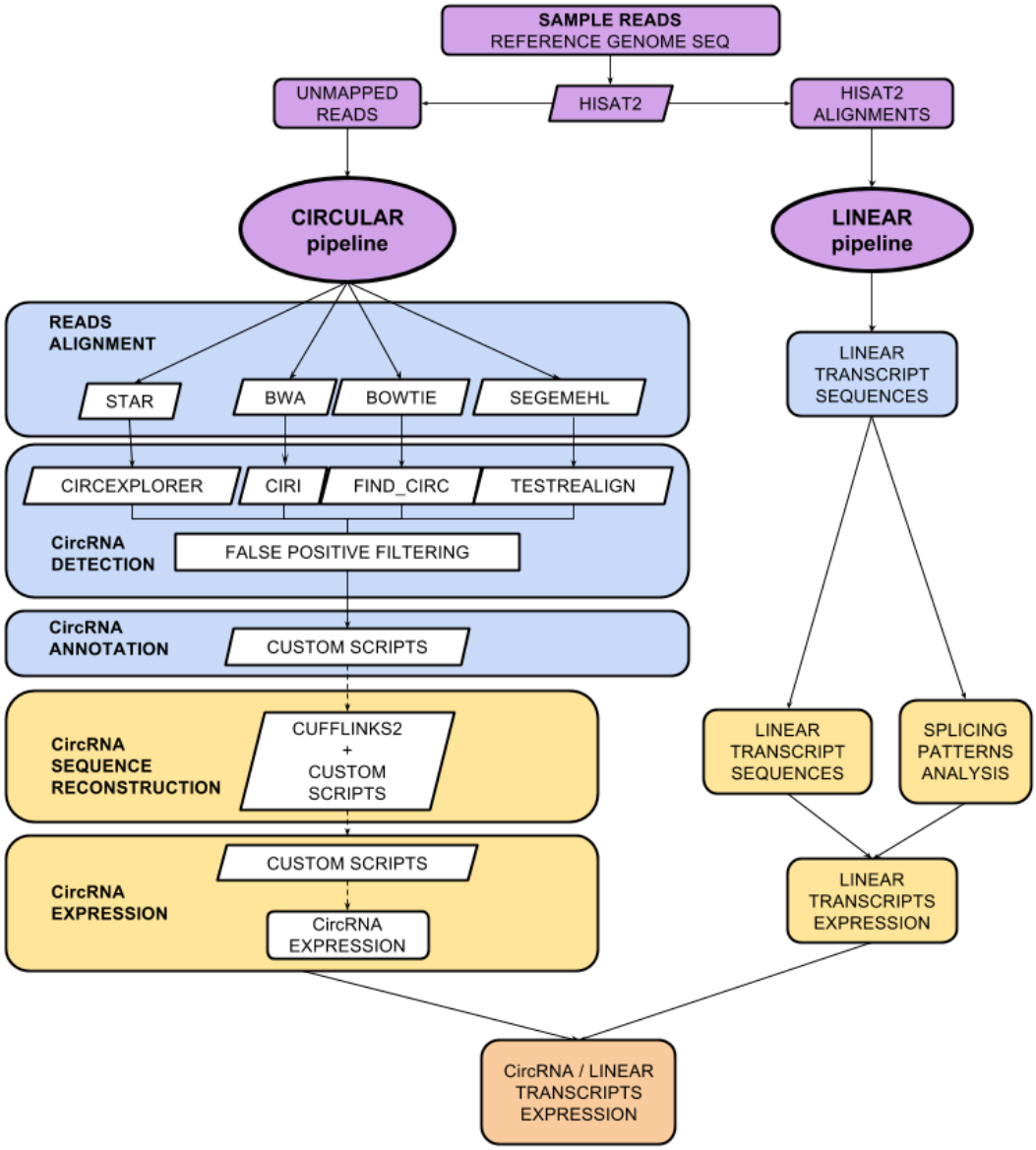



Figure 2.1: Architecture of the pipeline for circRNA and linear RNA analysis. The purple section indicates the preparation of the data, the blue section indicates the detection module of the pipeline and the yellow section indicates the analysis module, in orange the final output of the pipeline.

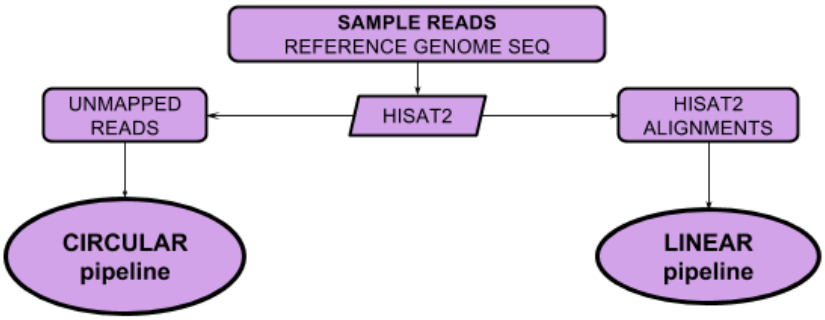



Figure 2.2: Before linear and circular RNA detection, a preliminary alignment step is performed to sort out reads aligning collinearly to the reference genome, to be used for linear transcriptome reconstruction, and reads not mapping, to be used for backsplice and circRNA detection.

of the reads is aligned only once to discriminate between mapped and unmapped reads, with relaxed filters. The subsequent alignment of the unmapped reads with parameters suitable for backsplice detection is performed on a smaller amount of reads with lower background noise in the alignment.

### 2.2.1 Linear pipeline

The preliminary alignment performed with HISAT2 produces a file containing mapped reads that is used by the 'linear pipeline' for the standard RNA-seq analyses on linear transcripts. The reconstruction of the sequence is performed by Cufflinks2 [105], which produces a list of transcripts containing various isoforms, and for each of them an expression level is calculated. The transcripts are associated to the gene of origin. The final step of the linear pipeline is the differential expression analysis.

## 2.3 circPipeline: detection module

The detection of the circRNAs is performed independently by the four different programs in parallel (Figure 2.3). The unmapped reads input to the pipeline are aligned by four different aligners (STAR, BWA, Bowtie and Segemehl), as each circRNA-detecting program requires a specific aligner in order to have a specific file format. Each aligner passes the mapped reads to the corresponding program for the detection of the putative circRNAs. The algorithms of the four chosen programs are described in section 2.1. The usage of four different programs increases the detection power and allows the possibility of a false positive filtering, by overlapping the four results. The output of each program is a file with different formats (one for each analysed sample), thus for the downstream comparisons only the fields that

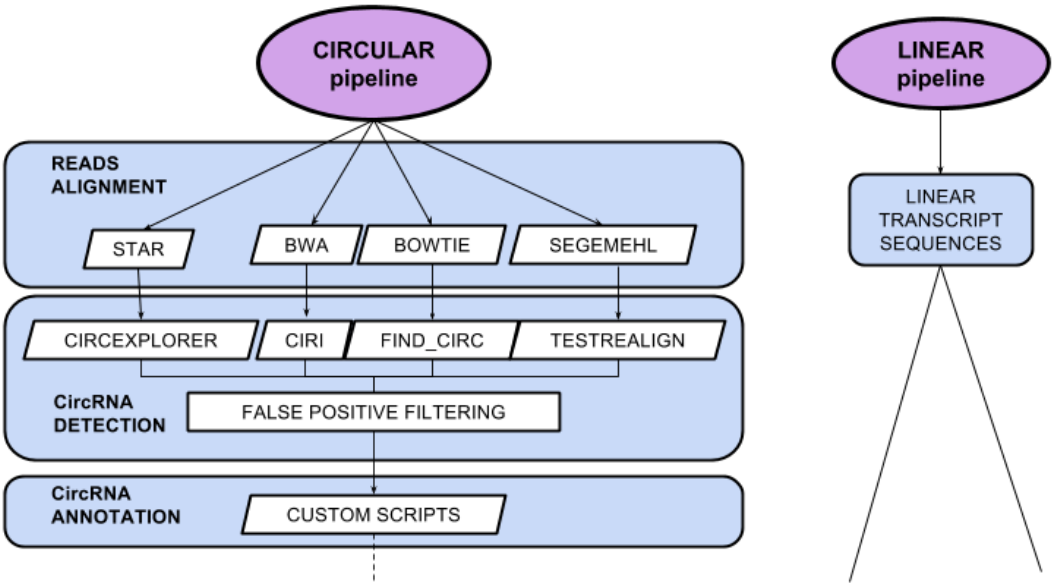



Figure 2.3: The detection module of the circPipeline. The left branch regards circular RNAs detection, in which the reads are alligned in parallel by four different reads aligners, and then processed by the relative circRNA detection tool; after that the results are filtered for false positives and annotated by custom scripts. The right side regards the linear branch of the pipeline, in which the reads are aligned for linear transcripts detection.

are common to all the files are maintained, such as the position of the backsplice and read counts.

The putative circRNAs are preliminarily filtered on the level of expression, keeping only those detected by at least two sequencing reads, as suggested previously [16], which is a technical filter to decrease the number of false positive results. The latter also has the effect of conforming the results to those of CIRI and Find_circ, which already have this filter implemented in their algorithm. We decided to put another technical filter on the length of the distance of the backsplice junction, to keep only circRNAs longer than 200 nucleotides. It has to be noted that exons shorter than 200 nucleotides do actually exist; this filter though was suggested by the work of Hansen *et al.* [85], as they pointed out that circRNAs shorter than 200 nucleotides were more likely to be false positives.

During the pipeline development we decided to overlap the results of the four used programs. The above mentioned article of Hansen *et al.* [85], in which they report five circRNA-detecting programs for specificity and sensibility, supports the validity of our choice. For their study they sequenced the RNA of four samples of human fibroblasts in two conditions each: one with only ribosomal RNA depletion and one with the addition of the RNAse R treatment before library preparation, which only digests linear molecules. The comparison of the circRNAs found in the two conditions gives a measure of how specific and sensitive each program

is. To perform this analysis the circRNAs are classified as 'enriched' if the fold change is greater than 5 in the treated samples, 'unchanged' if the expression level is unaffected by the treatment (fold change between −5 and 5), and 'depleted' if the fold change is lower than −5 in treated samples. They showed that some programs were more specific or sensitive than others, and most of all that using at least two programs together significantly lowered the false positive rate. This was measured by counting the number of circRNAs that were detected in the untreated samples and that could not be found in RNAse R treated samples, meaning that they were mistakenly indicated as circRNAs.

To allow the comparison between the four methods, the starting position of the backsplice had to be corrected for CIRI and Testrealign, where the output is '1-based', meaning that the first base is numbered as 1, while Find_circ and CIRCexplorer produce a '0-based' output, meaning that the first base is numbered as 0. We chose to subtract 1 to the starting base of the backsplice junction for CIRI and Testrealign, according to Hansen *et al.* [85].

Finally, to guarantee a higher robustness of the dataset we decided to select only putative circRNAs that were detected at least by two of the four programs used.

Together with the detection, some programs (CIRI and CIRCexplorer) do perform a simple annotation step, but we added a custom script in R to allow a more detailed annotation, indicating, among other information, the gene of origin and information regarding circRNA composition (exonic, intronic or intergenic).

The filtering and annotation of the circRNAs is performed with custom R scripts.

The first branch of the pipeline produces two files: a list of circRNAs identified by a unique key (built up as follows: `chr:start|stop:strand`) and a matrix with the expression levels of all detected circRNAs of all analysed samples. For each putative circRNA in the list we know the starting and final position of the backsplice junction on the chromosome, the strand, the number of reads detecting the specific backsplice junction, the genomic feature of the involved sequences and the gene of origin, the programs that detected it, the sample and cell type in which they were retrieved.

## 2.4 circPipeline: analysis module

The output matrix of the first branch is then used by the second branch of the pipeline (Figure 2.4) to perform the analyses on the sequence and expression levels of the putative circRNAs.

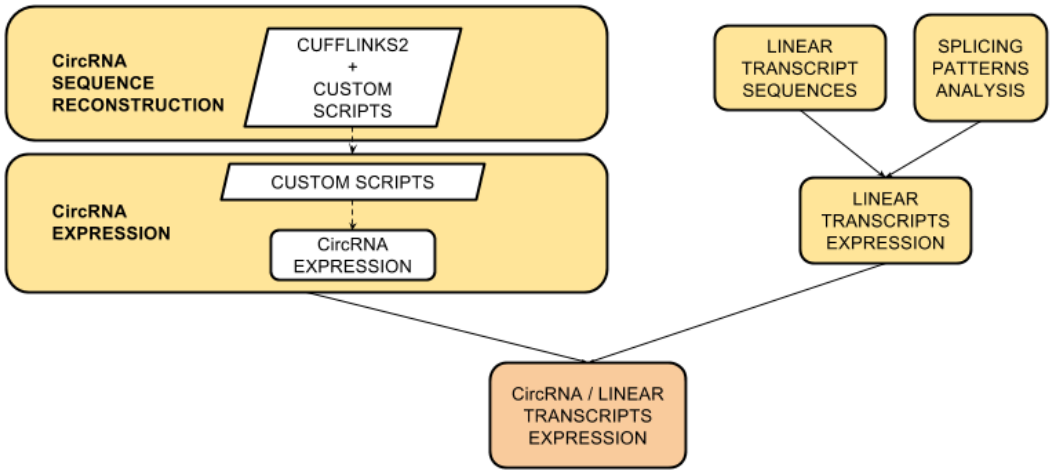



Figure 2.4: The analysis module of the circPipeline. The left side regards the branch for circRNAs analysis, that performs circRNAs sequence reconstruction and circRNAs expression levels analysis. The right side regards the linear branch of the pipeline, in which linear transcript reconstruction and expression level analysis are performed. In the and the information produced by the two branches is compared and integrated.

### 2.4.1 Sequence reconstruction

Sequence reconstruction can not be easily done for circRNAs because the only reads undoubtedly ascribable to circRNAs are those harbouring the backsplice junction. Other reads, deriving from the inner part of the molecule, such as those mapping completely on an exon or an intron can be equally attributed to the linear and circular isoform of that sequence. These are preliminarily mapped by HISAT2 to the reference genome, thus being fed to the linear pipeline and not to the circular pipeline.

The information available for each putative circRNA allows to know that the genomic features on which the two ends of the backsplice map are present in the circularized molecule. However, we cannot know if exons or introns included in the interested genomic region are actually present in the circularised molecule or if they have been excised. We can hypothesise that at least the features identified by the mapping are actually present, so at least that sequence can be inferred and used for downstream functional prediction. For example this is the case of a backsplicing junction involving a single exon or a single intron, which probably belongs to a mono-exonic circRNA or to a mono-intronic circRNA. Up to now the only accepted method to know the actual sequence of a circRNA is by Northern blotting after RNAse R digestion — using a probe that hybridises to the backsplice junction — that allows to know its actual length, and then to sequence it.

### 2.4.2 Expression levels and differential expression analysis

The amount of reads harbouring the backsplice junction is the most trustworthy measurement of expression level of a circRNA according to the current literature on the subject. The pipeline produces up to four different levels of expression for each circRNA in each sample. The raw reads counts are then normalized for each sample and for each program on the total number of reads mapped, to compare the results of the four programs. The expression level is then calculated as the median value of the normalized levels of expression detected by each program. The choice of using the median is dictated by the higher robustness of this value compared to the mean. The normalization and calculation of the median value are performed by custom R scripts. Once the expression level for each circRNA in each sample is defined, other custom R scripts are used to perform cluster analysis if requested for the specific experiment.

Once the expression levels are established, they are used to perform differential expression analyses using R, in particular with the package DESeq2 [106], to identify circRNAs and linear RNAs upregulated or downregulated amongst the different sample classes.

## 2.5 Discussion and conclusions

The first branch of the pipeline is fully automated and thanks to its modularity it can be expanded to add new programs for the detection of circRNAs, thus allowing to constantly keep up to date with the best and newest algorithms for circRNA retrieval.

The sequence reconstruction portion of the pipeline lags behind in performance to reliably infer the actual structure of the various circular molecules. Longer sequencing reads may be of use to overcome this issue.

Our pipeline performs expression level analysis for circRNAs considering the reads spanning the backsplice junction. The reads which belong to an internal portion of a circRNA, thus not harbouring the backsplicing junction, are originally mapped on the linear transcriptome. This produces a bias on the calculation of the actual expression level of the linear isoform. To overcome this issue we will implement a re-tuning of the expression level of those molecules basing on the expression level of the circular form and on the unbalance of the expression levels between the exons of the linear transcript.

CIRI provides also an extra useful information thanks to its ability to utilise the paired-end layout of the sequencing, which could be useful for a future improvement of the pipeline: its output contains in fact a field in which the number of mate-reads

mapping on a sequence compatible with a circular molecule is counted; for example the second read maps on the second of the two implicated exons. This piece of information might be of use to further refine both the quantification of circRNAs expression levels and the molecular structure. It may be useful to implement this kind of analysis irrespective of the program that detected the circRNAs.

Re-tuning the expression levels of both linear and circular isoforms will allow to improve the quality of the comparison between them, which is important information to be retrieved to speculate on circRNAs functions.

### 2.5.1 Perspectives

The putative circRNAs output of the pipeline can be used for functional predictions based on the retrieval of specific sequence motifs, such as those for RNA–RNA or RNA–protein interactions. These analyses have to be performed knowing the sequence of the molecule. An automated and dataset-wide analysis would be to search for sequence motifs in the reads spanning the backsplicing junction, as those hundred bases are the only actually available sequence. Exonic sequences are known to harbour different binding sites, like those for miRNAs, RNA-binding proteins, other RNAs; but it is of particular interest whether the new sequence produced by the backsplice junction creates a new motif or binding site.

Instead, for circRNAs with completely available sequence, such as the ones sequenced or those for which the sequence is inferred based on their putative structure, more in depth analyses can be performed. Internal ribosome binding sites (IRES) can be detected and the coding potential can be tested as well, for example by translating the circRNAs on all putative frames, to see if a looping protein or an entirely different polypeptidic chain from a well known coding exon could be expressed.

This pipeline allows to perform a fast analysis of the transcriptome in a refined fashion, by adding the analysis of the circRNA-ome that was up to now left behind in most of the published transcriptome studies. This will definitely allow the production of more reliable results in exploratory studies such as the linear-only RNA-seq studies, which will result in more complete and useful information when turning to functional studies.

# 3 Pilot study of circRNA expressed in lineage commitment of human blood progenitors

This chapter deals with a pilot study for circRNAs detection and quantification using publicly available RNA-seq data obtained with a polyA enrichment protocol. The RNA-seq data `EGAS00001000284` were originally obtained by Chen *et al.* (2014) and used in a study focused on gene expression programs driving or associated to lineage choices in haematopoietic cells differentiation [3], in a study focused on linear transcripts and linear isoforms expressed.

The dataset contains RNA-seq data from eight primary human haematopoietic progenitor populations representing the major myeloid commitment stages and the main lymphoid stage. In the original study Chen *et al.* identified cell-type-specific expression changes and classified differential expression patterns during lineage commitment. They described that each cell type also has a specific set of transcripts, with an over-representation of non-coding transcripts in stem cells compared to the over-representation of coding transcripts in more differentiated cells. Moreover, they highlighted that alternative splicing regulation is widely spread in key commitment stages. They reported at least 2000 differential alternative splicing events being significantly enriched in genes involved in regulatory processes. The study also highlighted the complexity of fating events in the progenitor populations, in terms of transcripts expression, but did not consider the existence of the circRNAs. This scholar study thus missed relevant information. Even if these data were originally obtained to study linear transcriptome and are polyA enriched, we were able to detect and analyse circRNAs from their RNA-seq data.

During the development of the circRNA pipeline we searched for a publicly available datasets to test the pipeline and to obtain preliminary results on the number of circRNA expressed by different cell types in the haematopoietic lineage. We selected the RNA-seq dataset published in 2014 by Chen *et al.* (from now on 'Lineage Commitment dataset') regarding haematopoiesis.

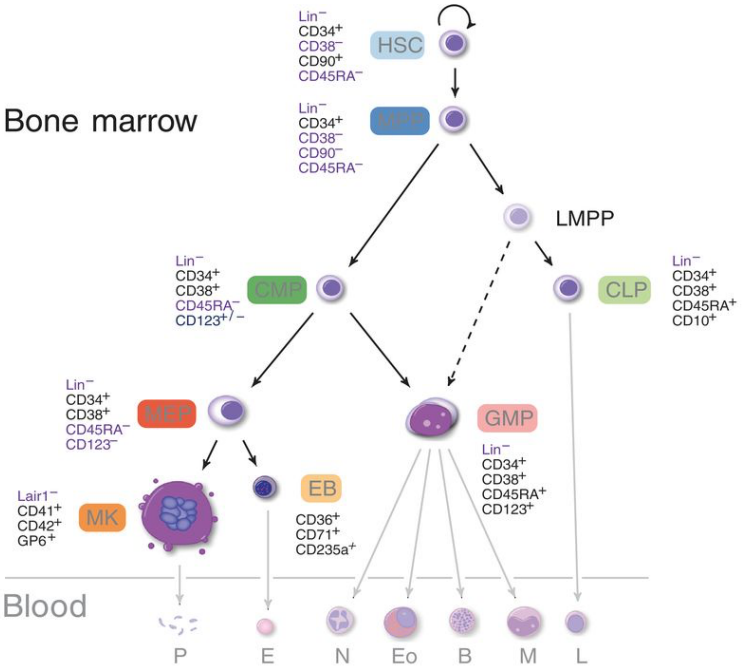



Figure 3.1: The cell populations of the haematopoietic tree analysed with RNA-seq sequenced by Chen *et al.*, 2014 [3].

## 3.1 Materials: description of the Lineage Commitment RNA-seq dataset

The Lineage Commitment dataset covers eight different cell types from the haematopoietic lineage, from stem cells to lineage restricted progenitor and to terminally differentiated cells: Haematopoietic Stem Cells (HSC), Multi Potent Progenitor cells (MPP), Common Lymphoid Progenitor cells (CLP), Common Myeloid Progenitor cells (CMP), Granulocyte–Monocyte Progenitor cells (GMP), Megakaryocyte–Erythroblast Progenitor cells (MEP), Megakaryocytes (MK) and Erythroblasts (EB) (Figure 3.1). For each cell type there are at least three biological replicates; furthermore for some samples also technical replicates are available. In the original study HSCs, MPPs, CLPs, CMPs, GMPs and MEPs were sorted from cord blood derived CD34+ cells. MKs and EBs were obtained by in vitro differentiation of CD34+ cells derived from cord blood, with thrombopoietin and erythropoietin respectively. TRIzol™ was used to extract total RNA from the samples; 100 pg of RNA were used to generate polyA+ RNA libraries with the SMARTer Ultra Low RNA and Advantage 2 PCR kits. The sequencing was performed on Illumina® HiSeq2000 instruments with TruSeq reagents, to obtain paired end reads 100 bp long. The layout of the dataset in terms of biological and technical replicates is summarised in Table 3.1 together with the number of reads sequenced and uniquely mapped.

Lineage Commitment dataset provides information on different cell types and

| | Biological replicates | | Technical replicates | |
|---|---|---|---|---|
| Cell type | Sample id | Unique reads | Sample id | Unique reads |
| HSC | EGAN00001076170 | 278 126 560 | EGAN00001095876_a | 39 116 780 |
| | EGAN00001068102 | 237 594 734 | EGAN00001095876_b | 40 937 346 |
| | EGAN00001076149 | 225 758 528 | EGAN00001095876_c | 35 254 708 |
| MPP | EGAN00001068106 | 232 892 714 | EGAN00001095875_a | 38 460 924 |
| | EGAN00001076171 | 269 943 210 | EGAN00001095875_b | 40 628 124 |
| | | | EGAN00001095875_c | 34 736 660 |
| CLP | EGAN00001076172 | 278 261 768 | EGAN00001095777_a | 21 769 286 |
| | | | EGAN00001095777_b | 24 012 542 |
| | | | EGAN00001095777_c | 25 234 278 |
| | | | EGAN00001096977_a | 34 320 484 |
| | | | EGAN00001096977_b | 36 059 028 |
| | | | EGAN00001096977_c | 31 242 500 |
| CMP | EGAN00001068103 | 285 898 104 | | |
| | EGAN00001068108 | 249 276 878 | | |
| | EGAN00001076151 | 252 893 490 | | |
| GMP | EGAN00001068104 | 269 897 054 | | |
| | EGAN00001076152 | 217 544 690 | | |
| | EGAN00001076173 | 280 676 782 | | |
| MEP | EGAN00001068109 | 233 678 804 | | |
| | EGAN00001076150 | 278 274 566 | | |
| | EGAN00001068105 | 257 474 094 | | |
| EB | | | EGAN00001095770_a | 30 228 364 |
| | | | EGAN00001095770_b | 28 677 874 |
| | | | EGAN00001095770_c | 31 323 350 |
| | | | EGAN00001095770_d | 28 346 756 |
| | | | EGAN00001095771_a | 23 854 044 |
| | | | EGAN00001095771_b | 22 379 286 |
| | | | EGAN00001095771_c | 22 661 582 |
| | | | EGAN00001095772_a | 24 828 780 |
| | | | EGAN00001095772_b | 26 457 926 |
| | | | EGAN00001095772_c | 25 134 586 |
| MK | | | EGAN00001095842_a | 37 761 680 |
| | | | EGAN00001095842_b | 40 156 756 |
| | | | EGAN00001095842_c | 38 257 390 |
| | | | EGAN00001095825_a | 32 698 630 |
| | | | EGAN00001095825_b | 34 783 176 |
| | | | EGAN00001095825_c | 33 111 900 |
| | | | EGAN00001091941_a | 32 493 332 |
| | | | EGAN00001091941_b | 34 822 618 |
| | | | EGAN00001091941_c | 32 868 522 |

Table 3.1: Details on the Lineage Commitment dataset in terms of represented cell types, biological and technical replicates available for each cell type, and sequencing depth per sample.

importantly includes several replicates per cell type. The latter feature of replicates per cell type was crucial for allowing us to develop and test the pipeline usage with multiple samples classes and biological replicates. Moreover, RNA-seq samples were obtained with Illumina® technology, thus allowing comparison with the original data we were producing in the meantime (see Chapter 4), and were sequenced with high depth (98 million uniquely mapped reads on average). Importantly, CIRI (one of the programs for circRNA detection used by the pipeline to detect backsplices and circRNAs) needs the raw input reads to be at least 50 bases long, thus the 100 nucleotides paired-end sequence reads design of the Lineage Commitment dataset resulted adequate for our purposes. Anyway, optimal RNA-seq data for circRNA detection are obtained with a library construction method involving ribosomal depletion but without polyA enrichment, since circRNAs are not polyadenylated.

The main drawback deriving from the choice of this dataset is due to the enrichment method used by the authors of the study: being circRNAs non-polyadenylated molecules the enrichment on sequences containing a polyA tail would in theory eliminate circRNAs. In practice the procedure is not 100 % effective thus the dataset is enriched in polyadenylated transcripts, but other types of RNA molecules, such as ribosomal RNAs and circRNAs, are still present after the enrichment and thus are sequenced.

It has to be kept in mind that circRNAs remaining in the sample after the polyA enrichment are randomly selected, so the most expressed are those with the higher probability of being carried over. Moreover, it is expected that the most expressed circRNAs in a cell type would be detected in the higher number of biological replicates. Furthermore we are fairly confident that if a circRNA is detected, it had been originally expressed, on the contrary we cannot exclude the possibility of false negative circRNA detections due to expressed circRNAs that were discarded by the polyA enrichment step. The pilot analysis we performed with only one of the four tools (CIRI), to deal with simplified data as our first try on circRNAs analyses, provided us with useful data to develop the custom scripts needed to handle the results with R.

In summary, the Lineage Commitment dataset allowed circRNA detection but limited to a qualitative evaluation of circRNAs isoform expression. The analysis of this dataset has been useful to test and refine the circRNA pipeline and produced also original, even though preliminary, results regarding the pervasiveness and the diversity of circRNA expression in key cell types and stages of the haematopoietic tree.

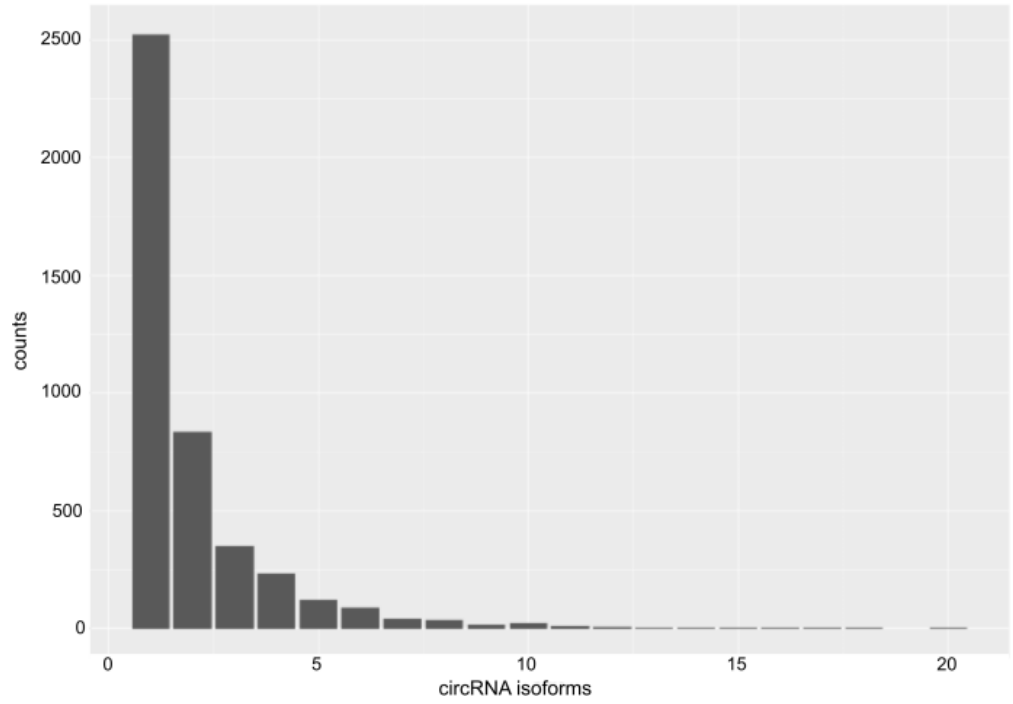



Figure 3.2: Distribution of the frequencies of the number of circular isoforms produced by each gene.

## 3.2 Results: circRNAs of the Lineage commitment RNA-seq dataset

This pilot study was performed with a pipeline version simpler than the definitive one that is described in Chapter 2, lacking also automated integration of circRNA detection and annotation results produced by different methods in parallel, that were implemented later as a series of R scripts.

Here we present a summary of circRNA discovery results grounding on CIRI method. 11 148 different circRNAs have been detected by CIRI throughout the whole dataset, expressed by 4317 different genes, meaning that more than one circRNA is expressed by the same gene of origin. In particular 58 % of the genes produce a single circRNA, whereas 837 genes (19 % of the total) produce two different circular isoforms each and 959 (22 %) three or more isoforms each (Figure 3.2).

### Technical and biological replicates in the Lineage Commitment dataset

By definition the technical replicates derive from different aliquots of the same biological sample processed in parallel. In this case enrichment, library preparation and sequencing were conducted in parallel on the same biological replicates, thus producing technical replicates useful to analyse the yield of the technical procedure. The biological replicates of this dataset are samples of a given cell type derived

from different individuals, informing also on the variability of the expression across individuals.

In our setting replicates analysis allowed us to explore the repeatability of circRNA expression from different points of view, considering specific cell populations separately, and to start filling a gap in the literature regarding circRNA expression variability. Most previous circRNA studies were largely discovery-driven and considered one or a limited number of cell types and, importantly, grounded on one sample per cell type [13, 19]. Only Memczak *et al.* in 2015 [5] compared, although in a quite preliminary way, circRNA expression similarity in two whole blood samples from healthy individuals. This study showed that out of 4300 unique circRNA candidates detected in each of the two samples at least one half are supported by at least two reads and are detected in both samples. Moreover, the expression levels in the two biological replicates were quite similar resulting in a high ($R = 0.80$) Spearman correlation between expression levels inter replicates, not very far from the inter-replicates correlation observed for linear transcripts.

### 3.2.1 HSC technical replicates analysis

We considered three replicates of the HSC cell population for circRNAs detection. The three selected technical replicates (`EGAN0001095876_a`, `EGAN0001095876_b`, `EGAN0001095876_c`) of the HSC sample `EGAN0001095876` were sequenced with average depth of 38 million reads and similar depth (39 116 780, 40 937 346 and 35 254 708 reads) in different replicates (Table 3.1).

In the three technical replicates, 349, 339 and 354 candidate expressed circRNAs were detected by CIRI, respectively. Classifying circRNAs in exonic, intronic and intergenic, according to their position with respect to genes and exons annotations, we observed that the distribution of the types of circRNAs found in the three replicates are similar (Figure 3.3), as the absolute number of circRNAs falling in each category is nearly the same in each sample.

This gives a strong indication of the consistency of the results, and it is confirmed when analysing the overlap of the putative circRNAs found in the three samples. As shown in the Venn diagram in Figure 3.4, 271 circRNAs are shared by the three samples, corresponding on average to 63 % of circRNAs expressed in each of them. Twenty-one percent of circRNAs on average is found in only one sample and the remaining 16 % is found in two of the three replicates. We speculated that the circRNAs detected in three technical replicates are the most expressed, those found only in two are less expressed and the remaining are the least expressed. Anyway, a fairly good consistency of the detected circRNAs amongst the technical replicates was observed.

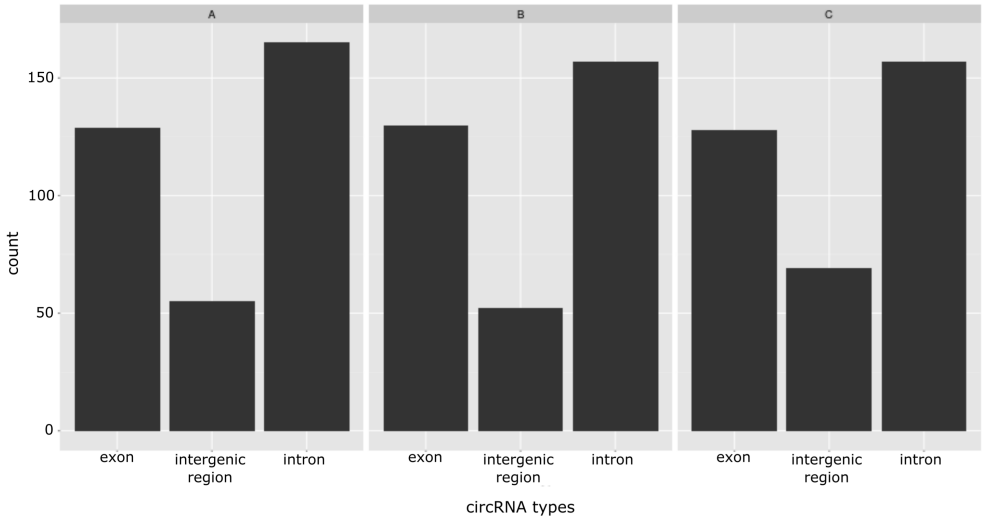



Figure 3.3: According to annotation of the corresponding genomic regions, detected circRNAs were classified into exonic, intronic and falling in unannotated (intergenic) regions; the bar plot shows the abundance of each circRNA class in the three technical replicates of HSC sample EGAN0001095876.

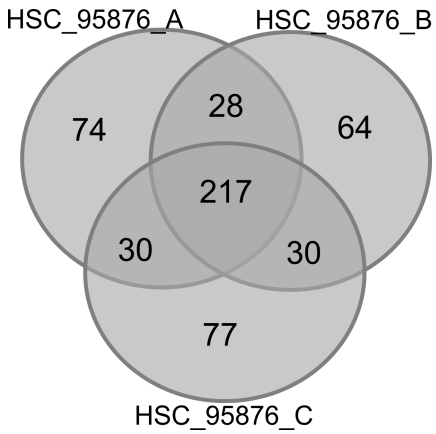



Figure 3.4: Venn diagram displaying the overlap of putative circRNAs in the three technical replicates of HSC sample EGAN0001095876.

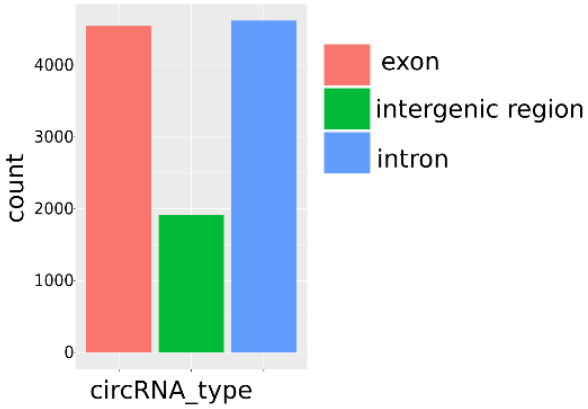



Figure 3.5: According to annotation of the corresponding genomic regions, detected circRNAs were classified into exonic, intronic and falling in unannotated (extragenic) regions; the bar plot shows the abundance of each circRNA class.

It has to be noted that the technical replicates have a lower sequencing coverage when compared to the biological replicates. That being the case, and considering that this analysis highlighted the consistency of the detected circRNAs amongst the technical replicates, we decided to collapse these replicates into a single sample to be handled as a single biological replicate. This allows a more meaningful and informative comparison of the results amongst the biological replicates of the same cell type.

### 3.2.2 Biological replicates analysis

Preliminary analyses performed with CIRI on the dataset allowed us to detect 11 148 different putative circRNAs expressed throughout the haematopoietic tree. Forty-one percent of these circRNAs are exonic, 42 % are intronic and the remaining map on unannotated regions of the genome (Figure 3.5).

Figure 3.6 and Table 3.2 give a detailed overview of the number of circRNAs detected in each sample and in particular the overlap within each cell type. The detected circRNAs are scarcely shared by the biological replicates of each cell type: only 0.77 % on average is present in every replicate, and only 4.49 % on average is present in at least two replicates.

### 3.2.3 Comparison of circRNA-ome expression in different cell types

In each cell population on average 15 % of the total circRNAs detected considering all the cell types in the tree is present. The large majority of circRNAs (87.5 %) was detected only in one of the eight cell types, thus every cell population has a specific

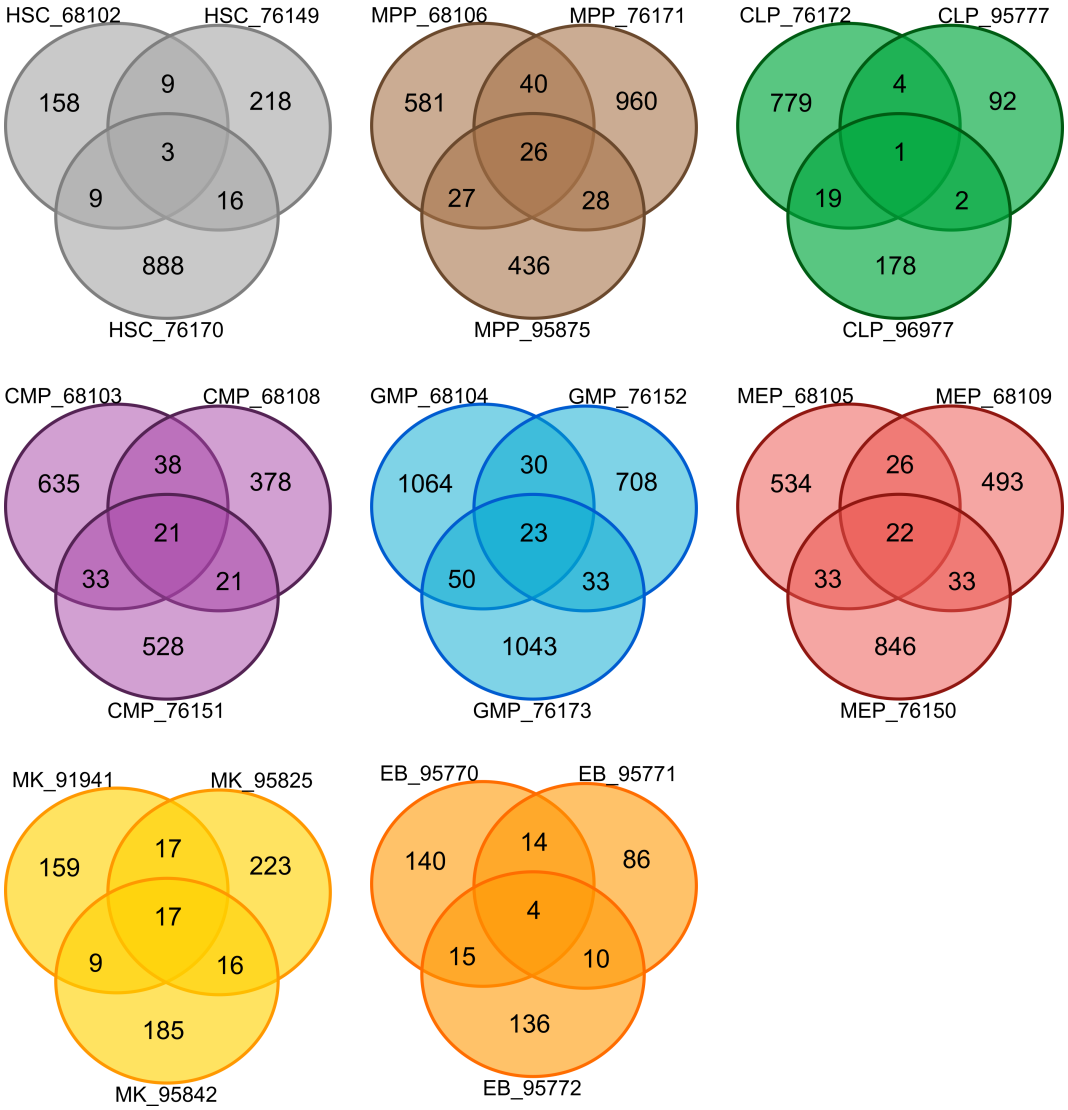



Figure 3.6: Overview of the overlap of the circRNAs retrieved in the three different biological replicates for each cell type. Samples are identified by the acronym of the corresponding cell type and the last digits of the sample ID, as they appear in Table 3.2; for example `HSC_68102` identifies sample `EGAN00001068102` of cell type HSC.

| Cell type | Sample ID | circRNAs | Total | Common to 3 | | Common to 2 | |
|---|---|---|---|---|---|---|---|
| HSC | EGAN00001068102 | 179 | 1301 | 3 | 0.23 % | 37 | 2.84 % |
| | EGAN00001076149 | 246 | | | | | |
| | EGAN00001076170 | 916 | | | | | |
| MPP | EGAN00001068106 | 674 | 2320 | 26 | 1.12 % | 121 | 5.22 % |
| | EGAN00001076171 | 1054 | | | | | |
| | EGAN00001095875 | 517 | | | | | |
| CLP | EGAN00001076172 | 803 | 1225 | 1 | 0.08 % | 26 | 2.12 % |
| | EGAN00001095777 | 99 | | | | | |
| | EGAN00001096977 | 200 | | | | | |
| CMP | EGAN00001068103 | 727 | 1654 | 21 | 1.27 % | 113 | 6.83 % |
| | EGAN00001068108 | 458 | | | | | |
| | EGAN00001076151 | 603 | | | | | |
| GMP | EGAN00001068104 | 1167 | 2951 | 23 | 0.78 % | 136 | 4.61 % |
| | EGAN00001076152 | 794 | | | | | |
| | EGAN00001076173 | 1149 | | | | | |
| MEP | EGAN00001068105 | 615 | 1987 | 22 | 1.11 % | 114 | 5.74 % |
| | EGAN00001068109 | 574 | | | | | |
| | EGAN00001076150 | 934 | | | | | |
| MK | EGAN00001095825 | 273 | 1491 | 17 | 1.14 % | 59 | 3.96 % |
| | EGAN00001095842 | 227 | | | | | |
| | EGAN00001091941 | 202 | | | | | |
| EB | EGAN00001095770 | 173 | 937 | 4 | 0.43 % | 43 | 4.59 % |
| | EGAN00001095771 | 114 | | | | | |
| | EGAN00001095772 | 165 | | | | | |
| Means | | | 1733 | | 0.77 % | | 4.49 % |

Table 3.2: Summary of circRNAs found in each considered cell type and sample. Total: circRNAs found in at least one biological replicate of each cell type. Common to 3: the circRNAs present in all biological replicates of a specific cell type and the percentage with respect to the total. Common to 2: the circRNAs present in at least 2 biological replicates of a specific cell type and the percentage with respect to the total.

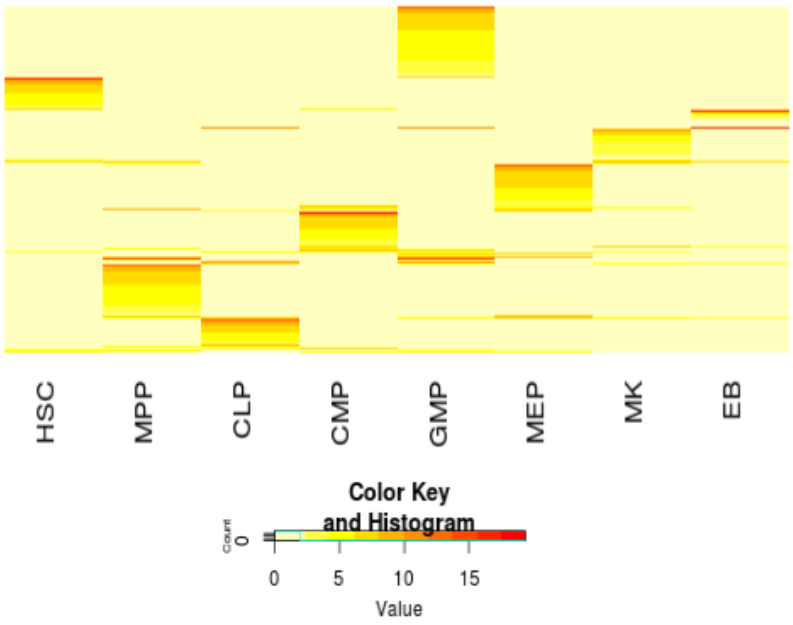



Figure 3.7: Heatmap of the expression levels (sum of the expression levels of the biological replicates of each cell type) of the 11 148 circRNAs in the dataset. For every cell type there is a well defined set of circRNAs, mostly not detect in the other cell types. Clustering is performed on the rows with Euclidean distance method. Darker colours indicate higher number of reads spanning the backsplicing junction.

set of expressed circRNAs.

Around 65 % of circRNAs detected in each cell type are 'unique', meaning that they are detected specifically in one cellular population only; thus the remaining 30–40 % is present in at least two cell types (Figure 3.7 and Table 3.3).

Considering previously reported poor overlap amongst circRNAs sets expressed by biological replicates, we reasoned that a more robust set of circRNAs in the dataset is represented by the circRNAs detected in all the biological replicates of all cell types: of the 11 148 detected circRNAs, 20 match this requirement.

Most of these circRNAs display very similar expression levels across different cell types (Figure 3.8). We can speculate that this subset of circRNAs could be part of a housekeeping set of circRNAs in the haematopoietic compartment. Therefore it would be interesting to check the expression of these circRNAs in cell types outside the haematopoietic compartment.

Amongst the 20 circRNAs in this short list, 14 derive from ten known genes. The other 6 derive from genomic regions annotated as intergenic. MALAT1 and MAN1A2 are associated to 3 circular isoforms in this subset: this is in line with the observed percentage of 31 % of genes expressing at least 2 circular isoforms (Figure 3.2). Table 3.4 summarises the genes of origin of the circRNAs displayed in Figure 3.8; all genes have ubiquitous expression amongst different tissues.

| Cell type | Total | Unique | | Shared | |
|---|---|---|---|---|---|
| HSC | 1301 | 970 | 74.56 % | 331 | 25.44 % |
| MPP | 2320 | 1655 | 71.34 % | 665 | 28.66 % |
| CLP | 1225 | 845 | 68.98 % | 380 | 31.02 % |
| CMP | 1654 | 1075 | 64.99 % | 579 | 35.01 % |
| GMP | 2951 | 2287 | 77.50 % | 664 | 22.50 % |
| MEP | 1987 | 1326 | 66.73 % | 661 | 33.27 % |
| MK | 1491 | 1038 | 69.62 % | 453 | 30.38 % |
| EB | 937 | 548 | 58.48 % | 389 | 41.52 % |

Table 3.3: Summary of the comparison between the different cell types. Total: all circRNAs found in each cell type (non redundant). Unique: circRNAs only found in the specific cell type and percentage with respect to the total. Shared: circRNAs of each cell type that are also found in other cell types and percentage with respect to the total.

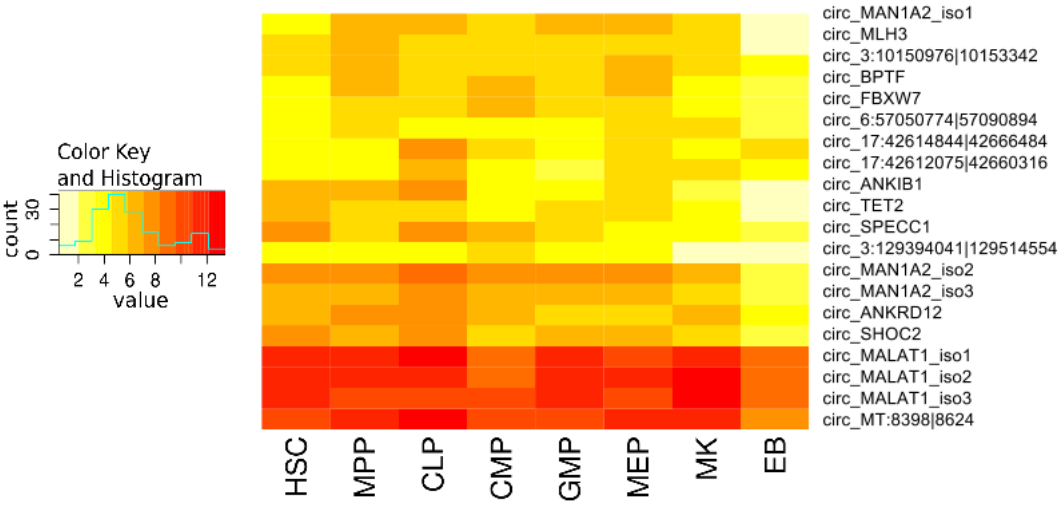



Figure 3.8: Heatmap of the expression levels (sum of the expression levels of the biological replicates of each cell type) of the 20 circRNAs shared by all cell types. Clustering is performed on the rows with Euclidean distance method. Darker colours indicate higher number of reads spanning the backsplicing junction.

| Gene | Description | Function |
|---|---|---|
| MAN1A2 | Mannosidase Alpha Class 1A Member 2 | N-glycan maturation in mammalian cells; |
| MLH3 | MutL-homologue (MLH) family member | DNA mismatch repair; frequently mutated in tumours with microsatellite instability. |
| BPTF | Bromodomain PHD Finger Transcription Factor | Regulation of transcription during proliferation; associated with Alzheimer's disease. |
| FBXW7 | F-Box And WD Repeat Domain Containing 7 | Ubiquitin-mediated degradation; potential role in the pathogenesis of human cancers. |
| ANKIB1 | Ankyrin Repeat And IBR Domain Containing 1 | Part of E3 Ligase ubiquitinating complex. |
| TET2 | Tet Methylcytosine Dioxygenase 2 | Epigenetic regulator involved in myelopoiesis; defects in this gene are associated with myeloproliferative disorders. |
| SPECC1 | Sperm Antigen With Calponin Homology And Coiled-Coil Domains 1 | Cytospin-A family, nuclear; highly expressed in some cancer cell lines. |
| ANKRD12 | Ankyrin Repeat Domain 12 | Inhibition the transcriptional activity of nuclear receptors. |
| SHOC2 | Leucine-Rich Repeat Scaffold Protein | Scaffold linking RAS to downstream signal transducers in the RAS/ERK cascade; mutations associated with Noonan-like syndrome. |
| MALAT1 | Metastasis Associated Lung Adenocarcinoma Transcript 1 (non-protein-coding, non-polyA) | Molecular scaffolds for ribonucleoprotein complexes; associated with proliferation and metastasis of tumour cells. |

Table 3.4: Functions of the ten genes from which the group of 20 circRNAs is derived.

## 3.3 Discussion and conclusions

The pilot study was designed to start handling RNA-seq data for circRNA retrieval, being this the first time for our group to investigate these molecules. Despite the polyA enrichment procedure to which the sequenced RNA had been subjected we were able to detect a large number of circRNAs. Their expression levels can not be definitely established because of the random selection of circRNAs that were actually sequenced.

The analysis of technical replicates highlighted that there is a high consistency in the detected circRNAs found by CIRI. The analysis of the biological replicates instead showed that there is high variability in the detected circRNAs in the samples of the same cell type. In this pilot analysis we exploited available data implementing innovative analyses with increased discovery power: we were able to retrieve circRNAs in specific cell types of interest, even if with quite simple analyses grounding on a suboptimal dataset. In the next chapter we will present more detailed and meaningful results regarding circRNA expression studied with an advanced version of the circPipeline thanks to a custom designed dataset, showing that studying circRNAs adds important elements to transcriptome definition at a better resolution. Obviously available RNA-seq datasets represent a rich mine for future studies to identify and characterize the circRNA-ome.

# 4 CircRNA expression in normal B cells, T cells and Monocytes

## 4.1 Methods

### 4.1.1 Samples collection and cell sorting

The samples were collected thanks to the collaboration with the Centro Trasfusionale of Padua Hospital, that granted us the access to peripheral blood of adult healthy donors.

Each whole blood sample was processed within 24 hours after the collection. Treatment with Lymphoprep™ by STEMCELL™ Technologies followed by haemolysis allowed the separation of the mononuclear cells with low red blood cells carry over. The cells were sorted with FACS Aria™ III of Becton Dickson after antibody marking (all used antibodies are produced by Beckman Coulter). Gating on CD45+ was used to isolate leukocyte cell populations and additional more specific markers (CD19+, CD3+ and CD14+) allowed to sort B cells, T cells and Monocytes respectively.

### 4.1.2 RNA extraction

The sorted cells were lysed with TRIzol™ by Invitrogen™ according to manufactural indications, and the RNA was extracted with chloroform, precipitated with isopropanol and washed with ethanol. The quality of the extracted RNA was assessed with Agilent 2100 Bioanalyzer and the final concentration with Qubit™ fluorometer. Only samples with RNA integrity measure (RIN) greater than 7 and with a concentration of at least 50 ng/µg were considered suitable for sequencing (Table 4.1).

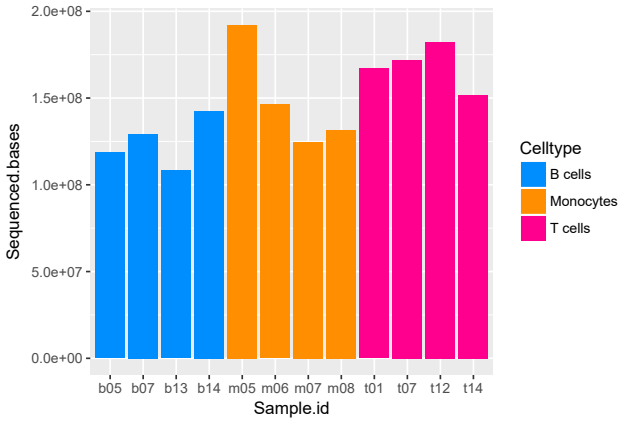
### 4.1.3 Library preparation and sequencing (RNA-seq)

RNA-seq was performed by IGA Biotechnology sequencing service. The library was prepared with TruSeq Stranded Total RNA Ribo-Zero Gold kit, to enrich the total RNA in non-ribosomal sequences. Mate-paired reads 100–125 bp long were produced with Illumina® HiSeq2000 instrument, with sequencing depth of 120 million reads per sample (Panel 4.1).

| Sample ID | Cell type | Sex | Sorted cells | RIN | RNA concentration (ng/µg) |
|---|---|---|---|---|---|
| b05 | CD19+ | M | 3 900 000 | 7.30 | 43.4 |
| b07 | CD19+ | M | 4 500 000 | 7.80 | 110.0 |
| b13 | CD19+ | M | 3 900 000 | 7.80 | 69.9 |
| b14 | CD19+ | F | 2 300 000 | 8.50 | 188.0 |
| m05 | CD14+ | M | 3 400 000 | 9.10 | 95.6 |
| m06 | CD14+ | F | 4 800 000 | 7.80 | 76.2 |
| m07 | CD14+ | M | 4 500 000 | 8.00 | 99.4 |
| m08 | CD14+ | M | 4 800 000 | 9.00 | 1180.0 |
| t01 | CD3+ | F | 7 800 000 | 8.60 | 63.6 |
| t07 | CD3+ | M | 5 000 000 | 8.30 | 151.0 |
| t12 | CD3+ | M | 7 500 000 | 8.40 | 232.0 |
| t14 | CD3+ | F | 2 500 000 | 8.40 | 306.0 |

Table 4.1: Description of collected samples, indicating cell type, sex of the donor, number of cells obtained after sorting, RIN integrity number (RIN) and final concentration of RNA.

| Sample ID | Reads |
|---|---|
| b05 | 118 585 032 |
| b07 | 129 235 360 |
| b13 | 108 236 716 |
| b14 | 142 565 832 |
| m05 | 192 086 930 |
| m06 | 146 374 148 |
| m07 | 124 539 144 |
| m08 | 131 493 682 |
| t01 | 166 967 662 |
| t07 | 172 067 828 |
| t12 | 182 155 352 |
| t14 | 151 858 242 |




Panel 4.1: Sequencing summary. Sequencing depth for each considered sample: sequenced reads are enumerated in the table and sequenced bases per sample are illustrated in the bar plot.

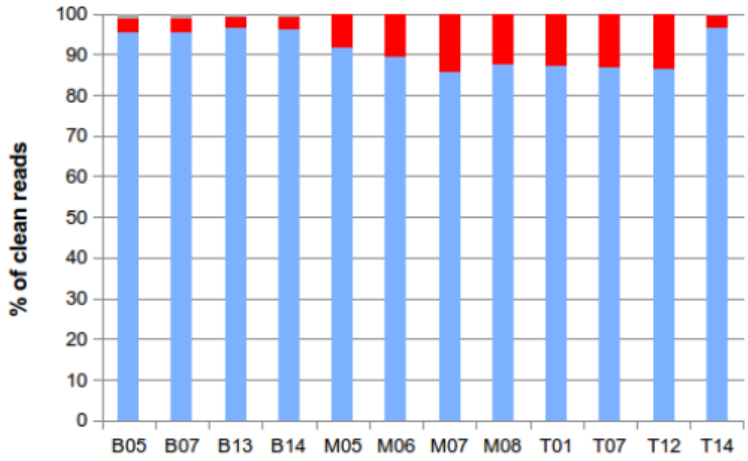



Figure 4.1: Proportions of reads collinearly mapped (blue) and unmapped (red) to the reference genome for each sample; unmapped reads contain the reads harbouring the backsplicing junction, which is used to detect circRNAs.

## 4.2 Results

### 4.2.1 CircRNAs detection and description

The preliminary HISAT2 reads alignment to reference genome (see Chapter 2 for additional details) is the first analysis step used to separate reads mapping collinearly to be fed to the linear branch of the pipeline and those unmapped to be fed to the circular branch of the pipeline. On average 91 % of the reads were mapped for each sequenced sample, the ratio between mapped and unmapped reads in different samples is displayed in Figure 4.1. B cells have a smaller fraction of reads unmapped to the linear portion, 4.2 million reads on average, compared to 11.5 millions for T cells and 10.7 millions for Monocytes, which could indicate a smaller circRNA-ome in this cell type.

CircRNAs detection grounded on the use of four different programs in parallel; each of them detected different amounts of backsplices, ranging from 1029 for CIRCexplorer to more than 330 000 for Testrealign (Table 4.2). The integration of the sets of circRNAs detected by the four programs produced a set of 350 299 putative circRNAs detected by at least one method. Of the total, 116 644 circRNAs were selected after filtering according to backsplice end distance (at least 200 nucleotides) and the number of reads spanning the backsplicing junction (at least 2 reads). The results are summarized in Table 4.2. Despite the algorithm differences between CIRI, Find_circ and Testrealign (that also find different numbers of circRNAs in the same data), their distributions of backsplice ends distances and reads numbers are similar (Figure 4.2). Both for backsplice ends distances and reads numbers CIRCexplorer marks an exception, which is probably due to the significantly lower

| Program | Total | > 200 nt | > 2 reads | Selected |
|---|---|---|---|---|
| CIRCexplorer | 1029 | 97.8 % | 51.4 % | 47.7 % |
| CIRI | 30 710 | 99.7 % | 100.0 % | 99.5 % |
| Find_circ | 26 223 | 98.8 % | 100.0 % | 97.9 % |
| Testrealign | 330 821 | 81.9 % | 40.5 % | 29.2 % |

Table 4.2: Summary of the number of unique circRNAs detected by each program in the whole set of considered cell types and samples. The second column reports the percentages of circRNAs longer than 200 nucleotides identified by each program; the third column displays the percentages of circRNAs covered by at least 2 backsplicing reads; the fourth contains the percentages of circRNAs passing both the applied filters and considered for further analyses; the percentages are calculated with respect to the initial total number of circRNAs.

number of circRNAs it detected. Figure 4.2 also shows that B cells tend to have lower backsplice ends distances than Monocytes and T cells, according to all used programs.

**Selection of a more robust subset of circRNAs** Panel 4.2 shows the overlap of circRNAs sets detected by the four programs. Find_circ and CIRI show high concordance: 83 % of circRNAs found by CIRI are also detected by Find_circ. Only 82 circRNAs (0.07 % of the total) have been detected by all the programs, while 10 321 (8.8 % of the total) are detected by at least three programs and 26 211 (22 %) are detected by at least two programs. On average 79 % of the circRNAs detected by one program is also found by at least one of the other programs.

The distribution of the distances of the backsplices is more symmetric for circRNAs detected by at least two programs, while those 'exotic' to one program display a more skewed distribution towards shorter backsplicing junctions. CIRI is the program that identifies the longest circRNAs not detected by other programs, thus displaying a distribution skewed on the other direction. In general the circRNAs detected by at least two programs are those with the highest numbers of reads spanning the backsplicing junction, with respect to the exotic ones. CIRCexplorer has the opposite trend, but this observation can not be considered significant because of the 10× lower amount of detected circRNAs compared to the other programs (Figure 4.3).

Summing up the observation that the 26 211 circRNAs detected by at least two programs were more expressed than exotic ones, and considering previous data from the study of Hansen *et al.* [85] we reasoned that this set of circRNAs probably represents a more robust set to carry on for further analysis.

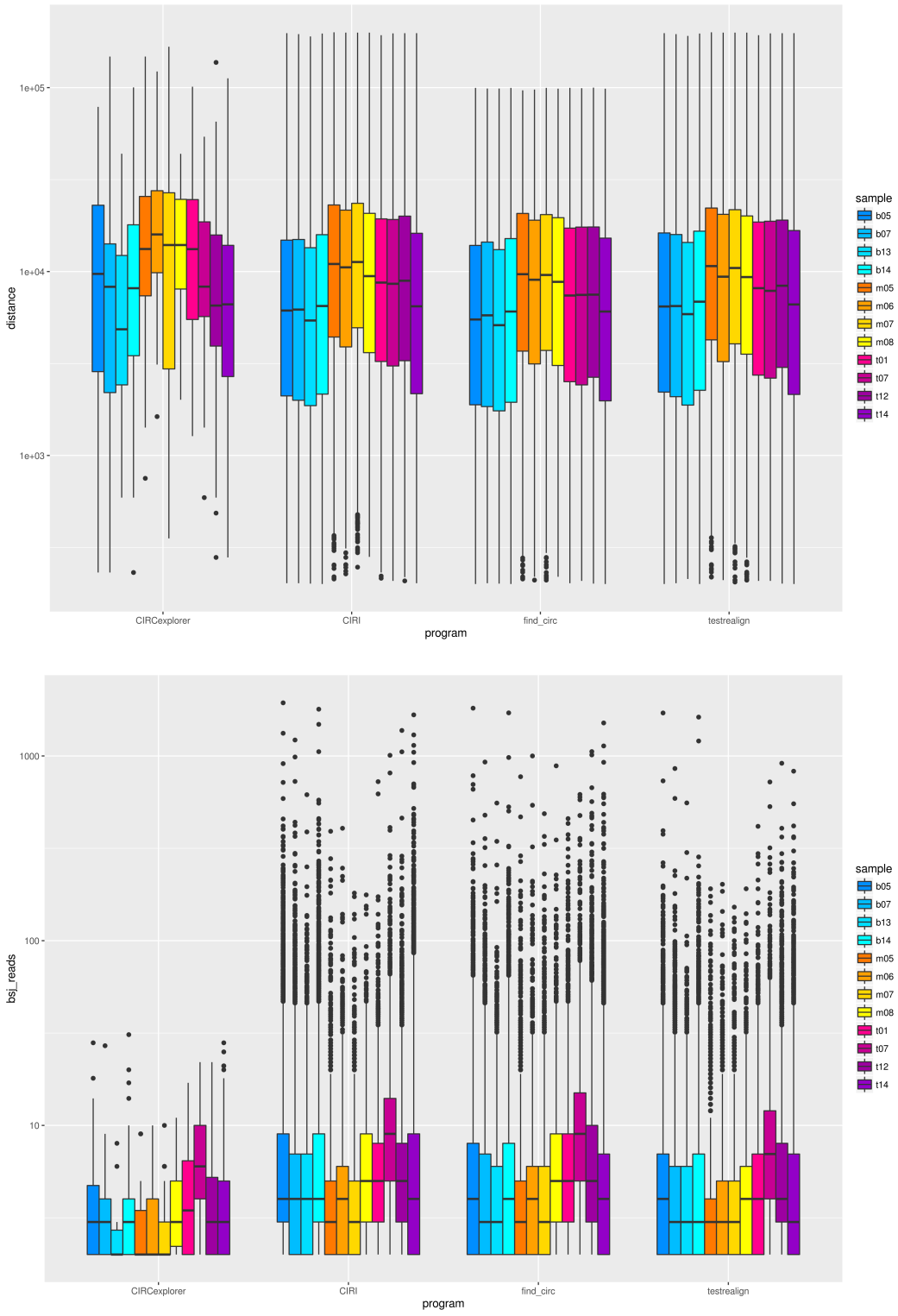
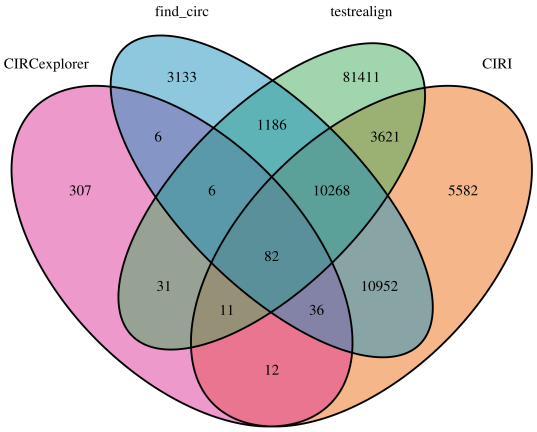
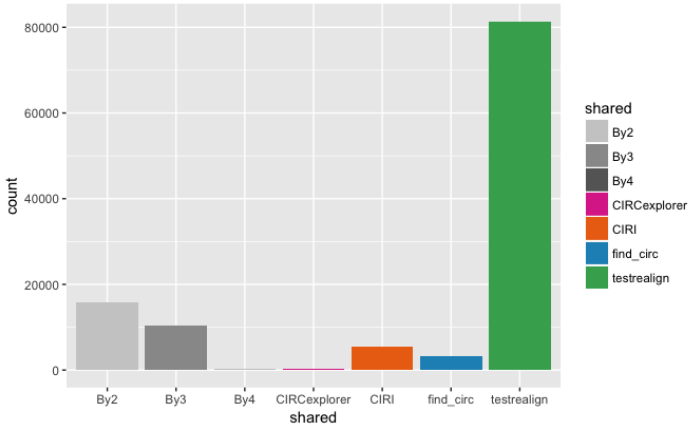
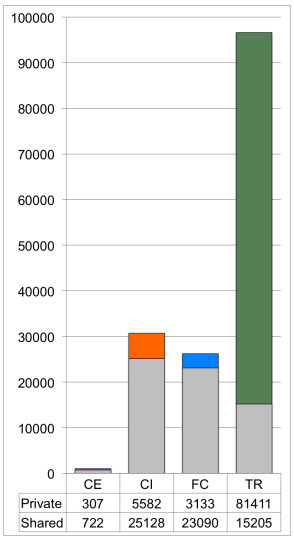



Figure 4.2: Distributions of the distances of the backsplice ends (top) and of the numbers of backsplice junction reads (bottom) for the circRNAs detected by each program in each sample.




| Category | CircRNAs |
|---|---|
| Shared by 2 | 15 808 |
| Shared by 3 | 10 321 |
| Shared by 4 | 82 |
| Only CIRCexplorer | 307 |
| Only CIRI | 5582 |
| Only Find_circ | 3133 |
| Only Testrealign | 81 411 |




Panel 4.2: Overlap of circRNA discovery results obtained with the four programs. The Venn diagram displays the overlap of circRNAs detected by different numbers of methods or only by one method. The table and the bottom left plot summarise the number of circRNAs present in the various categories deriving from the overlap. The bottom right plot displays for each program the proportion of circRNAs that are unique to that program or shared with at least another program.

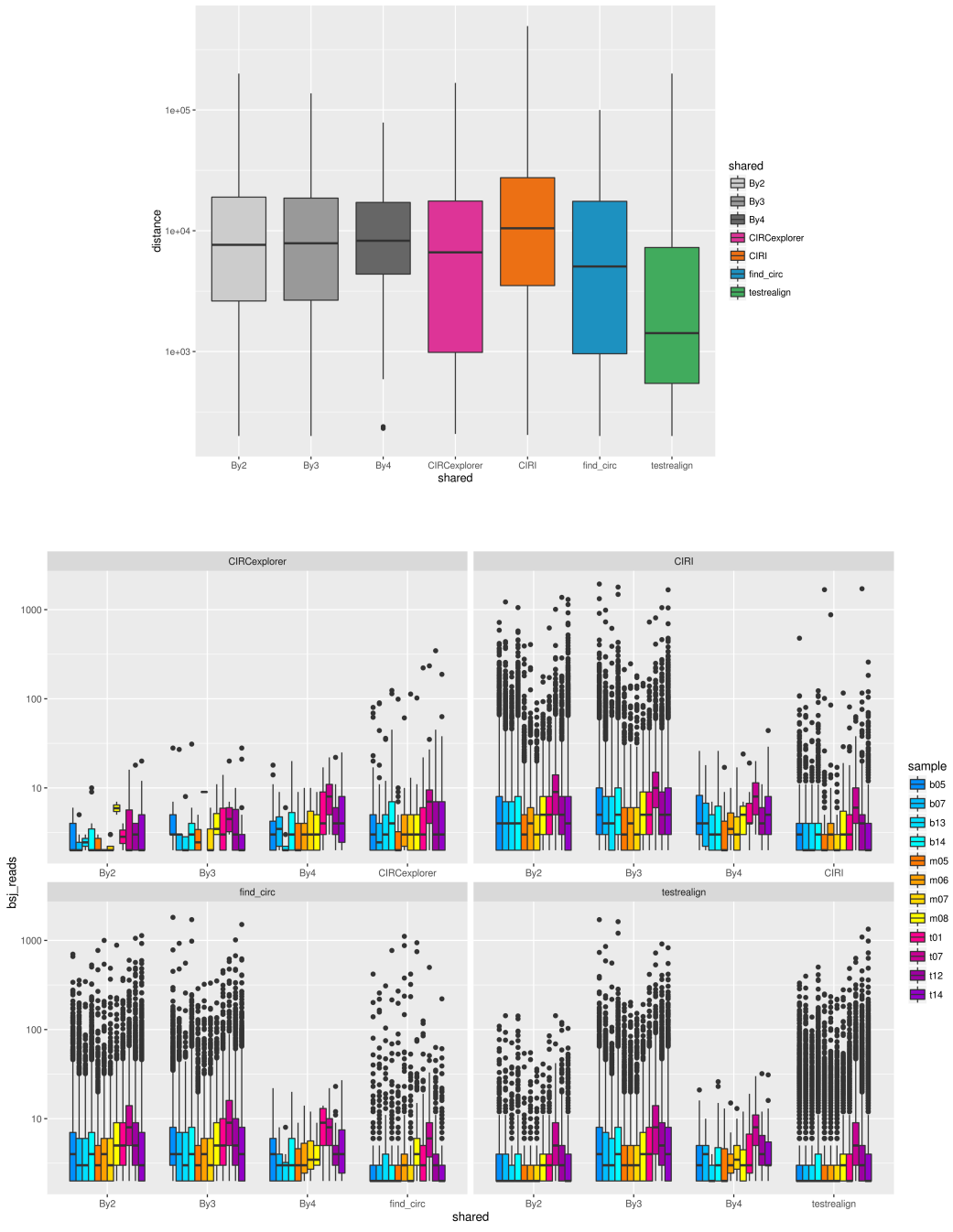



Figure 4.3: Distributions of the distances of the backsplice ends (top) and of the numbers of backsplice junction reads (bottom) for circRNAs detected by different numbers of methods or only by one method in each sample for each program.

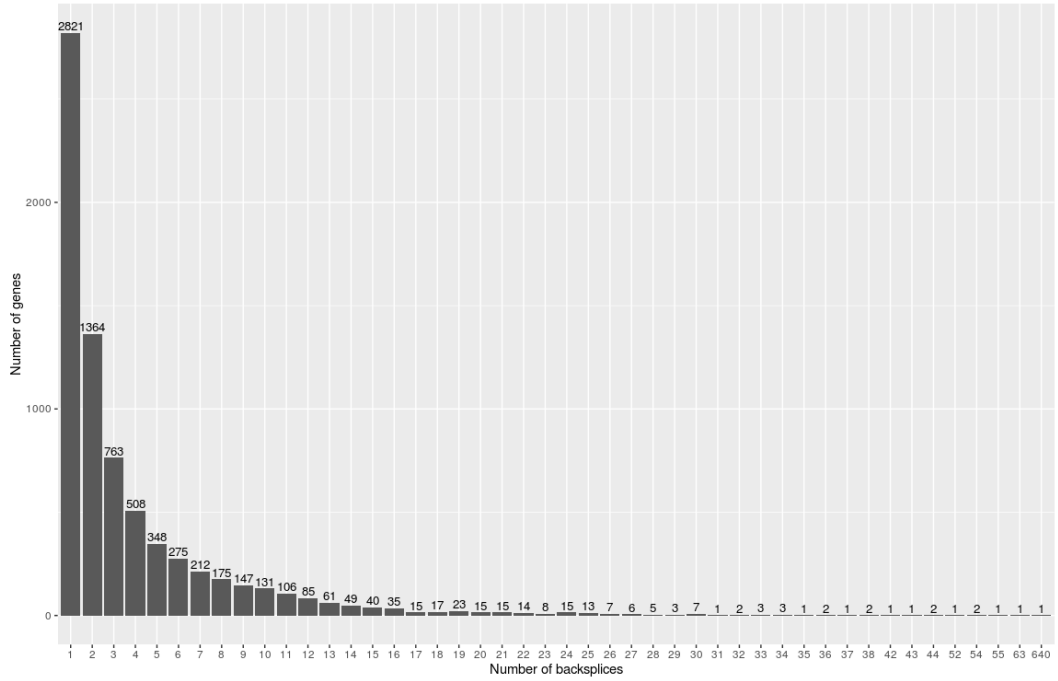



Figure 4.4: Distribution of the frequencies of circular isoforms per gene: on the x axis the number of isoforms produced by a certain gene, on the y axis the frequency of genes producing a certain number of circular isoforms. For example, the first bar indicates that there are 2821 genes producing 1 circular isoform. The last point on the x axis, 640, represents the circRNAs which have not been annotated to a specific gene, hence for which we can not establish if there are different circular isoforms.

**CircRNA isoforms deriving from the same gene in the selected subset** The subset of 26 211 circRNAs is expressed by 7307 different genes. Of these genes, 2821 (39 % of the total) produced only one circular isoform, while the remaining 61 % produced at least 2 circRNA isoforms: Figure 4.4 shows the distribution of the number of genes ranked according to the number of different circRNAs per gene. In particular 1364 genes (19 %) produce 2 circular isoforms and 1619 genes (22 %) express 3–5 different circRNAs each. Some examples of multiple circular isoforms expressed from a single gene are displayed in Figure 4.5. In these examples backsplicing positions prevalently overlap the boundaries of exons, and backsplicing ends fall in different exons. Circ_AFF1_iso1 and circ_MAN1A2_iso7 can contain at most the two exons on which the backsplicing ends map and the intron in between, while the other circular isoforms are potentially multi-exonic and multi-intronic because the backsplice ends map on non-contiguous exons, thus displaying a more complex pattern of exon combinations.

**CircRNA type classification of the selected subset** Of the selected 26 211 circRNAs the majority (22 562, 86 %) is 'exonic' (like those displayed in Figure 4.5), meaning that both the donor and acceptor site of its backsplice junction maps on

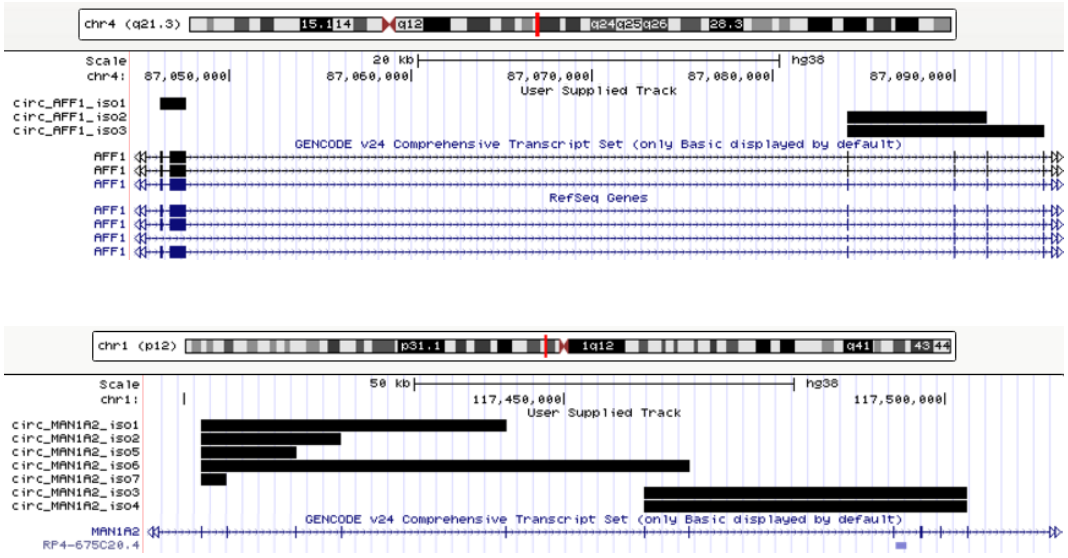
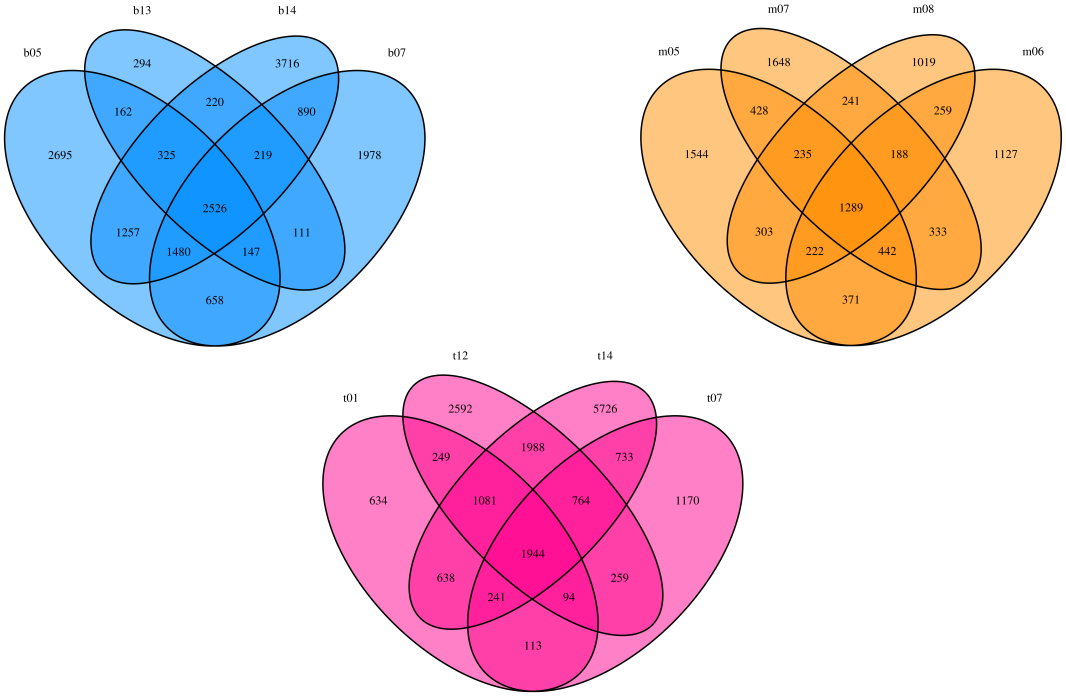



Figure 4.5: Examples of genes expressing more than one circular isoform. Black boxes represent the genomic region comprised between the detected backsplice ends of the three circRNAs deriving from the gene AFF1 on chromosome 4 (top) and of the seven circRNAs deriving from the gene MAN1A2 on chromosome 1 (bottom).

exonic sequences. 'Intronic' circRNAs — in which at least one of the ends of the backsplice maps on an intron — are 3010 (12 %), while 'intergenic' ones, mapping on unannotated regions of the genome, are 639 (2 %).

Considering exonic circRNAs, in 480 cases both backsplice ends map on the same exon, thus these circRNAs are probably produced by the circularisation of a single exon. CircRNAs of the latter subset for which the complete sequence can be retrieved are suitable candidates for *in silico* functional predictions. For circRNAs with backsplice ends far in the genome, spanning a region including two or more exons and the corresponding introns, the determination of the exact circRNA structure (included exons and retained introns) requires additional experimental analyses informing on circRNAs length or sequence.

**Indication of robustness of our results** The robustness of the selected set of detected circRNA has been tested by comparing our results with those of the previously mentioned article by Hansen *et al.* [85], referring to circRNAs detected in human fibroblast resulting 'enriched', 'unchanged' or 'depleted' by RNAse R treatment. Only about a hundred of circRNAs detected in blood cells were also detected in their samples, which was expected considering the very different cell types compared. Notably, 90 % of the overlapping circRNAs are categorized as 'enriched' in their study. This means that the exact same backsplicing junction was detected in human fibroblasts and blood cells and was not depleted by the treatment with RNAse R. This result gives a positive although indirect indication




| Cell type | Total | 4 samples | $\geq$ 3 samples | $\geq$ 2 samples | Private |
|---|---|---|---|---|---|
| B cells | 16 678 | 2526 | 4697 | 7995 | 8683 |
| Monocytes | 9649 | 1289 | 2376 | 4311 | 5338 |
| T cells | 18 226 | 1944 | 4124 | 8104 | 10 122 |

Panel 4.3: For each of the three considered cell types, the Venn diagrams and the table display and summarise the overlap of circRNAs detected in each biological replicate of the cell type.

that the majority of backsplicing junctions we detected are actually associated to circular molecules providing confidence about our results.

### 4.2.2 Comparison of circRNAs expressed in B cells, T cells and Monocytes

Panel 4.3 shows the overlap of the circRNAs detected in the three cell types: in B cells the pipeline detected 16 678 circRNAs, that is 64 % of the total circRNAs, in Monocytes 9649 (37 %) and in T cells 18 226 (70 %). In B cells 15 % of the detected circRNAs are detected in all of the four samples, 13 % of the circRNAs in Monocytes are shared by all four samples, and 10 % are shared by all T cells samples.

Of the total 26 211 detected circRNAs, 5293 (20 %) are expressed in each of the three considered cell types, 29.6 % are expressed in two cell types and 50.2 % are 'private' to one cell type at a time. It should be noted that the number of private circRNAs is highly variable across the twelve samples.


Figure 4.6: Venn diagram displaying the overlap of circRNAs expressed in the three cell types.

Regarding circRNAs specific of each cell type, 5067 (19 % of the total 26 211 circRNAs) are only detected in B cells, 6041 (23 %) are only detected in T cells and 2054 (8 %) are only detected in Monocytes (Figure 4.6). Monocytes, the cell type with less circRNAs detected — despite the highest number of sequenced bases and the highest ratio of mapped/unmapped reads — have nearly all of their circRNAs, 92 %, detected in at least another cell type. The number of circRNAs that are in common to the two lymphocytic population, 5454 (which is 32 % of their circRNAs on average), is much higher than the number of circRNAs that they share with Monocytes, that accounts for the 6 % of their total number of circRNAs.

### 4.2.3 CircRNAs expression profiles variability in considered cell types: descriptive analyses results

Descriptive analyses were performed to summarise the expression levels of the set of selected circRNAs, and to try and identify expression patterns throughout the different samples. For every circRNA, each of the backsplice detection methods provides an expression estimation. For circRNAs detected by two or more methods the expression level in each sample was calculated as the median value of the expression levels calculated by each program, after normalization of the values (see Section 2.4.2).

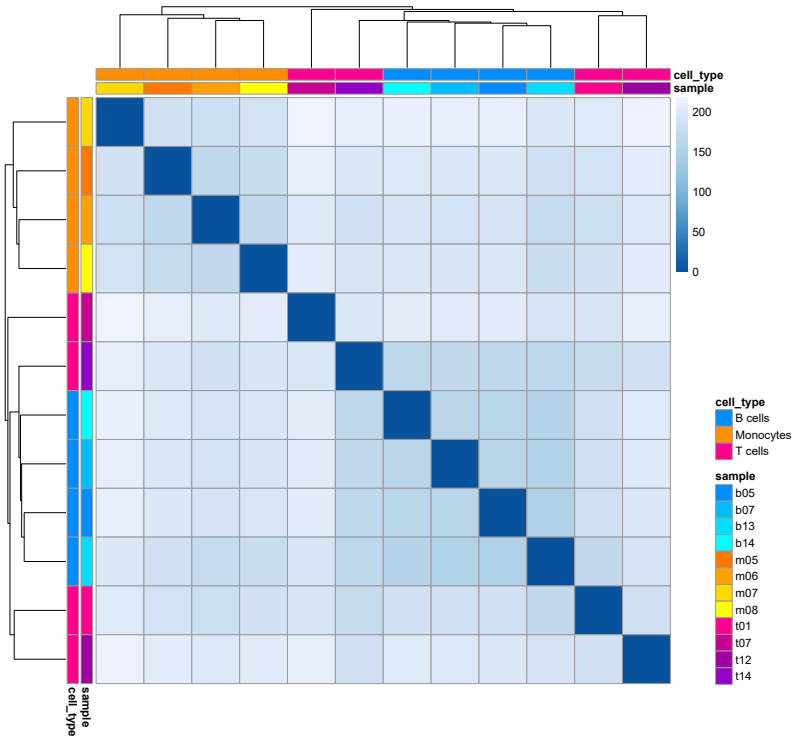



Figure 4.7: Distance plot of the different samples according to expression profiles of 26 211 circRNA detected by at least two methods (the more intense the blue colour, the more similar the samples pairwise, according to Euclidean distance).

Unsupervised clustering analysis was performed on the selected set of 26 211 circRNAs and revealed the tendency of the samples belonging to the same cell type to cluster together, except the separation is not perfect. Figure 4.7 displays the distance plot obtained by correlating the expression levels of the samples, that shows the above mentioned tendency of the various samples of each cell type to cluster together. Similar results are obtained by principal component analysis (PCA): Figure 4.8 shows the PCA plot obtained by considering the first two components describing the expression levels of the circRNAs. These two principal components, that account for less that 60 % of the total variance across sample, are incapable to segregate the cell types, meaning that the variance to be explained is to high.

This issue is probably due to the high level of background noise determined by circRNAs only present in one sample. The latter results in a high number of circRNAs with the same expression profile pattern: a certain number of sequencing reads is present in one sample only, but each of the remaining 11 samples has 0 sequencing reads.

**Biological replicates analysis** Taking the analysis to a deeper level we next examined the circRNAs expressed by the four replicates in each cell type. When considering only the circRNAs private to any given cell type (non-overlapping

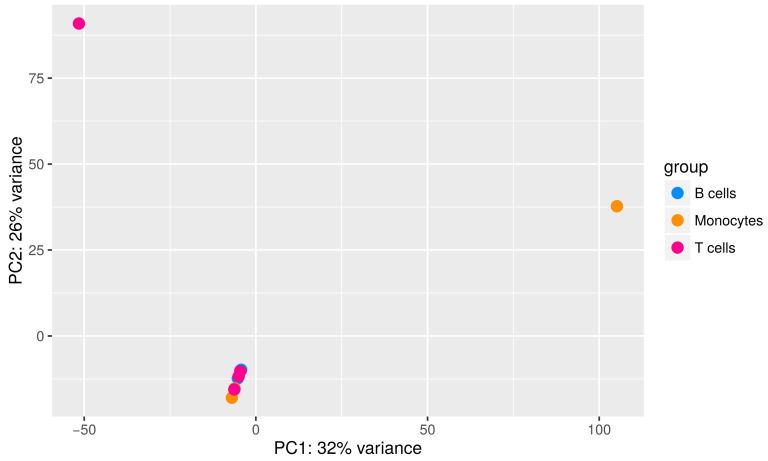



Figure 4.8: Principal component analysis of the different samples according to expression profiles of 26 211 circRNAs detected by at least two methods.

circRNAs in Figure 4.6), the overlap of the four samples allows to identify the circRNAs that are expressed in all samples of each cell type (Figure 4.9). Eighty-nine circRNAs are typical of B cells (2 %), 55 typical of Monocytes (3 %), and 43 typical of T cells (1 %). In this comparison the majority of circRNAs are private to one sample, 60 % for B cells, 78 % for Monocytes, and 80 % for T cells.

We reasoned that circRNAs with consistent expression in a given cell type will represent a more robust set to be used for a biologically relevant comparison across cell types. We thus selected the subset of 12 276 (47 % of the total) circRNAs detected in at least 2 samples of each cell type.

The unsupervised clustering analysis on this subset considering the correlation between the samples resulted in a more clear separation between the three cell types (Figure 4.10). The three cell types in fact do cluster almost perfectly, indicating that variations of circRNAs expression profiles reflect differences between cell types and that there actually are circRNAs expression specificities of each cell type, in terms of circRNAs. Moreover the PCA plot is highly significant with the two used principal components accounting for 59 % of the observed variance between the cell types. The two principal components used for this analysis are strong enough to clearly separate the samples according to the cell type.

The unsupervised clustering performed with Pearson's method on all 12 276 circRNAs perfectly separates the three cell types, but still the heatmap displays a high degree of background noise (Figure 4.11). This background noise is due to the presence of a high number of circRNAs with very low expression values throughout the samples (light blue in Figure 4.11) that do not allow to efficiently cluster the expression levels. The same analysis on the 1000 circRNAs with the highest variation coefficient (mean normalized variance) across all considered

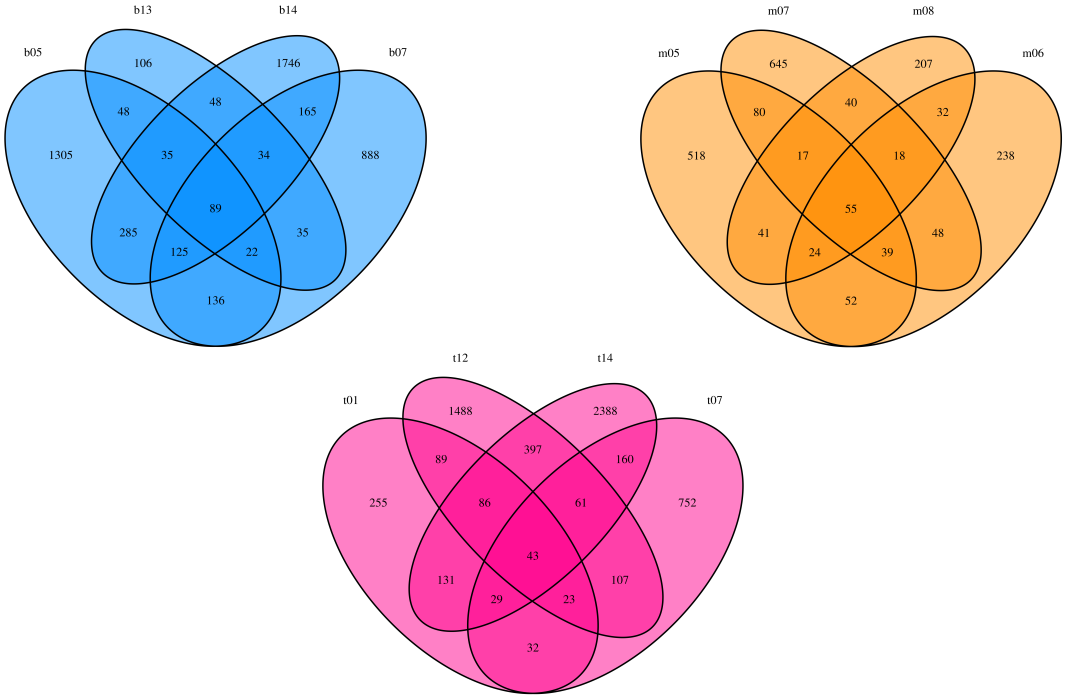



Figure 4.9: Venn diagrams displaying the overlap of circRNAs detected only in one of the three cell types when comparing the four samples of each cell type.

samples highlights the presence of subsets of circRNAs that are more specifically expressed in one cell type whereas some circRNAs are expressed in two cell types at distinct levels (Figure 4.12).

### 4.2.4 CircRNAs differential expression analysis

The above illustrated descriptive analysis gives a strong indication of the presence of differentially expressed circRNAs in the three cell types. One of the main aims of the study was the identification of expression specificities of considered differentiated blood cell types. Therefore differential expression analysis was performed on the whole set of 26 211 circRNAs detected by at least two methods. For each pairwise comparison (B cells vs. Monocytes, B cells vs. T cells and Monocytes vs. T cells) DESeq2 analysis with false discovery rate threshold of 0.05 identified a subset of circRNAs with significant differential expression: 2589, 168 and 977 respectively (see Table 4.3).

We also identified a subset of circRNAs differentially expressed in a certain cell type compared to both the other two. In particular 74 circRNAs are more expressed in B cells compared to Monocytes and T cells, 40 are more expressed in T cells, and 159 are more expressed in Monocytes, for a total of 273 circRNAs upregulated in one cell type compared to the other two.

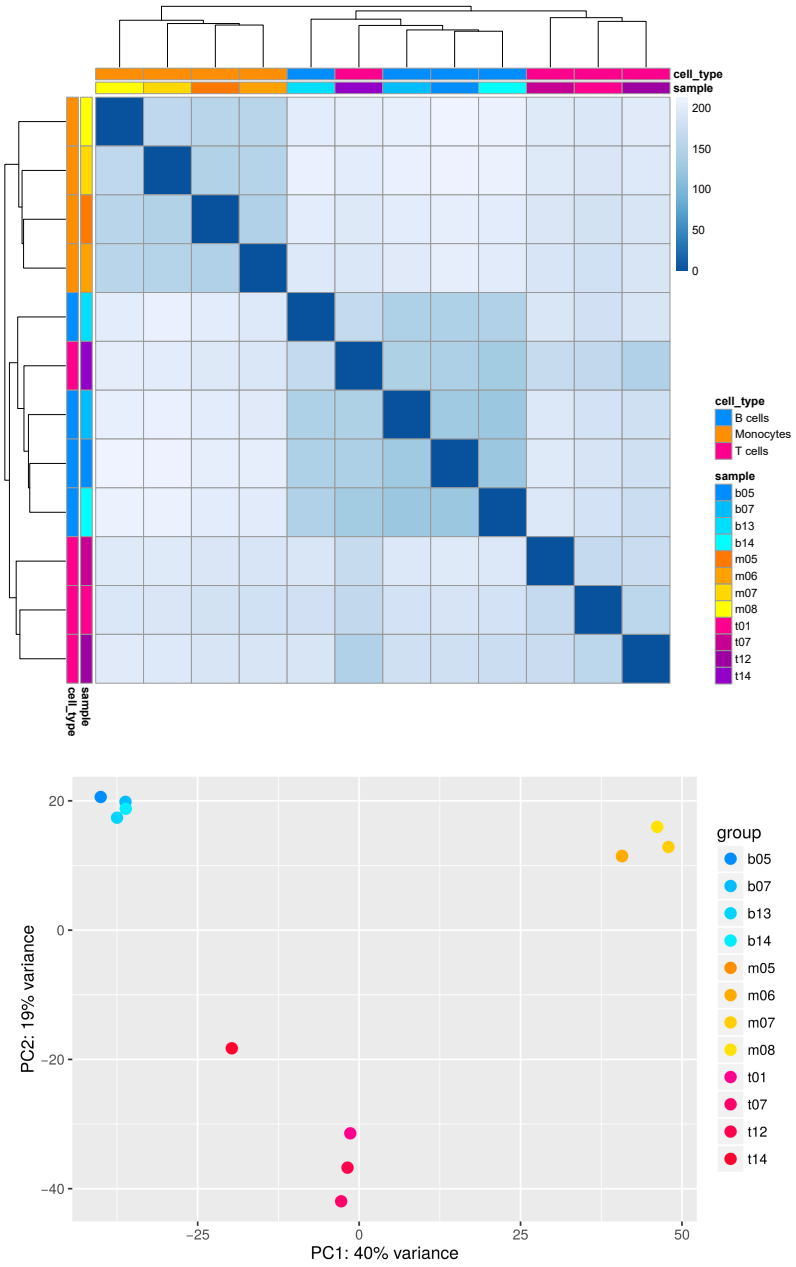



Figure 4.10: Unsupervised analyses describing the similarity of different samples according to expression profiles of 12 276 circRNAs detected by at least two methods and in at least 2 samples of each cell type: distance plot (top; the more intense the blue colour, the more similar the samples pairwise, according to Euclidean distance) and PCA plot (bottom) of the different samples according to selected circRNAs expression profiles.

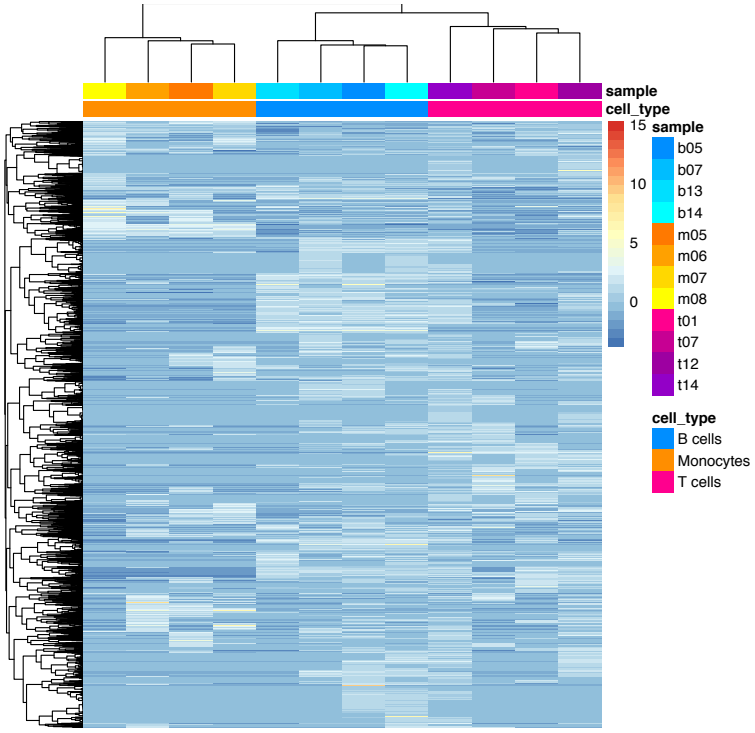



Figure 4.11: Heatmap of expression profiles of 12 276 circRNAs detected by at least two methods and in at least 2 samples of each cell type, showing also unsupervised clustering of samples and cell types.

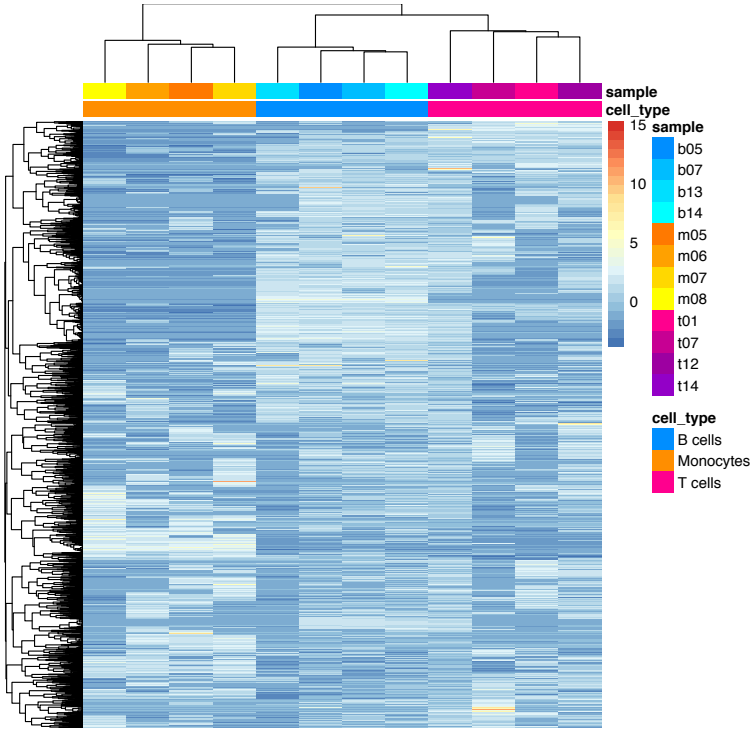



Figure 4.12: Heatmap of expression profiles of the 1000 circRNAs with the highest variation coefficient across considered samples among circRNA detected by at least two methods and in at least 2 samples of each cell type, showing also unsupervised clustering of samples and cell types.

| Comparison | Total | padj $< 0.05$ | | $\log_2$ FC $< 0$ | | $\log_2$ FC $> 0$ | |
|---|---|---|---|---|---|---|---|
| B vs. M | 19 666 | 2589 | 13.0 % | 2037 | 10.0 % | 552 | 3.0 % |
| B vs. T | 24 039 | 168 | 0.7 % | 99 | 0.4 % | 69 | 0.3 % |
| M vs. T | 20 693 | 977 | 5.0 % | 245 | 1.0 % | 732 | 4.0 % |

Table 4.3: Summary of the differential expression analysis. The first column displays the actual number of performed comparisons between sample groups, while the second column displays the number of significantly differentially expressed circRNAs (P-value adjusted for multiple testing lower than 0.05 detected). The last two columns display separately the number of down- and up-regulated circRNAs respectively in the comparison, and percentage over all the significantly differentially expressed circRNAs.

**Functional enrichment based on genes producing circRNAs** We decided to perform functional enrichment analysis to identify pathways or tissues whose genes are significantly enriched in a given sample, on the genes of the set of differentially expressed circRNAs. This analysis was performed with DAVID testing GO terms, UP_TISSUE and KEGG pathways databases. Many of the genes that produce circRNAs that are differentially expressed in at least one of the analysed cell types encode proteins that are involved in pathways and tissues related to haematopoiesis. Significantly enriched KEGG pathways are VEGF, MAPK, RAS signalling pathways, Lysine degradation pathway and Pathways in cancer; also genes expressed in T cells, Bone Marrow, Platelets, Lymphoblasts and Lymph nodes are enriched. Together with these very specific sets of genes also terms regarding more general purpose pathways are enriched, such as GO terms Protein binding, Membrane, ATP-binding, Cytosol and Cytoplasm. This analysis only gives an indirect indication of the possible processes in which the differentially expressed circRNAs are involved, and gives us an instrument to select a subset of circRNAs that should be deeper analysed.

**Short list of circRNAs based on differential expression and genes of origin** Amongst the circRNAs differentially upregulated in one cell type compared to the others, and considering the presence of their origin gene in enriched pathways we selected a short list of circRNAs with the highest $\log_2$ fold change (Table 4.4) to be further investigated.

Some of these circRNAs are produced by genes that were identified in the functional enrichment we performed. We checked for the presence of these circRNAs in circBase [40]. Interestingly 11 of the circRNAs were retrieved in the database, meaning they have already been detected in some of the datasets that are part of circBase. This is yet another indirect indication on the reliability of our results. The expression levels across the three cell types are displayed for some of the circRNAs

| B cells, T cells and Monocytes dataset | | | | | circBase | | |
|---|---|---|---|---|---|---|---|
| Up in | circRNA_ID | Gene | Single exon | Lin/circRNA correlation | circRNA | circRNAs/organism | Tissues |
| B cells | 5:88947807\|88948215:+ | MEF2C-AS1 | no | | | N. A. | |
| B cells | 9:37002647\|37020801:- | PAX5 | no | 0.635 595 7 | hsa_circ_0001857 | 2 hsa + 1 mmu | CD19+, Gm12878 |
| B cells | 18:44701174\|44707222:+ | SETBP1 | no | 0.415 045 63 | hsa_circ_0108445 | > 20 hsa + 1 lcm + 1 mmu | CD19+, Hs68, Gm12878, muscle, brain |
| B cells | 9:146101\|164037:- | CBWD1 | no | −0.030 706 71 | | 3 hsa | broad cell types |
| B cells | 7:90747680\|90863269:+ | CDK14 | no | 0.541 406 5 | hsa_circ_0080981 | > 30 hsa + 11 mmu | brain regions, broad cell types |
| B cells | 7:116110707\|116112038:- | TFEC | no | 0.631 188 1 | | N. A. | |
| B cells | 5:97101758\|97103094:+ | CTD-2215E18.1 | no | 0.584 849 8 | | N. A. | |
| T cells | 14:99231344\|99257839:- | BCL11B | no | 0.512 376 2 | hsa_circ_0103050 | 8 hsa + 2 mmu | brain regions |
| T cells | 16:30481438\|30484263:+ | ITGAL | no | 0.251 204 9 | | 20 hsa | Gm12878, CD19+, CD34+ |
| T cells | 6:130154824\|130184623:- | SAMD3 | no | 0.420 206 4 | | 10 hsa | Gm12878, Hs68, brain regions |
| T cells | 2:84870224\|84870785:- | TRABD2A | yes | −0.603 266 6 | | 1 hsa | brain regions |
| T cells | 4:48095142\|48114402:- | TXK | no | −0.254 012 9 | | 1 hsa | K562 |
| T cells | 9:6475390\|6500709:+ | UHRF2 | no | −0.225 130 6 | hsa_circ_0008560 | > 20 hsa + 3 mmu | brain regions, broad cell types |
| Monocytes | X:148661907\|148662768:+ | AFF2 | yes | −0.261 172 | hsa_circ_0001947 | 11 hsa + 2 mmu | brain regions |
| Monocytes | 4:105233896\|105237351:+ | TET2 | yes | 0 | hsa_circ_0070562 | 10 hsa + 1 mmu | K562, Mcf7, brain regions |
| Monocytes | 12:32598496\|32611283:+ | FGD4 | no | −0.002 214 399 | hsa_circ_0025843 | > 39 hsa + 5 mmu | brain regions, broad cell types |
| Monocytes | 12:66203710\|66228370:+ | IRAK3 | no | 0.491 502 9 | hsa_circ_0005505 | 12 hsa | Gm12878, Hs68 |
| Monocytes | 19:325634\|336173:- | MIER2 | no | −0.141 089 1 | hsa_circ_0002005 | > 20 hsa + 1 mmu | brain regions, broad cell types |
| Monocytes | 5:83537006\|83542268:+ | VCAN | yes | 0.606 438 2 | hsa_circ_0073237 | > 20 hsa + 5 mmu | CD19+, broad cell types |

Table 4.4: Short list of genes producing a circRNA that resulted differentially expressed in our study, indicating also if circRNAs were already detected and reported for the gene in circBase and the correlation between the expression profile of each circRNA and the corresponding gene in considered blood cell types. Up in: cell type in which the circRNA is upregulated; Gene: gene producing the circRNA detected in our dataset; Single exon: whether the circRNA is composed of a single circularised exon; Lin/circRNA correlation: correlation between the expression levels of the linear and circular transcript of the same gene of origin; circRNA: circRNA identifier in circBase; circRNAs/organism: how many circRNAs are present in circBase deriving from the same gene of origin, and in which organisms they were retrieved; Tissue: in which tissues or cell lines the circRNAs deriving from the selected genes and present in circBase were detected.

in the short list in Figure 4.13. Many of the analysed circRNAs are solely expressed in one of the cell types, thus the high fold change for which they were selected is expected, while some are also present in other cell types with a specific level of expression.

### 4.2.5 Relations among circRNAs and linear RNAs expressed from the same genes

We next turned to the data produced by the linear branch of the pipeline, to analyse the relationship between the expressed circRNAs and the linear transcripts produced by the same genes in our dataset. It is reported [13] that the ratio between the expression of circular and linear transcripts of the same gene is widely variable, as there are circRNAs more expressed than the linear transcripts and vice versa.

The pipeline detected 39 345 genes expressed in Monocytes, 49 651 genes expressed in T cells and 53 010 genes expressed in B cells. Of these genes, in B cells 3.4 % of the genes contribute to the 95 % of the expression, for Monocytes this is 3.6 % and for T cells it is 4.3 % of the genes, meaning that the great majority of genes have low expression levels. Figure 4.14 shows the distributions of expression levels of genes per cell type and per sample.

Figure 4.15 displays the distribution of the values of correlation between the expression levels of a circRNA and the expression level of the gene from which the circRNA is produced. The distribution has a normal-like shape centred very near to 0 on $R = 0.1$, meaning that the expression levels of circRNAs and genes only have a slight tendency to correlate. Correlations with $R > 0$ indicate that the linear and the circular transcript have a similar expression level, whereas $R < 0$ indicates that the expression levels of the circular and linear transcript are discordant, both with the linear more expressed than the circular and the opposite and more intriguing scenario of the circular more expressed than the linear transcript. Notably, most of the pairs display a weak correlation, as 6665 pairs have absolute level of R lower than 0.3 (69 % of the total pairs) or even display a strong but negative correlation: 975 (10 %) with $R < -0.3$.

## 4.3 Discussion

The analysis of our dataset is an important step further in the characterization and understanding of blood cell circRNA-ome, which up to now was only preliminarily studied. The only available study on circRNAs in blood cell with replicates to increase true discovery rate is the one by Memczak *et al.*, 2015 [5], which analysed

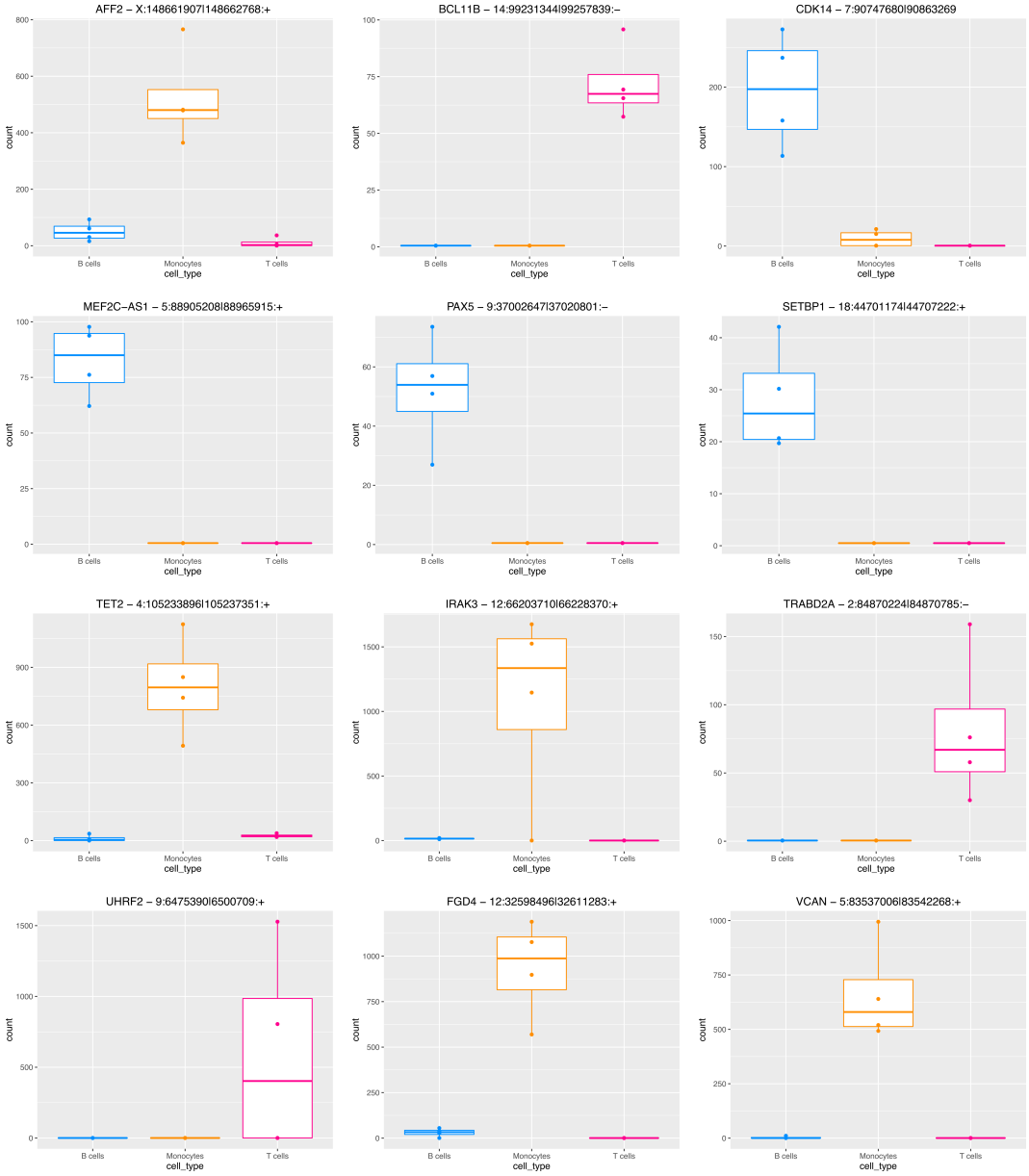



Figure 4.13: Expression levels across all cell types of 12 differentially expressed circRNAs in the short list summarised in Table 4.4.

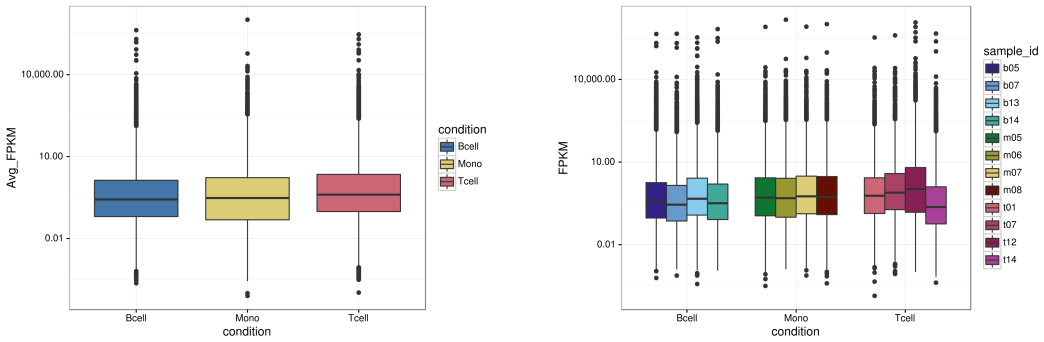



Figure 4.14: Distributions of expression levels of genes per cell type (left) and per sample (right).

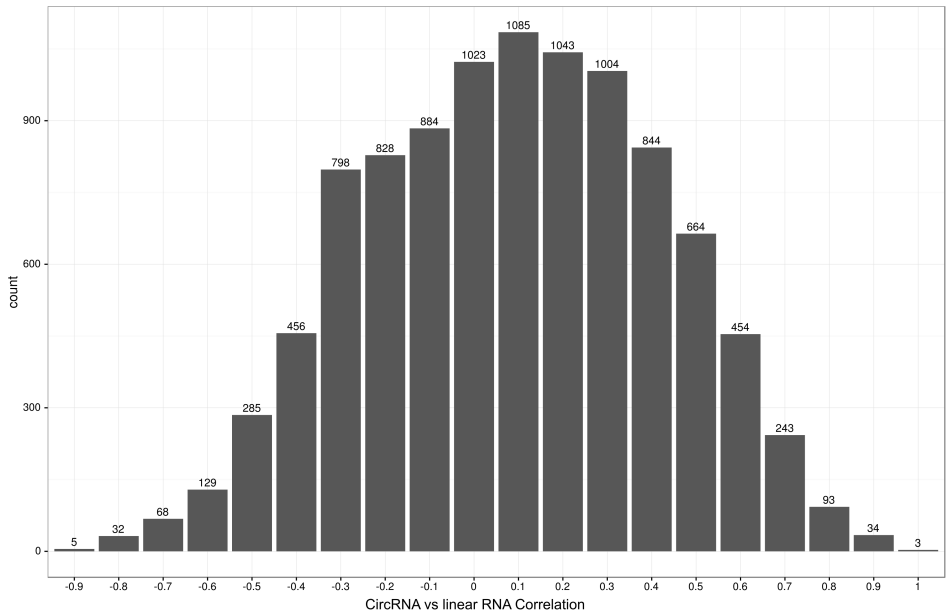



Figure 4.15: Distribution of the correlations between the expression levels of a gene and the circRNAs deriving from the same gene.

5 whole blood replicates. In this study they detected more than 28 000 circRNAs in total using a single tool for retrieval, which is comparable with our robust set of 26 211 circRNAs identified with the use of four tools in parallel.

The four programs have radically different circRNAs detection yields, which is due to the substantial differences of the algorithms they use. The different yields and the low percentage of overlapping circRNAs across the programs has already been pointed out by Hansen *et al.* in their 2015 study [85], suggesting the use of more than one program in parallel as a strategy to overcome this issue. Experimental technical validations are needed though to definitely assess the sensitivity and specificity of this approach to *in silico* detection.

Chen *et al.* in their 2014 study [3] analysed a high number of samples of 8 different cell types, but our re-analysis of their data only retrieved about one third of the amount of circRNAs in our dataset. Given the high number of samples and of different cell types a greater number of circRNAs would have been expected, except this dataset had been obtained with polyA enrichment instead of ribosomal RNA depletion. Ribosomal RNA depletion requires a higher amount of input RNA, compared to polyA enrichment, but we can confirm that it ensures better results on circRNAs detection, without the drawback of not allowing the calculation of a precise expression level. It should also be noted that Chen *et al.* could have slightly underestimated the expression levels of non-coding RNAs, because not all ncRNAs are polyadenylated. A thorough analysis of the samples we sequenced, to estimate the coding/non-coding transcripts ratio in different cell types, should be performed, to be compared with the results of the above mentioned study, that indicated an enrichment in coding transcripts in differentiated cell populations and of non-coding transcripts in progenitors. Anyway revisiting publicly available datasets treated with polyA enrichment can be useful for preliminary indications regarding the presence of circRNAs in malignancies too. On this subject a very interesting role is played by f-circRNAs, that are produced by fusion genes in genomic aberrations. Publicly available RNA-seq datasets of haematological malignancies harbouring genomic aberrations are thus perfect candidates for re-analysis focused on circRNAs detection.

Considering the differences observed in the analysed cell types the whole of these analyses will result in the definition of subsets of circRNAs putatively relevant in the biology of the analysed cells. These subsets of circRNAs can be compared with circRNAs described in literature or present in databases, to see if some of them have already been detected or at least if other circular isoforms from the same gene have been described. Preliminary analyses in this sense have been performed for circRNAs short-listed for their differential expression or for the function of

the gene producing them according to the functional enrichments performed, that highlighted that many of the genes involved in circRNAs production belong to haematopoiesis or cancer related pathways, such as *PAX5*, SETBP1, CDK14, BCL11Band AFF2. Interestingly enough, detected circRNAs of these genes have already been retrieved in other studies performed on cell lines, CD19+ cells and brain region samples. This piece of information strongly indicates the convenience of prioritising these circRNAs both for technical and biological validation, in order to start functional studies to characterise their role in the cell types in which are up significantly expressed.

Complete functional predictions though can also be performed for circRNAs whose sequence is known, as in the case of circularised single-exon, for example for TRABD2A, AFF2, TET2 and VCAN of our short list. Furthermore preliminary functional predictions can be performed on the sequence covered by backsplicing reads, that is a sequence not existing in the reference genome, which could harbour newly formed binding sites or sequence motifs. Given the interesting features of the short-listed set of circRNAs they will be prioritised for experimental validation.

As previously mentioned experimental validation is needed to complete this study and to give a start to functional experiments. First of all putative circRNAs should be technically validated, namely their actual expression in the samples in which they were retrieved should be experimentally tested, to obtain indications on the specificity of the four used programs and on the false discovery rate of the pipeline. To perform this technical validation circRNAs with diverse expression levels should be tested in the original samples. Furthermore to speculate on the sensitivity of the circRNAs detection of the circPipeline, some circRNAs with low expression levels should be considered for experimental validation. Next the validated circRNAs should be tested for detection in other samples of the same cell type to check if the expression pattern described in the samples we sequenced is reproducible, i.e. if a circRNAs is present in all samples of our dataset, it should be also present in any other sample of the same tested cell type.

### 4.3.1 Conclusions

This project allowed to acquire novel and deeper knowledge on circRNAs in healthy haematopoietic compartment, and besides on the methods useful to their detection, characterisation and analysis from RNA-seq data. The development of the semi-automated pipeline allows to perform these analyses in a reproducible way with increased true positive detection rate, compared to the methods previously available.

The identification of a high number of circRNAs, many of which with differential

expression in blood cell types, together with the great interest of some circRNAs because of the gene from which they derive, encourages to take the analyses a step further. Once the functional characteristics of these circRNAs have been described in normal haematopoiesis, the behaviour of these molecules in haematopoietic malignancies should be thoroughly analysed.

This more refined and complete analysis of the transcriptome in normal haematopoiesis will allow to unveil mechanisms underlying cellular processes in healthy condition and to better understand their dysregulation in haematologic malignancies.

# Credits

The work of this thesis has been carried out in the frame of a collaboration between the research groups led by prof. Geertruij te Kronnie and prof. Stefania Bortoluzzi.

In particular prof. Stefania Bortoluzzi, dr. Enrico Gaffo and dr. Andrea Bisognin collaborated to the development of the bioinformatics pipeline; prof. Geertruij te Kronnie, dr. Silvia Bresolin, dr. Luca Trentin and dr. Chiara Frasson were involved in sample collection and preparation for sequencing; prof. Stefania Bortoluzzi, dr. Enrico Gaffo and prof. Geertruij te Kronnie collaborated to the analyses of the datasets.

# References


1. Orkin, S. H. & Zon, L. I. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell* **132,** 631–644 (2008).
2. Iwasaki, H. *et al.* The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes & Development* **20,** 3010–3021 (2006).
3. Chen, L. *et al.* Transcriptional diversity during lineage commitment of human blood progenitors. *Science* **345.** doi:10.1126/science.1251033 (2014).
4. Lasda, E. & Parker, R. Circular RNAs: diversity of form and function. *RNA* **20,** 1829–1842 (2014).
5. Memczak, S., Papavasileiou, P., Peters, O. & Rajewsky, N. Identification and characterization of circular RNAs as a new class of putative biomarkers in human blood. *PloS one* **10,** e0141214 (2015).
6. Singh, R. K. & Cooper, T. A. Pre-mRNA splicing in disease and therapeutics. *Trends in Molecular Medicine* **18,** 472–482 (2012).
7. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478,** 64–69 (2011).
8. Wan, Y. & Wu, C. J. SF3B1 mutations in chronic lymphocytic leukemia. *Blood* **121,** 4627–4634 (2013).
9. Cazzola, M., Porta, M. G. D. & Malcovati, L. The genetic basis of myelodysplasia and its clinical relevance. *Blood* **122,** 4021–4034 (2013).
10. Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-type specific features of circular RNA expression. *PLoS Genet* **9,** e1003777 (2013).
11. Li, Z. *et al.* Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* **22,** 256–264 (2015).
12. Zhang, Y. *et al.* Circular intronic long noncoding RNAs. *Mol Cell* **51,** 792–806 (2013).
13. Rybak-Wolf, A. *et al.* Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol Cell* **58,** 870–885 (2015).
14. Dean, M., Fojo, T. & Bates, S. Tumour stem cells and drug resistance. *Nat Rev Cancer* **5,** 275–284 (2005).

15. Costa, D. *et al.* Metformin inhibition of neuroblastoma cell proliferation is differently modulated by cell differentiation induced by retinoic acid or overexpression of NDM29 non-coding RNA. *Cancer Cell Int* **14,** 59 (2014).

16. Bachmayr-Heyda, A. *et al.* Correlation of circular RNA abundance with proliferation-exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. *Sci Rep* **5,** 8057 (2015).

17. Li, P. *et al.* Using circular RNA as a novel type of biomarker in the screening of gastric cancer. *Clin Chim Acta* **444,** 132–136 (2015).

18. Bahn, J. H. *et al.* The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. *Clin Chem* **61,** 221–230 (2015).

19. Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS one* **7,** e30733 (2012).

20. Jeck, W. R. *et al.* Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19,** 141–157 (2013).

21. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495,** 333–338 (2013).

22. Westholm, J. O. *et al.* Genome-wide analysis of Drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep* **9,** 1966–1980 (2014).

23. Zhang, Z. *et al.* Discovery of replicating circular RNAs by RNA-seq and computational algorithms. *PLoS Pathog* **10,** e1004553 (2014).

24. Ivanov, A. *et al.* Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep* **10,** 170–177 (2015).

25. Zhang, Y. *et al.* The Biogenesis of Nascent Circular RNAs. *Cell Rep* **15,** 611–624 (2016).

26. Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J. & Kleinschmidt, A. K. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proc Natl Acad Sci U S A* **73,** 3852–3856 (1976).

27. Nigro, J. M. *et al.* Scrambled exons. *Cell* **64,** 607–613 (1991).

28. Cocquerelle, C., Mascrez, B., Hetuin, D. & Bailleul, B. Mis-splicing yields circular RNA molecules. *FASEB J* **7,** 155–160 (1993).

29. Chao, C. W., Chan, D. C., Kuo, A. & Leder, P. The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis. *Mol Med* **4,** 614 (1998).

30. Pasman, Z., Been, M. & Garcia-Blanco, M. Exon circularization in mammalian nuclear extracts. *RNA* **2,** 603 (1996).

31. Capel, B. *et al.* Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* **73,** 1019–1030 (1993).

32. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495,** 384–388 (2013).

33. Caldas, C. *et al.* Exon scrambling of MLL transcripts occur commonly and mimic partial genomic duplication of the gene. *Gene* **208,** 167–176 (1998).

34. Bailleul, B. During in vivo maturation of eukaryotic nuclear mRNA, splicing yields excised exon circles. *Nat Struct Mol Biol* **24,** 1015–1019 (1996).

35. Zaphiropoulos, P. G. Exon skipping and circular RNA formation in transcripts of the human cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in testis. *Mol Cell Biol* **17,** 2985–2993 (1997).

36. Li, X.-F. & Lytton, J. A circularized sodium-calcium exchanger exon 2 transcript. *J Biol Chem* **274,** 8153–8160 (1999).

37. Surono, A. *et al.* Circular dystrophin RNAs consisting of exons that were skipped by alternative splicing. *Hum Mol Genet* **8,** 493–500 (1999).

38. Burd, C. E. *et al.* Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet* **6,** e1001233 (2010).

39. Hansen, T. B. *et al.* miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J* **30,** 4414–4422 (2011).

40. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20,** 1666–1670 (2014).

41. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. & Yang, J.-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nat Struct Mol Biol* **42,** D92–D97 (2013).

42. Ghosal, S., Das, S., Sen, R., Basak, P. & Chakrabarti, J. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front Genet* **4,** 283 (2013).

43. Liu, Y.-C. *et al.* CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res* **44,** D209–D215 (2015).

44. Jeck, W. R. & Sharpless, N. E. Detecting and characterizing circular RNAs. *Nat Biotechnol* **32,** 453 (2014).

45. Ashwal-Fluss, R. *et al.* circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell* **56,** 55–66 (2014).

46. Starke, S. *et al.* Exon circularization requires canonical splice signals. *Cell Rep* **10,** 103–111 (2015).

47. Wang, Y. & Wang, Z. Efficient backsplicing produces translatable circular mRNAs. *RNA* **21,** 172–179 (2015).

48. Kelly, S., Greenman, C., Cook, P. R. & Papantonis, A. Exon skipping is correlated with exon circularization. *J Mol Biol* **427,** 2414–2417 (2015).

49. Dolci, S., Grimaldi, P., Geremia, R., Pesce, M. & Rossi, P. Identification of a promoter region generating Sry circular transcripts both in germ cells from male adult mice and in male mouse embryonal gonads. *Biol Reprod* **57,** 1128–1135 (1997).

50. Salzman, J. Circular RNA Expression: Its Potential Regulation and Function. *Trends Genet* **32,** 309–316 (2016).

51. Conn, S. J. *et al.* The RNA binding protein quaking regulates formation of circRNAs. *Cell* **160,** 1125–1134 (2015).

52. Lee, Y. & Rio, D. C. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu Rev Biochem* **84,** 291–323 (2015).

53. Liu, S. & Cheng, C. Alternative RNA splicing and cancer. *Wiley Interdiscip Rev RNA* **4,** 547–566 (2013).

54. Bisognin, A. *et al.* An integrative framework identifies alternative splicing events in colorectal cancer development. *Mol Oncol* **8,** 129–141 (2014).

55. Yoshida, K. & Ogawa, S. Splicing factor mutations and cancer. *Wiley Interdiscip Rev RNA* **5,** 445–459 (2014).

56. Rossi, D. *et al.* Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. *Blood* **118,** 6904–6908 (2011).

57. Martínez-Avilés, L. *et al.* Mutations in the RNA splicing machinery genes in myelofibrotic transformation of essential thrombocythaemia and polycythaemia vera. *Br J Haematol* **164,** 605–607 (2014).

58. Tefferi, A. *et al.* CALR vs JAK2 vs MPL-mutated or triple-negative myelofibrosis: clinical, cytogenetic and molecular comparisons. *Leukemia* **28,** 1472–1477 (2014).

59. Hou, H.-A. *et al.* Splicing factor mutations predict poor prognosis in patients with de novo acute myeloid leukemia. *Oncotarget* **7,** 9084–9101 (2016).

60. Haferlach, T. *et al.* Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report From the International Microarray Innovations in Leukemia Study Group. *J Clin Oncol* **28,** 2529–2537 (2010).

61. Makino, D. L., Halbach, F. & Conti, E. The RNA exosome and proteasome: common principles of degradation control. *Nat Rev Mol Cell Biol* **14,** 654–660 (2013).

62. Chen, L.-L. & Yang, L. Regulation of circRNA biogenesis. *RNA Biol* **12,** 381–388 (2015).

63. Thomas, L. F. & Sætrom, P. Circular RNAs are depleted of polymorphisms at microRNA binding sites. *Bioinformatics* **30,** 2243–2246 (2014).

64. Bretscher, M. S. Translocation in Protein Synthesis: A Hybrid Structure Model. *Nature* **218,** 675–677 (1968).

65. Chen, C.-Y. & Sarnow, P. Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. *Science* **268,** 415 (1995).

66. Abe, N. *et al.* Rolling circle translation of circular RNA in living human cells. *Scientific reports* **5,** 16435 (2015).

67. AbouHaidar, M. G., Venkataraman, S., Golshani, A., Liu, B. & Ahmad, T. Novel coding, translation, and gene expression of a replicating covalently closed circular RNA of 220 nt. *Proc Natl Acad Sci U S A* **111,** 14542–14547 (2014).

68. You, X. *et al.* Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat Neurosci* **18,** 603–610 (2015).

69. Qu, S. *et al.* Circular RNA: a new star of noncoding RNAs. *Cancer Lett* **365,** 141–148 (2015).

70. Li, F. *et al.* Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/β-catenin pathway. *Oncotarget* **6,** 6001–6013 (2015).

71. Guil, S. & Esteller, M. RNA–RNA interactions in gene regulation: the coding and noncoding players. *Trends in Biochemical Sciences* **40,** 248–256 (2015).

72. Huang, G. *et al.* cir-ITCH plays an inhibitory role in colorectal cancer by regulating the Wnt/β-catenin pathway. *PLoS One* **10,** e0131225 (2015).

73. Xie, H. *et al.* Emerging roles of circRNA_001569 targeting miR-145 in the proliferation and invasion of colorectal cancer. *Oncotarget,* Epub ahead of print 05 April 2016 (2016).

74. Wang, K. *et al.* A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223. *Eur Heart J,* Epub ahead of print 21 January 2016 (2016).

75. Ala, U. *et al.* Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc Natl Acad Sci U S A* **110,** 7154–7159 (2013).

76. Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol* **15,** 409 (2014).

77. Thomson, D. W. & Dinger, M. E. Endogenous microRNA sponges: evidence and controversy. *Nat Rev Genet* **17,** 272–283 (2016).

78. Du, W. W. *et al.* Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. *Nat Struct Mol Biol* **44,** 2846–2858 (2016).

79. Zuna, J. *et al.* Covert preleukemia driven by MLL gene fusion. *Genes Chromosomes Cancer* **48,** 98–107 (2009).

80. Gao, Y., Wang, J. & Zhao, F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol* **16,** 4 (2015).

81. Broadbent, K. M. *et al.* Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC genomics* **16,** 454 (2015).

82. Alhasan, A. A. *et al.* Circular RNA enrichment in platelets is a signature of transcriptome degradation. *Blood* **127,** e1–e11 (2016).

83. Londin, E. R. *et al.* The human platelet: strong transcriptome correlations among individuals associate weakly with the platelet proteome. *Biol Direct* **9,** 3 (2014).

84. Guarnerio, J. *et al.* Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations. *Cell* **165,** 289–302 (2016).

85. Hansen, T. B., Venø, M. T., Damgaard, C. K. & Kjems, J. Comparison of circular RNA prediction tools. *Nucleic Acids Res* **44,** e58 (2015).

86. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nat Struct Mol Biol* **38,** e178 (2010).

87. Szabo, L. *et al.* Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol* **16,** 126 (2015).

88. Chuang, T.-J. *et al.* NCLscan: accurate identification of non-co-linear transcripts (fusion, trans -splicing and circular RNA) with a good balance between sensitivity and precision. *Nucleic Acids Res* **44,** e29 (2015).

89. Izuogu, O. G. *et al.* PTESFinder: a computational method to identify post-transcriptional exon shuffling (PTES) events. *BMC Bioinformatics* **17,** 1–11 (2016).

90. Tuck, A. C. & Tollervey, D. RNA in pieces. *Trends Genet* **27,** 422–432 (2011).

91. Schamberger, A., Sarkadi, B. & Orbán, T. I. Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner. *RNA Biol* **9,** 1177–1185 (2012).

92. Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. & Blelloch, R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev* **22,** 2773–2785 (2008).

93. Bortoluzzi, S., Biasiolo, M. & Bisognin, A. MicroRNA–offset RNAs (moRNAs): by-product spectators or functional players? *Trends Mol Med* **17,** 473–474 (2011).

94. Bortoluzzi, S. *et al.* Characterization and discovery of novel miRNAs and moRNAs in JAK2V617F-mutated SET2 cells. *Blood* **119,** e120–e130 (2012).

95. Guglielmelli, P. *et al.* Small RNA Sequencing Uncovers New miRNAs and moRNAs Differentially Expressed in Normal and Primary Myelofibrosis CD34+ Cells. *PloS one* **10,** e0140445 (2015).

96. Maute, R. L. *et al.* tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. *Proc Natl Acad Sci U S A* **110,** 1404–1409 (2013).

97. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144,** 296–309 (2011).

98. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351.** doi:10 . 1126 / science . aab2116 (2016).

99. Shi, L. *et al.* Developmental transcriptome analysis of human erythropoiesis. *Hum Mol Genet* **23,** 4528–4542 (2014).

100. Morlando, M., Ballarino, M. & Fatica, A. Long non-coding RNAs: new players in hematopoiesis and leukemia. *Front Med (Lausanne)* **2,** 23 (2015).

101. O'Connell, R. M. *et al.* MicroRNAs enriched in hematopoietic stem cells differentially regulate long-term hematopoietic output. *Proc Natl Acad Sci U S A* **107,** 14235–14240 (2010).

102. Zhang, L., Sankaran, V. G. & Lodish, H. F. MicroRNAs in erythroid and megakaryocytic differentiation and megakaryocyte-erythroid progenitor lineage commitment. *Leukemia* **26,** 2310–2316 (2012).

103. Zhang, X.-O. *et al.* Complementary Sequence-Mediated Exon Circularization. *Cell* **159,** 134–147 (2014).

104. Hoffmann, S. *et al.* A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol* **15,** R34 (2014).

105. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7,** 562–578 (2012).

106. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15.** doi:10.1186/s13059-014-0550-8 (2014).