From www.bloodjournal.org by guest on July 7, 2016. For personal use only.
Blood First Edition Paper, prepublished online July 5, 2016; DOI 10.1182/blood-2015-11-680843

Long non-coding RNAs in B cell development and activation

Tiago F. Brazão[1]*, Jethro S. Johnson[2]*, Jennifer Müller[1], Andreas Heger[2], Chris P. Ponting[2], Victor L. J. Tybulewicz[1,3]

[1]The Francis Crick Institute, London NW7 1AA, UK;
[2]MRC Computational Genomics Analysis and Training Centre, MRC Functional Genomics Unit, Department of Physiology Anatomy and Genetics, University of Oxford, Oxford, OX1 3QX, UK;
[3]Imperial College, London W12 0NN, UK

*Authors contributed equally to this work

Running title: LncRNAs in B cell development and activation

Word count main text: 4000
Word count abstract: 165
Figure count: 5
Reference count: 48
Scientific category: Immunobiology

Key point:
4516 long non-coding RNAs identified across multiple stages of B cell development and activation.

Correspondence to:
Victor L. J. Tybulewicz,
The Francis Crick Institute
The Ridgeway
London NW7 1AA
UK

Tel: +44 20 8816 2184
Email: Victor.T@crick.ac.uk

Copyright © 2016 American Society of Hematology

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

ABSTRACT

Long noncoding RNAs (lncRNAs) are potentially important regulators of cell differentiation and development, but little is known about their roles in B-lymphocytes. Using RNA-seq and *de novo* transcript assembly, we identified 4516 lncRNAs expressed in 11 different stages of B cell development and activation. Most of these lncRNAs have not been previously detected, even in the closely related T cell lineage. Comparison with lncRNAs previously described in human B cells identified 185 mouse lncRNAs that have human orthologues. Using ChIP-seq we classified 20% of the lncRNAs as either enhancer-associated (eRNA) or promoter-associated (pRNA) RNAs. We identified 126 eRNAs whose expression closely correlated with the nearest coding gene, thereby indicating the likely location of numerous enhancers active in the B cell lineage. Furthermore, using this catalogue of newly discovered lncRNAs we show that PAX5, a transcription factor required to specify the B cell lineage, bound to and regulated the expression of 109 lncRNAs in pro-B and mature B cells and 184 lncRNAs in acute lymphoblastic leukemia.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

INTRODUCTION

Long noncoding RNAs (lncRNAs) have emerging roles in innate and adaptive immunity. For example, *lnc-DC* is required for normal dendritic cell differentiation and function[1], *IL1β-eRNA* and *IL1β-RBT46* are required for LPS-induced pro-inflammatory responses in monocytes[2], and *NRAV* modulates cellular responses to viral infections[3]. In T cells, an intronic lncRNA *NRON* abrogates the nuclear transport of NFAT, and hence modulates expression of IL-2[4]. In B cell lymphomas, the lncRNA *Fas-AS1* modulates expression of soluble Fas receptor mRNA, an important regulator of apoptosis[5]. Thus lncRNAs have the potential to influence both normal and pathological immune cell development and function.

LncRNAs may operate via a variety of molecular mechanisms[6]. For example, enhancer-associated lncRNAs (eRNAs) act in *cis* and originate from transcribed extragenic or intragenic enhancer regions, whereas promoter-associated lncRNAs (pRNAs) can act in *trans* and originate from canonical promoter-derived transcriptional activity[7,8]. These two broad lncRNA categories are distinguished by the ratio of mono- versus tri-methylation of histone 3 lysine 4 (H3K4me1/3)[8]. Compared with pRNAs, eRNAs tend to exhibit more restricted expression and their RNA sequences show less constraint[8].

Advances in sequencing technology have enabled the identification of large numbers of putative lncRNA loci[9,10]. However, the proportion of lncRNAs with clearly defined function is small[11,12], caused in part by poor annotation of lncRNAs expressed in a tissue of interest, making it difficult to select candidate lncRNAs for targeted studies. This is a consequence of the expression patterns of lncRNAs which are often restricted to one, or very few tissues or cell types[9]. Recent studies have addressed this limitation by surveying lncRNA expression in a number of organisms and tissues[13-17], including murine T cells[18]. However, there have been no comparable

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

attempts to use sequencing technologies to describe the murine B cell lncRNA repertoire.

To facilitate the study of lncRNA biology in B cells, we describe here a catalogue of 4516 *de novo* assembled high-confidence lncRNAs expressed in 11 mouse B cell populations. We identify human lncRNAs that may be orthologues of the mouse genes. Furthermore, we classify subsets of eRNAs and pRNAs, and perform an unsupervised clustering analysis to associate lncRNAs with mRNAs at key stages of B cell development. Finally, we utilize the lncRNA catalogue to show that PAX5, a transcription factor required to specify the B cell lineage[19], binds to and regulates expression of lncRNA loci in both pro-B and mature B cells, as well as in acute lymphoblastic leukemia.

MATERIALS AND METHODS

Mice

All RNA-seq and ChIP-seq experiments were performed with female C57BL/6JNimr mice aged 7-9 weeks, except for RNAseq of plasmablasts and plasma cells which were obtained from 12-14 week old Blimp1-GFP mice[20].

Cell Sorting

Gating strategies for cell sorting are shown in Supplemental Figure 1.

RNA-seq

Sorted populations of cells were re-suspended in Trizol (Life Technologies), and RNA was purified using the RNeasy Mini Kit (Qiagen). RNA quality was assessed using the 2100 Expert Agilent Bioanalyser. For all samples except plasmablasts and plasma cells, stranded polyA-enriched libraries were made using the Stranded TruSeq RNA Sample Preparation Kit (Illumina) and sequenced on the HiSeq 2500

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

(Illumina), collecting 100 base paired-end reads. For plasma cells and plasmablasts, unstranded non-rRNA-enriched libraries were made using the SMARTer Ultra Low Input RNA Kit for Sequencing v3 (Clontech) and sequenced collecting 50 base paired-end reads. All RNA-seq and ChIP-seq data (see below) has been deposited at NCBI Gene Expression Omnibus (accession number GSE72019).

ChIP-seq

Chromatin immunoprecipitation-sequencing was performed in triplicate for all stages of B cell development, except plasmablasts and plasma cells, as described previously[21]. For details see Supplemental Methods.

RNA-seq read alignment and transcript assembly

RNA-seq reads were aligned to the C57BL/6J mouse reference genome (mm10, GRCm38) using STAR[22] v2.3.0e. Transcript assembly was performed separately for all samples except plasmablasts and plasma cells using Cufflinks[23] v2.2.0. Non-uniquely mapping reads were retained during assembly and potential transcripts (transfrags) were discarded when they contained fewer than five successfully mapped reads. Individual assemblies were subsequently compared and transfrags were discarded if they were not assembled in at least two samples.

Identification of long non-coding RNAs

The lncRNA discovery pipeline is depicted in Supplemental Figure 2A. Following assembly, transcripts <200bp in length were discarded. Remaining transcripts were filtered against existing databases (Ensembl v72 and NCBI RefSeq), and discarded if they intersected (>=1bp, on the same strand) intervals annotated as anything other than non-coding RNA. Transcripts were also discarded if they were found to intersect nuclear mitochondrial DNA (Numts) or pseudogenes predicted using Exonerate[24], if

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

they were in close proximity downstream of a protein coding gene, or if they were classified as coding by both Coding Potential Calculator[25] and PhyloCSF[26]. Further details, including a description of lncRNA nomenclature, are provided in Supplemental Methods. LncRNAs classed as intergenic (>5kb from a protein-coding gene) were identified as eRNAs or pRNAs using H3K4me1 and H3K4me3 ChIP-seq data in conjunction with a stringent classification pipeline (Supplemental Figure 2B), in which successfully classified loci were required to intersect a called chromatin peak and to show four-fold greater coverage of reads arising from their characteristic chromatin mark.

UCSC public hub

Genomic data has been visualized as a UCSC public hub, accessible at https://www.cgat.org/downloads/public/projects/proj010/UCSC_track_hub/hub.txt.

RESULTS

4516 lncRNA loci expressed during B cell development

In order to identify and characterize lncRNAs expressed during B cell development and activation, we used fluorescence-activated cell sorting to obtain 5 replicate cell samples for each of 8 purified B cell populations derived from the bone marrow, spleen, and peritoneal cavity of adult female C57BL/6JNimr mice (Figure 1A, Supplemental Figure 1A-D). These populations represent the major subsets of developing B cells in the bone marrow (pro-B cells, pre-B cells and immature B cells) and subsets of mature B cells in the bone marrow, spleen (follicular and marginal zone B cells), and peritoneal cavity (B1a cells). In addition we also sorted germinal center (GC) B cells, an activated B cell subset. From each of these 40 samples we extracted and sequenced the polyA+ RNA fraction in a strand-specific manner to an average depth of 42 million uniquely mapped reads, using a 2x100 base paired-end sequencing protocol.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

To identify lncRNAs we used a *de novo* transcript assembly pipeline (Supplemental Figure 2A). Stringency of selection of potential lncRNA loci was assured by discarding transcripts whose assembly was not replicated in at least 2 of the 40 sample libraries, by discarding loci represented in public databases of protein-coding annotations (ENSEMBL & RefSeq), and by discarding loci containing transcripts with predicted coding potential (Supplemental Figure 3A-C). Using this process we identified 4516 lncRNA loci that were expressed in at least one of the 8 B cell populations, and comprised, on average, 1.8 transcripts per locus (Supplemental Table 1). To validate the assembly process, we used PCR to check for expression of predicted lncRNAs. Of 53 lncRNAs with expression values ranging between 0.9 and 667 FPKM, we were able to successfully validate 47 (89%), demonstrating that the assembly pipeline had generated reliable transcripts (Supplemental Figure 4).

Of the predicted lncRNA loci, 3025 (67%) showed no evidence of splicing (single-exon loci), whereas the remaining 1491 showed evidence of at least one spliced transcript (multi-exon loci). Approximately half of all single and multi-exon loci were intergenic, with the remainder being either flanking (<5kb from a protein-coding gene) or overlapping a protein-coding gene on the antisense strand (Figure 1B, Supplemental Table 2). The single-exon and multi-exon lncRNAs showed similar coding potential, expression levels and mean exon sizes, supporting the view that they belong to the same gene class (Supplemental Figure 5A-D).

Intergenic B cell lncRNAs show little overlap with existing lncRNA catalogues

We compared the 2349 intergenic lncRNA loci identified in this study with intergenic lncRNAs reported in mouse T cell subsets[18], and with the latest Ensembl (v74) long intergenic non-coding RNA (lincRNA) annotations. 1829 (78%) of the intergenic lncRNAs in our B cell catalogue did not overlap on the relevant strand with either the

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

T cell or Ensembl lincRNAs (Figure 1C), reinforcing the notion that many lncRNAs are tissue- or cell type-specific in their expression[9,12,18].

The highly tissue-specific expression of B cell lncRNAs was further confirmed by comparison of our catalogue with the FANTOM5 study of transcriptional start sites (TSS) in 128 mouse primary cell types[27], which included a single B cell sample. As FANTOM annotation required a TSS to be detected in two or more samples, we did not expect this dataset to include TSS for lncRNAs expressed only in B cells. Accordingly, only 299 (13%) of our intergenic lncRNAs were within 500bp of a FANTOM annotated TSS (Figure 1D,E). Notably, a greater overlap with FANTOM annotated TSS was observed for multi-exon transcripts than for single-exon transcripts. The same trend was observed in comparisons with previous lncRNA annotations (Supplemental Figure 5E,F). We therefore conclude that expression at a lncRNA locus is more likely to be recapitulated in different cell types if the locus contains evidence of splicing.

A previous study had shown that many lncRNAs overlap with transposable elements (TEs)[28]. In agreement with this, we find that around 40% of the lncRNAs identified here overlap with TEs, and a comparison of these with the genomic content of TEs shows an enrichment of endogenous retroviruses and a relative depletion of LINE elements, similar to the previous report (Figure 1F, Supplemental Table 3).

In summary, we have generated a novel, high-confidence catalogue of lncRNAs expressed across key stages of B cell development. To facilitate the use of this catalogue, we have visualized our genomic data as a UCSC public hub, which can be accessed at https://www.cgat.org/downloads/public/projects/proj010/UCSC_track_hub/hub.txt.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

Chromatin signatures identify intergenic lncRNAs with enhancer-associated and promoter-associated expression

In order to identify intergenic lncRNAs within our catalogue that may be either eRNAs or pRNAs, we determined the genome-wide abundance of H3K4me1 and H3K4me3 chromatin marks known to distinguish these two lncRNA subtypes[8] (Figure 2A,B). We quantified the relative abundance of these marks at the proposed TSS of 2349 intergenic lncRNAs across all 8 B cell populations, and identified eRNAs and pRNAs using stringent selection criteria (Supplemental Figure 2B).

702 eRNAs and 192 pRNAs were predicted in one or more B cell populations (Figure 2C, Supplemental Figure 6A-C, Supplemental Table 2), and were evenly distributed between single-exon (320 eRNAs, 96 pRNAs) and multi-exon loci (382 eRNAs, 96 pRNAs, Figure 2D). H3K4me1 and H3K4me3 marks at all lncRNA loci remained broadly consistent across B cell populations (Figure 2E). Furthermore, lncRNA classified as eRNAs or pRNAs in one B cell population displayed a comparable chromatin signature in other cell populations (Supplemental Figure 6D,E), suggesting that these classifications represent two largely mutually-exclusive lncRNA classes.

In agreement with earlier studies[8], B cell eRNAs were distinguishable from pRNAs by a significantly higher correlation of expression with their proximal protein-coding gene (Figure 2F), by lower expression (Fig 2G), and by expression in fewer cell subsets (Figure 2H). Thus lncRNAs with a high H3K4me1:H3K4me3 ratio at their TSS show characteristics typical of eRNAs and their expression tends to be associated with the expression of adjacent protein-coding genes.

LncRNAs are expressed at discrete stages during B cell development

Previous studies in the T cell lineage had shown that intergenic lncRNA expression is highly restricted even between closely related T cell subsets[17,18]. If this were also the case for B cells, then B cell subsets should be distinguishable on the basis of lncRNA

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

expression. Indeed, using both hierarchical clustering and principal component analysis on either coding gene or lncRNA expression, all 8 B cell populations were distinguishable and replicate samples clustered closely (Figure 3A,B, Supplemental Figure 7A-D). For both types of genes, follicular and marginal zone B cells were the most closely related, whereas B1a cells and GC B cells showed the greatest distinction from other cell types.

A greater proportion of B cell lncRNAs were differentially expressed between cell populations than protein-coding genes (Supplemental Figure 8A). Whilst cell type specificity was, as expected, dependent on expression level, even at equivalent expression levels, lncRNAs displayed greater cell-type restriction than coding genes (Figure 3C, Supplemental Figure 8B,C), suggesting that lncRNAs have very restricted spatio-temporal roles during B cell development.

The trend for lncRNAs to be more restricted in expression than protein-coding genes was observed across all stages of B cell development. At a defined expression threshold (FPKM>1), a greater proportion of protein-coding genes than lncRNAs were expressed in all bone marrow-derived B cell populations, in all splenic and peritoneal cavity mature B cell populations, and in both naïve and activated B cell populations (Figure 3D,E).

To extend this analysis further we examined expression of the lncRNAs in antibody-secreting cells. We performed RNA-seq on splenic plasmablasts and plasma cells and on bone marrow plasma cells (Supplemental Figure 1E,F Supplemental Figure 7E) and determined expression of lncRNAs and coding genes in these three additional subsets (Supplemental Table 2). Principal component analysis showed that the 3 subsets clustered closely together (Supplemental Figure 7F,G) and lncRNAs were again more cell type-specific (Figure 3D,E).

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

Generating hypotheses for lncRNA functions through correlation with protein-coding gene expression

One genome-scale approach to generate hypotheses regarding lncRNA function is to exploit a 'guilt by association' approach by correlating expression of lncRNAs with coding genes of known biological function[9,29]. We predicted the biological processes to which lncRNAs might contribute by clustering protein-coding loci expressed in at least one B cell type (median FPKM>1) into one of 17 clusters, based on correlation of expression across B cell development[30] (Supplemental Figure 9A,B, Supplemental Table 2). We subsequently correlated individual lncRNAs with the representative gene expression profile (eigengene) for each protein-coding cluster (Figure 4A), and defined an association between a lncRNA and a cluster of protein-coding genes at a Pearson correlation coefficient of $|\rho| > 0.8$ (Figure 4B). At this threshold, 987 (22%) lncRNAs associated with protein-coding gene clusters, suggesting that these lncRNAs contribute to the same biological processes as the protein-coding genes within the respective cluster. While some clusters were characteristic of individual cell stages (Figure 4A), Gene Ontology enrichment analysis[31] identified others that were associated with cellular processes (Supplemental Table 4). For example the lightpink4 cluster contains genes up- or down-regulated in MZ B cells, whereas the floralwhite cluster is enriched for genes associated with the cell cycle, and its eigengene is highest in pro-B cells, pre-B cells and GC B cells, the 3 subsets with the highest rates of cell division.

We previously showed an increased correlation of expression between eRNAs and their proximal protein-coding gene compared to pRNAs (Figure 2F), and the association of lncRNAs with clusters of protein-coding genes allowed us to test this hypothesis further. Permutation tests demonstrated that eRNAs showed strongest

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

correlation with the cluster containing their proximal protein-coding gene (p = 0.04), whilst no such trend was observed for pRNAs (p = 0.51).

Combining information about genomic location, chromatin state, and expression allowed us to identify eRNAs that may function *in cis* with genes of distinctive biological function. For example, Figure 4C shows a single cluster of protein-coding genes that exhibit large changes of expression within marginal zone B cells. Of the lncRNAs correlated with this cluster, we show one (*LNCGme02323*) that is proximal to cluster member *Zc3h12c*, encoding a zinc finger protein implicated in the regulation of pro-inflammatory activation in macrophages[32]. Using our catalogue, *LNCGme02323* is identifiable as an eRNA that is co-expressed with *Zc3h12c* in marginal zone B cells, and which lies ~67 kb upstream of the coding gene TSS (Figure 4D,E), suggesting that this lncRNA may be transcribed from an enhancer controlling the expression of *Zc3h12c*. In total we were able to identify 126 eRNA loci whose expression was most strongly correlated with that of a WGCNA module containing an adjacent protein-coding gene, including 48 with a correlation of $|\rho| > 0.8$ (Supplemental Table 5).

Human lncRNA orthologues

To extend the utility of this catalogue of mouse lncRNAs, we sought to identify potential orthologous human lncRNAs by comparing the mouse genes to lncRNAs previously identified in human B cells[17,33-35]. Based on pairwise genomic alignments we identified 185 mouse lncRNAs that have a human syntenic orthologue (Supplemental Table 6). Taking advantage of the fact that eRNAs are a distinct class of lncRNA throughout B cell development (Supplemental Figure 6D) we compared mouse eRNAs with human eRNAs reported in CD19+ B cells[34]. Pairing eRNAs adjacent to orthologous human and mouse coding genes identified a further 228

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

mouse eRNAs that have one or more human eRNA counterparts (Supplemental Table 6). Such conserved location indicates that the human/mouse pairs may be marking enhancers of orthologous coding genes.

PAX5 regulates lncRNA expression during normal and malignant B cell development

The transcription factor PAX5 is expressed throughout B cell development and is crucial for commitment to the B cell lineage as well as for maintaining B cell homeostasis[19,36]. PAX5 mediates these functions through the regulation of protein-coding gene expression, but it may also regulate lncRNA expression. Using previously published PAX5 binding sites[36], we identified 784 and 717 lncRNAs in pro-B cells and mature B cells, respectively, whose loci either overlapped, or were <1 kb downstream of a PAX5 binding site that had no comparable association with a protein-coding gene. This represented a statistically significant enrichment compared to a null model of PAX5 binding (pro-B cells: 6.1-fold, $p < 10^{-4}$, mature B cells: 10.5-fold, $p < 10^{-4}$). Using published RNA-seq data for wild-type or *Pax5*-deficient pro-B cells or mature B cells[36], we further identified 28 and 87 of these lncRNA loci, respectively, for which there is evidence that PAX5 both binds and regulates their expression (Supplemental Figure 10A-D, Supplemental Table 7).

Disruption of PAX5 binding during early B cell development is associated with development of B-progenitor acute lymphoblastic leukemia (B-ALL)[37]. For this reason we extended our analysis of PAX5 binding and lncRNA expression using RNA-seq data from a mouse model of B-ALL in which *Pax5* expression can be regulated by doxycycline[38]. Induction of *Pax5* expression within leukemic cells causes the tumor cells to differentiate and the leukemia to regress. Comparison of B-ALL cells with and without doxycycline-induced *Pax5* expression revealed 331 differentially expressed lncRNAs ($q < 0.05$), of which a large majority (91%) were up-regulated as a result of

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

Pax5 expression (Supplemental Table 7). Furthermore, most (184) of these lncRNAs originated from loci bound by PAX5 in pro-B cells, or mature B cells (Figure 5A,B). As an example, we show two previously annotated lncRNA loci (*LNCGme00432* and *LNCGme00344*) together with a novel locus (*LNCGme00345*), all of which are eRNAs, and are bound by PAX5, differentially expressed in the absence of PAX5, and lie downstream of the B cell lymphoma 11a gene (*Bcl11a*) whose expression is also PAX5-dependent in B-ALL cells (Figure 5C). We note that *Bcl11a* does not have PAX5 binding sites near its TSS, and yet its expression is PAX5-dependent. Taking these observations together, we hypothesize that *LNCGme00432*, *LNCGme00344* and *LNCGme00345* may be transcribed from PAX5-induced enhancers that in turn regulate expression of *Bcl11a* and thus explain how expression of *Bcl11a* is indirectly PAX5-dependent. Since BCL11A has been proposed to activate expression of the *Rag1* and *Rag2* genes and hence to promote VDJ recombination[39], it may contribute to PAX5-dependent differentiation of B-ALL cells and subsequent regression of the leukemia. Furthermore, BCL11A is essential for B cell development[40,41], and thus these lncRNAs may also mediate the effect of PAX5 on expression of *Bcl11a* during normal B cell development. Interestingly, *LNCGme00344* and *LNCGme00345* have human eRNA orthologues (*BMThy_chr2_0943* and *BMThy_chr2_0945*, Supplemental Table 6), suggesting that this transcriptional regulation may be conserved in human B cells. In total we report 199 lncRNAs (including 73 eRNAs) that are bound by PAX5, show PAX5-dependent expression, and are neighboring a protein-coding gene that also shows PAX5-dependent expression (Supplemental Table 7).

DISCUSSION

To aid in the identification of lncRNAs active within the immune system, we present a comprehensive catalogue of 4516 lncRNAs, including 2349 intergenic loci, expressed across 11 murine B cell populations. Translating genome-scale lncRNA identification

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

into *in vivo* functional studies remains a major challenge that necessitates the use of model organisms. Previous work has identified lncRNAs expressed in a subset of these populations in humans[17,33-35]. However, given the high tissue-specificity and rapid evolution of lncRNAs, a murine catalogue is therefore a prerequisite to successful lncRNA studies in the mouse. In addition, we defined patterns of expression, chromatin modification, and transcription factor binding that provide insight into the regulation and function of these transcripts, and we identified a subset of lncRNAs that have human orthologues and thus are of direct relevance to human research.

The majority (67%) of lncRNA loci present in our catalogue contained no evidence of splicing. While single-exon transcripts had to be replicated in two or more samples to be included in this catalogue, their loci were less likely to be in published datasets. However, the high rate of PCR validation of these single-exon lncRNAs (85%) and their comparable intersection with clearly defined chromatin peaks confirms the reliability of the transcript annotation, and the demonstrated functions of single-exon lncRNAs such as NEAT1[42] and Paupar[43] highlight the importance of including these transcripts in the catalogue.

LncRNA genomic locations have the potential to provide insight into function. A substantial proportion of multi-exon (48%) and single-exon (48%) loci were antisense to or flanking (<5kb) protein-coding genes. Previous studies have classified as many as 60% of detected lncRNAs as antisense transcripts arising from bidirectional transcription at protein-coding promoters[44], that could regulate expression of genes encoding transcriptional regulators[45]. These are included within the lncRNAs classified as flanking in this study. Full-length antisense lncRNAs have also been implicated in a variety of cellular functions[46,47], most notably FAS-AS1, which regulates alternative splicing in its antisense coding gene *Fas*[5,48].

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

Half of all lncRNAs detected in B cells (52%) were intergenic (>5kb from a protein-coding gene). The use of H3K4me1 and H3K4me3 chromatin marks to distinguish eRNAs from pRNAs has been previously demonstrated[8], and we were able to corroborate many of the features of these transcripts. The consistent classification of these loci across B cell development strongly suggests that these markers identify two distinct and mutually exclusive classes of lncRNA.

Comparison with lncRNAs expressed in human B cells identified 185 mouse lncRNAs that had a human orthologue, and a further 228 mouse eRNAs that have human eRNA counterparts. This offers the possibility of using the power of mouse genetics to study the *in vivo* function of these human lncRNAs by genetic modification of the corresponding mouse genes.

The use of expression correlation as a means of associating lncRNA function with that of protein-coding genes has been successfully demonstrated in previous studies[17,18]. By clustering protein-coding genes on the basis of co-expression, we were able to identify discrete groups whose patterns of expression transcend individual B cell populations and whose members are associated with, for example, cell cycle function. Post hoc association of lncRNAs with these clusters implicated a substantial proportion (22%) of the lncRNAs with these specific processes. In addition this analysis supported the hypothesis that eRNAs may function to regulate expression of adjacent protein-coding genes[7].

Using external data we identified loci regulated by PAX5, a crucial mediator of B cell development, whose loss contributes critically to the development of B-ALL[38]. After incorporating our classification of lncRNAs based on chromatin state, we identified 73 eRNAs situated proximal to protein-coding genes that also show *Pax5*-dependent

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

expression in B-ALL cells. The roles of such transcripts remain to be investigated; however, their discovery demonstrates the utility of our catalogue when used in conjunction with external datasets to probe the complexities of B cell development and disease.

In conclusion, RNA-seq followed by *de novo* transcript assembly, has the potential to identify thousands of novel long non-coding transcripts. While a growing number of these transcripts have been assigned to diverse cellular processes, as yet, the majority have no known function. Given the large numbers of these transcripts, a major challenge when seeking to identify and characterize lncRNA function lies in the prioritization of plausible candidates for *in vivo* functional studies. We anticipate that this catalogue will serve as a valuable resource for such future research, for example by allowing systematic genetic screens of high-confidence lncRNAs to identify those with key functions in B cell development and activation.

ACKNOWLEDGEMENTS

We thank Stephen Nutt for Blimp1-GFP mice, and the Flow Cytometry and Advanced Sequencing Facilities of the MRC National Institute for Medical Research (now the Francis Crick Institute) for support. The MRC Computational Genomics Analysis and Training Centre was supported by the Medical Research Council as was VLJT (MRC programme number U117527252).

AUTHORSHIP CONTRIBUTIONS

TB, JJ and JM performed research; TB, JJ, JM and AH analyzed data; TB, JJ, AH, CP and VT designed research and wrote the paper.

CONFLICT OF INTEREST DISCLOSURE

The authors have no conflicts of interest

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

REFERENCES

1. Wang P, Xue Y, Han Y*, et al.* The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science.* 2014;344(6181):310-313.
2. Ilott NE, Heward JA, Roux B*, et al.* Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nat Commun.* 2014;5:3979.
3. Ouyang J, Zhu X, Chen Y*, et al.* NRAV, a long noncoding RNA, modulates antiviral responses through suppression of interferon-stimulated gene transcription. *Cell Host Microbe.* 2014;16(5):616-626.
4. Willingham AT, Orth AP, Batalov S*, et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science.* 2005;309(5740):1570-1573.
5. Sehgal L, Mathur R, Braun FK*, et al.* FAS-antisense 1 lncRNA and production of soluble versus membrane Fas in B-cell lymphoma. *Leukemia.* 2014;28(12):2376-2387.
6. Gardini A Shiekhattar R. The many faces of long noncoding RNAs. *FEBS J.* 2015;282(9):1647-1657.
7. De Santa F, Barozzi I, Mietton F*, et al.* A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol.* 2010;8(5):e1000384.
8. Marques A, Hughes J, Graham B, Kowalczyk M, Higgs D, Ponting C. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol.* 2013;14(11):R131.
9. Derrien T, Johnson R, Bussotti G*, et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* 2012;22(9):1775-1789.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

10. Yue F, Cheng Y, Breschi A*, et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature.* 2014;515(7527):355-364.

11. Bassett AR, Akhtar A, Barlow DP*, et al.* Considerations when investigating lncRNA function in vivo. *eLife.* 2014;3.

12. Fatica A Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet.* 2014;15(1):7-21.

13. Alvarez-Dominguez JR, Hu WQ, Yuan BB*, et al.* Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood.* 2014;123(4):570-581.

14. Guttman M, Garber M, Levin JZ*, et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotech.* 2010;28(5):503-510.

15. Peng L, Paulson A, Li H*, et al.* Developmental Programming of Long Non-Coding RNAs during Postnatal Liver Maturation in Mice. *PLoS One.* 2014;9(12):e114917.

16. Ramos AD, Diaz A, Nellore A*, et al.* Integration of Genome-wide Approaches Identifies lncRNAs of Adult Neural Stem Cells and Their Progeny In Vivo. *Cell Stem Cell.* 2013;12(5):616-628.

17. Ranzani V, Rossetti G, Panzeri I*, et al.* The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nat Immunol.* 2015;16:318-325.

18. Hu G, Tang Q, Sharma S*, et al.* Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat Immunol.* 2013;14(11):1190-1198.

19. Nutt SL Kee BL. The transcriptional regulation of B cell lineage commitment. *Immunity.* 2007;26(6):715-725.

20. Kallies A, Hasbold J, Tarlinton DM*, et al.* Plasma cell ontogeny defined by quantitative changes in blimp-1 expression. *J Exp Med.* 2004;200(8):967-977.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

21. Kagey MH, Newman JJ, Bilodeau S*, et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature.* 2010;467(7314):430-435.

22. Dobin A, Davis CA, Schlesinger F*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2012;29:15-21.

23. Trapnell C, Williams BA, Pertea G*, et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech.* 2010;28(5):511-515.

24. Slater G Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* 2005;6(1):31.

25. Kong L, Zhang Y, Ye Z-Q*, et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 2007;35(suppl 2):W345-W349.

26. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics.* 2011;27(13):i275-i282.

27. Fantom Consortium. A promoter-level mammalian expression atlas. *Nature.* 2014;507(7493):462-470.

28. Kelley D Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* 2012;13(11):R107.

29. Rinn JL Chang HY. Genome Regulation by Long Noncoding RNAs. *Annu Rev Biochem.* 2012;81(1):145-166.

30. Langfelder P Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9(1):559.

31. Young M, Wakefield M, Smyth G, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 2010;11(2):R14.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

32. Liang J, Wang J, Azfer A*, et al.* A novel CCCH-zinc finger protein family regulates proinflammatory activation of macrophages. *J Biol Chem.* 2008;283(10):6337-6346.

33. Bonnal RJ, Ranzani V, Arrigoni A*, et al.* De novo transcriptome profiling of highly purified human lymphocytes primary cells. *Sci Data.* 2015;2:150051.

34. Casero D, Sandoval S, Seet CS*, et al.* Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. *Nat Immunol.* 2015;16(12):1282-1291.

35. Petri A, Dybkaer K, Bogsted M*, et al.* Long Noncoding RNA Expression during Human B-Cell Development. *PLoS One.* 2015;10(9):e0138236.

36. Revilla-i-Domingo R, Bilic I, Vilagos B*, et al.* The B-cell identity factor Pax5 regulates distinct transcriptional programmes in early and late B lymphopoiesis. *EMBO J.* 2012;31(14):3130-3146.

37. Mullighan CG, Goorha S, Radtke I*, et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature.* 2007;446(7137):758-764.

38. Liu GJ, Cimmino L, Jude JG*, et al.* Pax5 loss imposes a reversible differentiation block in B-progenitor acute lymphoblastic leukemia. *Gene Dev.* 2014;28(12):1337-1350.

39. Heger A, Webber C, Goodson M, Ponting CP, Lunter G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics.* 2013;29(16):2046-2048.

40. Liu P, Keller JR, Ortiz M*, et al.* Bcl11a is essential for normal lymphoid development. *Nat Immunol.* 2003;4(6):525-532.

41. Yanai I, Benjamin H, Shmoish M*, et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics.* 2005;21(5):650-659.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

42. Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. MEN ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* 2009;19(3):347-359.

43. Vance KW, Sansom SN, Lee S*, et al.* The long non-coding RNA Paupar regulates the expression of both local and distal genes. *EMBO J.* 2014;33(4):296-311.

44. Sigova AA, Mullen AC, Molinie B*, et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *P Natl Acad Sci USA.* 2013;110(8):2876-2881.

45. Lepoivre C, Belhocine M, Bergon A*, et al.* Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics.* 2013;14(1):914.

46. Feng J, Bi C, Clark BS, Mady R, Shah P, Kohtz JD. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Gene Dev.* 2006;20(11):1470-1484.

47. Pandey RR, Mondal T, Mohammad F*, et al.* Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation. *Mol Cell.* 2008;32(2):232-246.

48. Yan M-D, Hong C-C, Lai G-M, Cheng A-L, Lin Y-W, Chuang S-E. Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas. *Hum Mol Genet.* 2005;14(11):1465-1474.

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

FIGURE LEGENDS

Figure 1. Identification of lncRNAs expressed in B cells. (A) Schematic representation of the ontogenetic relationships between B cell populations used to generate the lncRNA catalogue. Solid arrows indicate developmental progression through B cell stages or activation (to germinal center B cells). Dashed line indicates recirculation of follicular B cells back to the bone marrow. Pro-B cells (PRO), pre-B cells (PRE), immature B cells (IMM), mature B cells (MAT), follicular B cells (FO), marginal zone B cells (MZ), germinal center B cells (GC), B1a B cells (B1A). (B) Genomic distribution of the 1491 multi-exon and 3025 single-exon lncRNAs identified by this study. Positions are described relative to ENSEMBL v72 protein-coding gene annotations as antisense (overlapping a coding gene on antisense strand), flanking (< 5 kb from coding gene) and intergenic (> 5 kb from coding gene). (C) Overlap between the 2349 intergenic lncRNAs identified by this study (B Cell), with those identified in T lineage cells[18] (T Cell), and those annotated in Ensembl v78. (D, E) Kernel density plots representing the distribution of distance between each multi-exon (D) and single-exon (E) intergenic lncRNA TSS and the nearest annotated TSS on the same strand that appeared in two or more of the 128 mouse cell lines considered by the FANTOM5 consortium. Shaded regions indicate a null distribution as measured by distance to the nearest FANTOM5 annotated TSS on the opposing strand. Grey vertical dashed line indicates a distance of 500bp. (F) Coverage of the genome and of lncRNA exons by the indicated transposon elements.

Figure 2. Identification of intergenic lncRNAs with enhancer-like and promoter-like characteristics. (A, B) Examples of intergenic lncRNA loci with chromatin signatures in marginal zone B cells that are characteristic of (A) enhancer-like RNA (eRNA) loci (LNCGme01103), and (B) promoter-like (pRNA) loci (LNCGme01293). The former are distinguished by high H3K4me1 read coverage across the TSS and

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

the absence of a corresponding H3K4me3 peak. The latter are distinguished by high H3K4me3 coverage, the presence of which excludes H3K4me1 from the TSS. Also shown in (B) is a second lncRNA (LNCGme01244) that arises as a result of bi-directional transcription from the *Scyl1* promoter, but this is not classified as an eRNA or pRNA due to its proximity to a coding gene. (C) The proportion of the 2349 intergenic lncRNAs identified in this study that could be classified as either eRNAs or pRNAs on the basis of their chromatin state. Remaining lncRNAs are either classified as unassigned (insufficient read coverage/fold-change to determine chromatin state) or conflicted (classified as eRNA in one B cell stage and pRNA in another). (D) The proportion of eRNA and pRNA loci that are classified as multi-exon or single-exon. (E) Pairwise comparisons showing the consistency of chromatin signatures across B cell populations. Within each B cell population intergenic lncRNAs are ranked on the ratio of H3K4me1:H3K4me3 coverage across their TSS. Individual plots show the local regression (loess) of rank order between two B cell populations. (F) Distribution of the Pearson's correlation coefficient between the expression of an eRNA or pRNA and expression of the more highly correlated of either its nearest upstream or downstream protein-coding gene. (G) Distribution of median expression values (rlog-transformed read counts) calculated across all cell stages. (H) Distribution of cell-stage specificity of expression of eRNAs and pRNAs. (***Mann-Whitney U test: $p < 0.0001$).

Figure 3. Cell-stage specific expression at lncRNA and protein-coding loci. Principal component analysis of regularized log-transformed expression patterns at (A) protein-coding loci and (B) the 4516 lncRNA loci identified in this study. Grey dashed lines indicate groups identified by unsupervised hierarchical clustering (Supplemental Figure 7A, B). (C) Box plots of specificity of expression of all coding and lncRNA loci separated into quartile bins on the basis of their median expression across all eight B cell stages. Numbers below each quartile indicate the number of

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

lncRNA and protein-coding loci that fall into each category. (D, E) Venn diagrams showing the number of protein-coding (D) and lncRNA (E) loci that are either expressed in multiple cell populations (blue) or expressed in a single cell population (red) at an FPKM threshold of 1.0. Splenic plasmablasts (PB) and plasma cells (PC(SP)), and bone marrow plasma cells (PC(BM)). Numbers adjacent to each plot depict the proportion of loci falling into each category.

Figure 4. Association between lncRNAs and protein-coding genes based on correlation of expression across B cell development. (A) Weighted gene co-expression network analysis (WGCNA) identifies clusters of protein-coding genes with comparable expression profiles across B cell development: plots show a representative expression profile (eigengene) for each of 10 protein-coding gene clusters. (B) Stacked bar charts showing the number of lncRNAs whose expression is strongly correlated ($|\rho| > 0.8$) with the adjacent eigengene. Colors depict lncRNA classification on the basis of chromatin state. (Results are shown for WGCNA clusters with >10 associated lncRNAs.) (C) Heat map showing the normalized expression of cluster lightpink4 containing 31 protein-coding genes identified as upregulated or downregulated in marginal zone B cells. (D) The normalized expression profile of a single gene (*Zc3h12c*) from this cluster identified as upregulated in marginal zone B cells and a single eRNA (*LNCGme02323*) identified as strongly correlated with the respective WGCNA cluster. (E) Genome plots showing the location of *Zc3h12c* and *LNCGme02323*, as well as H3K4me1 and H3K4me3 chromatin signatures in marginal zone B cells, and RNA-seq read coverage in all 8 B cell populations considered in this study.

Figure 5. Identification of lncRNAs with PAX5-dependent expression in B-ALL cells. (A) Venn diagram depicting overlap between lncRNAs with sufficient read

From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

coverage to be included in this analysis (black), lncRNAs differentially expressed (DE) between B-ALL cells with and without doxycycline-induced *Pax5* expression (red), and PAX5 transcription factor binding sites (TFBS) annotated in either pro-B cells or mature B cells (blue) that could not be associated with a protein-coding gene. A subset of lncRNAs is both differentially expressed and has PAX5 bound within the gene body or promoter region (gold). (B) Volcano plot depicting the fold-change in lncRNA expression between B-ALL cells versus B-ALL cells with doxycycline-induced *Pax5* expression (see Supplemental Table 7) plotted against the probability that this difference had occurred by chance (q-value). Each dot represents a single lncRNA and is colored black unless it was differentially expressed ($q < 0.05$) and either near or not near a PAX5 binding site (gold or red respectively). (C) Genome plot showing PAX5-bound eRNA loci (*LNCGme00432*, *LNCGme00344 and LNCGme00345*), together with their proximal protein-coding gene, the zinc-finger protein gene B cell lymphoma 11a (*Bcl11a*). All are differentially expressed in B-ALL cells upon induction of *Pax5* expression. PAX5 binding sites in Pro-B cells are indicated in gold. The other annotated lncRNA (*LNCGme00346*) is also an eRNA that is differentially expressed on induction of *Pax5* expression, but it shows no evidence of PAX5 binding.

Figure 1

From www.bloodjournal.org by guest on July 7, 2

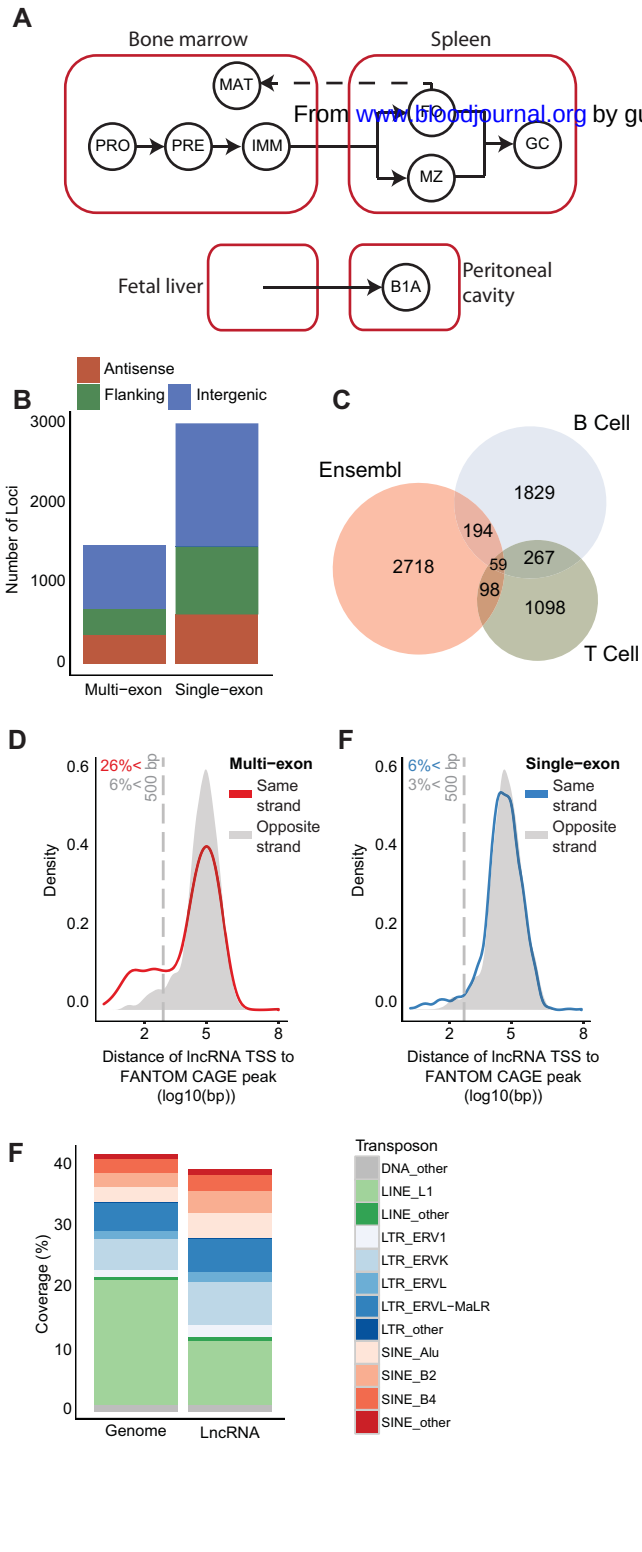


Figure 2

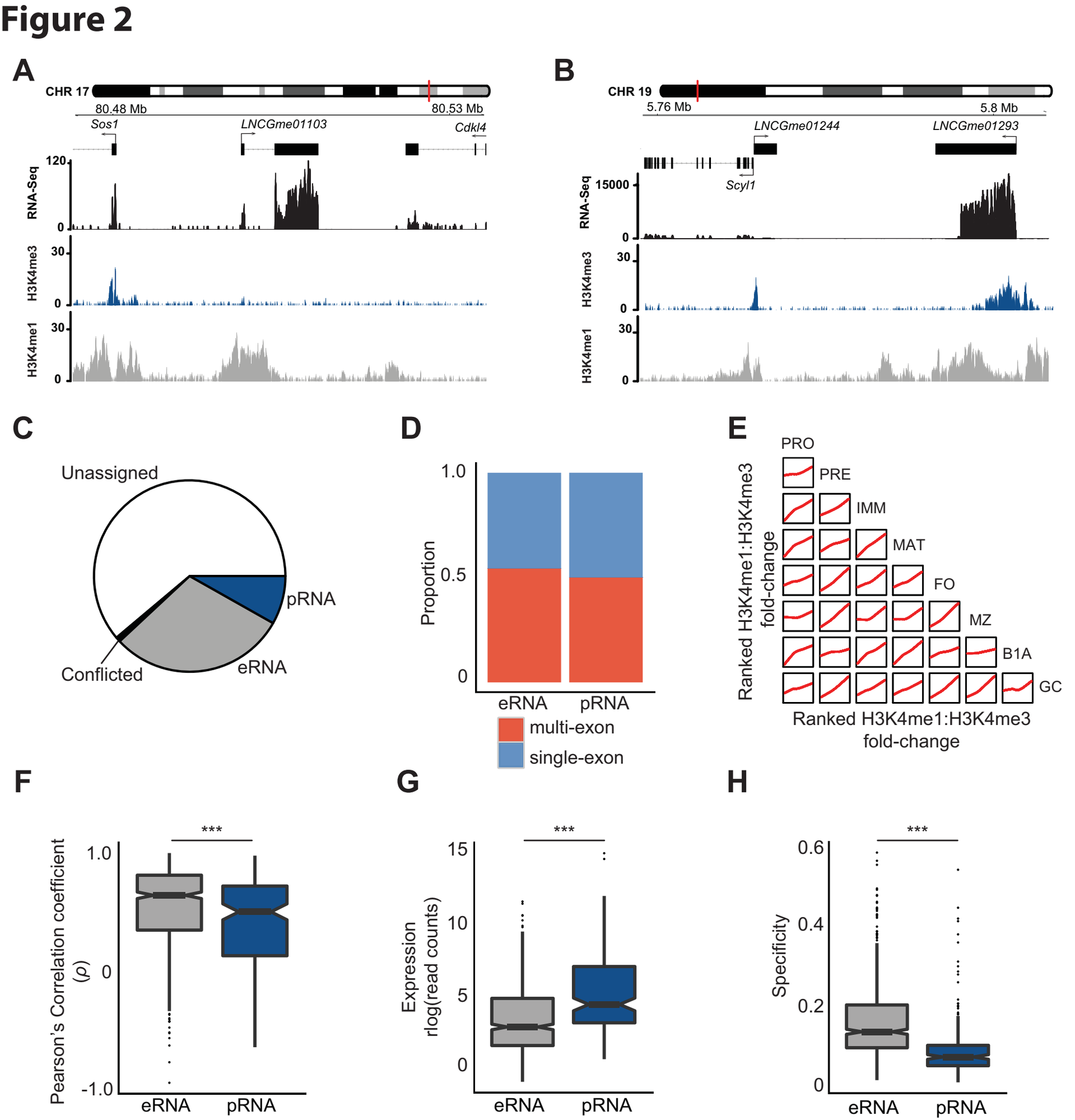

A
CHR 17
80.48 Mb
80.53 Mb
Sos1
LNCGme01103
Cdkl4
RNA-Seq
120
0
H3K4me3
30
0
H3K4me1
30
0
B
CHR 19
5.76 Mb
5.8 Mb
LNCGme01244
LNCGme01293
Scyl1
RNA-Seq
15000
0
H3K4me3
30
0
H3K4me1
30
0
C
Unassigned
pRNA
eRNA
Conflicted
D
Proportion
1.0
0.5
0
eRNA
pRNA
multi-exon
single-exon
E
PRO
PRE
IMM
MAT
FO
MZ
B1A
GC
Ranked H3K4me1:H3K4me3 fold-change
Ranked H3K4me1:H3K4me3 fold-change
F
***
Pearson's Correlation coefficient (ρ)
1.0
0
-1.0
eRNA
pRNA
G
***
Expression rlog(read counts)
15
10
5
0
eRNA
pRNA
H
***
Specificity
0.6
0.4
0.2
0
eRNA
pRNA

Figure 3

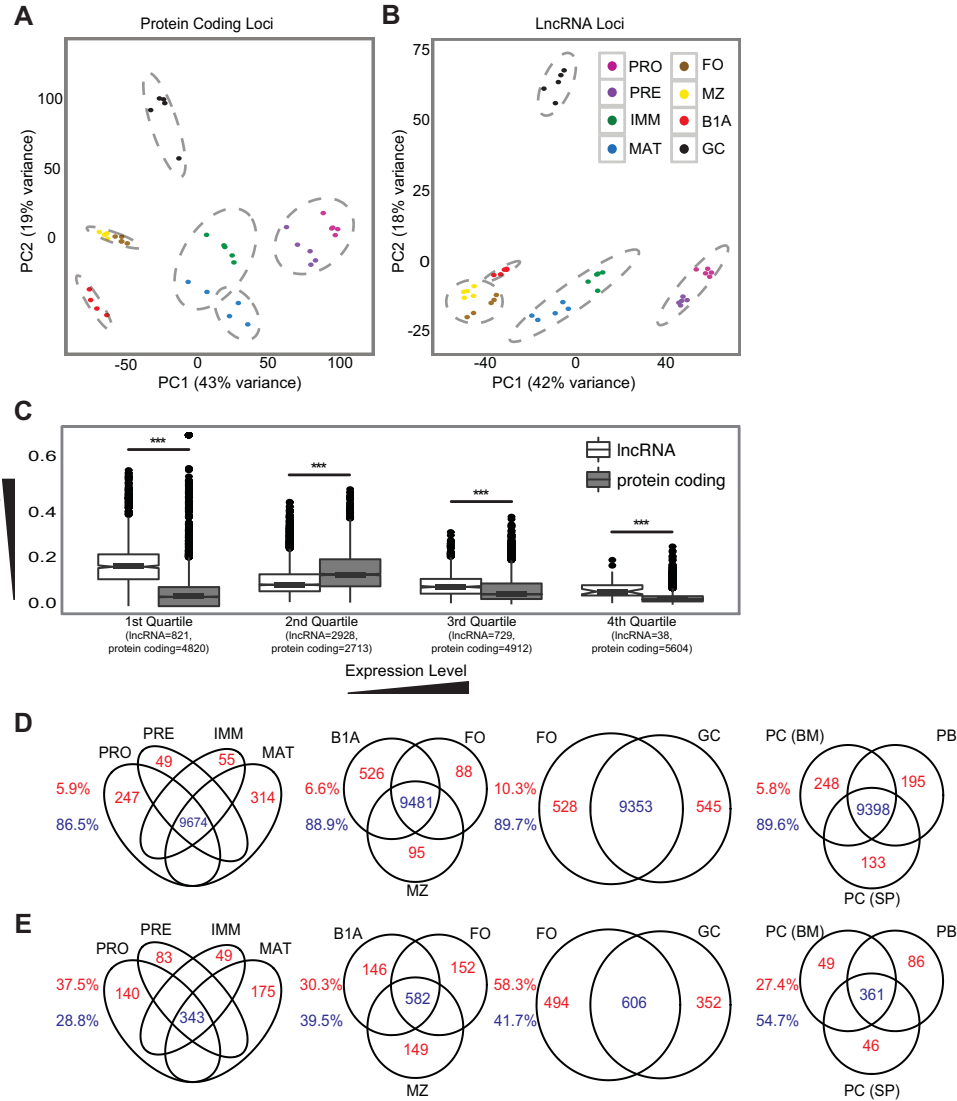
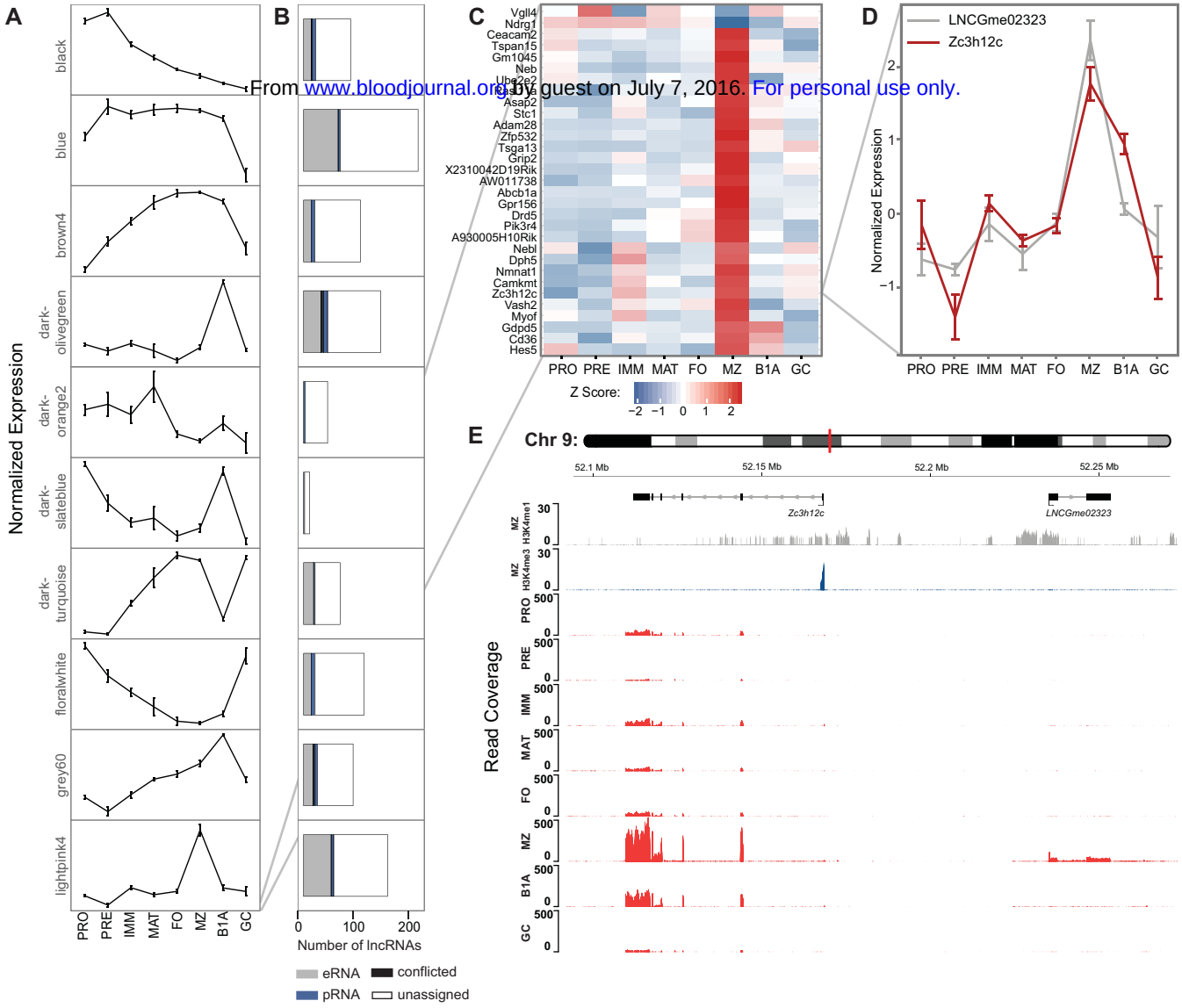

A
Protein Coding Loci
PC2 (19% variance)
PC1 (43% variance)
B
LncRNA Loci
PRO
PRE
IMM
MAT
FO
MZ
B1A
GC
PC2 (18% variance)
PC1 (42% variance)
C
lncRNA
protein coding
Specificity
***
1st Quartile
(lncRNA=821,
protein coding=4820)
2nd Quartile
(lncRNA=2928,
protein coding=2713)
3rd Quartile
(lncRNA=729,
protein coding=4912)
4th Quartile
(lncRNA=38,
protein coding=5604)
Expression Level
D
PRO
PRE
IMM
MAT
5.9%
86.5%
49
55
247
314
9674
B1A
FO
MZ
6.6%
88.9%
526
88
9481
95
FO
GC
10.3%
89.7%
528
9353
545
PC (BM)
PB
PC (SP)
5.8%
89.6%
248
195
9398
133
E
PRO
PRE
IMM
MAT
37.5%
28.8%
83
49
140
175
343
B1A
FO
MZ
30.3%
39.5%
146
152
582
149
FO
GC
58.3%
41.7%
494
606
352
PC (BM)
PB
PC (SP)
27.4%
54.7%
49
86
361
46

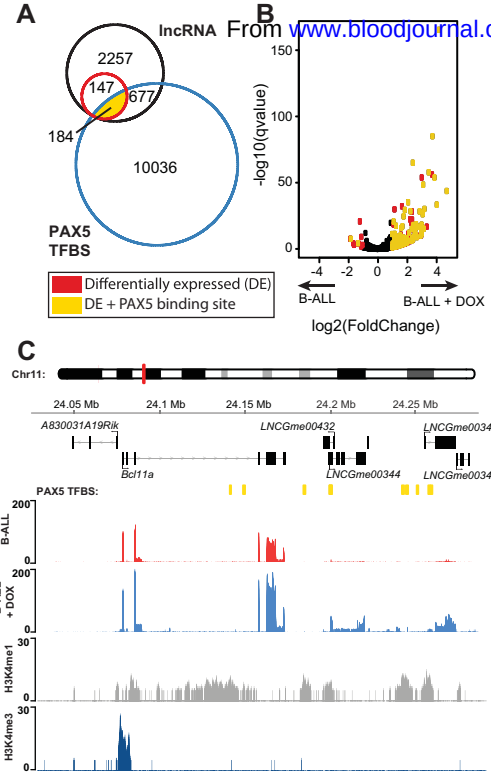
Figure 4



From www.bloodjournal.org by guest on July 7, 2016. For personal use only.

Figure 5

From www.bloodjournal.org by guest on July



From www.bloodjournal.org by guest on July 7, 2016. For personal use only.



Prepublished online July 5, 2016;
doi:10.1182/blood-2015-11-680843

Long non-coding RNAs in B cell development and activation

Tiago F. Brazão, Jethro S. Johnson, Jennifer Müller, Andreas Heger, Chris P. Ponting and Victor L.J. Tybulewicz

Information about reproducing this article in parts or in its entirety may be found online at:
http://www.bloodjournal.org/site/misc/rights.xhtml#repub_requests

Information about ordering reprints may be found online at:
http://www.bloodjournal.org/site/misc/rights.xhtml#reprints

Information about subscriptions and ASH membership may be found online at:
http://www.bloodjournal.org/site/subscriptions/index.xhtml

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include digital object identifier (DOIs) and date of initial publication.

Blood (print ISSN 0006-4971, online ISSN 1528-0020), is published weekly by the American Society of Hematology, 2021 L St, NW, Suite 900, Washington DC 20036.
Copyright 2011 by The American Society of Hematology; all rights reserved.