SUPPLEMENTAL METHODS

Immunization

Female 7 week-old C57BL/6JNimr mice were injected i.p. with 10^9 sheep red blood cells (Innovative Research) in endotoxin–free PBS. Mice were sacrificed 10 d post-immunization and splenocytes were sorted as described below.

Cell Sorting

Spleens were excised and mashed through a 70 μm cell strainer to generate single cell suspensions. Red blood cells (RBC) were lysed for 2 min in ACK Lysis Buffer and remaining cells were stained with appropriate antibodies and a near-IR dead cell stain (LIVE/DEAD fixable Dead Cell Stain, Life Technologies) in PBS at 0°C for 15 min. To sort germinal center (GC) B cells, RBC-lysed splenocytes were depleted of $CD43^+$ cells using anti-CD43-biotin and streptavidin Dynabeads (Life Technologies), followed by antibody staining in PBS. Gating strategies are shown in Supplemental Figure 1. Cells were sorted on a BD FACSAria II, BD Influx or Beckman Coulter MoFlo XDP cell sorter. Germinal center B cells were sorted from immunized mice. For plasmablasts and plasma cells isolated from C57BL/6J-*Prdm1*^{tm1Nutt} (Blimp1-GFP) mice[1], cells were sorted for GFP^+ which indicates expression of Blimp1, a marker of these subsets.

ChIP-seq

Sorted cells from the 8 B cell subsets (2 – 5 x 10^6 cells/sample) were cross-linked with 1% formaldehyde for 10 min at room temperature, after which 1/20 vol. 2.5M glycine was added to stop the crosslinking reaction. Cells were washed twice with PBS at 0°C, lysed with SDS-lysis buffer (1% SDS, 10mM EDTA, 50mM Tris-HCl pH 8) on ice for 10 min and chromatin was sheared by sonicating samples in a Diagenode Bioruptor in ice-cold water, alternating 30 s of sonication with 30 s rest for

a total of 18 min. Anti-H3K4me1 (8895, Abcam) or anti-H3K4me3 (49-1005, Life Technologies) antibodies were added to samples and incubated overnight at 4°C. Protein A Dynabeads (Life Technologies 10002D) were added to samples and incubated 2 h at 4°C. Beads were washed 5 times: once with low salt wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl pH 8, 150mM NaCl), once with high salt wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl pH 8, 500mM NaCl), once with LiCl buffer (0.25mM LiCl, 1% NP-40, 1% Na-deoxycholate, 1mM EDTA, 10mM Tris-HCl pH 8) and twice with 10mM Tris-HCl pH 8, 1mM EDTA. Chromatin was eluted with SDS lysis buffer and incubated at 65°C overnight to reverse crosslinks. RNA was degraded with RNAseA (0.2 mg/ml) for 2 h at 37°C followed by protein degradation with proteinase K (0.2 mg/ml) for 1 h at 55°C. DNA was purified by phenol/chloroform extraction and ethanol precipitation using standard methods, and DNA quality was assessed using the 2100 Expert Agilent Bioanalyser. Sequencing libraries were made using the ChIP Sample Preparation Library Kit (Illumina) and sequenced in a Genome Analyser IIx (Illumina), collecting 20 - 50 million single-end reads of 50 bases per sample.

PCR validation

RNA purified from sorted B cell subsets as described above was reverse transcribed to cDNA according to the manufacturer's instructions (SuperScript III First Strand Synthesis SuperMix for qRT-PCR – Life Technologies). PCR was performed as indicated by the manufacturer (Taq DNA Polymerase, Life Technologies) in a T100 Thermal Cycler (BIO-RAD). PCR products were analyzed by electrophoresis in a 2% agarose gel by standard methods. Primers used for lncRNA validations are indicated in Supplemental Table 8.

RNA-seq read alignment

Mapped library sizes ranged between 22 and 89 million read pairs (median 41 million) and a median of 90% of reads were uniquely mapped (Supplemental Table 9). RNA-seq reads were aligned using STAR[2] v2.3.0e, with a genome/suffix array that incorporated a splice junction database derived from Ensembl v72 reference annotations. In addition to default mapping options, parameters were adjusted to ensure that only high confidence spliced reads were retained (--outFilterType BySJout, --outFilterIntronMotifs RemoveNoncanonicalUnannotated).

Identification of long non-coding RNAs

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[3]. Further steps were taken to filter transcripts in close proximity to protein-coding genes, including removal of loci containing transcripts that were sense intronic, removal of loci <5kb sense downstream of an Ensembl annotated protein-coding gene, which were considered to potentially be unannotated 3' UTRs, removal of loci within 20kb sense downstream of a protein-coding gene if >60% of the intergenic space was covered by one or more reads in at least two samples, and removal of loci connected to a protein-coding gene via one or more spliced reads. Finally, the protein-coding potential of remaining transcripts was calculated for concatenated exons using the Coding Potential Calculator[4] (CPC) v0.9-r2 and PhyloCSF[5] v1.0. Loci containing transcripts classified as coding by both methods were discarded (Supplemental Figure 3). Computational pipelines developed for lncRNA identification and subsequent analysis are publicly available (www.cgat.org/downloads/public/projects/proj010/pipelines).

LncRNA nomenclature

LncRNA loci containing one or more spliced transcripts were classified as multi-exon and given the prefix ‘LNCGme’. Loci containing no spliced transcripts were classified as single-exon and given the prefix ‘LNCGse’. Any lncRNAs arising from our pipeline that were <200bp apart on the same strand were merged (i.e. assigned the same gene identifier), and merged loci containing one or more spliced transcript were given the prefix ‘LNCGmm’, whereas merged loci containing no spliced transcripts were given the prefix ‘LNCGsm’. The transcriptional start site (TSS) of each lncRNA was defined as the most 5’ position within all the transcripts assigned to a locus.

UCSC public hub

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

The hub, termed ‘LncRNAs in B cells’ has 19 dropdown menus. The first eight menus with the prefix ChipSeq_*, show normalized ChIPseq read coverage for each of the eight B cell populations in this study. By default, overlapping displays are shown for the two chromatin marks, which are colored to correspond to the Figures in the paper. Clicking on the blue title above each dropdown menu in the panel gives access to the Track Settings page, where it is possible to toggle on/off the display for each chromatin mark. The next two dropdown menus with the prefix H3K4me*, show intervals covered by the ChIP peaks called as statistically significant in this study, again color-coded to match the Figures in the paper. By default a single set of intervals is displayed, which represents a merged consensus of the intervals called in the eight principal B cell populations. However, by accessing the Track Settings (as above), it is possible to toggle on/off displays of the peak intervals called for each individual cell type. Next is a single dropdown menu, entitled LncRNA_Loci, which shows the collapsed gene model for each lncRNA identified in this study. In Track

Settings it is possible to toggle on/off a display of the full set of transcript models for each locus. Finally, eight dropdown menus with the prefix RNASeq*, show the normalized RNAseq read coverage for each of the eight B cell populations in this study. By default a single display is shown, which represents the pooled coverage across five replicate RNAseq samples; however, by entering the Track Settings it is possible to toggle on/off the display for each individual replicates. As a rule, the suffix '-R0' on track names indicates that the track displayed is comprised of pooled replicates, whereas the suffix '-R[1-9]' indicate that the track displayed represents a single sample.

Comparison with FANTOM5 CAGE data

Cap Analysis of Gene Expression (CAGE) data produced by the FANTOM consortium[6] were downloaded from their website (http://fantom.gsc.riken.jp/5/), and CAGE tags, where possible, assigned to ENSEMBL reference annotations (build 72). The distances between unassigned CAGE tags and TSS of collapsed gene models for lncRNA were calculated, and following the threshold used by the FANTOM consortium, a locus was assigned to a CAGE peak if the distance was <500 bp.

LncRNA overlap with transposable elements.

The locations of transposable elements (TEs) within the mm10 genome were identified by downloading UCSC RepeatMasker annotations classified as 'DNA', 'LINE', 'SINE', 'LTR', or 'Transposon'. Subsequently, transcripts at each mouse lncRNA locus were merged to create a single 'consensus' transcript spanning all the exons within a locus, the number of bases overlapping each TE class at each locus was calculated, and overlap was reported as a proportion of the consensus transcript length.

ChIP-seq read alignment and peak calling

ChIP-seq reads for chromatin marks (H3K4me3 and H3K4me1) were trimmed to a common length (40bp) and aligned to the mouse reference genome (mm10) using BWA[7] v0.7.8. Mapping was carried out using a seed length of 28 bases, with a maximum of two mismatches allowed within the seed region and a maximum of four mismatches allowed across the entire read. Peak calling employed the ENCODE Irreproducibility Discovery Rate (IDR) framework[8] to filter libraries responsible for poor peak consistency within each cell type. Briefly, non-uniquely mapping and duplicated alignments were removed from libraries prior to peak calling. Peaks were then called against a single, pooled input for each cell type using macs2[9] v2.0.10 with a p-value threshold of $5x10^{-3}$ and all other parameters set as default. A maximum of 150,000 peaks were subsequently retained for IDR analysis and self-consistency/inter-replicate consistency was determined at an IDR threshold of $2x10^{-2}$, whilst pooled-consistency was determined at a threshold of $5x10^{-3}$. In accordance with the IDR framework, samples with either low self-consistency or low inter-replicate consistency were discarded and those samples that remained were pooled, enabling a consensus set of peaks to be called for each cell type. After filtering, a minimum of two ChIP-seq libraries remained for each chromatin mark/cell type combination.

Identification of intergenic lncRNAs with enhancer-like and promoter-like expression

Intergenic lncRNAs were further classified on the basis of their relative enrichment of H3K4me1 vs. H3K4me3 read coverage across proposed TSS (Supplemental Figure 2B). However, to avoid the confounding effect of chromatin signatures associated with protein coding genes, this analysis was restricted to lncRNA loci classified as intergenic (i.e. >5kb from the nearest Ensembl protein coding gene annotation). Mapped ChIP-seq libraries that passed IDR filtering (described above) were filtered to remove duplicate and non-uniquely mapping reads, and replicates were pooled to

create a single library for each cell stage/chromatin mark combination. In order to obtain comparable fold-change estimates across cell stages, the larger of the two chromatin libraries for each cell stage was down-sampled to equal the read depth of the smaller library. Read coverage was then quantified independently for each cell stage across a 4kb window centered on the TSS of the collapsed gene model for each lncRNA locus. For lncRNA TSS within 500bp of a robust FANTOM5 CAGE annotation, the FANTOM5 CAGE peak was used to center the 4kb window, as it was assumed to provide a more accurate estimation of the true TSS. Loci with low read coverage (<100 reads) in both libraries were discarded and remaining loci ranked based on their respective H3K4me1:H3K4me3 read ratios, before an empirical fold-change threshold (>4-fold difference in normalized read coverage) was applied to identify subsets of loci with either enhancer-like (high H3K4me1:H3K4me3) or promoter-like (low H3K4me1:H3K4me3) chromatin signatures. To increase the robustness of lncRNA classification, loci associated with either enhancer-like, or promoter-like chromatin signatures in one or more cell stages were further filtered to remove instances where read coverage for the dominant chromatin mark was not significantly enriched above background (i.e. input). This was achieved by discarding any locus above/below the specified fold change thresholds that did not also have a supporting H3K4me1/H3K4me3 IDR peak, respectively, within 2kb of its TSS.

Following classification of lncRNA independently within each cell stage, results were pooled across all eight cell stages and intergenic lncRNAs classified as either i) eRNA (enhancer-like chromatin signature in one or more cell stages), ii) pRNA (promoter-like chromatin signature in one or more cell stages), iii) conflicted (enhancer-like chromatin signature in one cell stage, and promoter-like signature in another), or iv) unassigned (no definable chromatin signature in any cell stage).

Within each cell stage, the relative strength of signal in the two ChIP-seq experiments (H3K4me1 vs. H3K4me3) was variable (Supplemental Figure 6A), which in turn affected the number of eRNAs/pRNAs called at the empirically selected fold-change threshold (Supplemental Figure 6B-C). However, the relative strength of the H3K4me1:H3K4me3 chromatin signature at each lncRNA TSS (represented by ranked fold-change estimates) was broadly consistent across cell stages (Figure 2E).

This observation was further assessed using gene set enrichment analysis (GSEA) implemented in the R Bioconductor package Piano. For each cell population, the 2349 intergenic lncRNA loci in our catalogue were ranked on the ratio of H3K4me1:H3K4me3 coverage across the TSS. The position of lncRNA classified as eRNA in each of the other cell populations was then calculated within this rank using GSEA. Loci classified as eRNA in one cell population were consistently positioned highly within the rank order for other cell populations, indicating that eRNAs always maintained a high H3K4me1:H3K4me3 chromatin signature (Supplemental Figure 6D). Reciprocally, loci classified as pRNA in one cell population were consistently positioned lowly within the ranked order for all other cell populations, indicating that pRNAs always maintained a low H3K4me1:H3K4me3 chromatin signature (Supplemental Figure 6E). This result therefore supported the observation that the classification of a lncRNA locus as an eRNA or pRNA remained broadly consistent during B cell development.

Identifying human orthologues of mouse lncRNAs.

Human lncRNA annotations were retrieved from two sequence-based studies of human lymphocytes[10,11] and one microarray-based study[12]. For each sequence-based study genomic intervals for lncRNA exons were converted from human (hg19) to mouse (mm10) coordinates using the UCSC liftOver tool in conjunction with reciprocal best chain files (http://hgdownload-

test.cse.ucsc.edu/goldenPath/hg19/vsMm10/reciprocalBest/) generated from pairwise genomic alignment. For the microarray-based study re-mapped probesets for the Human Exon 1.0 Affymetrix microarray were downloaded from the BrainArray database (http://brainarray.mbni.med.umichedu) and probesets containing fewer than four probes were discarded. Remaining probes were matched to human gene annotations using microarray metadata and exons annotated as lncRNAs by Petri et al.[12] were converted to mm10 coordinates as described for the sequence-based studies. Human lncRNAs were classified as syntenic orthologues if one or more human exon overlapped (>1bp) a mouse lncRNA locus identified in the current study.

To account for redundancy between human datasets, human genomic coordinates for lncRNAs from all three human studies were intersected and cases where different lncRNA annotations originated from the same locus were recorded. In Supplemental Table 6, all human lncRNAs from the same locus are reported if one or more of those lncRNAs was found to intersect a mouse lncRNA annotation.

Human eRNAs were identified using histone methylation data for $CD19^{+}$ cells (H3K4me1 and H3K4me3) reported by Casero et al.[11]. Following the stringent criterion used in the current study, a human lncRNA was classified as an eRNA if there was a >4-fold ratio of H3K4me1:H3K4me3 coverage across the lncRNA TSS. The nearest protein-coding gene upstream and downstream of each human eRNA was identified and details of mouse protein coding orthologues were downloaded from Ensembl BioMart (www.ensembl.org/biomart). Supplemental Table 6 reports cases where human and mouse protein-coding orthologues are both flanked by a lncRNA locus that could be identified as an eRNA in human[11] and mouse (the current study).

Tissue Specificity of Expression

Read counts across loci were regularized log (rlog) transformed using the R package DESeq2[13]. Median rlog-transformed values were calculated for each B cell stage and used to calculate tissue specificity of expression as described[14], where a value of 1 indicates expression in a single cell stage (high specificity), and a value of 0 indicates even expression across all cell stages (low specificity).

Sample Clustering and Principal Components Analysis

Hierarchical clustering of samples was carried out using the R package hclust, on a Euclidean distance matrix derived from regularized log (rlog) transformed RNAseq read counts across annotated protein and lncRNA loci. Principal components analysis of the same data was performed using the R package prcomp.

Weighted Gene Co-expression Network Analysis

Protein coding genes were filtered to select only loci expressed above a median threshold of FPKM = 1 in one or more B cell populations. Regularized log-transformed (rlog) read count data for individual samples (n=40) were then analyzed using the R package WGCNA[15]. An unsigned, weighted matrix of pairwise correlations between genes was constructed, raised to the power 12 (soft-threshold), and converted into a topological overlap matrix (TOM). The resulting TOM was converted to a dissimilarity measure and clustered using average-linkage hierarchical clustering within the R package flashClust to produce a tree that described relatedness between genes based on their co-expression profiles. We then used the dynamic tree cut algorithm[16] (without recursion) to identify modules of co-expressed genes and compared the similarity of resulting modules by calculating the Pearson correlation between module eigengenes. Modules with an eigengene correlation >0.7 were subsequently merged to produce a final set of 17 WGCNA modules (Supplemental Figure 9). Modules of protein coding genes identified through WGCNA were tested for enriched/depleted representation of Gene Ontology (GO)

categories using the R package GOSeq (Supplemental Table 4). LncRNAs were subsequently tested for an association with WGCNA modules by correlating lncRNA gene expression (FPKM) with module eigengenes, and the strongest eigengene correlation for each lncRNA locus was recorded (Supplemental Table 2).

To test the significance of the frequency at which loci classified as eRNAs/pRNAs showed strongest correlation in expression with a WGCNA module containing one of their neighboring protein coding genes, a Boolean list was constructed which recorded the occurrence of this event across the 2349 intergenic lncRNAs identified in this study. In total 126 eRNAs and 30 pRNAs showed strongest correlation with a WGCNA module containing one of their neighboring protein coding genes. The Boolean list of coincidence in module association was then randomly permuted 10000 times to provide a null distribution of the expected frequency at which this association would occur for each lncRNA class. P values from these permutation tests are reported in the main text.

PAX5 regulation of lncRNA expression

RNA-seq data relating to PAX5-dependent regulation of gene expression in pro-B cells and mature B cells originated from a published study[17] and were downloaded from the Gene Expression Omnibus (GEO, accession GSE38046). Published intervals for PAX5 binding in pro-B cells and mature B cells were downloaded from the same study. RNA-seq reads were aligned to the mouse reference genome (mm10) using STAR v2.3.0e and read coverage across lncRNAs was quantified using FeatureCounts[18]. Tests for differential expression of lncRNAs between wild-type and Pax5-deficient cells were performed using the R Bioconductor package DESeq2.

Published intervals for PAX5 binding in pro and mature B cells were filtered to remove intervals overlapping or <1kb upstream of an annotated protein coding locus (Ensembl V72), where overlap was defined as an intersection of ≥ 1bp. Intervals that remained were considered to have no association with a protein coding gene. They were subsequently tested for enrichment of overlaps across lncRNA loci by expanding lncRNA loci 1kb upstream and subsequently testing for genomic overlap using the Genomic Association Tester[19] (GAT). The GAT workspace was obtained by dividing the ungapped genome into 8 equally sized regions of equivalent GC content (isochores).

RNA-seq data relating to gene expression regulated by Pax5 expression in a mouse model of B-ALL, originated from a published study by Liu *et al.*[20] and were downloaded from GEO (accession GSE52868, GSE52870). Read mapping, quantification, and differential expression testing were performed as described above.

SUPPLEMENTAL TABLES

Supplemental Table 1. Genomic co-ordinates of lncRNA loci and transcripts identified in the current study. Data are in 12 column bed format (bed12). The name column contains the lncRNA gene ID and lncRNA transcript ID separated by an underscore.

Supplemental Table 2. LncRNA and protein coding gene expression. Table lists expression (FPKM) of lncRNAs and coding genes across all B cell subsets, showing either the expression in each replicate sample, or the median across replicates. Also shown is the WGCNA module assignment for both lncRNAs and protein coding genes, lncRNA classification on the basis of chromatin state, and the location of lncRNAs relative to nearest protein coding gene.

Supplemental Table 3. Intersection between lncRNAs and transposable elements. Table shows the percent overlap between lncRNA exons and different families of transposable elements.

Supplemental Table 4. Gene Ontology (GO) enrichment analysis results for WGCNA modules identified on the basis of protein coding gene expression. Results are shown for modules with significantly over or underrepresented GO terms ($p <= 0.05$).

Supplemental Table 5. Correlation in expression between eRNAs and their proximal protein coding genes. The first sheet lists correlations for eRNAs that show strongest correlation in expression with a WGCNA module containing a

neighboring protein coding gene, highlighting those with | ρ | > 0.8. The second sheet lists all eRNAs with correlations to their nearest 5' or 3' protein coding gene.

Supplemental Table 6. Human-mouse lncRNA orthologues. The first sheet lists mouse lncRNAs with orthologous human lncRNAs identified based on conserved genomic location. The second sheet lists pairs of mouse and human eRNAs that are adjacent to orthologous coding genes.

Supplemental Table 7. PAX5 regulation of lncRNA expression. Analysis of differential expression of lncRNAs between wild-type and PAX5-deficient pro-B cells and mature B cells, and between B-ALL cells with and without *Pax5* expression.

Supplemental Table 8. Details of primers used for lncRNA validation.

Supplemental Table 9. RNA-seq library sizes and proportion of reads successfully mapped.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1. Flow cytometric gating strategy for cell sorting of B cell subsets. In all cases cells were gated on live cells (negative for dead cell stain) and gated using a forward scatter/side scatter gate known to include most lymphocytes. B cell subsets were isolated using the following further gating strategies: (A) follicular (Fo, $B220^{+}AA4.1^{-}Cd23^{+}IgM^{intermediate}$) and marginal zone (MZ, $B220^{+}AA4.1^{-}Cd23^{-}IgM^{high}$) B cells from the spleen; (B) germinal center (GC, $B220^{+}Cd38^{low}PNA^{high}Fas^{+}GL7^{+}$) B cells from spleen; (C) mature B cells (MAT, $B220^{+}Cd19^{+}IgM^{intermediate}IgD^{+}$), pro-B cells (PRO, $B220^{+}Cd19^{+}IgD^{-}IgM^{-}Cd2^{-}$), pre-B cells (PRE, $B220^{+}Cd19^{+}IgD^{-}IgM^{-}Cd2^{+}$) and immature B cells (IMM, $B220^{+}Cd19^{+}IgD^{-}IgM^{+}Cd2^{+}$) from bone marrow; (D) B1a ($B220^{intermediate}IgM^{high}Cd5^{+}Cd23^{-}$) cells from the peritoneal cavity; (E) plasmablasts (PB, $CD4^{-}CD8^{-}NK1.1^{-}CD138^{+}Blimp1^{+}B220^{intermediate}$) and plasma cells (PC, $CD4^{-}CD8^{-}NK1.1^{-}CD138^{+}Blimp1^{+}B220^{-}$) from spleen; and (F) plasma cells (PC, $CD11b^{-}CD138^{+}Blimp1^{+}B220^{-}$) from bone marrow.

Supplemental Figure 2. Bioinformatic pipelines for analysis of RNA-seq and ChIP-seq data. (A) Identification of lncRNAs based on *de novo* assembly of transcripts from RNA-seq data. (B) Identification of lncRNAs with either enhancer-associated (eRNA), or promoter-associated (pRNA) chromatin marks based on ChIP-seq data for H3K4me3 and H3K4me1.

Supplemental Figure 3. Filtering of potential lncRNA loci based on transcript coding potential. (A) Cumulative distribution function (CDF) plot showing the distribution of phyloCSF scores obtained for lncRNA transcripts identified by the current study (blue) and, for comparison, transcripts annotated as long intergenic noncoding RNAs (lincRNAs) in ENSEMBL reference annotation v72 (red). (B) CDF

plot showing the distribution of CPC scores obtained for lncRNA transcripts identified by the current study (blue) and ENSEMBL annotated lincRNA transcripts (red). (C) Smoothed scatter plot showing the correlation between phyloCSF score and CPC score for potential lncRNA transcripts identified in the current study. Dashed lines denote thresholds at which transcripts were flagged as having coding potential by the respective classification method. Loci containing one or more transcripts identified as coding by both phyloCSF and CPC methods (panel C, top right quadrant) were discarded from the final lncRNA catalogue.

Supplemental Figure 4. Validation of lncRNAs by PCR. PCR products of 53 lncRNAs were run on a 2% agarose gel. 47 were successfully validated (88.7% validation rate). The label on each indicates the lncRNA ID, expected PCR product size (in bp) and B cell subset from which cDNA was made. GC, germinal center B cells; PRE, pre-B cells; PRO, pro-B cells; Fo, follicular B cells; MZ, marginal zone B cells.

Supplemental Figure 5. Comparison of single-exon and multi-exon lncRNAs. Comparison between single-exon and multi-exon lncRNAs and coding genes for (A) coding potential expressed as a PhyloCSF score, (B) level of expression in follicular B cells, (C) mean exon size, and (D) mean transcript size. Overlap between the (E) 799 multi-exon and (F) 1550 single-exon intergenic lncRNAs identified by this study (B Cell), with those identified by Hu *et al.* (T Cell), and those annotated as lincRNAs in Ensembl v74. FPKM, fragments per kilobase per million reads.

Supplemental Figure 6. Classification of eRNAs and pRNAs on the basis of chromatin marks. (A) Heat maps depicting read coverage across a 4kb window centered on the TSS of collapsed lncRNA gene models for all intergenic lncRNA loci. For each cell stage, data represent pooled ChIP-seq libraries and windows are

ranked vertically on the basis of the ratio H3K4me1:H3K4me3 read coverage. Numbers at the top and bottom of each plot represent the H3K4me1:H3K4me3 ratio for the first and last ranked window, respectively. (B) Histograms depicting the number of cell stages in which a lncRNA locus was classified as either eRNA or pRNA (data only represent loci where a classification was consistent across cell stages). (C) Bar chart depicting the number of eRNA or pRNA loci that were successfully classified using ChIP-seq libraries derived from each B cell population. (D, E) Gene-set enrichment analysis (GSEA) demonstrates that, in spite of the variable number of lncRNA loci classified as eRNA or pRNA across B cell populations (see panel C), the chromatin signatures identifying these lncRNA types remain consistent across B cell development. For each B cell population, the 2349 intergenic lncRNAs identified in this study were ranked on the ratio of H3K4me1:H3K4me3 coverage across their TSS. Subsequently, sets of (D) eRNAs and (E) pRNAs identified within each B cell population were assessed using GSEA to determine enrichment of each set within the ranked gene lists produced for that population (diagonal panels), and enrichment within the ranked gene lists produced for every other B cell population (off-diagonal panels). Heatmaps show the enrichment score for each set of eRNAs or pRNAs identified with one B cell population (y-axis) within the gene list ranked by H3K4me1:H3K4me3 ratio for its own and all other B cell populations (x-axis). Scores for all eRNA sets were positive across all ranks, whereas scores for all pRNA sets were negative across all ranks, indicating that chromatin signatures for these gene sets were generally consistent across the eight B cell populations.

Supplemental Figure 7. Hierarchical clustering and principal component analysis (PCA) of protein coding gene and lncRNA expression data. Unsupervised hierarchical clustering of regularized log-transformed (A) protein coding and (B) lncRNA expression data. Grey dashed lines depict a 60% similarity

threshold that defines clusters depicted in the PCA plots shown in Figure 3A,B. (C) Principal component analysis and (D) unsupervised hierarchical clustering of regularized log-transformed expression data for the 2349 intergenic lncRNA loci identified in this study, demonstrating that clustering of samples on the basis of lncRNA expression (Figure 3B) is not an artifact arising from proximal protein coding gene expression. (E) Schematic representation of the ontogenetic relationships between B cell populations considered in this study. Subsets labeled as in Figure 1A with the addition of antibody-secreting subsets: plasmablasts (PB) and plasma cells (PC) in the spleen and PC in the bone marrow. Solid arrows indicate developmental progression through B cell stages or activation. Dashed line indicates recirculation of follicular B cells or plasma cells back to the bone marrow. (F, G) Principal component analysis of regularized log-transformed expression patterns of (F) protein coding loci and (G) the 4516 lncRNA loci identified in this study, including antibody-secreting subsets. Grey dashed lines indicate groups identified by unsupervised hierarchical clustering.

Supplemental Figure 8. Differential expression of lncRNAs and protein coding genes across 8 B cell subsets. (A) Heat maps depicting the proportion of loci that are differentially expressed (FDR < 0.05%) between B cell subsets in pairwise comparisons of expression at protein coding loci and all 4516 lncRNA loci. (B, C) Heat maps for protein coding (B) and lncRNA (C) loci binned into quartiles on the basis of median expression across all B cell samples. Expression bins are the same as those used in Figure 3C. Numbers above each heat map depict the number of loci falling within a particular expression bin.

Supplemental Figure 9. Modules of protein coding gene expression identified through weighted gene co-expression network analysis (WGCNA). (A) Dendrogram depicting clustering of 11,772 protein coding genes with sufficient

expression (FPKM > 1) to be included in analysis. Lower panel depicts clusters identified using the dynamic tree cutting approach[16] (B) Heat maps depict normalized expression values (z-scores) for genes assigned to each WGCNA cluster (see Supplemental Table 2). Line plots above each heat map depict the cluster expression profile (eigengene), which was used to identify lncRNAs with similar expression patterns.

Supplemental Figure 10. Identification of lncRNAs with PAX5-dependent expression in pro-B cells and mature B cells. Venn diagrams for pro-B cells (A) and mature B cells (C), depicting overlap between lncRNAs with sufficient read coverage in PAX5-expressing and PAX5-deficient data sets[17] to be included in this analysis (black), lncRNAs differentially expressed between wild-type and PAX5-deficient cells (red), and PAX5 transcription factor binding sites (TFBS, blue). A subset of lncRNAs is both differentially expressed and has PAX5 bound within the gene body or promoter region (gold). Volcano plots for pro-B cells (B) and mature B cells (D), showing change in expression between wild-type (WT) versus PAX5-deficient (KO) cells 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).

REFERENCES

1. 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.
2. Dobin A, Davis CA, Schlesinger F*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2012;29:15-21.
3. Slater G Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* 2005;6(1):31.
4. 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.
5. 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.
6. Fantom Consortium. A promoter-level mammalian expression atlas. *Nature.* 2014;507(7493):462-470.
7. Li H Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760.
8. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat.* 2011;5(3):1752-1779.
9. Zhang Y, Liu T, Meyer C*, et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.
10. Bonnal RJ, Ranzani V, Arrigoni A*, et al.* De novo transcriptome profiling of highly purified human lymphocytes primary cells. *Sci Data.* 2015;2:150051.
11. 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.
12. Petri A, Dybkaer K, Bogsted M*, et al.* Long Noncoding RNA Expression during Human B-Cell Development. *PLoS One.* 2015;10(9):e0138236.

13. Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

14. 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.

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

16. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R. *Bioinformatics.* 2007.

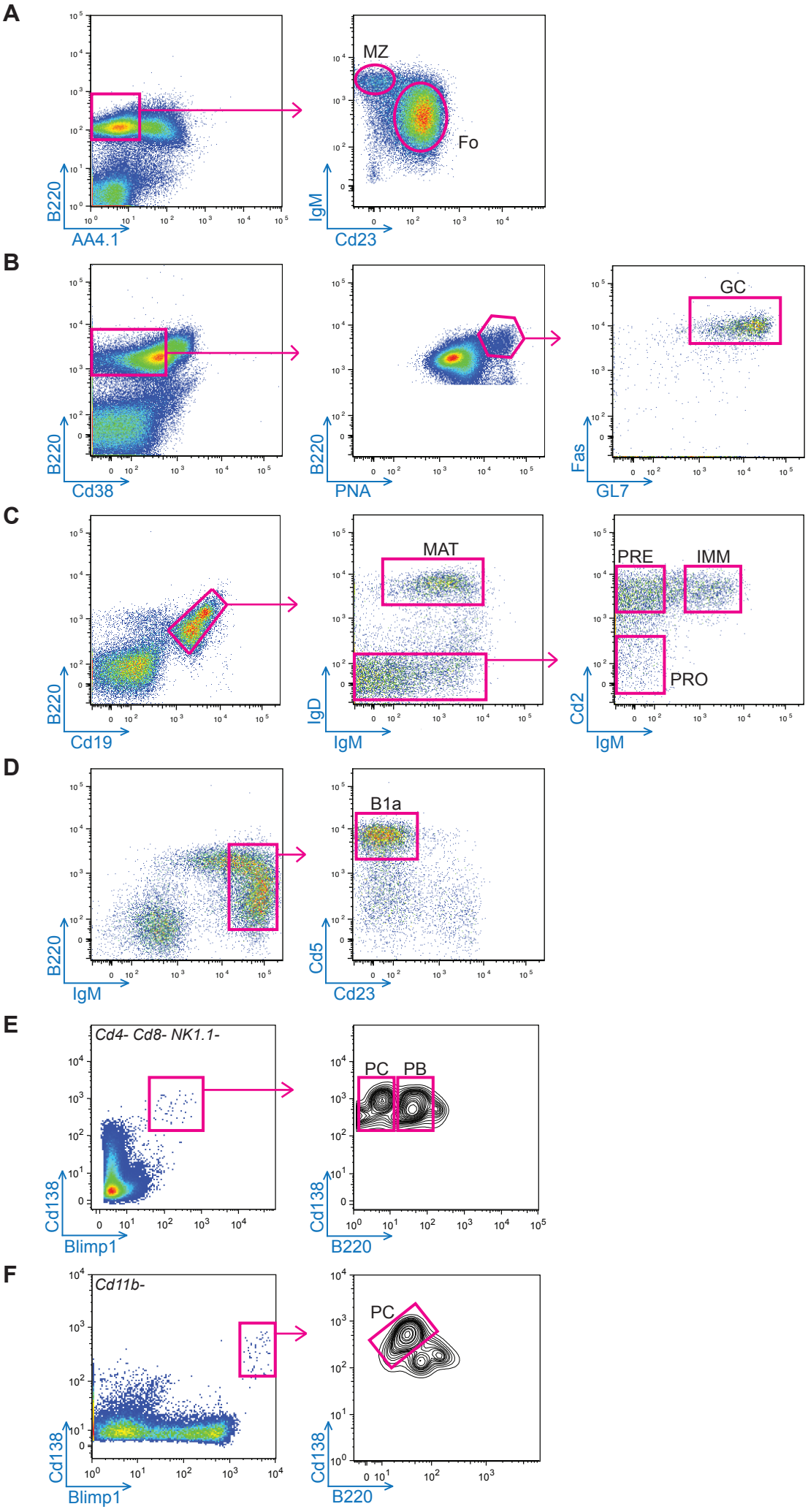
17. Revilla-i-Domingo R, Bilic I, Vilagos B*, et al.* The B-cell identity factor Pax5 regulates distinct transcriptional programmes in early and late B lymphopoiesis. *EMBO J.* 2012;31(14):3130-3146.

18. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2013.

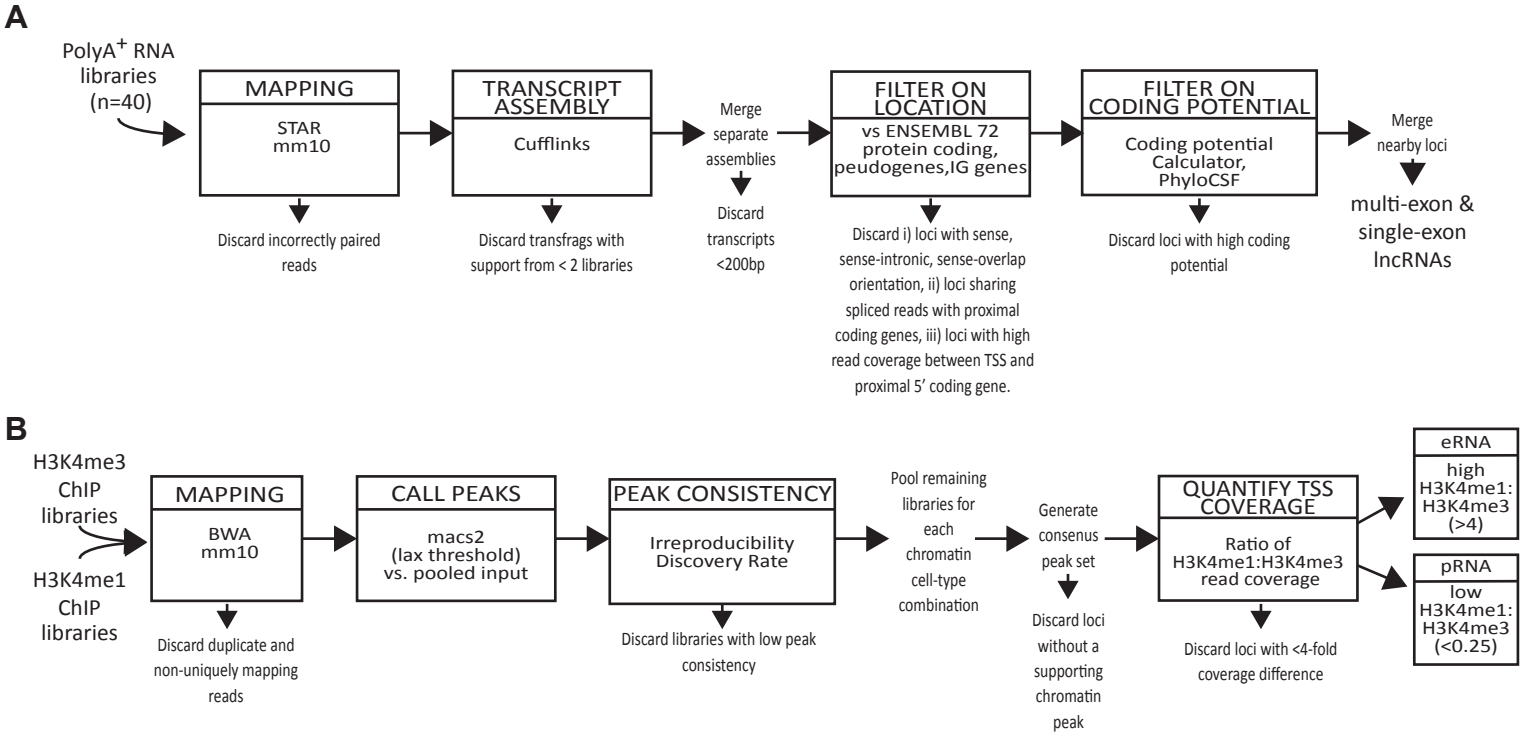
19. Heger A, Webber C, Goodson M, Ponting CP, Lunter G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics.* 2013;29(16):2046-2048.

20. 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.

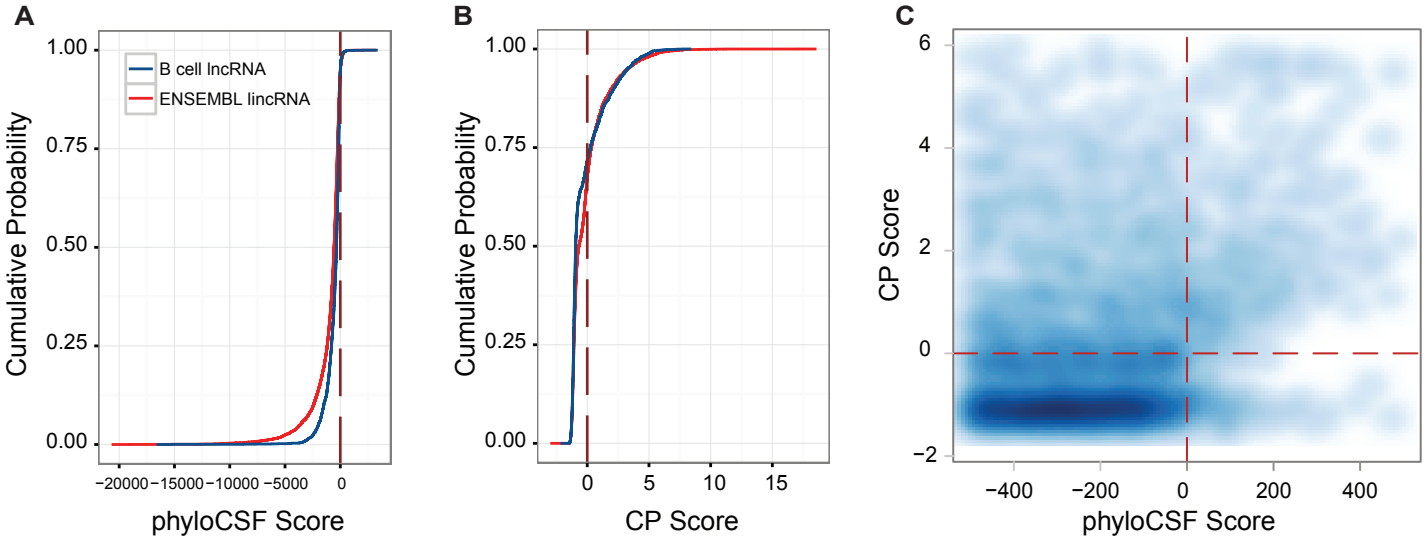
Supplemental Figure 1



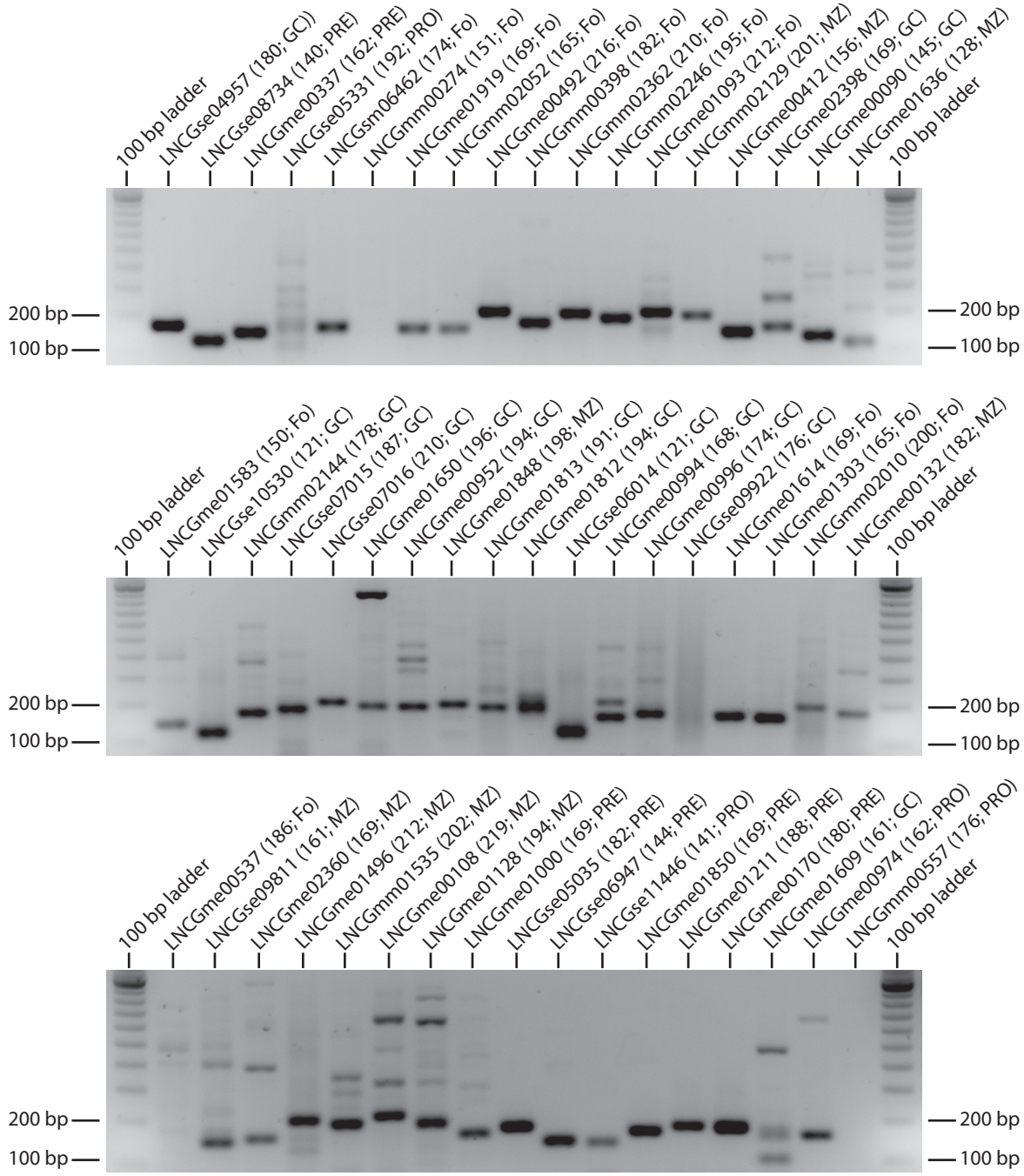
Supplemental Figure 2



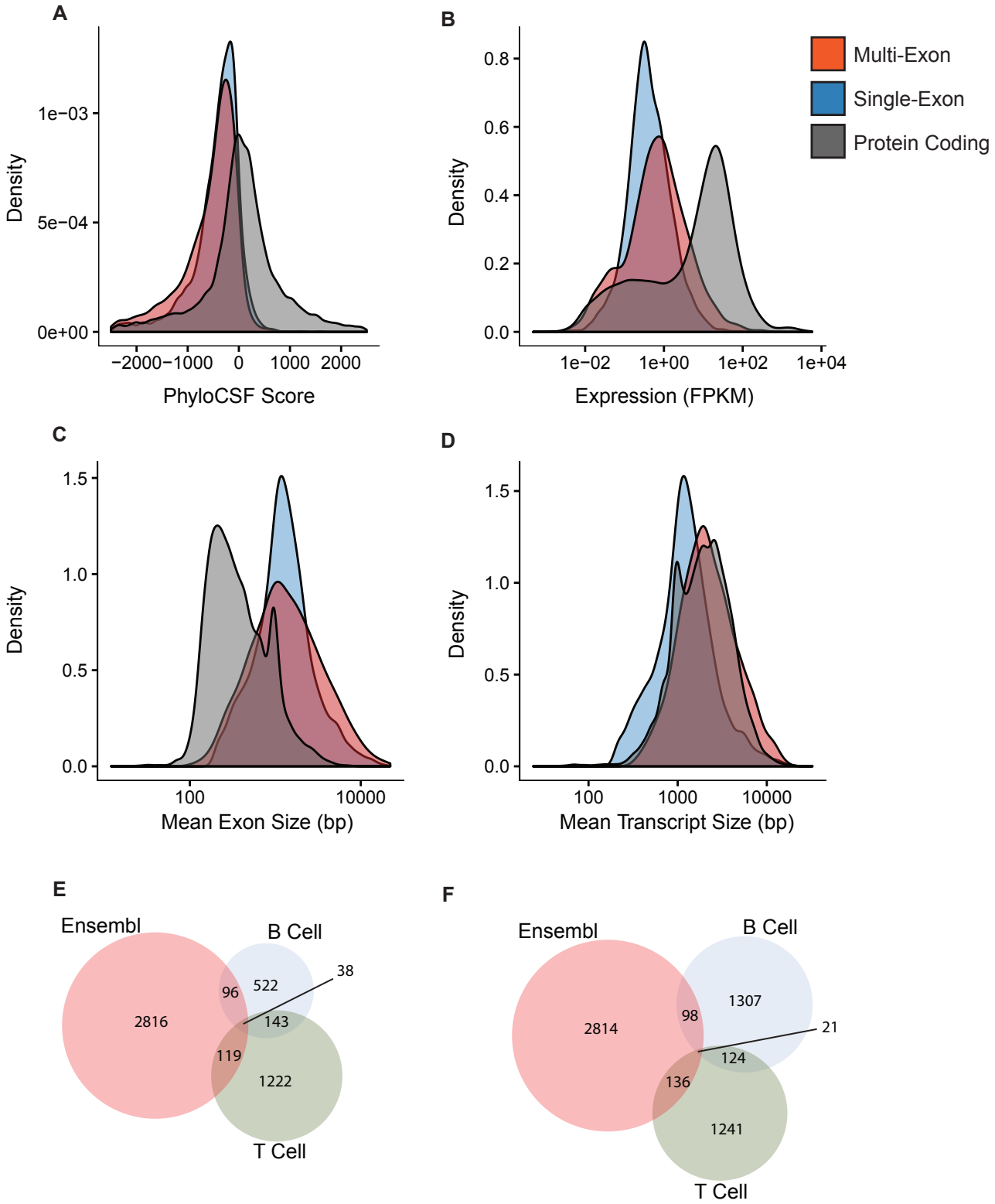
Supplemental Figure 3



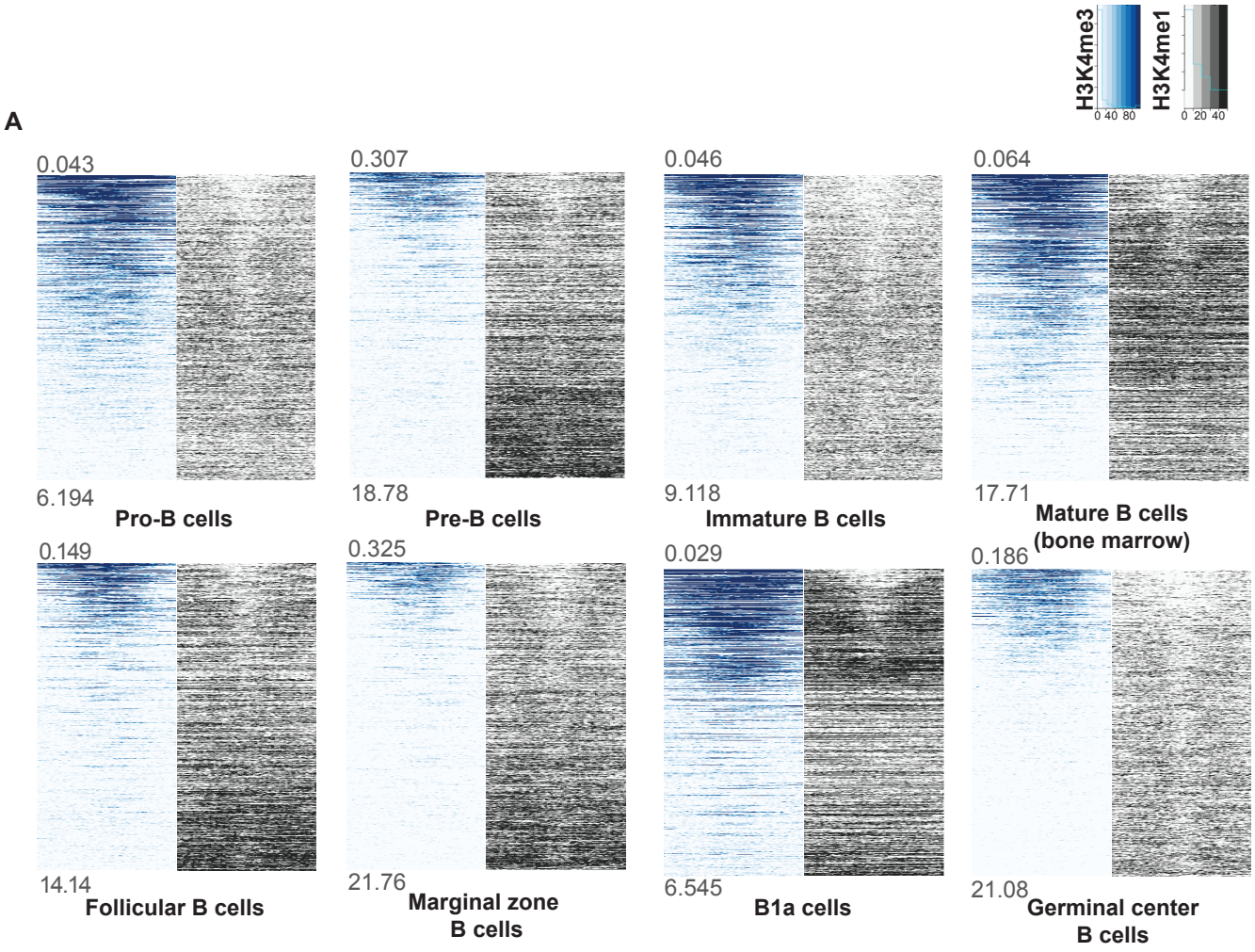
Supplemental Figure 4



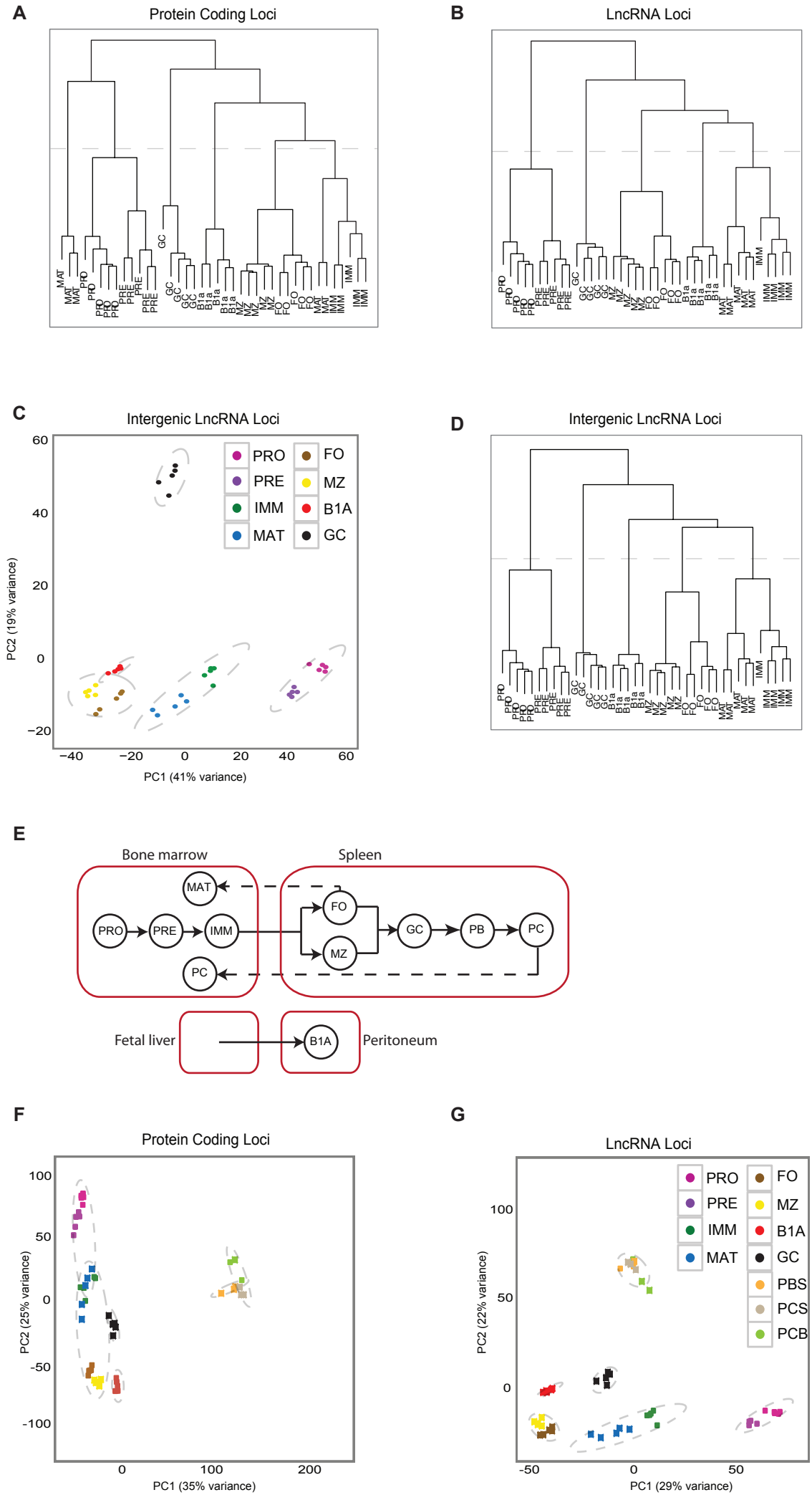
Supplemental Figure 5



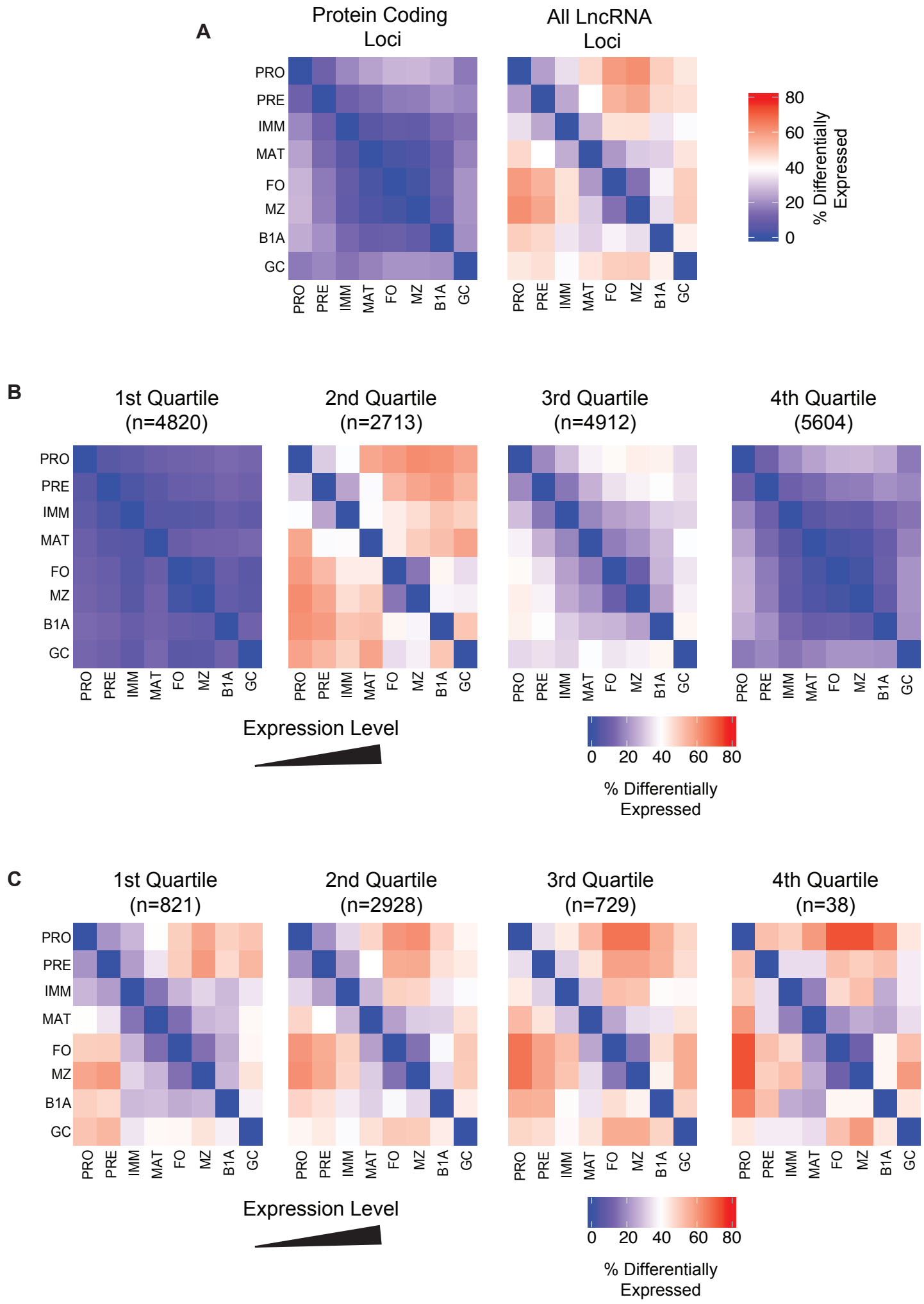
Supplemental Figure 6



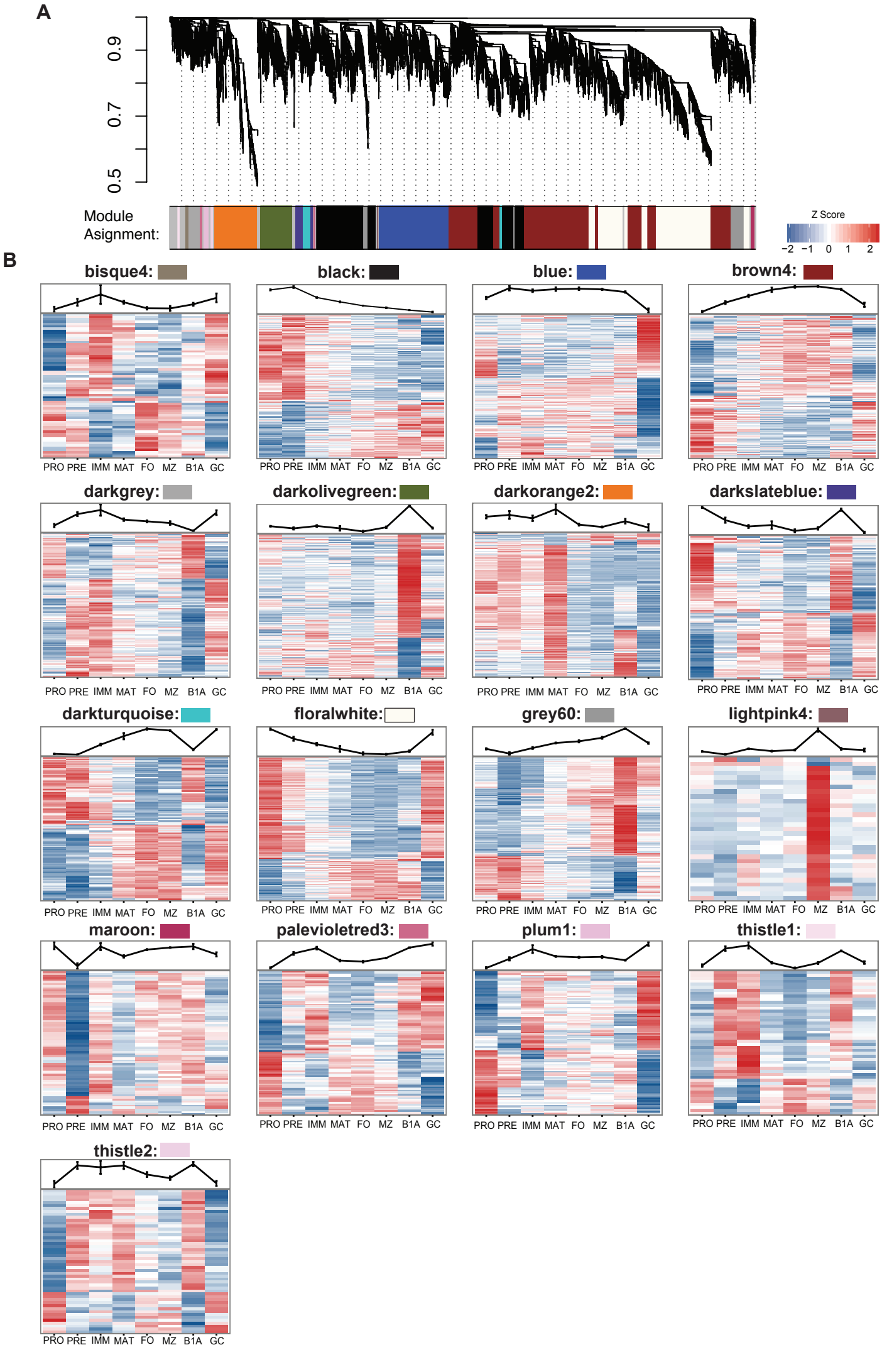
Supplemental Figure 7



Supplemental Figure 8



Supplemental Figure 9



Supplemental Figure 10