Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation.

LiQi Li[1]*, Johannes Freudenberg[2]*, Kairong Cui[3], Ryan Dale[4], Sang-Hyun Song[4], Ann Dean[4], Keji Zhao[3], Raja Jothi[2]#, and Paul E. Love[1]#

SUPPLEMENTAL METHODS

Template preparation for ChIP-Seq analysis

Template preparation was performed essentially as described [1]. Total bone marrow cells were isolated from C57BL/6 mice by flushing bone marrows with serum free medium (Mediatech, Inc.). Cells were crosslinked with formaldehyde, and genomic DNA was fragmented by sonication and subjected to chromatin immunoprecipitation (ChIP) with antibodies specific for Ldb1 (Santa Cruz, sc-11198), Tal1 (Santa Cruz, sc-12984), Gata1 (Abcam, ab11852) or control IgG. DNA sequencing was performed as described [2]. Briefly, approximately 200 ng of ChIP DNA was end-repaired using the Epicentre DNA END-Repair kit followed by treatment with Taq polymerase to generate a protruding 3-prime A base used for adaptor ligation. Following ligation of a pair of Solexa adaptors to the repaired ends, the ChIP DNA was amplified using adaptor primers for 17 cycles and the fragments of approximately 220 bp (mononucleosome + adaptors) were isolated from an agarose gel. The purified DNA was used directly for cluster generation

and sequence analysis using the Solexa 1G Genome Analyzer according to the manufacturers protocols.

ChIP-Seq data analysis

Sequences were aligned to the reference genome (mouse NCBI36/mm8 assembly) using Bowtie 0.12.2 [3] with default settings, aligning all 36 bases allowing for up to 2 mismatches. Only those sequence reads that mapped to unique genomic locations were retained for further analysis. Next, SISSRs [4,5] with default settings was used to identify binding sites (or peak-calling) for Ldb1, Tal1, Gata1. Klf1 ChIP-Seq data from primary erythroid cells were obtained from GSE20478 [6]. Genes were defined using Entrez gene IDs in the union of the UCSC Genome Browser's known Genes dataset and refGene table (mouse NCBI36/mm8 assembly) [7]. If a gene ID mapped to multiple Refseq transcript IDs (i.e. multiple entries in the joint table), the ID of the longest transcript was selected. Genes were considered bound/targeted by a transcription factor if they have at least one transcription factor binding site overlapped within 5kb upstream of the transcription start site (TSS) or within the gene body. Conversely, if multiple genes were in the vicinity of a transcription factor binding site, the closest gene was considered to be associated with the binding site. Binding sites for two transcription factors (identified using the SISSRs peak-finder) [4] were defined as overlapping if the centers of the binding sites are within 200 nucleotides. Using this definition, an "Ldb1-complex site" is an Ldb1 binding site that overlaps with both a Gata1 binding site and a Tal1 binding site. A "Gata1-

only" site is a Gata1 binding site that does not overlap with an Ldb1 site nor with a Tal1 binding site. Fold enrichment of ChIP DNA over input DNA was computed using SISSRs [4]. To facilitate comparison and visualization of our ChIP-Seq data with those published by other groups, we re-mapped out sequence reads to the mm9 genome assembly. The UCSC genome browser [8] was used to display and capture PDF shots of ChIP-Seq tag densities.

Identification of erythroid genes

Erythroid-specific "fingerprint" genes were obtained from Chambers et al. [9]. The erythroid gene list was generated by performing an NCBI Gene search with the following search terms: Erythro* OR Mega* OR Hemoglobin OR Heme OR Red blood cell OR platelet OR Mast cell NOT HSC NOT leukocyte NOT Macrophage NOT Monocyte NOT Dendritic cell NOT Lymphocyte NOT Thymo* NOT Natural killer cell NOT Granulocyte NOT Neutrophil. The resulting list of 652 genes was then screened by performing PubMed and BioGPS searches for each gene. Only genes that were highly expressed or exhibited restricted expression in erythroid lineage cells and/or that had been shown to expressed in or functional in the erythroid lineage in publication(s) were retained. The resulting list of 533 genes is included in Supplemental Table 3.

Microarray gene expression data analysis

Microarray experiments were performed as described [10,11]. Using applicable R/Bioconductor packages (affy, genefilter, limma) [12] raw microarray data files

("CEL files", Affymetrix array type Mouse 430 2.0) were processed applying the RMA [13] methodology as well as Entrez gene based custom chip definition files (CDFs), version 13, [14] http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/genomic_curated_CDF.asp. This approach results in a single expression intensity measure per sample per Entrez gene ID per sample. All subsequent analyses were carried out on the $\log_2$ scale. An intensity filter was applied to discard probesets whose intensity value was within the bottom 25% in each array. This filter removes approximately 20% of the probesets. To determine differentially expressed genes, a family-wise moderated *t*-test (Ldb1 KD vs. control) [15] was performed followed by a multiple testing correction procedure to control the false discovery rate (FDR) [16]. Genes were considered differentially expressed if they had an FDR of 0.1 or less and were at least 1.5 fold up- or down-regulated unless otherwise stated (e.g., FDR cutoffs of 0.01 were used for the designation of down-regulated genes annotated in Table 1 and Supplemental Table 5). Cluster analysis of per-gene normalized expression levels was performed using the CLEAN software package [17] where genes were normalized by subtracting the per-gene arithmetic mean of the respective control/untreated samples. Hematopoietic fingerprint gene lists [9] were mapped to Entrez gene IDs using an organism annotation package from Bioconductor [12]. Statistical significance of enrichment of fingerprint genes in a gene list (e.g. list of Ldb1 bound genes) was determined using the hypergeometric test. Entrez gene IDs were used to map mRNA expression data ("probesets") and genes in the vicinity of ChIP-Seq

defined transcription factor binding sites (as defined above). If multiple transcription factor binding sites were near a gene with a probeset on the array, each was mapped to that probeset. Functional enrichment analysis was performed using the CLEAN software package [17]. The Gata1 microarray data were from GSE18042 [18].

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1. Representative genes with non-Ldb1/Tal-associated "Gata1-only" binding sites. UCSC Genome Browser shots of data from Ldb1, Gata1, or Tal1 ChIP-Seq runs are shown for: A) *Ddx24*, B) *Hscb*, C) *Mfsd2b*, D) *Diap3*. Y axis shows number of sequence reads. Mammalian conservation tracks are shown at the bottom.

Supplemental Figure 2. Representative genes with Ldb1-complex binding sites within the first intron. A) *Alad*, B) *Alas2*, C) *Urod*, D) *Sox6*, E) *Ank1*, F) *Tmod1*. UCSC Genome Browser shots of data from Ldb1, Gata1, or Tal1 ChIP-Seq runs. Y axis shows number of sequence reads. Mammalian conservation tracks are shown at the bottom.

Supplemental Figure 3. Ldb1-complex binding sites at or near erythroid genes. A) *Nfe2*, B) *Epb4.1*, C) *Spna1*, D) *Car2*, E) *Hebp1*, F) *Prdx2*. UCSC Genome

Browser shots of data from Ldb1, Gata1, or Tal1 ChIP-Seq runs. Y axis shows number of sequence reads. Mammalian conservation tracks are shown at the bottom.

Supplemental Figure 4. Ldb1-complex binding sites at known or presumed erythroid enhancers. A) *Slc25a37 (-20, -34, -37.5)*, B) *Lmo2 (-75)*, C) *Tal1 (+40)*, D) *Gata1 (-3.5, -25)*, E) *Cbfa2t3 (-22)*, F) *Pabpc1 (presumed)*. Tracks for H3K27ac, H3K4me1 and p300 in MEL cells and mammalian conservation tracks are shown at the bottom.

Supplemental Figure 5. A) Conserved motifs identified by de novo motif search (MEME) within Ldb1, Tal1, Gata1, Ldb1/Tal1, Ldb1/Gata1, Tal1/Gata1 or Ldb1/Tal1/Gata1 ChIP-Seq peaks. B) Known motifs identified within Ldb1/Tal1/Gata1 ChIP-Seq peaks by MAST search. Percentage of peaks that contain each motif are shown.

Supplemental Figure 6. A) Overlap of the entire set of 840 Klf1 binding sites from Tallack et al [6] with Ldb1-complex (Ldb1/Tal1/Gata1) binding sites or with non-Ldb1/Tal1-associated Gata1 binding sites. B) Overlap of Ldb1-complex sites or non-Ldb1/Tal1-associated Gata1 binding sites with PU.1 ChIP-Seq binding sites from Wontalal et al [19]. C) Genomic distribution of Ldb1-complex-binding sites (left) and Klf1 binding sites (right).

Supplemental Figure 7. Representative genes with overlapping Ldb1-complex and Klf1 binding sites. A)*Slc4a1*, B) *Nfe2*, C) *Ank1*, D) *Pcx*, E) *Epb4.9*, F) *Hbb*. UCSC Genome Browser shots of data from Ldb1, Gata1, Tal1 or Klf1 ChIP-Seq runs. Y axis shows number of sequence reads. Tracks for H3K27ac, H3K4me1 and p300 in MEL cells and mammalian conservation tracks are shown at the bottom.

Supplemental Figure 8. Ldb1-complexes directly regulate the expression of genes in induced MEL cells. Stable clones of MEL cells expressing Ldb1 shRNA or control shRNA were treated with 1.5% DMSO to induce erythroid differentiation. Total RNA was isolated and gene expression was assayed by microarray. A) Highly expressed genes are the most strongly down-regulated in Ldb1 Knockdown (KD) MEL cells. Genes bound by Ldb1-complexes are shown in yellow, unbound genes are shown in gray. Zero line is indicated (dotted line). B) Genes that are most strongly induced in differentiated control MEL cells are the most strongly down-regulated in Ldb1 KD MEL cells. Genes bound by Ldb1-complex are shown in yellow, unbound genes are shown in gray. Zero lines are indicated (dotted lines). X-axis histograms show stronger induction of total Ldb1-complex-bound genes relative to unbound genes ($p<1.0e-32$). Y axis histograms show stronger knockdown of total Ldb1-complex-bound genes relative to all unbound genes ($p<4.4e-15$). Significance for all comparisons was calculated by Mann-Whitney U. C) Confirmation of transcriptional down-regulation of erythroid

gene expression in MEL cells stably transfected with Ldb1 shRNA by quantitative RT-PCR. Cont (control MEL cells), KD (Ldb1 Knockdown MEL cells). *p<.05, **p<.01. D. Log2 fold change in expression of Ldb1-complex bound (yellow) non-Ldb1-complex associated Gata1 (Gata-only) bound (red) and Unbound genes (gray) in induced vs uninduced MEL cells. Unbound genes are those lacking Ldb1 complexes or Gata1 binding sites. Only significantly down-regulated (Log2 fold <-1.5) or up-regulated (Log2 fold >1.5) genes are shown. Ldb1 complex bound genes are mainly up-regulated in terminally differentiated MEL cells. In contrast, similar numbers of non-Ldb1-complex associated Gata1 bound genes are up-regulated and down-regulated in differentiated MEL cells.

SUPPLEMENTAL REFERENCES

1. Li L, Jothi R, Cui K, et al. Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat Immunol*. 2011;12(2):129-136.
2. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007;129(4):823-837.
3. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
4. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from 2-Seq data. *Nucleic Acids Res*. 2008;36(16):5221-5231.
5. Narlikar L, Jothi R. ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder. *Methods Mol Biol*. 2012;802:305-322.
6. Tallack MR, Whitington T, Yuen WS, et al. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res*. 2010;20(8):1052-1063.
7. Fujita PA, Rhead B, Zweig AS, et al. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*. 2011;39(Database issue):D876-882.

8. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006.
9. Chambers SM, Boles NC, Lin KY, et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell*. 2007;1(5):578-591.
10. Song SH, Kim A, Dale R, Dean A. Ldb1 regulates carbonic anhydrase 1 during erythroid differentiation. *Biochim Biophys Acta*. 2012;1819(8):885-891.
11. Song SH, Hou C, Dean A. A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. *Mol Cell*. 2007;28(5):810-822.
12. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.
13. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185-193.
14. Dai M, Wang P, Boyd AD, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 2005;33(20):e175.
15. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.
16. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001;125(1-2):279-284.
17. Freudenberg JM, Joshi VK, Hu Z, Medvedovic M. CLEAN: CLustering Enrichment ANalysis. *BMC Bioinformatics*. 2009;10:234.
18. Fujiwara T, O'Geen H, Keles S, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell*. 2009;36(4):667-681.
19. Wontakal SN, Guo X, Will B, et al. A large gene network in immature erythroid cells is controlled by the myeloid and B cell transcriptional regulator PU.1. *PLoS Genet*. 2011;7(6):e1001392.

Supplemental Table 2. Gene ontology analysis shows an enrichment in erythopoiesis-relevant categories for genes with Ldb1-complex binding sites.

ID	Description	p-value	FDR
GO:0006778	porphyrin metabolic process	6.2E-14	1.4E-10
GO:0033013	tetrapyrrole metabolic process	6.2E-14	1.4E-10
GO:0006779	porphyrin biosynthetic process	2.4E-12	2.7E-09
GO:0033014	tetrapyrrole biosynthetic process	2.4E-12	2.7E-09
GO:0042168	heme metabolic process	1.3E-10	1.2E-07
GO:0006783	heme biosynthetic process	4.7E-10	3.6E-07
GO:0042440	pigment metabolic process	2.5E-08	1.6E-05
GO:0051188	cofactor biosynthetic process	4.3E-08	2.5E-05
GO:0046148	pigment biosynthetic process	1.4E-07	7.0E-05
GO:0051186	cofactor metabolic process	1.3E-06	6.0E-04
KEGG:mmu00860	Porphyrin and chlorophyll metabolism	2.7E-06	5.7E-04
GO:0020027	hemoglobin metabolic process	1.2E-05	5.0E-03
GO:0030863	cortical cytoskeleton	1.7E-05	6.5E-03
GO:0044271	cellular nitrogen compound biosynthetic process	3.2E-05	0.011
GO:0046483	heterocycle metabolic process	3.2E-05	0.011
GO:0018130	heterocycle biosynthetic process	4.2E-05	0.013
GO:0030218	erythrocyte differentiation	1.0E-04	0.029
GO:0034101	erythrocyte homeostasis	1.7E-04	0.045
GO:0015669	gas transport	2.4E-04	0.059
GO:0005372	water transmembrane transporter activity	2.4E-04	0.059
GO:0044448	cell cortex part	2.8E-04	0.065
GO:0032012	regulation of ARF protein signal transduction	3.9E-04	0.086
GO:0055072	iron ion homeostasis	4.5E-04	0.094

Supplemental Table 4. Binding of Ldb1-complexes at known erythroid enhancers.

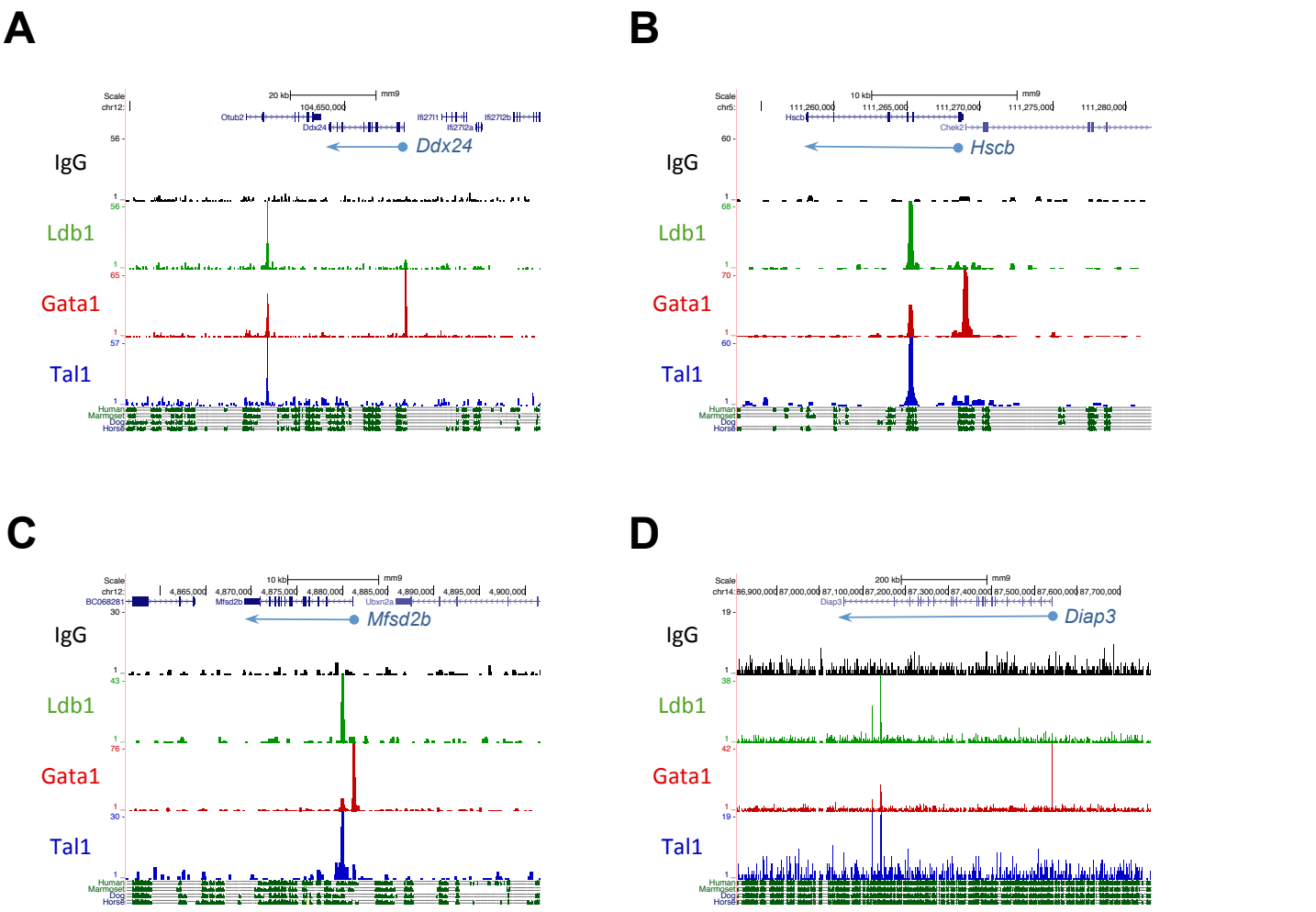
Gene with enhancer	Reference
Runx1 (+23)	Nottingham (2007), Blood 110:4188-97; Wilson (2009), Blood 113:5456-65.
Gata1 (-3.5; -25)	McDevitt (1997), PNAS 94:7976-81; Onodera (1997), PNAS 94:4487-92; Drissen (2010) Blood, 115:3463-71.
Klf1 (-1)	Chen (1998), JBC 273:25031-40.
MARE	Gourdon (1995), Blood 86:766-75.
Beta-globin LCR	Hardison (1997), Gene 205:73-94.
Cbfa2t3h (-22)	Wilson (2009), Blood 113:5456-65.
E2F2 (intron 1)	Tallack (2009), JBC 284:20966-74.
E2F4 (intron 5)	Tallack (2010), Genome Res. 20:1052-63.
Slc25a37 (-20, -34, -37.5)	Amigo (2011), MCB 31:1344-56.
Alas2 (intron 8)	Surinya (1998), JBC 273:16798-809.
Tal1 (+40)	Ogilvy (2007), MCB 27:7206-19.
Gfi1b (+16)	Wilson (2009), Blood 113:5456-65.
Fog (+2.7)	Wilson (2009), Blood 113:5456-65.
Tox2 (+4)	Wilson (2009), Blood 113:5456-65.
Nfe2 (-7)	Wilson (2009), Blood 113:5456-65.
Cebpe (+6)	Wilson (2009), Blood 113:5456-65.
Lmo2 (-75)	Landry (2009), Blood 113:5783-92.
Sfpi1 (-14)	Li (2001), Blood 98:2958-65.
Klf2 (-51)	Wilson (2009), Blood 113:5456-65.
Myb (-36, -61, -109)	Stadhouders (2012), EMBO J 31:986-999.

Supplemental Table 5. Cellular response genes bound by Ldb1-complexes (± 5kb of gene body).

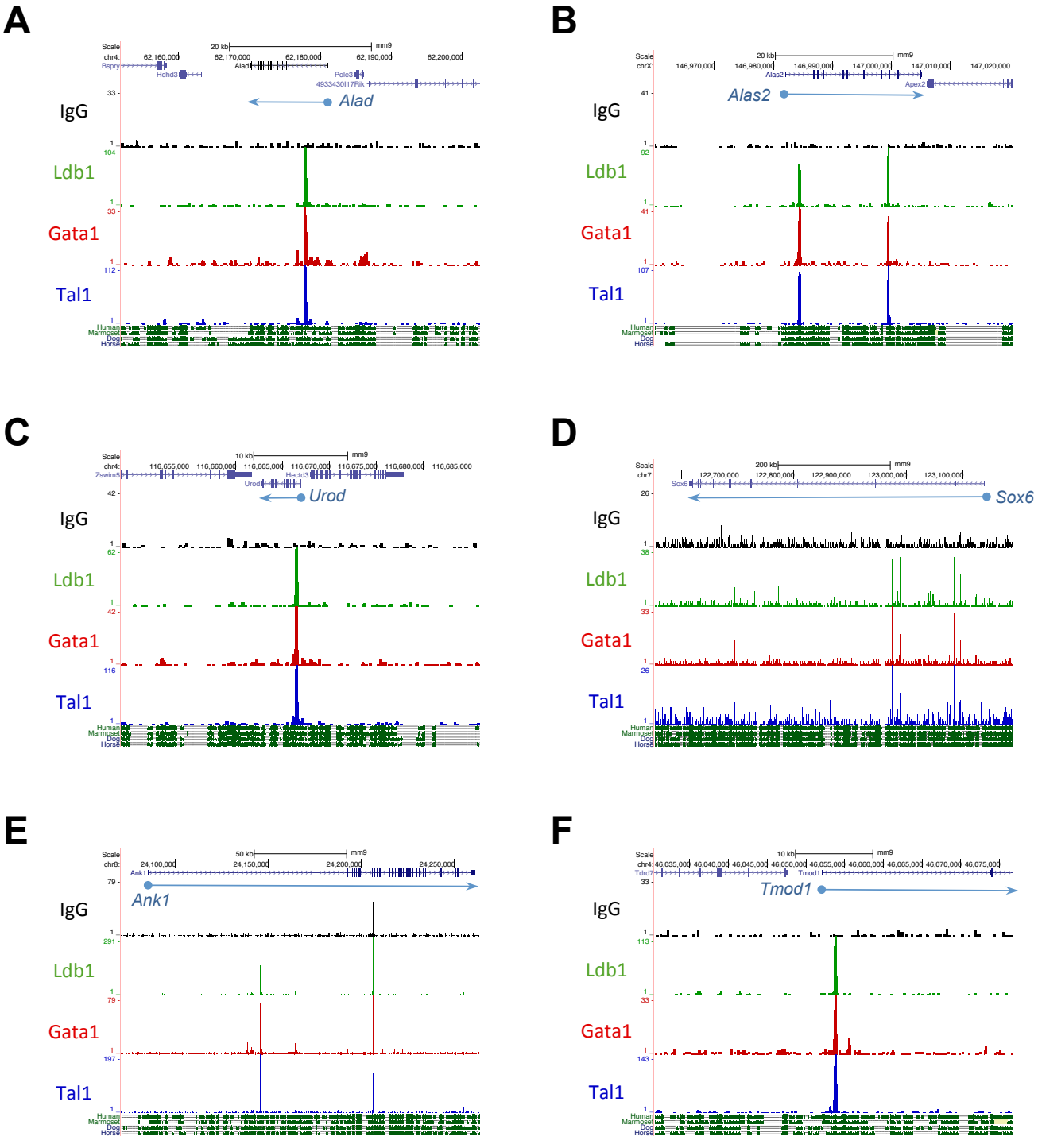
Autophagy/ Mitophagy/ Enucleation	**Atg4d, Atg7, Atg10, Birc5*, Bnip3l*, Dapk2*, Diap3*, Foxo3, Map1lc3b*, Mapk14, Myh10, Mxi1*, Rb1***
Survival/ Apoptosis	**Bcl2l1, Bcl2l11, Bcl2l13*, Bid*, Bnip1, Dapk2*, Epor, Sox6*, Trp53**
Stress Response	**Cldn13*, Gpx1*, Nfe2, Nfe2l1, Nfe2l2, Prdx2*, Ptk2, Ppp1r15a/Myd116***

*** Significantly down-regulated (FDR<0.01) in Ldb1 shRNA MEL cells (see text).**

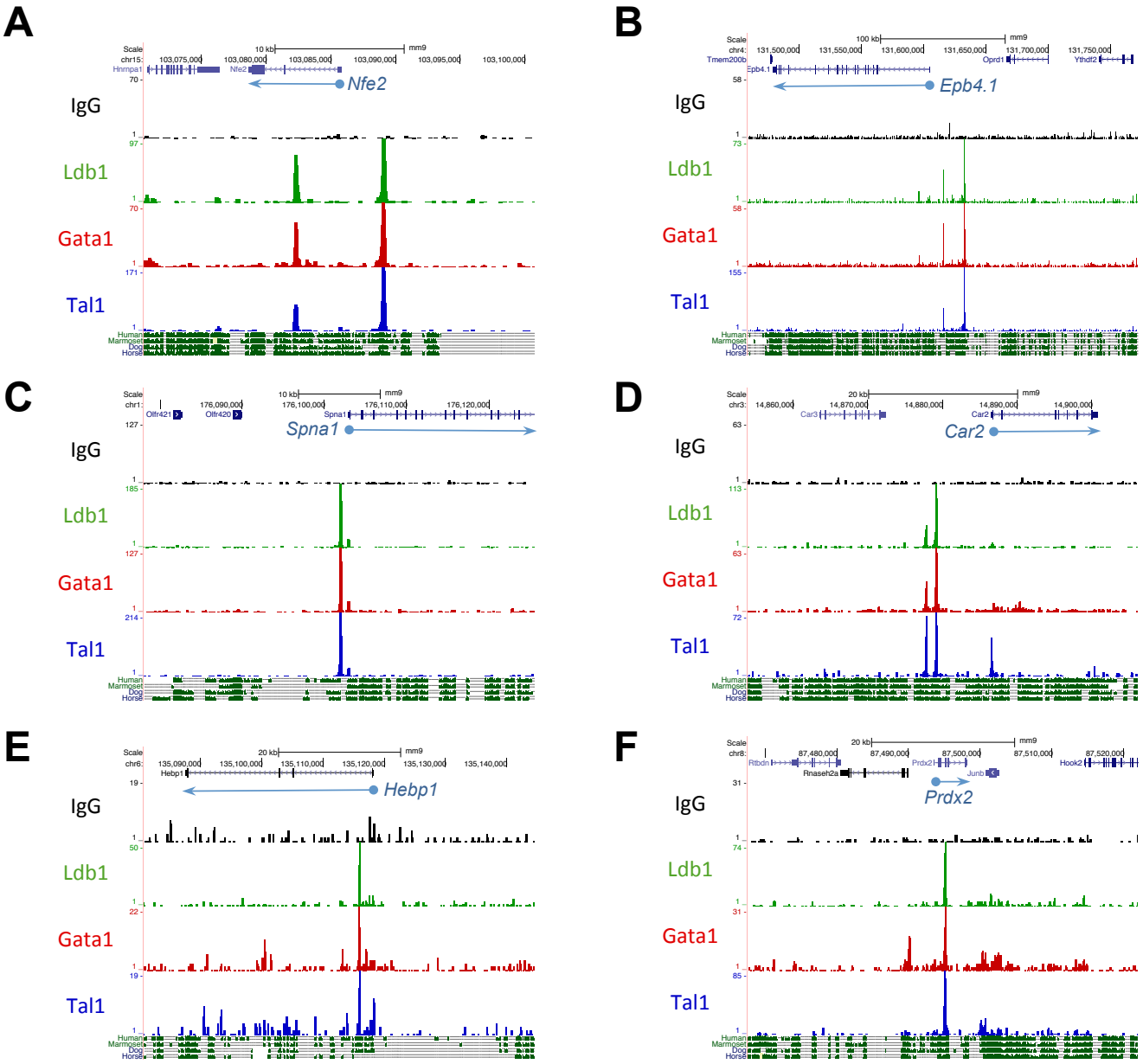
Supplemental Figure 1 Li et al.



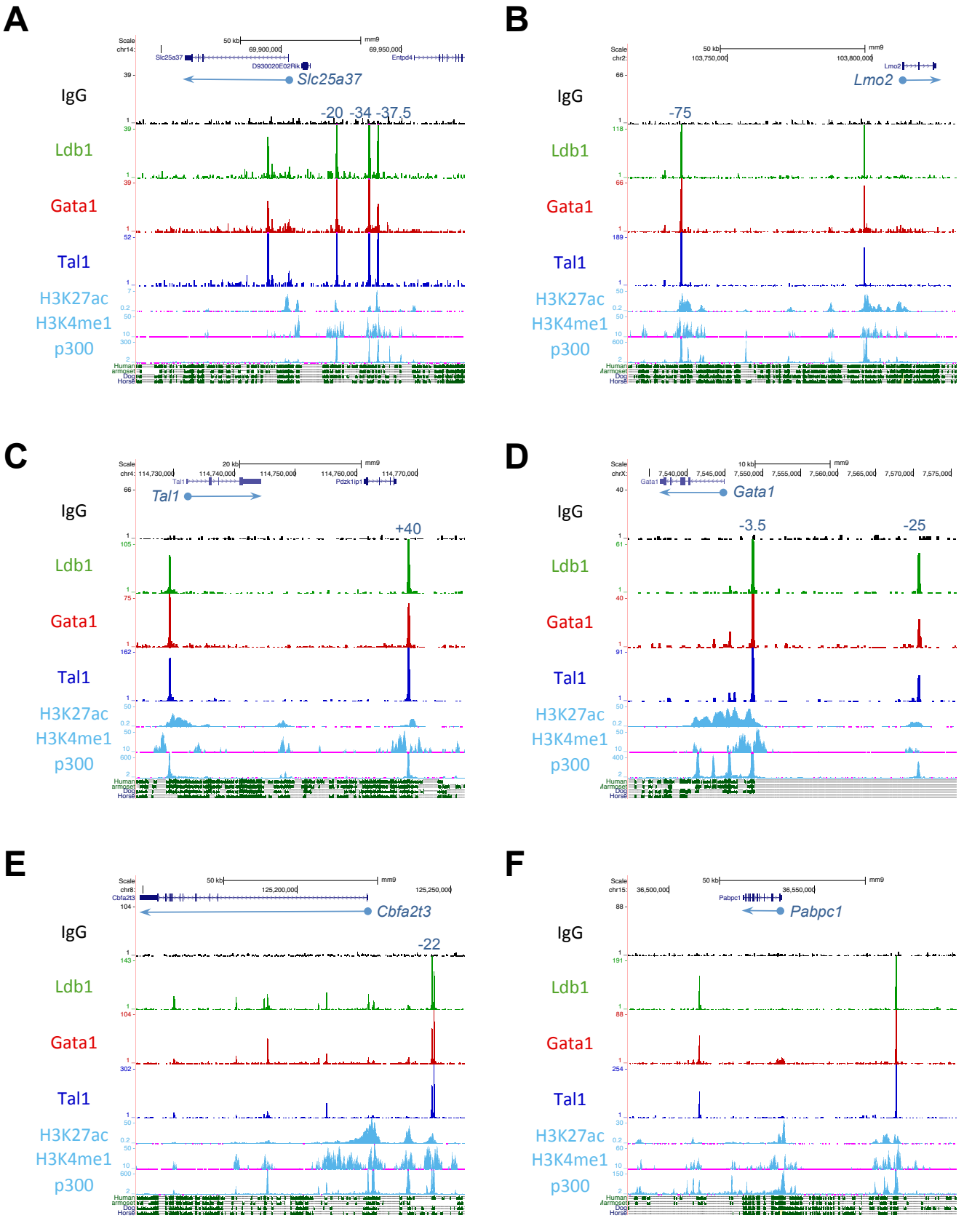
Supplemental Figure 2 Li et al.



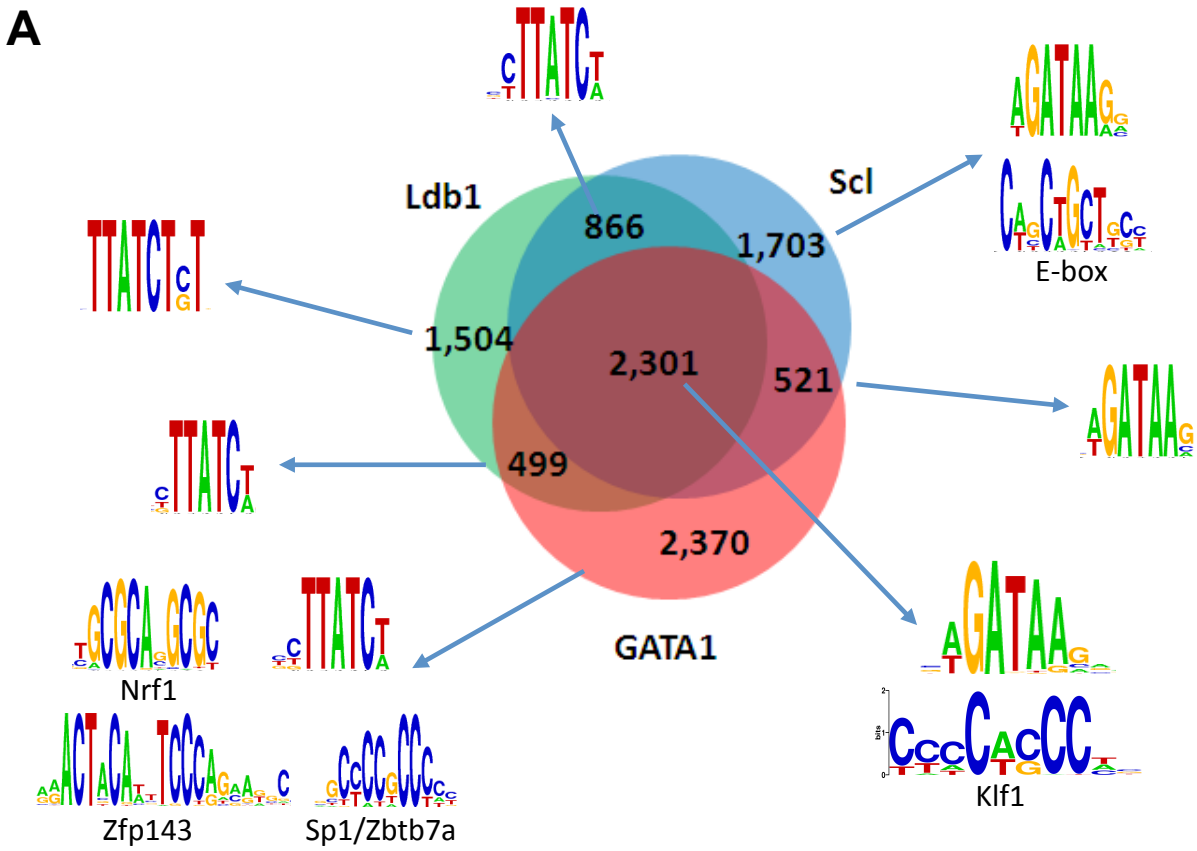
Supplemental Figure 3 Li et al.



Supplemental Figure 4 Li et al.



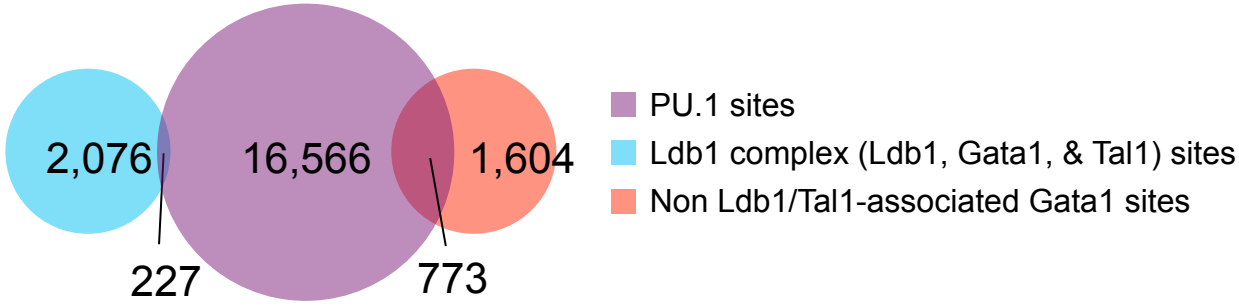
Supplemental Figure 5 Li et al.



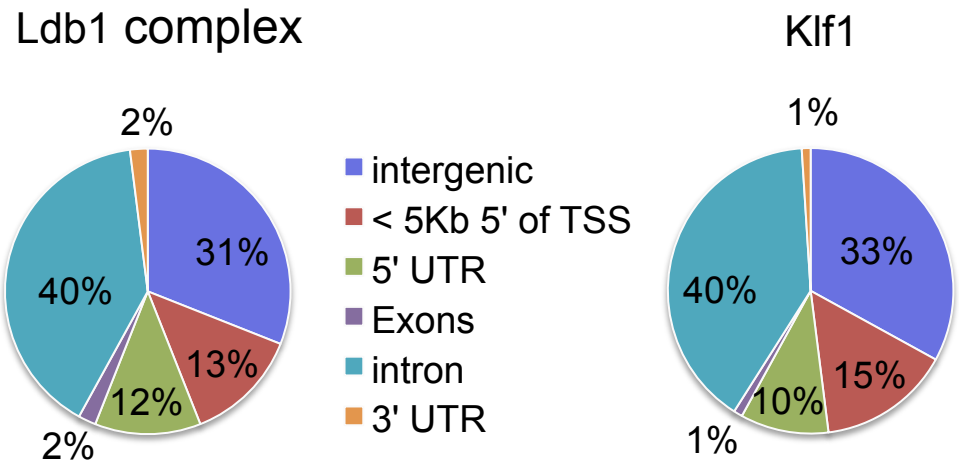
Supplemental Figure 6 Li et al.

A

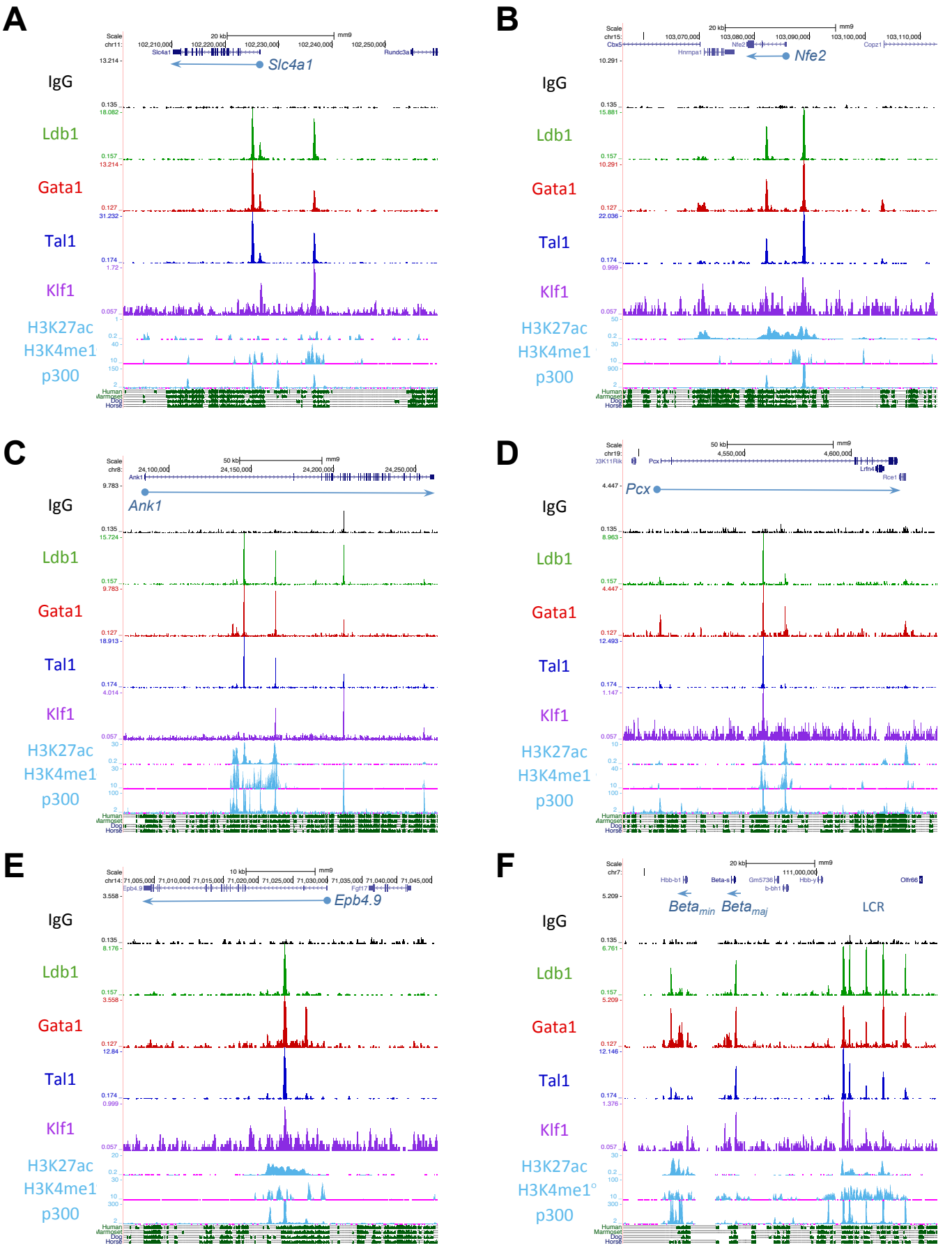
B



C



Supplemental Figure 7 Li et al.



Supplemental Figure 8 Li et al.