Conserved DNA-binding motif loci reflect functional transcription factor binding sites

Akihiro Kuno (akihiro.kuno@gmail.com)

Department of Anatomy and Embryology, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575

https://orcid.org/0000-0002-4674-6882

Satoru Takahashi

Tsukuba Daigaku Igaku Iryokei

Research article

Keywords: Conserved sequence, DNA-binding motif, Regulatory DNA element, Transcription factor

Posted Date: October 9th, 2019

DOI: https://doi.org/10.21203/rs.2.15866/v1

Abstract

Background Transcription factors (TF) regulate gene expression by binding to specific DNA elements called DNA-binding motifs, or motifs. The motifs at a certain genomic locus can offer clues for predicting TFs that control target genes. However, since motifs are short nucleotide sequences and many TFs share similar motifs, they are distributed across the whole genome, making it difficult to investigate candidate functional DNA elements. Because previous studies have reported that TF binding sites are highly conserved among mammals, we focused on the conserved motif loci between humans and mice to identify functional DNA elements. Results Our results showed that conserved motif loci overlapping considerably with TF binding sites at both promoter and intergenic regions compared to nonconserved motif loci. In addition, conserved motif loci were significantly enriched in enhancer regions and enabled us to predict GATA4 binding to the heart-specific enhancer. Moreover, the integration of ChIP-seq and RNA-seq glioblastoma data revealed that genes with ASCL1 binding at the conserved ASCL1-related motifs were associated with NOTCH signaling, which is a critical pathway in glioblastoma. Conclusions These results suggest that conserved motif loci can reflect gene regulatory elements and can be utilized to predict candidate TFs that regulate genes of interest. The conserved motif data are publicly available at https://osf.io/jhfnb/?view_only=1c82bc9963ee45b799be42512abb08d1.

Background

Transcription factors (TFs) are a large family of DNA-binding proteins that play critical roles in regulating gene transcription. TFs typically contain a DNA-binding domain (DBD) and control the activity of regulatory DNA elements such as promoters and enhancers through their binding. One of the important functions of TFs is recognizing specific DNA elements, which are called DNA-binding motifs, or motifs. Because the structure of the DBD is divergent among TFs, each TF has a particular preference of motifs to a greater or lesser degree[1].

Recent technological developments have allowed us to characterize the specificity of motifs in each TF using high-throughput methods. The analysis by protein-binding microarrays (PBMs)[2], high-throughput SELEX (HT-SELEX)[3] and ChIP-seq[4] revealed high-resolution motifs in a large proportion of mammalian TFs[5]. These results have provided a high-quality motif catalog, and several well-maintained databases are available for obtaining motif information[3, 6–9]. In addition, by applying the motif information, the prediction of motif sites across the whole genome is accessible with databases[6] or can be performed by software[10].

Although these resources for motif sites throughout the whole genome are undoubtedly important for investigating DNA elements that can regulate transcription, it is difficult to infer functional motif loci from the predicted motif sites. Because the motifs are extremely short (typically ten nucleotides in length)[11] and several motifs allow flexibility to match DNA sequences, the predicted motif sites are almost ubiquitously distributed throughout the whole genome. Moreover, many TFs recognize similar DNA sequences because they share similar DBD structures, which makes motif sites too redundant and enormous to handle and visualize the data.

Therefore, we aimed to extract candidates for functional motif loci. Previous studies have reported that the gene expression patterns of each tissue are highly conserved in several vertebrates[12, 13] and that the primary TF binding sites are also stable among mammalian species[14, 15]. These results suggest that a major proportion of regulatory DNA elements is evolutionarily conserved. Thus, we focused on the motifs localized at DNA sequences that are conserved between humans and mice. Here, we demonstrated that the conserved motif loci showed a higher overlap with actual TF binding sites and that they were enriched in enhancer loci compared to nonconserved motif loci. Furthermore, the genes near TF binding sites on the conserved motif loci reflected a critical pathway in human primary glioblastoma. These results suggested that the conserved motif loci can capture functional DNA elements.

Results And Discussion

Conserved motif loci captured actual TF binding sites

We first extracted conserved DNA-binding motifs from the JASPAR UCSC track hub[6] (see details in Materials and Methods). To visualize the number of motif loci in a certain genomic region, we first aligned motif sites from raw data to conserved motif loci in the NANOG promoter region in the human and mouse genomes (Fig. 1). Fig. 1A shows a narrow segment that is 60 nucleotides in length in the NANOG promoter region. Original data from the JASPAR database contained a bewildering amount of motif loci. The number of motif loci significantly decreased after filtration using a motif matching score greater than 400. Because the filtered motif loci included redundant motifs that localized to the same genomic loci, we next merged these motif loci into one locus according to the TF family[16]. For example, the nine motifs named Dmbx1, GSC, Pitx1, PITX3, OTX1, OTX2 and RHOXF1 were located at the same genomic locus in the human NANOG promoter (chr12:7,941,908-7,941,924 in hg19) because these TFs share a similar DBD and belong to the same TF family[16]. Therefore, these nine motif loci were summarized into one region named “Paired-related HD factors” referring to the TFClass[16]. Finally, we aligned sequences from human and mouse and acquired conserved motif loci between the two species. The conserved motif loci included POU domain factors, SOX-related factors, and paired-related HD factors in the NANOG promoter (Fig. 1A). Furthermore, Fig. 1B displays a wider range of NANOG promoters of 1,200 nucleotides in length. The conserved motif loci specifically overlapped with actual POU5F1, SOX2, and OTX2 binding sites (Fig. 1B). These results indicated that the conserved motif loci can reflect actual TF binding sites.

Importantly, the original data on genome-wide binding site predictions contain more than ten billion motif loci, and the file size is very large (hg19: 65 GB in gzip compression), which makes the data difficult to handle and visualize. With the procedures, we ultimately extracted 9,281,940 conserved motif loci (0.09% of the original data) in the human genome. The conserved motif loci enable easy data processing and visualization using the genome browser because of its reduced file size (hg19: 305MB in gzip compression) (Table 1).

Conserved motif loci were localized in TSS sites

We next evaluated the genomic distribution of conserved and nonconserved motif loci. Although both conserved and nonconserved motif loci were frequently located near TSS regions, conserved motif loci showed sharper localization at TSS sites compared with nonconserved motif loci (Supplementary Fig. S1A). Moreover, the distribution of genomic features displayed conserved motif loci that were highly enriched in promoter-TSS and exon sites (conserved: 3.7% and 7.0%, nonconserved: 1.2% and 0.7%, respectively) (Supplementary Fig. S1B). The results indicated that conserved motif loci showed higher enrichment in TSS sites compared to nonconserved motif loci. In addition, the higher motif matching score was correlated with the higher localization of TSS and exon sites, which can reflect the evolutionarily conserved sequence at these loci. (Supplementary Fig. S1B).

Conserved motif loci were enriched in TF binding sites and histone marked regions at both promoter and intergenic regions

To evaluate how conserved motif loci are enriched in regulatory DNA elements compared to nonconserved motif loci, we counted the percentage of motif sites that overlapped with TF binding sites and histone marks. We utilized publicly available ChIP-seq peak data for TFs, H3K27ac (active promoter and enhancer mark) and H3K4me1 (active enhancer mark), which were downloaded from ChIP-Atlas[17].

We found that the conserved motif loci showed significantly higher overlap with TF binding sites in every genomic feature compared to nonconserved motif loci (all regions; conserved: 10.5%, nonconserved: 3.8%) (Fig. 2A). Interestingly, in the intergenic region, conserved motif loci also showed significant enrichment compared with nonconserved motif loci (Intergenic; conserved: 8.2%, nonconserved: 3.1%), as well as in the promoter regions (Fig. 2A). Likewise, the conserved motif loci displayed enrichment at the active histone marks H3K27ac (all regions; conserved: 36.9%, nonconserved: 23.1%) (Fig. 2B) and H3K4me1 (all regions; conserved: 31.6%, nonconserved: 19.0%), which indicated that conserved motif loci were preferentially activated (Fig. 2C).

Since the cooperation of multiple TFs is important for triggering target gene transcription[18], we next hypothesized that genomic loci that accumulated with multiple conserved motif loci are related to actual TF binding. The results showed the significant dependency of the number of motif accumulations and TF bindings in conserved motif loci (number of accumulations = 1: 46.1%; number of accumulations ≥ 6: 67.9%). In contrast, nonconserved motif loci did not clearly show the dependency of motif accumulation (number of accumulations = 1: 34.1%; number of accumulation ≥ 6: 37.0%) (Fig. 2D). Furthermore, the

accumulation of conserved motif loci was significantly associated with H3K27ac and H3K4me1 marks. On the other hand, nonconserved motif loci did not show dependency (Fig. 2E and F).

Next, viewing of the SOX2 genomic locus revealed that pluripotent stem cell (PSC)-related TFs, such as NANOG, POU5F1 and SOX2, actually bound the corresponding conserved motif loci (NK-related, POU domain, and SOX-related motif sites, respectively) at the SOX2 promoter and distal regions in PSCs. Moreover, the PSC-related TF binding sites localized at accumulated motif loci (Fig. 2D). The results demonstrated that the conserved motif loci were enriched in functional DNA elements at both the promoter and distal regulatory regions.

Conserved motif loci were enriched in enhancer loci

Enhancers are one of the major distal regulatory DNA regions and are essential for the precise control of gene transcription[19, 20]. Because the conserved motif loci showed higher enrichment at TF binding sites and active histone marked regions in distal regulatory regions as well as promoter regions, we next focused on the association between conserved motif loci and enhancer regions. We utilized the FANTOM5 enhancer atlas, which describes promising active enhancer regions according to bidirectional enhancer RNA expression[21]. The percentage of overlapping conserved motif loci with FANTOM5 enhancer sites demonstrated that the conserved motif loci were highly enriched in the enhancer region compared to nonconserved motif loci (all regions; conserved: 1.8%, nonconserved: 0.7%) (Fig. 3A). Next, we examined the genomic locus, including the enhancer DNA element (VISTA Enhancer element ID: hs1862), that showed reproducible enhancer activity in heart tissue[22]. At the VISTA enhancer locus, we found the conserved motif locus “GATA-type zinc fingers” overlapped with the FANTOM5 enhancer. From this finding, we processed publicly available ChIP-seq data for GATA4[23], a transcription factor that plays essential roles in heart development[24, 25]. Indeed, we observed GATA4 binding at the conserved motif locus and enhancer region in iPSC-derived cardiomyocytes (Fig. 3B). These results indicated that conserved DNA loci can preferentially be utilized as enhancers.

Conserved motif loci were functionally exploited in glioblastoma

To assess whether TF binding at conserved motif loci can trigger the expression of critical genes in a certain cellular system, we focused on the transcription factor Acheate-scute like 1 (ASCL1), which plays a key role in neuronal differentiation in glioblastoma by NOTCH signaling inhibition[26, 27]. We utilized the ChIP-seq data for ASCL1 and found that the conserved ASCL1-related motif loci were highly overlapped in actual ASCL1 binding sites compared to nonconserved motif loci (conserved: 1%, nonconserved: 0.3%) (Fig. 4A). Next, we analyzed the RNA-seq data for WT and ASCL1 KO human primary glioblastoma cells[27] and integrated the ChIP-seq and motif data to predict possible ASCL1 target genes. We found that possible target genes, such as *JAG2*, *DLL1* and *DLL3*, that included ASCL1 binding on the conserved ASCL1-related motif loci were significantly associated with the NOTCH signaling pathway (Fig. 4B). We confirmed that the ASCL1 binding sites corresponded to the conserved ASCL1-related motif loci near *DLL3* and *DLL1*. In addition, these ASCL1 binding sites localized at accumulated motif loci (Fig. 4C). In summary, these results suggest that conserved motif loci can be functionally exploited to regulate the expression of key genes in glioblastoma cells.

Conclusions

In this study, we extracted the conserved DNA-binding motif loci between humans and mice and demonstrated that they reflect a proportion of functional TF binding sites compared with nonconserved motif loci. Particularly, the conserved motif loci enabled us to predict GATA4 binding to heart-specific enhancer region and GATA4 ChIP-seq data validated the prediciton (Fig. 3B). Moreover, we demonstrated that the conserved motif loci can be preferencially utilized as functional TF binding sites (Fig. 4). These result indicated that our data of the conserved motif loci provides a useful clue for predicting functional TF binding sites.

Because the data of conserved motif loci contains only 0.1% of the motif sites from raw data, it allows us to easily visualize them in a genome browser and recognize the candidate TFs that can regulate a gene of interest (Table 1 and Fig. 1B). Although we recognized that there are still a considerable number of false positives, we believe that the conserved motif data can allow researchers to predict possible candidate TFs that control the transcription of genes of interest.

Methods

Extracting conserved DNA-binding motif loci

We downloaded BED files of hg19 and mm10 genomic data from the JASPAR UCSC track hub (http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/; last modified on January 16, 2018)[6, 28]. The BED files contain chromosomal loci of predicted motif sites, JASPAR motif names, and motif matching scores (transformed p-values, which indicate the significance of matches between genomic sequences and known motif profiles)[29]. We first extract motif sites with a motif matching score greater than 400 (p-value < 10-4). To exclude redundant motif loci, we merged motifs that included the same TF family. For example, the motifs of GATA1, GATA2, GATA3 and GATA4 can be located at the same genomic locus because their binding motifs are quite similar to each other. Thus, these motif loci were merged into one locus as “GATA-related motifs”. To merge original motifs into the family, we referred to human TFClass (draft version, October 05, 2018)[16] that provides the hierarchical classification of eukaryotic TFs based on their DBDs. We used bedtools (version 2.28.0)[30] to merge genomic regions that included motif loci of the same TF family. The merged genomic loci were excluded when the sequence length was extremely short (less than four nucleotides) or long (more than thirty nucleotides). Then, we culled the conserved motif loci that existed in the corresponding locus between the hg19 and mm10 genomes. To do this, we utilized hg19-mm10 alignment sequence data that can be accessed from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/vsHg19/axtNet/; last modified on March 08, 2012)[31]. The genomic annotation of motif locations in hg19 was conducted by HOMER annotatePeaks.pl (version 4.10.4)[10]. The motif sites located 1-2 kb and 2-3 kb upstream of the transcription start site (TSS) were annotated as “promoter-1-2 kb” and “promoter-2-3 kb”, respectively.

ChIP-Atlas

The data on TF binding sites in pluripotent stem cells (PSCs) near the NANOG promoter region were obtained from the ChIP-Atlas Peak Browser (accessed on July 01, 2019)[17]. The BED files containing the TF, H3K27ac, and H3K4me1 ChIP-seq peaks were also downloaded from the ChIP-Atlas Peak Browser (accessed at July 01, 2019)[17]. We selected the peaks with a MACS2 score greater than 100.

Motif accumulation

Any motif loci that overlap or are within 10 bp of one another were counted and merged by bedtools merge (version 2.28.0)[30]. Because the merged motif loci with a high number of motif accumulations appear to be longer genomic loci, every motif locus was transformed to a uniform 200 bp length to remove the sequence length dependency when counting the overlapping motif sites and ChIP-seq bindings.

Overlapping motif loci and ChIP-seq data

Any motif loci that overlapping with ChIP-seq data (TFs, H3K27ac, and H3K4me1 from the ChIP-Atlas21) were counted using bedtools intersect (version 2.28.0)[30]. In terms of TFs, the overlap of motif loci and TF binding sites that belong to the same TF family according to TFClass[16] was counted.

FANTOM5 enhancer

The BED file of human permissive enhancers was downloaded from FANTOM5 datafiles (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz; last modified at February 12, 2015)[21].

ChIP-seq

The data for GATA4, H3K27ac and input DNA in induced pluripotent stem cell (iPSC)-derived cardiomyocytes were downloaded from NCBI SRA (accession number: GSM2280004, GSM2280005 and GSM2280009)[23]. The data for ASCL1 and input DNA in glioblastoma were obtained from NCBI SRA (accession number: GSM2335531, GSM2335532,GSM2335533, GSM2335534 and

GSM2335535)[27]. The FASTQ files were mapped to hg19 by BWA-MEM (version 0.7.17-r1188)[32] with default parameters. The output SAM files were converted and sorted to BAM files using samtools (version 1.9)[33]. To visualize the expression by the Integrative Genomics Viewer[34], the BAM files were converted to BigWig files by deepTools (version 3.2.0)[35] with CPM normalization, and the bin size was set to 10. Peak calling was conducted by MACS2 (version 2.1.1)[36] with the q-value < 0.01.

RNA-seq

Data on human glioblastoma cells were obtained from NCBI SRA (accession number: GSE87615, GSE87617)[27]. The FASTQ files were mapped to the hg19 reference genome by STAR (version 2.7.1a)[37] with default parameters. The output SAM files were converted and sorted to BAM files using samtools (version 1.9)[33]. To visualize the expression with the Integrative Genomics Viewer[34], the BAM files were converted to BigWig files by deepTools (version 3.2.0)[35] with CPM normalization, and the bin size was set to 10. Differentially expressed genes were identified by DESeq2 (version 1.24.0)[38] with an adjusted p-value < 0.01 and $\log_2$ fold change > 1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data and code generated in this study are publicly available at a repository of Open Science Framework (https://osf.io/jhfnb/?view_only=1c82bc9963ee45b799be42512abb08d1).

Competing interests

The authors declare no conflicts of interest associated with this manuscript.

Funding

This study is supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan, KAKENHI, No. 19H03142.

Authors' contributions

AK designed the study. AK performed experiments and wrote the paper. ST edited and approved the manuscript. ST supervised the work. All authors discussed the results and commented on the manuscript text.

Acknowledgments

We would like to thank Dr. Michito Hamada for the critical discussion and comments.

References

1. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. DBD—taxonomically broad transcription factor predictions: New content and functionality. Nucleic acids research. 2007;36 suppl_1:D88–92.

2. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep III PW, Bulyk ML. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology. 2006;24:1429.
3. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152:327–39.
4. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-dna interactions. Science. 2007;316:1497–502.
5. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. 2018;172:650–65.
6. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, Lee R van der, et al. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic acids research. 2017;46:D260–6.
7. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein–dna interactions. Nucleic acids research. 2014;43:D117–22.
8. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.
9. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module transcompel: Transcriptional gene regulation in eukaryotes. Nucleic acids research. 2006;34 suppl_1:D108–10.
10. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. Molecular cell. 2010;38:576–89.
11. Stewart AJ, Hannenhalli S, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. Genetics. 2012;192:973–85.
12. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343.
13. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science. 2012;338:1593–9.
14. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010;328:1036–40.
15. Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, et al. Principles of regulatory information conservation between mouse and human. Nature. 2014;515:371.
16. Wingender E, Schoeps T, Haubrock M, Krull M, Dönitz J. TFClass: Expanding the classification of human transcription factors to their mammalian orthologs. Nucleic acids research. 2017;46:D343–7.
17. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, et al. ChIP-atlas: A data-mining suite powered by full integration of public chip-seq data. EMBO reports. 2018;19:e46255. doi:10.15252/embr.201846255.
18. Todeschini A-L, Georges A, Veitia RA. Transcription factors: Specific dna binding and specific gene regulation. Trends in genetics. 2014;30:211–9.
19. Blackwood EM, Kadonaga JT. Going the distance: A current view of enhancer action. Science. 1998;281:60–3.
20. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: Five essential questions. Nature Reviews Genetics. 2013;14:288.
21. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455.
22. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA enhancer browser—a database of tissue-specific human enhancers. Nucleic acids research. 2006;35 suppl_1:D88–92.
23. Ang Y-S, Rivas RN, Ribeiro AJS, Srivas R, Rivera J, Stone NR, et al. Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. Cell. 2016;167:1734–1749.e22. doi:10.1016/j.cell.2016.11.033.
24. Kuo CT, Morrisey EE, Anandappa R, Sigrist K, Lu MM, Parmacek MS, et al. GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. Genes & development. 1997;11:1048–60.

25. Molkentin JD, Lin Q, Duncan SA, Olson EN. Requirement of the transcription factor gata4 for heart tube formation and ventral morphogenesis. Genes & development. 1997;11:1061–72.
26. Somasundaram K, Reddy SP, Vinnakota K, Britto R, Subbarayan M, Nambiar S, et al. Upregulation of ascl1 and inhibition of notch signaling pathway characterize progressive astrocytoma. Oncogene. 2005;24:7073.
27. Park NI, Guilhamon P, Desai K, McAdam RF, Langille E, O'Connor M, et al. ASCL1 reorganizes chromatin to direct neuronal fate and suppress tumorigenicity of glioblastoma stem cells. Cell stem cell. 2017;21:209–24.
28. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the ucsc genome browser. Bioinformatics. 2013;30:1003–5.
29. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.
30. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
31. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human–mouse alignments with blastz. Genome research. 2003;13:103–7.
32. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
34. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011;29:24.
35. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research. 2014;42 Web Server issue:W187–91. doi:10.1093/nar/gku365.
36. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008;9:R137. doi:10.1186/gb-2008-9-9-r137.
37. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal rna-seq aligner. Bioinformatics. 2013;29:15–21.
38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology. 2014;15:550.

Tables

Table 1

	Human (hg19)			Mouse (mm10)		
	# of motif loci	% of motif loci	file size (gzipped)	# of motif loci	% of motif loci	file size (gzipped)
All JASPAR motif loci	10464676413	100.00	65GB	9005684579	100.00	56GB
Motif loci with score > 400	373319759	3.57	2.9GB	348610133	3.87	2.7GB
Merged motif loci	165206259	1.58	1.6GB	145618340	1.62	1.4GB
Conserved motif loci	9281940	0.09	305MB	9584474	0.11	310MB

Summary of the number of motif loci, the percentage from raw data, and the gzip compressed file size during procedures.

Figures

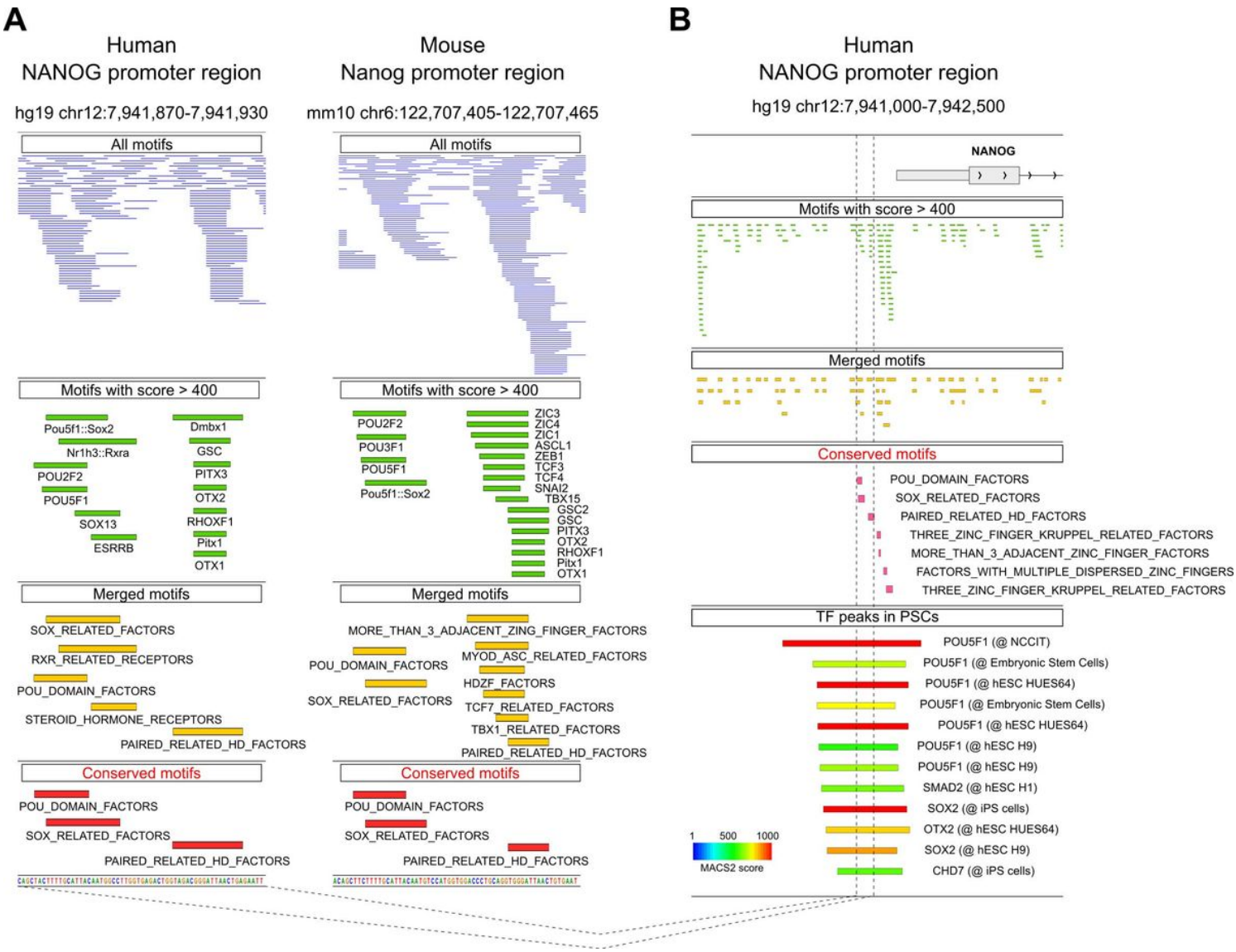

Figure 1

Example genomic viewing of DNA-binding motif loci from raw data to conserved motif loci (A) An example of a genomic locus illustrating the NANOG promoter region in the human (hg19) and mouse (mm10) genomes. All JASPAR motif sites (blue) were filtered by a score greater than 400 (green). Then, the motif sites that belonged to the same TF family were merged (yellow). The conserved motif loci (red) were extracted by aligning the sequence between hg19 and mm10. (B) Expanded genomic view at the human NANOG promoter. The ChIP-seq peaks of TF binding sites in pluripotent stem cells (PSCs) near the NANOG promoter region were obtained from ChIP-Atlas Peak Browser21. The colors of each peak region showed the MACS2 score. All visualized peaks in the figure had a MACS2 score greater than 500.

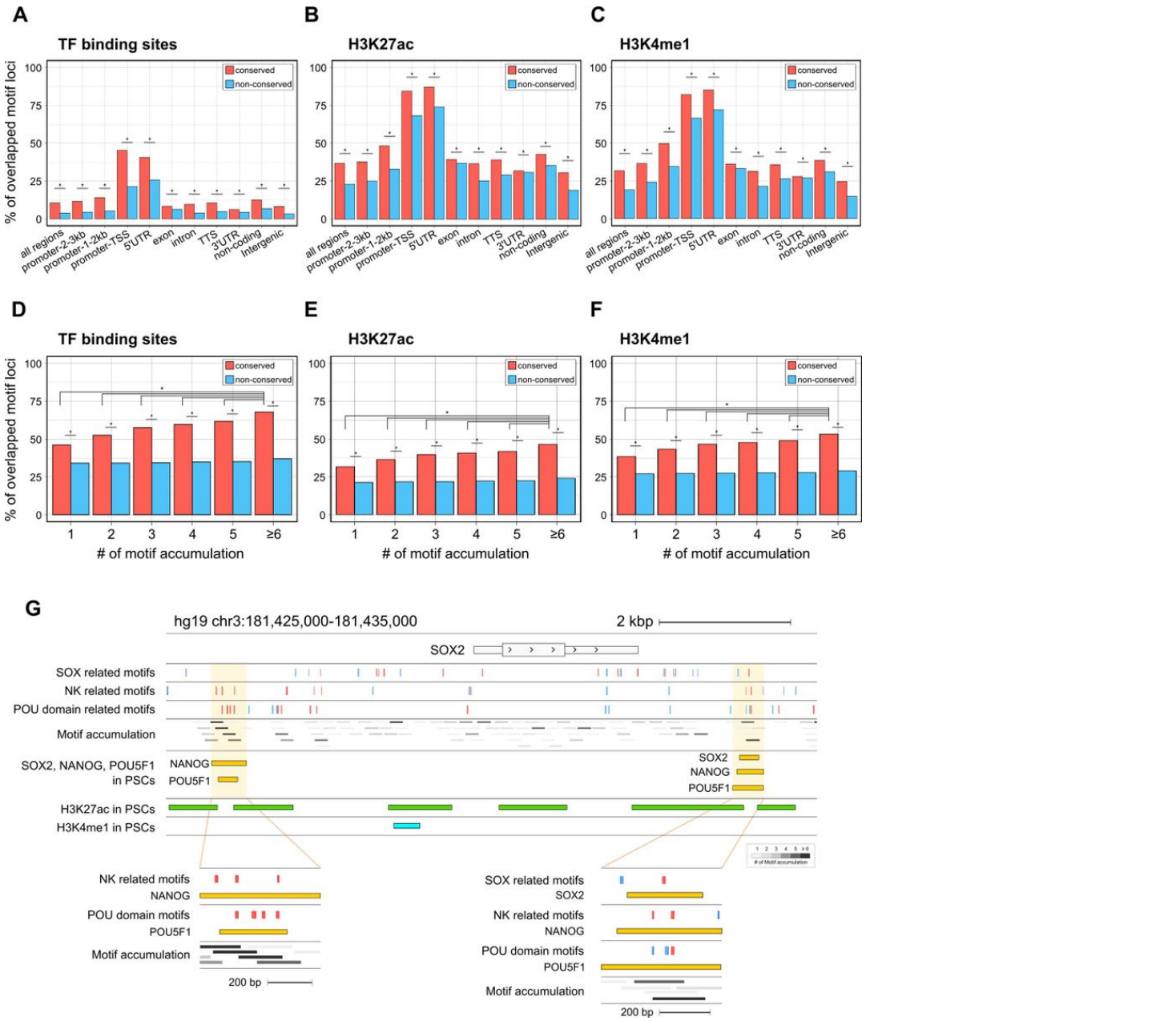

Figure 2

Comparative analysis of overlapping motif loci with ChIP-seq binding sites (A-C) Percentage of the number of overlapping motif loci with TFs and histone marks (H3K27ac and H3K4me1) from the ChIP-Atlas. : p-value < 0.01 (Fisher's exact test). (D-F) The number of motifs accumulated and the percentage of overlapping motif loci with TF binding and histone marked sites (H3K27ac and H3K4me1) from the ChIP-Atlas. : p-value < 0.01 (Fisher's exact test). Holm p-value adjustment was conducted in multiple comparisons. (G) Aligned data for SOX, NK, and POU domain-related motif loci (conserved: red, nonconserved: blue), the accumulation of conserved motif sites (grayscale) and ChIP-seq data for SOX2, NANOG, POU5F1 (yellow), H3K27ac (green), and H3K4me1 (light blue) in PSCs near the SOX2 locus. The bottom panel shows the enlargement of yellow highlighted regions.

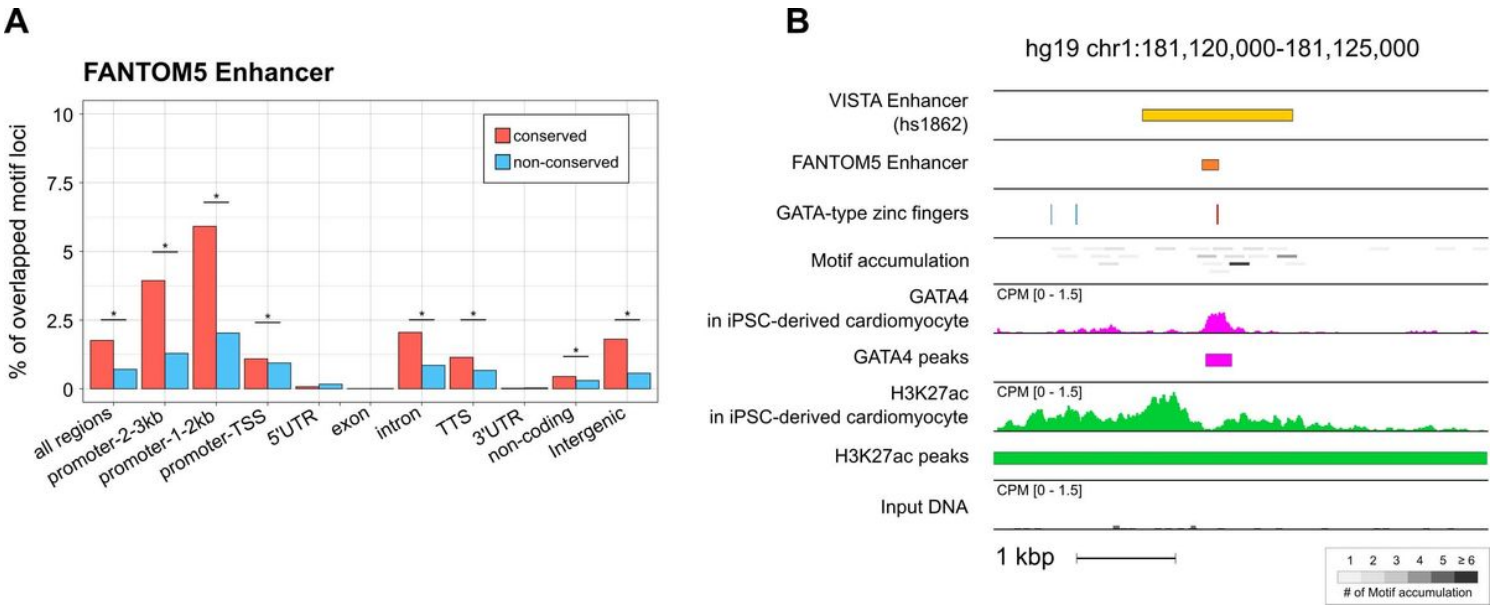

Figure 3

Comparative analysis of overlapping motif loci with enhancer loci (A) Percentage of the number of overlapping motif loci with FANTOM5 enhancer loci. *: p-value < 0.01 (Fisher's exact test). (B) Aligned data showed VISTA enhancer (yellow), FANTOM5 enhancer (orange), conserved motif loci (red) and nonconserved motif loci (blue) associated with GATA-type zinc fingers and the accumulation of conserved motif sites (grayscale). The ChIP-seq data of GATA4 (pink) and H3K27ac (green) binding regions of iPSC-derived cardiomyocytes at the heart-specific enhancer locus are displayed.

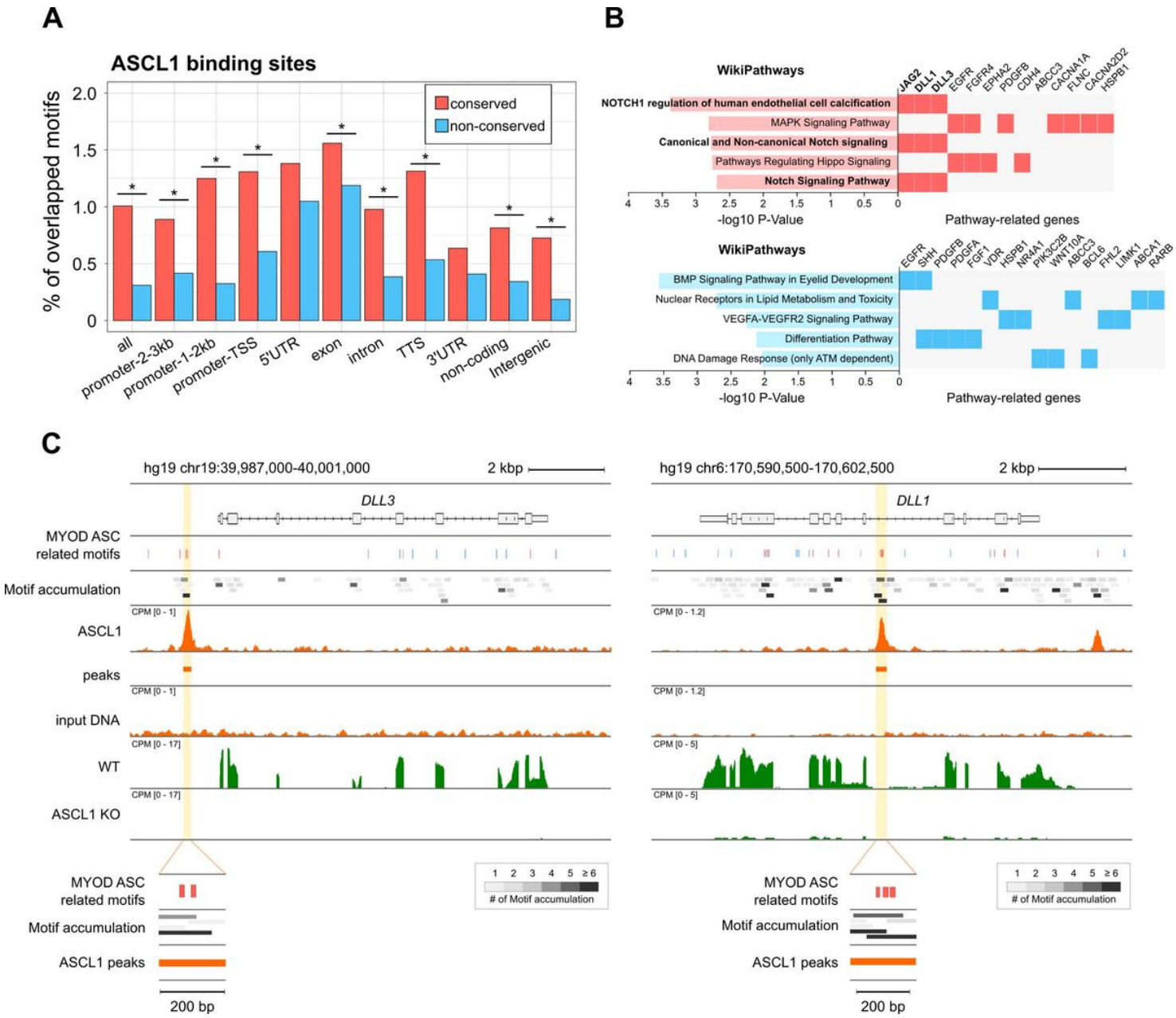

Figure 4

ASCL1 binding to the conserved ASCL1 motif sites near NOTCH signaling-related genes in human primary glioblastoma (A) Percentage of the number of overlapping motif loci with ASCL1 binding sites. *: p-value < 0.01 (Fisher's exact test). (B) Pathway analysis of differentially expressed genes between ASCL1 WT and KO glioblastoma cells with ASCL binding at conserved motif loci (red) and nonconserved motif loci (blue). NOTCH signaling-related genes are highlighted in bold. (C) Aligned data for ASC-related conserved motif loci (red) and nonconserved motif loci (blue), the accumulation of conserved motif sites (grayscale), ASCL1 binding (orange) and RNA-seq (green) of ASCL1 WT and KO cells near the DLL3 and DLL1 loci. The bottom panel shows the enlargement of yellow highlighted regions.