D186–D193 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 23 October 2017*
doi: 10.1093/nar/gkx948

OverGeneDB: a database of 5′ end protein coding overlapping genes in human and mouse genomes

Wojciech Rosikiewicz[1], Yutaka Suzuki[2] and Izabela Makałowska[1,*]

[1]Department of Integrative Genomics, Institute of Anthropology, Faculty of Biology, Adam Mickiewicz University in Poznań, 61-712 Poznań, Poland and [2]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, 272-8562, Japan

Received August 14, 2017; Revised September 20, 2017; Editorial Decision October 04, 2017; Accepted October 20, 2017

ABSTRACT

Gene overlap plays various regulatory functions on transcriptional and post-transcriptional levels. Most current studies focus on protein-coding genes overlapping with non-protein-coding counterparts, the so called natural antisense transcripts. Considerably less is known about the role of gene overlap in the case of two protein-coding genes. Here, we provide OverGeneDB, a database of human and mouse 5′ end protein-coding overlapping genes. The database contains 582 human and 113 mouse gene pairs that are transcribed using overlapping promoters in at least one analyzed library. Gene pairs were identified based on the analysis of the transcription start site (TSS) coordinates in 73 human and 10 mouse organs, tissues and cell lines. Beside TSS data, resources for 26 human lung adenocarcinoma cell lines also contain RNA-Seq and ChIP-Seq data for seven histone modifications and RNA Polymerase II activity. The collected data revealed that the overlap region is rarely conserved between the studied species and tissues. In ∼50% of the overlapping genes, transcription started explicitly in the overlap regions. In the remaining half of overlapping genes, transcription was initiated both from overlapping and non-overlapping TSSs. OverGeneDB is accessible at http://overgenedb.amu.edu.pl.

INTRODUCTION

Gene overlap in eukaryotes, which is defined here as sharing of the same DNA sequence by at least two different genes (1,2), was discovered over 30 years ago (3–5). For a long time, this phenomenon was believed to be rare. However, over the last three decades, increasing examples of gene overlap were reported in various animal, plant and fungal species (6–19). It is estimated that >30% of human and mouse genes overlap with another gene (20–22). Large-scale projects, such as those from the FANTOM Consortium, revealed that 72% of transcription events might proceed in both directions (23).

Genes may overlap in various manners (1), including complete overlap when one gene is nested within the other or partial overlap when only the 3′ or 5′ end(s) of genes are overlapping. Gene overlap is currently intensively studied in the context of protein coding genes regulated by their antisense non-protein coding counterparts, i.e. natural antisense transcripts (NATs), which exhibit various regulatory functions. Briefly, NATs were suggested to regulate protein coding gene expression levels during transcription by various mechanisms of transcriptional interference (TI), including promoter competition, occlusion, 'sitting duck' interference or polymerase collisions (24). Presence of the antisense RNA may also regulate gene expression post-transcriptionally via double-stranded RNA formation (25), leading to RNA editing (26), interference (27–31) or masking (32–39). NATs also regulate protein-coding gene expression levels epigenetically by inducing repressive chromatin modifications within the sense gene promoters or even the entire genomic *loci* and downregulating the expression levels of neighboring genes (40–43). Nevertheless, the extent to which gene expression is regulated by antisense transcription remains a matter of debate and needs to be further investigated (25,44–48), especially given that numerous NATs were connected with various pathological states, such as Parkinson's or Alzheimer's diseases (34,49), cancer (50,51) and numerous other disorders (52,53). Researchers are interested in many of these NATs as therapeutic targets given that their artificial up- or downregulation directly influences the expression levels of the sense, protein-coding genes (42).

Although numerous researchers are working on the gene overlap phenomenon, relatively few databases dedicated to overlapping genes are available. Currently, the most comprehensive database is PlantNATsDB (54), which focuses on antisense transcripts in plants and predicts >2 million NATs in 70 plant species. Until recently, the best equivalent for animal species was NATsDB (55), in which authors deposited thousands of antisense transcripts identified by mapping EST sequences for 11 model organisms. Unfortu-

*To whom correspondence should be addressed. Tel: +48 618295835; Email: izabel@amu.edu.pl

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

nately, NATsDB and other databases, such as EVOG (56), antiCODE (57) or the database created by Veeramachaneni *et al*. (16), are no longer maintained. Nevertheless, the abovementioned databases were primarily dedicated to protein-coding genes that overlap with non-protein coding counterparts, and none of these databases are suitable for large-scale tissue-specific studies of gene overlap. As described in this paper, OverGeneDB is a database of 5′ end(s) protein coding overlapping genes in human and mouse genomes. OverGeneDB contains information regarding 582 human and 113 mouse overlapping protein-coding gene pairs that were identified based on the exact genomic coordinates of the alternative transcription start sites (TSSs) in 73 human and 10 mouse TSS-Seq libraries from various organs, tissues and cell lines. For 26 human lung adenocarcinoma tissue samples, studies of overlapping genes were strengthen by RNA-Seq and ChIP-Seq data analyses of seven histone modifications and RNA Polymerase II activity studies. OverGeneDB is a platform that offers easy access and visualization of all identified overlapping gene pairs and associated data. In addition, these data can be downloaded for further analysis.

MATERIALS AND METHODS

Representative gene coordinates

Coordinates of all known RefSeq transcripts (58) for human GRCh38/hg38 and mouse NCBI37/mm9 genome versions were downloaded from the UCSC database using *Table Browser* (59). Coordinates of splice variants were further used to determine side-to-side gene positions as shown in Figure 1A. TSS coordinates were downloaded for a total of 73 human and 4 mouse libraries from DBTSS database versions 9 (60) and 8 (61) for human and mouse, respectively. TSS coordinates from six additional mouse organs were sequenced for the purpose of this study and processed using the same protocol used for other data from the DBTSS database (60). All human and mouse libraries are listed in Supplementary Table S1. The downloaded data were filtered to ensure that only TSSs with the 'confident' status, in which the normalized expression level is ≥5 parts per million (ppm), were considered, as suggested by Yamashita and coworkers after the detailed validation of the TSS-Seq method (62). Additionally, the maximum distance between a TSS and the closest gene on the same DNA strand was limited to 5000 bp upstream of the gene's annotated 5′ end. Next, each gene was analyzed separately in every TSS library to identify representative gene coordinates. The 3′ end was based on the RefSeq annotations, whereas the 5′ end was determined based on the position of the TSS in a given library. If more than one TSS was assigned to a gene, the coordinates of the distal TSS were considered to represent the 5′ end of the gene (Figure 1B).

Overlapping genes detection and characterization

The identification of genes overlapping at 5′ end(s) was performed for genes expressed in a given library based on the representative gene coordinates. The genes were required to overlap by at least one base. The procedure was performed for each library independently.

Genes may be simultaneously expressed using one or more TSS. In numerous cases, this phenomenon results in genes that only overlap in relation to a subset of alternative TSSs. To determine to what degree the gene is transcribed from the overlapping TSSs, the overlap ratio (OR) was developed. This value is simply a fraction of the total gene expression assigned to the overlap region (Figure 2). Consequently, to estimate to what extent transcripts in the gene pair are transcribed from the overlapping region, the JoinedOR ratio was calculated. JoinedOR is a product of the OR values of genes in an overlapping pair (Figure 2). OR and JoinedOR values equal to 1 indicate that all transcripts originated from the overlapping TSSs in the gene and pair, respectively. The lower the values, the smaller the subset of transcripts that originated from the overlapped region. A value of 0 indicates that no expression was assigned to the overlapping TSSs.

Expression level estimation using RNA-seq data

Raw Illumina RNA-Seq paired-end reads from 26 lung adenocarcinoma samples, which were the same samples used for TSS sequencing, were downloaded from the ENA database (63) where these data are stored under accession number *PRJDB2256*. All reads were subjected to quality filtering using the Trimmomatic program (version 0.36) (64) with the following parameters: *-phred33; ILLUMINACLIP: adapters/TruSeq3-PE.fa:2:30:10; LEADING: 20; TRAILING: 20; SLIDINGWINDOW:5:20;* and *MINLEN:50.* Quality control before and after filtering was performed using FastQC (version 0.11.5) (65). Filtered reads were aligned to the human reference hg38 genome downloaded from UCSC (66) using the HISAT2 program (version 2.0.5) (67) with the—*downstream-transcriptome-assembly* parameter. The numbers of mapped and unmapped reads for each library are presented in supplementary Table ST2. Next, all SAM files were sorted and converted to BAM format using SAMtools (version 1.3.1) (68). Expression levels of individual transcripts were further estimated in FPKM (fragments per kilobase of exon per million fragments mapped) using StringTie (version 1.3.1c) (69) with *–e* and *–B* flags guided by GENCODE (version 24) genome reference annotations (70). The expression levels of individual transcripts were summed to represent the total expression levels of genes. No minimal expression level was required in this step.

Histone modifications and RNA polymerase II activity studies

Pre-aligned reads from ChIP-Seq experiments for RNA Polymerase II and seven histone modification types, including H3ac, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3, with their controls were downloaded from the DBTSS' (version 9) (60) FTP server in BED3+1 format. The additional column represents the number of tags mapped in a certain position. Data were available for 26 tissue samples from patients with lung adenocarcinoma. Each library was converted to standard BED6 format and subsequently screened for peak enrichment using the MACS2 program (version 2.1.0) (71) with

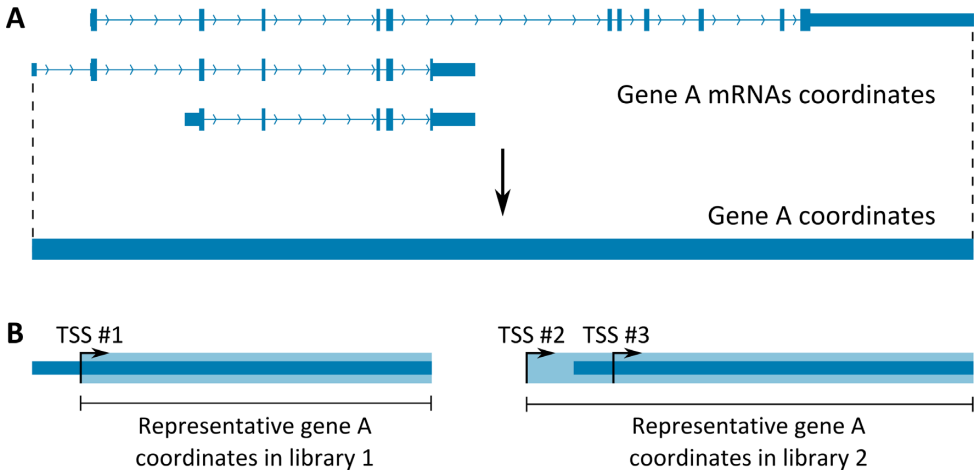


Figure 1. Representative gene coordinates computation strategy. (**A**) Gene A coordinates are based on the distal 5′ and 3′ annotated coordinates of all gene's mRNAs; (**B**) Representative gene coordinates in different libraries are based on the annotated gene A 3′ end and distal alternative transcription start site.

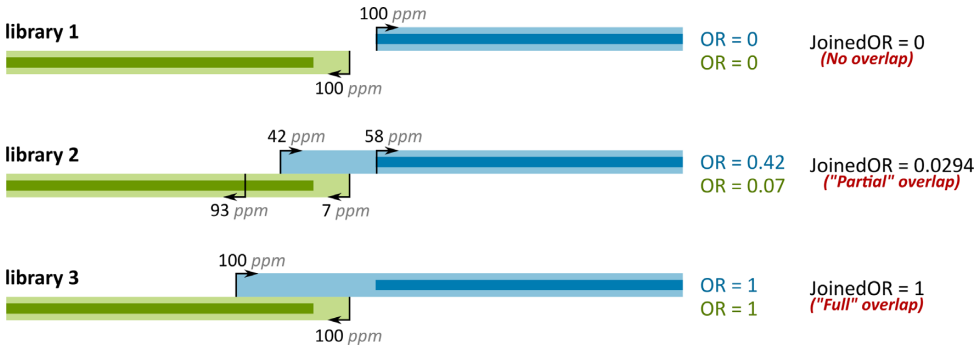


Figure 2. OR and JoinedOR values for the example gene pair expressed from the non-overlapping TSS in library 1 and the overlapping TSS in libraries 2 and 3. Blue and green narrow solid boxes represent the annotated coordinates of genes on plus and minus strand, respectively. Wider light blue and green boxes indicate representative gene coordinates in particular library, whereas arrows represent alternative transcription start sites accompanied by assigned to them normalized to parts per million (ppm) expression levels.

the *–nomodel* flag used for all libraries. In addition, parameters *–broad* and *–broad-cutoff 0.1* were used for data regarding seven histone modifications. For the purpose of visualization in the web browser, MACS2 output peaks were converted to the bigBed format using the *bedtobigbed* program from UCSC (72). All BED6 files representing coordinates of mapped reads were converted to bigWig format using *genomecov* from the BEDTools package (version 2.25.0) (73) with *–bg* flag and *bedGraphToBigWig* program from UCSC (72).

Association of transcription factors with TSSs

To associate transcription factors (TFs) with particular promoters, transcription start sites were first clustered across all human and separately all mouse TSS-Seq libraries. In this step, all TSSs assigned to the same gene and located <300 bp from each other were joined in a single cluster. Next, nucleotide sequences of all clusters together with up to 500 bp in both directions were screened for the presence of transcription factor binding site (TFBS) motifs obtained from the JASPAR database (74). The search was conducted using the *searchSeq* function from TFBSTools (version 1.14.0) (75) with the minimal score set to 95%. All TFs potentially associated with promoters were subsequently filtered based on their expression, i.e. only TFs that were expressed in a particular library were considered as potentially regulating a certain promoter. Finally, hierarchical clustering of all potential TFs in all libraries was performed for each promoter.

Database implementation

The OverGeneDB database was implemented in MySQL (https://www.mysql.com/). The publicly accessible interface was generated using HTML, PHP and JavaScript. The detailed overlapping gene pair view was additionally equipped with an embedded Dalliance (76) genome browser and interactive charts from plot.ly (https://plot.ly/).

DATABASE COMPOSITION AND USAGE

Data stored in OverGeneDB may be accessed via three different methods: Browse, Search or sequence similarity search. The Browse page lists all human and mouse gene pairs that overlap in at least one library. This page also provides information regarding the number of libraries in

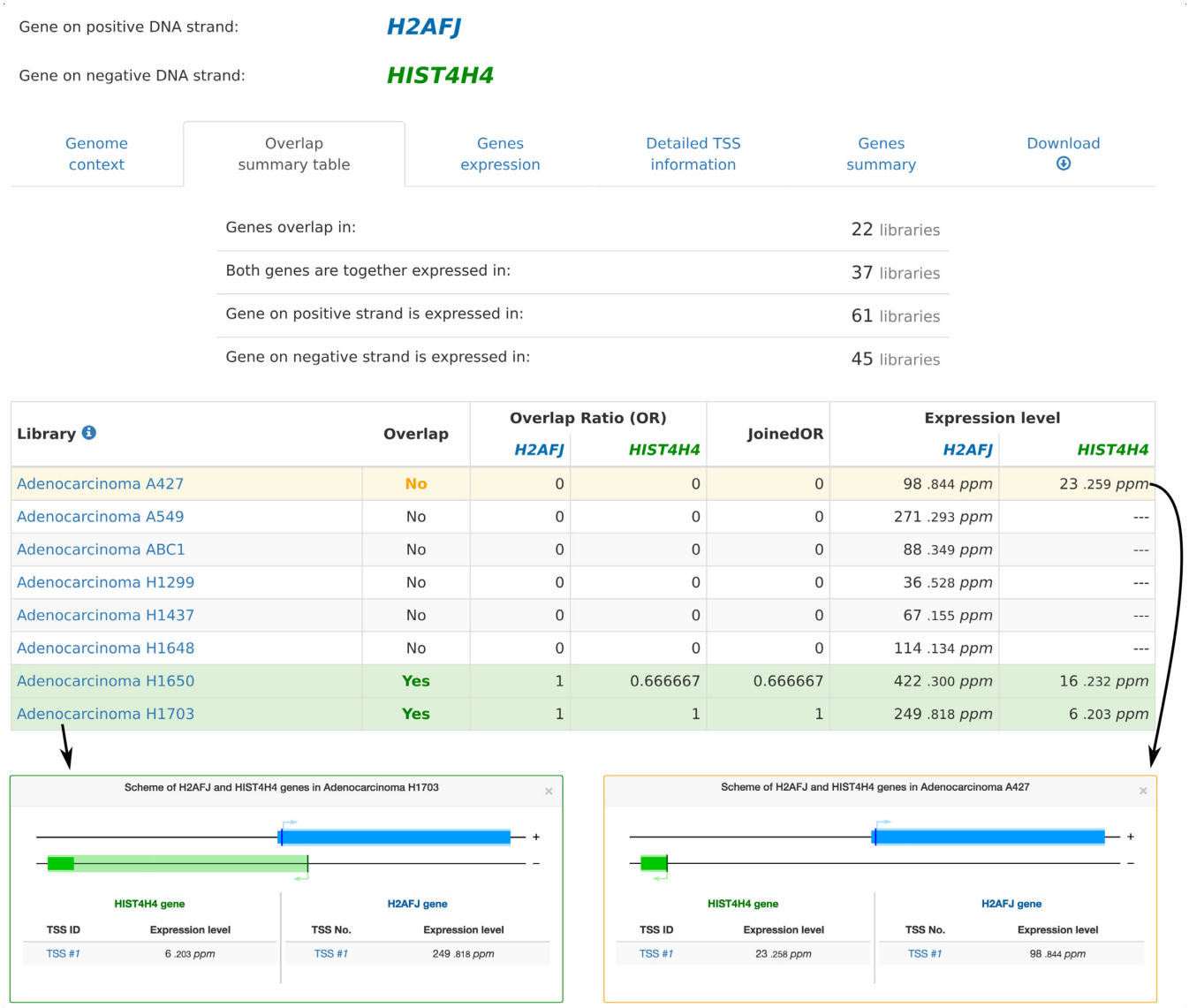


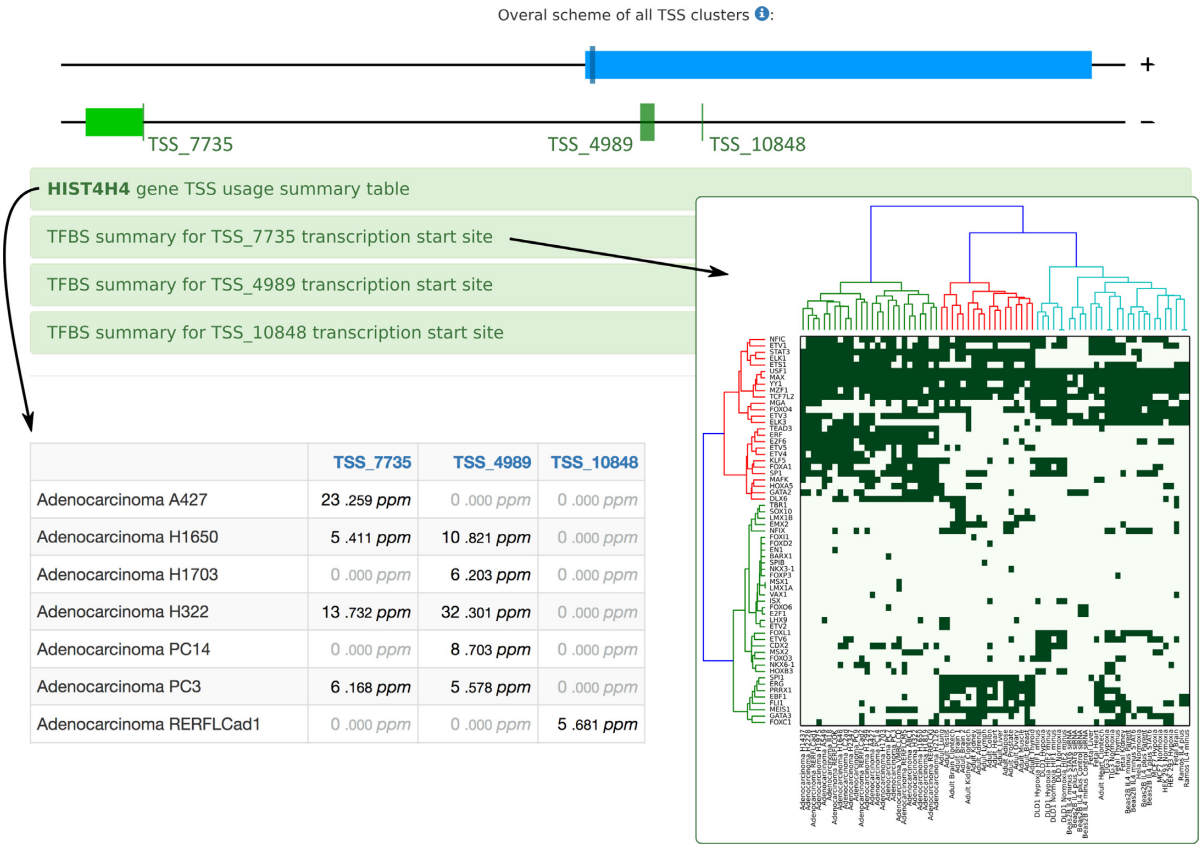
Figure 3. Overlap summary table for the *H2AFJ* and *HIST4H4* gene pair. Clicking on library names opens a small pop-up window with a scheme of genes arrangement within selected library.

which genes in a given pair are identified as overlapping and whether both or only one gene from a pair is expressed. The Search option allows the user to specify the libraries or the number of libraries in which genes overlap, only one or none of genes from a pair is expressed, or both genes are expressed regardless of their overlap status. It is also possible to perform a sequence-based similarity search using the BLAST program (77,78) against the overlapping regions or the representative gene sequences in all or selected libraries.

The above described Search and Browse methods generate lists of gene pairs meeting the specified criteria. Detailed information about a given pair can be obtained by clicking on the 'Details' button. The overlapping gene pair view is separated into six sections displayed in tabs as follows:

i. Genome context— annotated genes and transcripts can be examined using a built-in dalliance web browser (76). Additional tracks may be selected for human and mouse libraries using a button above the browser. These tracks contain alternative TSSs and overlap regions displayed as blocks and positions of predicted TFBS. For the 26 human lung adenocarcinoma libraries, it is also possible to display tracks of raw BAM files of mapped RNA-Seq reads and raw ChIP-Seq signals in bigWig format and peaks for RNA Polymerase II and seven types of histone modifications.

ii. Overlap summary table (Figure 3)—this table displays detailed information about overlapping genes, including OR and JoinedOR ratios and TSS-Seq-based expression levels. Upon clicking the library name, a simple visualization of the gene overlap in a selected library is displayed where one may also inspect the expression levels of individual TSSs.



	TSS_7735	TSS_4989	TSS_10848
Adenocarcinoma A427	23 .259 *ppm*	0 .000 *ppm*	0 .000 *ppm*
Adenocarcinoma H1650	5 .411 *ppm*	10 .821 *ppm*	0 .000 *ppm*
Adenocarcinoma H1703	0 .000 *ppm*	6 .203 *ppm*	0 .000 *ppm*
Adenocarcinoma H322	13 .732 *ppm*	32 .301 *ppm*	0 .000 *ppm*
Adenocarcinoma PC14	0 .000 *ppm*	8 .703 *ppm*	0 .000 *ppm*
Adenocarcinoma PC3	6 .168 *ppm*	5 .578 *ppm*	0 .000 *ppm*
Adenocarcinoma RERFLCad1	0 .000 *ppm*	0 .000 *ppm*	5 .681 *ppm*

Figure 4. Detailed TSS information for the *HIST4H4* gene example. 'HIST4H4 gene TSS usage summary table' panel displays expression levels assigned to individual TSS in various libraries. This summary lists only libraries, in which inspected gene was expressed. 'TFBS summary table' displays TFs for which binding sites were identified in the region <500 bp from a given TSS. These TFs were hierarchically clustered based on information if it is expressed (marked in green) or non-expressed (marked in white) in a given library.

iii. Gene expression—this tab provides detailed information about the TSS-Seq and RNA-Seq expression levels of genes in an overlapping gene pair when available.
iv. Detailed TSS information (Figure 4)—this tab provides a visualization of the gene pair and positions of TSS clusters. For each gene in a pair, a summary table with the TSS usage is provided. The user may easily inspect the libraries in which a particular TSS was utilized and what normalized expression level was assigned to it. In addition, for each promoter, a hierarchically clustered summary of the TFs potentially responsible for the regulation of this particular promoter is displayed. The user may download this summary in PNG, SVG or Tab Separated Value (TSV) formats.
v. Gene summary—this tab displays basic gene information with links to cross database references, gene references into functions (Gene RIF), Gene Ontology and PubMed literature references associated with NCBI Gene IDs, which were all downloaded from the NCBI Gene database (79).
vi. Download—user may download the most essential information associated with selected overlapping gene pair in TSV, BED or FASTA formats.

DISCUSSION

To the best of the authors' knowledge, OverGeneDB is the first database strictly dedicated to protein coding genes overlapping at their 5′ end(s). The database contains 582 human and 113 mouse gene pairs that were identified as overlapping with a minimum of one TSS pair in at least one library. These gene pairs included 1150 human and 225 mouse genes, among which 14 genes overlapped with more than one other gene. A total of 4075 promoters in human and 518 promoters in mouse were assigned to these genes. The average overlapping region size is 1570 bp long, and the longest overlap region, which was ∼50 kb, was reported for gene pair *RUNX2* and *SUPT3H*. The collected data revealed that the overlap region is rarely conserved among the studied species, organs, tissues and cell lines. Surprisingly, for up to 300 human and mice gene pairs, a >100-bp difference in the overlap region length across libraries was identified. In the case of 159 gene pairs, this difference is >1000 bp. Moreover, in 85 human overlapping pairs, the difference is not only in the overlap length but also it is significantly shifted and covers a different genomic region. The *HTRA2* and *AUP1* gene pair serves as an example. In this gene pair, the overlapping region shifts between libraries from exclusively covering the *HTRA2* annotated gene body in *Adenocarcinoma VMRCLCD* to an overlap that is mainly lo-

cated within the *AUP1* gene in *Adenocarcinomas PC3* and *PC14*. Among all identified overlapping gene pairs, 90 human and mouse pairs were identified as always overlapping whenever both genes were expressed. The remaining 605 pairs were occasionally expressed from the overlapping and non-overlapping promoters. In total, 203 human and mouse gene pairs were overlapping only in one library, whereas both genes where expressed without gene overlap in almost all other libraries.

Genes often utilize multiple alternative promoters simultaneously, among which only a subset may be overlapping. Therefore, if regulation by transcriptional interference or double stranded RNA formation occurs, only transcripts initiated within the overlap region may be subjected to regulation via these mechanisms. To assess this phenomenon, OR and JoinedOR ratios were introduced, and these values represent the frequency of the expression initiated within the overlap region in gene and gene pair, respectively. In 57% human and 44% mouse overlap events, transcription started explicitly in the overlap regions, as reflected by a JoinedOR value equal to one. In contrast, in the remaining cases, transcription was initiated both from overlapping and non-overlapping TSSs by at least one of the genes in a pair. Human genes *FBXL15* and *PSD*, which are *inter alia* overlapping in brain tissue, serve as a great example. Both genes simultaneously utilize two alternative TSSs, among which only distal TSSs are overlapping. Assuming equal expression level of both genes and utilizing a JoinedOR ratio, <13% of *FBXL15* and *PSD* gene transcripts were estimated to be possibly subjected to overlap-related regulation on transcriptional or post-transcriptional levels.

Studies of the overlapping gene pairs in OverGeneDB were strengthened by the large-scale *in silico* prediction of TFBS within the overlapping gene promoter regions. Moreover, overlapping gene expression levels in 26 lung adenocarcinoma samples were independently estimated based on RNA-Seq data. Finally, the same twenty six tissue samples were also studied using ChIP-Seq experiment results aimed at RNA Polymerase II activity and seven histone modifications, which exhibit significant potential in overlapping promoter studies (44). Taking all of these features together, the OverGeneDB is a very valuable source of data for anyone interested in antisense transcription, the regulation of promoter usage, and the mechanisms of gene expression regulation.

AVAILIBILITY

OverGeneDB database is freely available at the URL http://overgenedb.amu.edu.pl. Scripts for automated overlapping gene pairs' identification are available at https://github.com/forrest1988/OverGeneDB.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Science Centre [NCN 2013/11/N/NZ2/02524], EC FP7 People: Marie Curie Actions Grant 'EVOLGEN' [PIRSES-GA-2009–247633]; KNOW Poznan RNA Centre [01/KNOW2/2014]. Funding for open access charge: National Science Centre [NCN 2013/11/N/NZ2/02524].
Conflict of interest statement. None declared.

REFERENCES

1. Makalowska,I., Lin,C.F. and Makalowski,W. (2005) Overlapping genes in vertebrate genomes. *Comput. Biol. Chem.*, **29**, 1–12.
2. Faghihi,M.A. and Wahlestedt,C. (2009) Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol.*, **10**, 637–643.
3. Spencer,C.A., Gietz,R.D. and Hodgetts,R.B. (1986) Overlapping transcription units in the dopa decarboxylase region of Drosophila. *Nature*, **322**, 279–281.
4. Williams,T. and Fried,M. (1986) A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. *Nature*, **322**, 275–279.
5. Henikoff,S., Keene,M.A., Fechtel,K. and Fristrom,J.W. (1986) Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands. *Cell*, **44**, 33–42.
6. van Duin,M., van Den Tol,J., Hoeijmakers,J.H., Bootsma,D., Rupp,I.P., Reynolds,P., Prakash,L. and Prakash,S. (1989) Conserved pattern of antisense overlapping transcription in the homologous human ERCC-1 and yeast RAD10 DNA repair gene regions. *Mol. Cell. Biol.*, **9**, 1794–1798.
7. Mol,J.N., van der Krol,A.R., van Tunen,A.J., van Blokland,R., de Lange,P. and Stuitje,A.R. (1990) Regulation of plant gene expression by antisense RNA. *FEBS Lett.*, **268**, 427–430.
8. Quesada,V., Ponce,M.R. and Micol,J.L. (1999) OTC and AUL1, two convergent and overlapping genes in the nuclear genome of Arabidopsis thaliana. *FEBS Lett.*, **461**, 101–106.
9. Osato,N., Yamada,H., Satoh,K., Ooka,H., Yamamoto,M., Suzuki,K., Kawai,J., Carninci,P., Ohtomo,Y., Murakami,K. *et al.* (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol.*, **5**, R5.
10. Wang,H., Chua,N.H. and Wang,X.J. (2006) Prediction of trans-antisense transcripts in Arabidopsis thaliana. *Genome Biol.*, **7**, R92.
11. Steigele,S. and Nieselt,K. (2005) Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. *Nucleic Acids Res.*, **33**, 5034–5044.
12. David,L., Huber,W., Granovskaia,M., Toedling,J., Palm,C.J., Bofkin,L., Jones,T., Davis,R.W. and Steinmetz,L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 5320–5325.
13. Lehner,B., Williams,G., Campbell,R.D. and Sanderson,C.M. (2002) Antisense transcripts in the human genome. *Trends Genet.*, **18**, 63–65.
14. Shendure,J. and Church,G.M. (2002) Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.*, **3**, research0044.1–0044.14.
15. Zhou,C. and Blumberg,B. (2003) Overlapping gene structure of human VLCAD and DLG4. *Gene*, **305**, 161–166.
16. Veeramachaneni,V., Makalowski,W., Galdzicki,M., Sood,R. and Makalowska,I. (2004) Mammalian overlapping genes: the comparative perspective. *Genome Res.*, **14**, 280–286.
17. Ge,X., Rubinstein,W.S., Jung,Y.C. and Wu,Q. (2008) Genome-wide analysis of antisense transcription with Affymetrix exon array. *BMC Genomics*, **9**, 27.
18. Misener,S.R. and Walker,V.K. (2000) Extraordinarily high density of unrelated genes showing overlapping and intraintronic transcription units. *Biochim. Biophys. Acta*, **1492**, 269–270.
19. Lee,S., Bao,J., Zhou,G., Shapiro,J., Xu,J., Shi,R.Z., Lu,X., Clark,T., Johnson,D., Kim,Y.C. *et al.* (2005) Detecting novel low-abundant transcripts in Drosophila. *RNA*, **11**, 939–946.
20. Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, **21**, 379–386.
21. Chen,J., Sun,M., Kent,W.J., Huang,X., Xie,H., Wang,W., Zhou,G., Shi,R.Z. and Rowley,J.D. (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.*, **32**, 4812–4820.

22. Zhang,Y., Liu,X.S., Liu,Q.R. and Wei,L. (2006) Genome-wide *in silico* identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.*, **34**, 3465–3475.
23. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.
24. Shearwin,K.E., Callen,B.P. and Egan,J.B. (2005) Transcriptional interference–a crash course. *Trends Genet.*, **21**, 339–345.
25. Rosikiewicz,W. and Makalowska,I. (2016) Biological functions of natural antisense transcripts. *Acta Biochim. Pol.*, **63**, 665–673.
26. Peters,N.T., Rohrbach,J.A., Zalewski,B.A., Byrkett,C.M. and Vaughn,J.C. (2003) RNA editing and regulation of Drosophila 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. *RNA*, **9**, 698–710.
27. Yu,D., Meng,Y., Zuo,Z., Xue,J. and Wang,H. (2016) NATpipe: an integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes. *Sci. Rep.*, **6**, 21666.
28. Watanabe,T., Totoki,Y., Toyoda,A., Kaneda,M., Kuramochi-Miyagawa,S., Obata,Y., Chiba,H., Kohara,Y., Kono,T., Nakano,T. *et al.* (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **453**, 539–543.
29. Werner,A., Cockell,S., Falconer,J., Carlile,M., Alnumeir,S. and Robinson,J. (2014) Contribution of natural antisense transcription to an endogenous siRNA signature in human cells. *BMC Genomics*, **15**, 19.
30. Ling,K.H., Brautigan,P.J., Moore,S., Fraser,R., Cheah,P.S., Raison,J.M., Babic,M., Lee,Y.K., Daish,T., Mattiske,D.M. *et al.* (2016) Derivation of an endogenous small RNA from double-stranded Sox4 sense and natural antisense transcripts in the mouse brain. *Genomics*, **107**, 88–99.
31. Zhang,X., Xia,J., Lii,Y.E., Barrera-Figueroa,B.E., Zhou,X., Gao,S., Lu,L., Niu,D., Chen,Z., Leung,C. *et al.* (2012) Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. *Genome Biol.*, **13**, R20.
32. Wang,G.Q., Wang,Y., Xiong,Y., Chen,X.C., Ma,M.L., Cai,R., Gao,Y., Sun,Y.M., Yang,G.S. and Pang,W.J. (2016) Sirt1 AS lncRNA interacts with its mRNA to inhibit muscle formation by attenuating function of miR-34a. *Sci. Rep.*, **6**, 21865.
33. Modarresi,F., Faghihi,M.A., Patel,N.S., Sahagan,B.G., Wahlestedt,C. and Lopez-Toledano,M.A. (2011) Knockdown of BACE1-AS nonprotein-coding transcript modulates beta-amyloid-related hippocampal neurogenesis. *Int. J. Alzheimers Dis.*, **2011**, 929042.
34. Scheele,C., Petrovic,N., Faghihi,M.A., Lassmann,T., Fredriksson,K., Rooyackers,O., Wahlestedt,C., Good,L. and Timmons,J.A. (2007) The human PINK1 locus is regulated *in vivo* by a non-coding natural antisense RNA during modulation of mitochondrial function. *BMC Genomics*, **8**, 74.
35. Stazic,D., Lindell,D. and Steglich,C. (2011) Antisense RNA protects mRNA from RNase E degradation by RNA-RNA duplex formation during phage infection. *Nucleic Acids Res.*, **39**, 4890–4899.
36. Wery,M., Descrimes,M., Vogt,N., Dallongeville,A.S., Gautheret,D. and Morillon,A. (2016) Nonsense-mediated decay restricts LncRNA Levels in yeast unless blocked by double-stranded RNA structure. *Mol. Cell*, **61**, 379–392.
37. Portal,M.M., Pavet,V., Erb,C. and Gronemeyer,H. (2015) Human cells contain natural double-stranded RNAs with potential regulatory functions. *Nat. Struct. Mol. Biol.*, **22**, 89–97.
38. Beltran,M., Puig,I., Pena,C., Garcia,J.M., Alvarez,A.B., Pena,R., Bonilla,F. and de Herreros,A.G. (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev.*, **22**, 756–769.
39. Ebralidze,A.K., Guibal,F.C., Steidl,U., Zhang,P., Lee,S., Bartholdy,B., Jorda,M.A., Petkova,V., Rosenbauer,F., Huang,G. *et al.* (2008) PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. *Genes Dev.*, **22**, 2085–2092.
40. Tufarelli,C., Stanley,J.A., Garrick,D., Sharpe,J.A., Ayyub,H., Wood,W.G. and Higgs,D.R. (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.*, **34**, 157–165.
41. Li,K. and Ramchandran,R. (2010) Natural antisense transcript: a concomitant engagement with protein-coding transcript. *Oncotarget*, **1**, 447–452.
42. Halley,P., Khorkova,O. and Wahlestedt,C. (2013) Natural antisense transcripts as therapeutic targets. *Drug Discov. Today Ther. Strateg.*, **10**, e119–e125.
43. Modarresi,F., Faghihi,M.A., Lopez-Toledano,M.A., Fatemi,R.P., Magistri,M., Brothers,S.P., van der Brug,M.P. and Wahlestedt,C. (2012) Inhibition of natural antisense transcripts *in vivo* results in gene-specific transcriptional upregulation. *Nat. Biotechnol.*, **30**, 453–459.
44. Conley,A.B. and Jordan,I.K. (2012) Epigenetic regulation of human cis-natural antisense transcripts. *Nucleic Acids Res.*, **40**, 1438–1445.
45. Struhl,K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, **14**, 103–105.
46. Werner,A. and Sayer,J.A. (2009) Naturally occurring antisense RNA: function and mechanisms of action. *Curr. Opin. Nephrol. Hypertens.*, **18**, 343–349.
47. Cui,I. and Cui,H. (2010) Antisense RNAs and epigenetic regulation. *Epigenomics*, **2**, 139–150.
48. Nishizawa,M., Okumura,T., Ikeya,Y. and Kimura,T. (2012) Regulation of inducible gene expression by natural antisense transcripts. *Front. Biosci. (Landmark Ed)*, **17**, 938–958.
49. Faghihi,M.A., Modarresi,F., Khalil,A.M., Wood,D.E., Sahagan,B.G., Morgan,T.E., Finch,C.E., St Laurent,G. 3rd, Kenny,P.J. and Wahlestedt,C. (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.*, **14**, 723–730.
50. Morris,K.V., Santoso,S., Turner,A.M., Pastori,C. and Hawkins,P.G. (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.*, **4**, e1000258.
51. Yu,W., Gius,D., Onyango,P., Muldoon-Jacobs,K., Karp,J., Feinberg,A.P. and Cui,H. (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature*, **451**, 202–206.
52. Michael,D.R., Phillips,A.O., Krupa,A., Martin,J., Redman,J.E., Altaher,A., Neville,R.D., Webber,J., Kim,M.Y. and Bowen,T. (2011) The human hyaluronan synthase 2 (HAS2) gene and its natural antisense RNA exhibit coordinated expression in the renal proximal tubular epithelial cell. *J. Biol. Chem.*, **286**, 19523–19532.
53. Khalil,A.M., Faghihi,M.A., Modarresi,F., Brothers,S.P. and Wahlestedt,C. (2008) A novel RNA transcript with antiapoptotic function is silenced in fragile X syndrome. *PLoS One*, **3**, e1486.
54. Chen,D., Yuan,C., Zhang,J., Zhang,Z., Bai,L., Meng,Y., Chen,L.L. and Chen,M. (2012) PlantNATsDB: a comprehensive database of plant natural antisense transcripts. *Nucleic Acids Res.*, **40**, D1187–D1193.
55. Zhang,Y., Li,J., Kong,L., Gao,G., Liu,Q.R. and Wei,L. (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res.*, **35**, D156–D161.
56. Kim,D.S., Cho,C.Y., Huh,J.W., Kim,H.S. and Cho,H.G. (2009) EVOG: a database for evolutionary analysis of overlapping genes. *Nucleic Acids Res.*, **37**, D698–D702.
57. Yin,Y., Zhao,Y., Wang,J., Liu,C., Chen,S., Chen,R. and Zhao,H. (2007) antiCODE: a natural sense-antisense transcripts database. *BMC Bioinformatics*, **8**, 319.
58. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
59. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
60. Suzuki,A., Wakaguri,H., Yamashita,R., Kawano,S., Tsuchihara,K., Sugano,S., Suzuki,Y. and Nakai,K. (2015) DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res.*, **43**, D87–D91.
61. Yamashita,R., Sugano,S., Suzuki,Y. and Nakai,K. (2012) DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. *Nucleic Acids Res.*, **40**, D150–D154.
62. Yamashita,R., Sathira,N.P., Kanai,A., Tanimoto,K., Arauchi,T., Tanaka,Y., Hashimoto,S., Sugano,S., Nakai,K. and Suzuki,Y. (2011) Genome-wide characterization of transcriptional start sites in

humans by integrative transcriptome analysis. *Genome Res.*, **21**, 775–789.
63. Toribio,A.L., Alako,B., Amid,C., Cerdeno-Tarraga,A., Clarke,L., Cleland,I., Fairley,S., Gibson,R., Goodgame,N., Ten Hoopen,P. *et al.* (2017) European Nucleotide Archive in 2016. *Nucleic Acids Res.*, **45**, D32–D36.
64. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
65. Andrews,S. (2010) FastQC: a quality control tool for high throughput sequence data. Availible on-line at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
66. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
67. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
68. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
69. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
70. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
71. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
72. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
73. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
74. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
75. Tan,G. and Lenhard,B. (2016) TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, **32**, 1555–1556.
76. Down,T.A., Piipari,M. and Hubbard,T.J. (2011) Dalliance: interactive genome viewing on the web. *Bioinformatics*, **27**, 889–890.
77. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
78. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
79. Coordinators,N.R. (2017) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **45**, D12–D17.