D50–D54 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 24 October 2018*
doi: 10.1093/nar/gky986

CoevDB: a database of intramolecular coevolution among protein-coding genes of the bony vertebrates

Xavier Meyer[1,2,*], Linda Dib[3] and Nicolas Salamin[1,3,*]

[1]Department of Computational Biology, University of Lausanne, Biophore, 1015 Lausanne, Switzerland, [2]Department of Integrative Biology, University of California, 3060 Valley Life Sciences Bldg, Berkeley, CA 94720-3140, USA and [3]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

Received August 14, 2018; Revised September 17, 2018; Editorial Decision October 05, 2018; Accepted October 10, 2018

ABSTRACT

The study of molecular coevolution, due to its potential to identify gene regions under functional or structural constraints, has recently been subject to numerous scientific inquiries. Particular efforts have been conducted to develop methods predicting the presence of coevolution in molecular sequences. Among these methods, a few aim to model the underlying evolutionary process of coevolution, which enable to differentiate the shared history of genes to coevolution and thus improve their accuracy. However, the usage of such methods remains sparse due to their expensive computational cost and the lack of resources alleviating this issue. Here we present CoevDB (http://phylodb.unil.ch/CoevDB), a database containing the result of a large-scale analysis of intramolecular coevolution of 8201 protein-coding genes of bony vertebrates. The web interface of CoevDB gives access to the results to 800 millions of statistical tests corresponding to all the pairs of sites analyzed. Several type of queries enable users to explore the database by either targeting specific genes or by discovering genes having promising estimations of coevolution.

INTRODUCTION

Molecular coevolution is the evolutionary process by which interactions between distant sites of one or multiple molecules (RNA or proteins) are maintained such as to preserve advantageous functional or structural constraints (1). For instance, these interactions occur within the 16S ribosomal RNA gene to preserve its structure stability (2,3) and within proteins to maintain binding specificity and folding constraints (4). Studying coevolution at the molecular level has hence shown to produce valuable information on key mutations having known roles in genetic diseases (4) or viruses (5).

Predicting coevolution from multiple sequence alignments has therefore large potentials and numerous approaches have been developed. The bulk of these methods relies on looking for patterns of coevolution in amino acids sequences by using statistic tests to predict coevolving residues (6–10). Several web services have been built to facilitate the use of these predictive methods, compare their results and/or provide highly detailed informations such as the mapping of the predicted coevolving residues on a reference structure provided as PDB file (11–13). However, these methods do not consider the underlying (co)evolutionary process and therefore the shared ancestry of genes, which negatively impact their accuracy (14,15).

A smaller subset of methods incorporate this crucial information by explicitly modeling the evolutionary process of coevolution along a phylogenetic tree provided as input data (16–18). The goal is to differentiate double substitutions due to coevolution from those due to the common evolutionary history of the sequences. These approaches, however, come with the drawback of increasing the computational cost of the analyses. A solution could be to provide access to remote computing resources as it was recently done for the Coev method of Dib *et al.* (18) by providing a web service for the analyses (19). However, this approach falls short to provide an efficient tool for a large-scale study of coevolution with a phylogeny-aware method.

In this article, we present CoevDB, a database containing the result of a large-scale analysis of intramolecular coevolution in 8201 protein-coding genes of the bony vertebrates with the Coev method. To our knowledge, CoevDB is the first database to contain the results of a systematic analysis of pair-wise coevolution estimations obtained using a phylogeny-aware method. These characteristics makes it unique with respect to previously existing databases on coevolution such as the InterEvol database (20), which contains predictions of interfaces coevolution inferred from the known structure of protein complexes, or the Prolinks database (21), which contains prediction of inter-protein coevolution inferred using simple indicators (e.g. gene neighborhood). In addition of being unique with respect to its

*To whom correspondence should be addressed. Tel: +1 510 365 0057; Email: xav.meyer@gmail.com
Correspondence may also be addressed to Nicolas Salamin. Tel: +41 21 692 4154; Fax: +41 21 692 4165; Email: nicolas.salamin@unil.ch

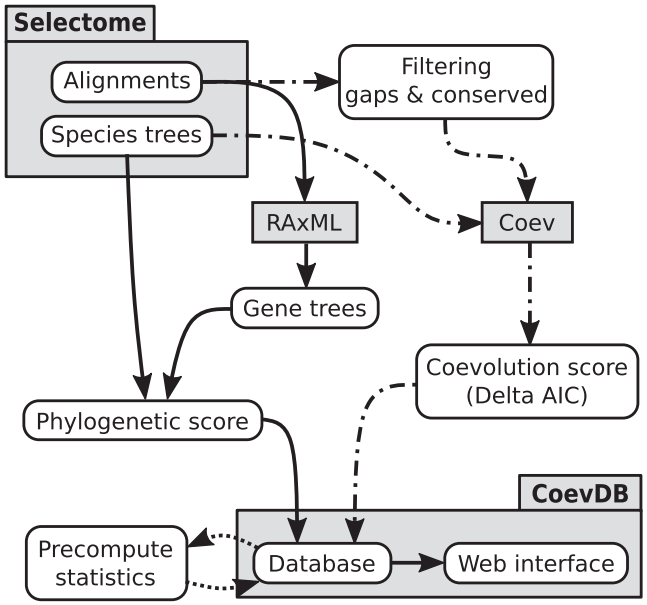

Figure 1. CoevDB flowchart.

content, CoevDB interface is designed such as to ease the browsing of the 800 millions of pairs of sites tested and to provide an informative display of these sites along their phylogenetic tree.

DATA AND METHODS

CoevDB contains predictions of co-evolution within genes obtained for 8201 protein coding genes from the Euteleostomi clade (Figure 1). The alignments and phylogenies employed for this large-scale analysis have been obtained from the Selectome database release 6 (22), itself based on the Ensembl 2012 database (23). We selected all the alignments having between 100 and 20 000 nucleotides per sequence in Selectome, which resulted in alignments containing from 21 to 757 sequences (median alignment length of 325 nucleotides after filtering of the gaped and fully conserved sites).

We tested for coevolution all the possible pairs of sites within each of these alignments using the approach implemented in the Coev software (18). Briefly, this method compares two statistical models. The first hypothesis assumes that sites along a sequence evolve independently according to the Jukes and Cantor model (24). The second hypothesis tested assumes that both sites are dependent and co-evolve according to the Coev model, such that the nucleotides at both sites remain within a predefined set of nucleotides combinations called the coevolution profile (18). A substitution in one site of a coevolving pairs is expected to trigger a substitution at the other site to maintain the coevolution profile. After maximizing the likelihood of the data observed at the pair of sites, the two models are compared using the difference in their respective value of the Akaike Information Criterion (DAIC), which indicates the support for the hypothesis of coevolution against the one of independent evolution.

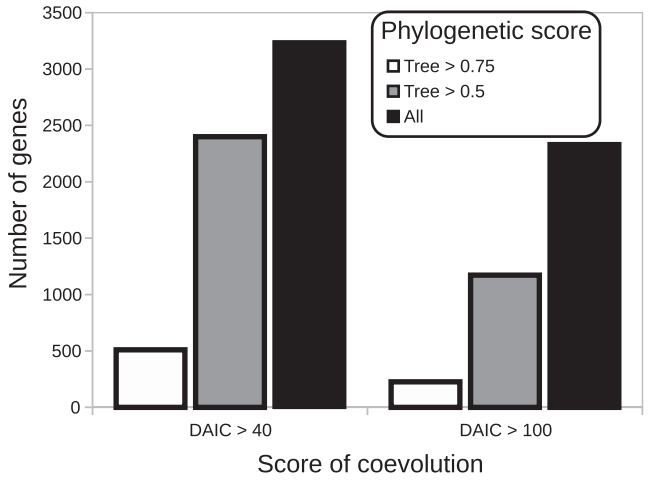

Figure 2. Number of genes having one or more pairs of sites with strong support for the coevolution hypothesis.

In addition to the alignment, the Coev method assumes that the phylogenetic tree is known. We employed the species trees from the Ensembl database for the genes considered. These trees were inferred using gene concatenation (23) and were employed for the analyses of positive selection conducted for the Selectome database (22). Employing these species trees, instead of gene trees, had the advantage of extending the range of genes for which we could estimate coevolution. Indeed, genes, for which the amount of molecular data wouldn't allow to adequately infer gene trees, should have then been discarded from our analyses. The downside of employing species trees is that they might not represent accurately the evolutionary history of each of these individual genes, which can negatively impact the prediction of coevolution.

We therefore compared the species trees from Ensembl to each individual gene trees inferred using RAxML (25). We defined a score to measure the similarity between the species tree and the gene tree. This phylogenetic score combines the normalized Robinson–Foulds distance d (26) between both trees and the average bootstrap support s for the gene tree as $(1 - d)*s$. We report this score for each gene in the web interface along coevolution estimations as an indication on the degree of confidence to put in these results with respect to the species tree employed.

DATABASE AND WEB INTERFACE

CoevDB contains the estimation of coevolution for ∼800 millions pairs of sites under all their potential coevolution profiles (which can range from 1 to 192). The resulting 12 billions DAIC scores are stored in a MySQL database. The computing time of these analyses amounted to the equivalent of 650 years of computations on a single processor, but the analyses were done in more than one year on the BlueGen/Q infrastructure of CADMOS (www.cadmos.org). Among the millions of pairs tested for coevolution, approximately 400 000 showed a strong support for the Coev model (DAIC > 40). These pairs were found in 3224 different genes (Figure 2). The phylogenetic score revealed that 16% (510) of these genes had a strong support for their

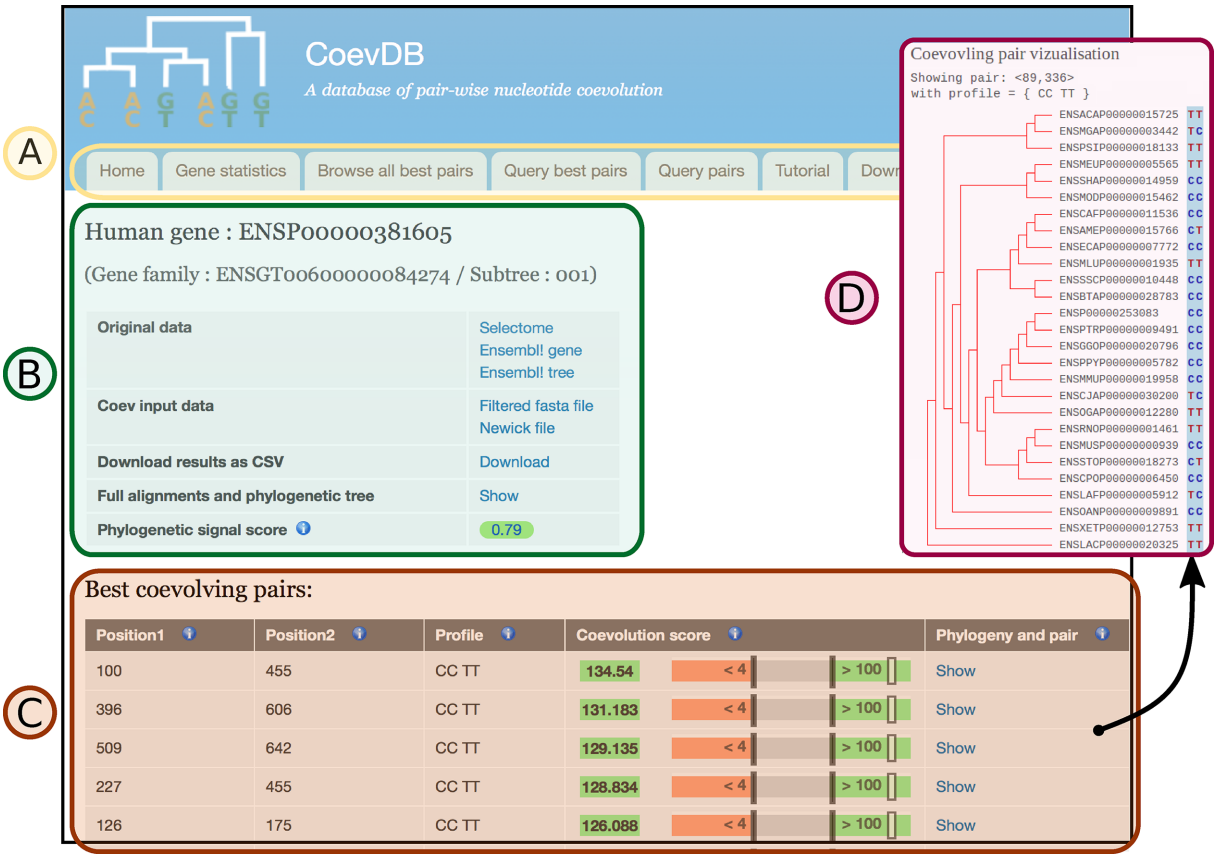

Figure 3. Overview of the result of a query on the ENSP00000381605 gene. (**A**) Access to the different type of queries and pages of the website. (**B**) Information about gene including the phylogenetic score and links to Ensembl and Selectome. (**C**) Estimations of coevolution with visual indicators for the scores. (**D**) Pop-up visualization of the phylogeny with a pair of sites resulting from a user action.

species tree (score > 0.75) and 74% (2400) had a good support for their species tree (score > 0.5). Furthermore, over the 2330 genes that had pairs of sites with highly significant DAIC score (DAIC > 100), 226 genes had a strong similarity between their species and gene trees. In summary, 29% of the 8201 genes analyzed had at least one pair of sites with strong support for the coevolution hypothesis inferred with an adequate phylogenetic tree and 3% of them had a highly significant support for both the coevolution hypothesis and the phylogeny employed.

Among the 29% of genes having high phylogenetic score (score > 0.75) and strong estimations of coevolution (DAIC > 40), some are coding for proteins having an important role in the bony vertebrates such as the establishment and function of cell-cell neural connections (Human gene name in Ensembl and CoevDB: ENSP00000023113; (27)) or the olfactory system (ENSP00000323606; (28)). Some others are known to be linked with mutations leading to human disease or disability. For instance, the BTB domain containing seven genes is known to play a role in various cancer (ENSP00000335615; (29)), the huntingtin interacting protein 1 gene is associated with worse survival in some lymphoma patients (ENSP00000253083; (30)) and the lysine demethylase 5C is associated with X-linked cognitive disability (ENSP00000364550; (31)).

These estimations of coevolution are accessible through the web interface designed for CoevDB (Figure 3). In addition of some tutorials (e.g. about molecular coevolution), this web interface enables the user to browse the database according to two different use-cases: (i) by searching for genes having significant estimations of coevolution (*blind query*) and (ii) by reporting the coevolution estimations for a specific gene (*targeted query*). A *blind query* can be achieved by accessing the *Gene statistics* page where all the genes are ranked by their amount of coevolving pairs of sites having either a medium or strong DAIC score. A *targeted query* requires the user to input the Ensembl name of the gene of interest (e.g. ENSP00000231134) in either the *Browse all best pairs*, the *Query best pairs* or the *Query pairs* pages. The first page returns the estimations of coevolution potentially significant (DAIC > 25) under the best profile of coevolution for all the pairs of sites. The second page provides filtering options (site or pair of sites) to the user on the same potentially significant estimations of coevolution. The last page returns the coevolution estimations regardless of their statistical significance for all the coevolving profiles analyzed for a given set of sites pairs.

After a query, the user is directed to the result page that contains information about the queried gene in addition to the estimations of the coevolving sites. These information include the phylogenetic score, link to the Selectome and

Ensembl original data as well as the alignment and phylogenetic tree employed for the analyses with Coev. On this page, the phylogenetic tree and the whole alignment can be simultaneously displayed upon request by the user. The display of these information is achieved using a modified version of SnipViz (32). Finally, the pairs of position corresponding to the query are listed along with their coevolution profile and DAIC score. As for the whole alignment, a specific pair of sites can be displayed along with the phylogenetic tree.

CONCLUSION

CoevDB provides access to results of intramolecular coevolution in 8,201 protein-coding genes of the bony vertebrates and is, to our knowledge, the first database containing a large-scale analysis of coevolution with a method modeling the coevolutionary process (Coev; (18)). The use of the Coev method makes it a unique tool with respect to existing coevolution database (20,21) that differs in the method employed and the type of coevolution estimated (intramolecular versus intermolecular). The robustness of the statistical tests employed in the Coev method and its modeling of the (co)evolutionary process along the phylogeny results in a more accurate but computationally expensive analysis. CoevDB addresses this issue by providing coevolution estimations for 800 millions of pairs of sites, or the equivalent of 650 years of computations on a single processor, that are directly available through a custom web interface. This web interface accommodates for multiple types of query on the database, each tailored for a specific usage, and enables to visualize in the results along the phylogenetic tree representing the shared history of the gene using the SnipViz framework (32).

We expect CoevDB to be useful both for functional and evolutionary studies. This large-scale analysis provides an overall estimation on the abundance of coevolution within protein-coding genes of the bony vertebrates. Furthermore, ranking genes by their precomputed estimates of coevolution score enables any user to rapidly identify genes that have potentially been the target of functional or structural constraints through their evolution. Finally, CoevDB has been designed to cover molecular dataset containing human genes and its web interface enable a user to rapidly assess if specific region of a human gene known to play a role in disabilities or diseases are likely to be subject to intramolecular coevolution.

The future development of CoevDB will follow two main directions. The first will be to broaden the range of analyses conducted by employing alignments from recent version of Ensembl (e.g. (33)), extending the range of species and considering the analysis of intergenic coevolution. The second is to further increase the confidence of the coevolution estimations contained in CoevDB. This goal could be achieved by taking advantage of known protein structures for the coding genes analyzed in CoevDB. The distance between residues of a protein could be compared with the coevolution score to validate potential structural constraints. Both directions would further contribute to make of CoevDB a central resources containing readily available and reliable estimations of coevolution based on the analysis of the underlying (co)evolutionary process.

ACKNOWLEDGEMENTS

The computations were performed on the BlueGen/Q infrastructure of CADMOS (www.cadmos.org).

FUNDING

Swiss National Science Foundation [CR32I3_143768, P2GEP2_178032]. Funding for open access charge: Swiss National Science Foundation.
Conflict of interest statement. None declared.

REFERENCES

1. Carmona,D., Fitzpatrick,C.R. and Johnson,M.T.J. (2015) Fifty years of co-evolution and beyond: integrating co-evolution from molecules to species. *Mol. Ecol.*, **24**, 5315–5329.
2. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Müller,K.M. *et al.* (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
3. Dutheil,J.Y., Jossinet,F. and Westhof,E. (2010) Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Mol. Biol. Evol.*, **27**, 1868–1876.
4. Dib,L. and Carbone,A. (2012) Protein Fragments: Functional and structural roles of their coevolution networks. *PLoS One*, **7**, e48124.
5. Douam,F., Fusil,F., Enguehard,M., Dib,L., Nadalin,F., Schwaller,L., Hrebikova,G., Mancip,J., Mailly,L., Montserret,R. *et al.* (2018) A protein coevolution method uncovers critical features of the Hepatitis C Virus fusion mechanism. *PLoS Pathog.*, **14**, e1006908.
6. Dunn,S.D., Wahl,L.M. and Gloor,G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.
7. Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1293–E1301.
8. Jones,D.T., Buchan,D. W.A., Cozzetto,D. and Pontil,M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.
9. Dib,L. and Carbone,A. (2012) CLAG: an unsupervised non hierarchical clustering algorithm handling biological data. *BMC Bioinformatics*, **13**, 194.
10. Champeimont,R., Laine,E., Hu,S.-W., Penin,F. and Carbone,A. (2016) Coevolution analysis of *Hepatitis C* virus genome to identify the structural and functional dependency network of viral proteins. *Sci. Rep.*, **6**, 26401.
11. Iserte,J., Simonetti,F.L., Zea,D.J., Teppa,E. and Marino-Buslje,C. (2015) I-COMS: Interprotein-COrrelated mutations server. *Nucleic Acids Res.*, **43**, W320–W325.
12. Oteri,F., Nadalin,F., Champeimont,R. and Carbone,A. (2017) BIS2Analyzer: a server for co-evolution analysis of conserved protein families. *Nucleic Acids Res.*, **45**, W307–W314.
13. Colell,E.A., Iserte,J.A., Simonetti,F.L. and Marino-Buslje,C. (2018) MISTIC2: comprehensive server to study coevolution in protein families. *Nucleic Acids Res.*, **46**, W323–W328.
14. Dutheil,J.Y. (2012) Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinformatics*, **13**, 228–243.
15. Talavera,D., Lovell,S.C. and Whelan,S. (2015) Covariation is a poor measure of molecular coevolution. *Mol. Biol. Evol.*, **32**, 2456–2468.
16. Dutheil,J., Pupko,T., Jean-Marie,A. and Galtier,N. (2005) A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.*, **22**, 1919–1928.
17. Yeang,C.-H. and Haussler,D. (2007) Detecting coevolution in and among protein domains. *PLoS Comput. Biol.*, **3**, e211.
18. Dib,L., Silvestro,D. and Salamin,N. (2014) Evolutionary footprint of coevolving positions in genes. *Bioinformatics*, **30**, 1241–1249.
19. Dib,L., Meyer,X., Artimo,P., Ioannidis,V., Stockinger,H. and Salamin,N. (2015) Coev-web: a web platform designed to simulate

and evaluate coevolving positions along a phylogenetic tree. *BMC Bioinformatics*, **16**, 394.
20. Faure,G., Andreani,J. and Guerois,R. (2012) InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res.*, **40**, D847–D856.
21. Bowers,P.M., Pellegrini,M., Thompson,M.J., Fierro,J., Yeates,T.O. and Eisenberg,D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.
22. Moretti,S., Laurenczy,B., Gharib,W.H., Castella,B., Kuzniar,A., Schabauer,H., Studer,R.A., Valle,M., Salamin,N., Stockinger,H. *et al.* (2014) Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Res.*, **42**, D917–D921.
23. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
24. Jukes,T.H. and Cantor,C.R. *et al.* (1969) Evolution of protein molecules. *Mamm. Protein Metab.*, **3**, 132.
25. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
26. Robinson,D. and Foulds,L. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.
27. Frank,M. and Kemler,R. (2002) Protocadherins. *Curr. Opin. Cell Biol.*, **14**, 557–562.
28. Rouquier,S., Blancher,A. and Giorgi,D. (2000) The olfactory receptor gene repertoire in primates and mouse: evidence fo reduction of the functional fraction in primates. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 2870–2874.
29. Fang,L.-Z., Zhang,J.-Q., Liu,L., Fu,W.-P., Shu,J.-K., Feng,J.-G. and Liang,X. (2017) Silencing of Btbd7 Inhibited Epithelial-Mesenchymal Transition and Chemoresistance in CD133+ Lung Carcinoma A549 Cells. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, **25**, 819–829.
30. Wong,K.K., Ch'ng,E.S., Loo,S.K., Husin,A., Muruzabal,M.A., Møller,M.B., Pedersen,L.M., Pomposo,M.P., Gaafar,A., Banham,A.H. *et al.* (2015) Low HIP1R mRNA and protein expression are associated with worse survival in diffuse large B-cell lymphoma patients treated with R-CHOP. *Exp. Mol. Pathol.*, **99**, 537–545.
31. Peng,Y., Suryadi,J., Yang,Y., Kucukkal,T.G., Cao,W. and Alexov,E. (2015) Mutations in the KDM5C ARID domain and their plausible association with syndromic Claes-Jensen-Type disease. *Int. J. Mol. Sci.*, **16**, 27270–27287.
32. Jaschob,D., Davis,T.N. and Riffle,M. (2014) SnipViz: a compact and lightweight web site widget for display and dissemination of multiple versions of gene and protein sequences. *BMC Res. Notes*, **7**, 468.
33. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G *et al.* (2017) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.