MATTERS ARISING
https://doi.org/10.1038/s42256-021-00371-6

nature machine intelligence

Reusability report: Compressing regulatory networks to vectors for interpreting gene expression and genetic variants

Wanwen Zeng[1,2,7], Jingxue Xin[2,7], Rui Jiang[3] ✉ and Yong Wang[4,5,6] ✉

ARISING FROM Qin Cao et al. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-020-0205-2 (2020).

Integrating multi-omics data to better interpret transcription control and reveal regulatory mechanisms is of fundamental importance. Usually, high-dimensional data are mathematically represented and modelled in a biological network in which nodes represent biological units and edges represent the interactions between the units. Recent progress in representation learning has demonstrated the possibility of embedding heterogeneous networks with multiple types of nodes and links in low-dimensional vector space[1]. In particular, Cao et al. have utilized a state-of-the-art embedding method, GEEK ('Gene Expression Embedding frameworK'), to combine biological networks and omics data with the metapath concept[1], and have produced interpretable biological knowledge such as gene function, protein complex, chromatin domain and replication timing[2].

To demonstrate the robustness and re-usability of the embedding framework, we carried out two different downstream tasks that are complementary to the GEEK study: (1) integrating the regulatory information embedded in vectors generated by GEEK to regress the gene expression level in K562 cells using DeepExpression[3] and (2) incorporating an attention score based on GEEK embedding vectors to prioritize genetic variants for high-altitude adaptation around the *EPAS1* region in human umbilical vein endothelial cells (HUVECs), as also identified by vPECA ('variants interpretation method by paired expression and chromatin accessibility') in a previous publication[4]. Briefly, DeepExpression is a densely connected convolutional neural network for integrating DNA sequence information and enhancer–promoter interaction data to model gene expression, and vPECA is a variant interpretation method for identifying active selected regulatory elements (REs) and the associated regulatory network. Our objective is to evaluate the regulatory information in GEEK embedding vectors by investigating whether the performance of those methods can be improved with the incorporation of the vectors. The results show that GEEK embedding vectors are informative for predicting gene expression and potentially useful in prioritizing genetic variants. Applications using the embedding vector from GEEK should be carefully interpreted with consideration of their context-specific and non-specific information.

Predicting gene expression using DeepExpression

We first systematically compared the performances of GEEK embedding vectors, sequence-based embedding vectors and three-dimensional (3D) HiChIP interactions in the task of regressing gene expression level in K562 cells using the DeepExpression model[3]. Given feature vector $\mathbf{x}$ and gene expression value $\mathbf{y}$, we solved the following optimization model (1) to fit a complex hierarchical function $f(\mathbf{x}_i;\mathcal{W})$ by determining the collection of weights $\mathcal{W}$:

$$\min_{\mathcal{W}}\{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\mathcal{W}))^2 + \lambda J(\mathcal{W})\} \quad (1)$$

$J(\mathcal{W})$ is a non-negative term on elements of $\mathcal{W}$ with various types of regularization[3] and $\lambda \geq 0$ is a tuning parameter.

Feature vector $\mathbf{x}$ can be constructed in three ways (Fig. 1) (throughout, for information on the code and data used here, see the 'Code availability' and 'Data availability' sections). For the GEEK embedding vectors ('Embedding' in Table 1), we retrained GEEK without gene expression level to obtain a 96D real vector for each gene as input for fair comparison. We confirmed that the pre-trained embedding vectors were obtained without gene expression data to avoid possible information leakage. This is equivalent to an ablation study with GEEK using the N + A strategy with network information (N) and DNase I hypersensitivity attributes (A). For sequence-based embedding vectors ('Sequence' in Table 1) we extracted the DNA fragments of 2,000 base pairs (bp) around the transcription start site (TSS) of a gene as its promoter region, utilized the encoding layer to encode the nucleotide in each position as a 4D one-hot binary vector, and obtained a 4 × 2,000 binary matrix as input. For 3D HiChIP interactions ('HiChIP' in Table 1), following DeepExpression, we took signals (loop counts in HiChIP experiments) for bins to form a 400D input vector. As shown in Table 1, all three features (Sequence, Embedding and HiChIP) from the different approaches are predictive for gene expression in K562 cells (Pearson correlation coefficient (PCC) score of >0.4 for all cross-validation experiments; comparisons were carried out

[1]College of Software, Nankai University, Tianjin, China. [2]Department of Statistics, Department of Biomedical Data Science, Bio-X Program, Stanford University, Stanford, CA, USA. [3]Ministry of Education Key Laboratory of Bioinformatics, Research Department of Bioinformatics at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing, China. [4]CEMS, NCMIS, HCMS, MDIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. [5]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. [6]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou, China. [7]These authors contributed equally: Wanwen Zeng, Jingxue Xin. ✉e-mail: ruijiang@tsinghua.edu.cn; ywang@amss.ac.cn

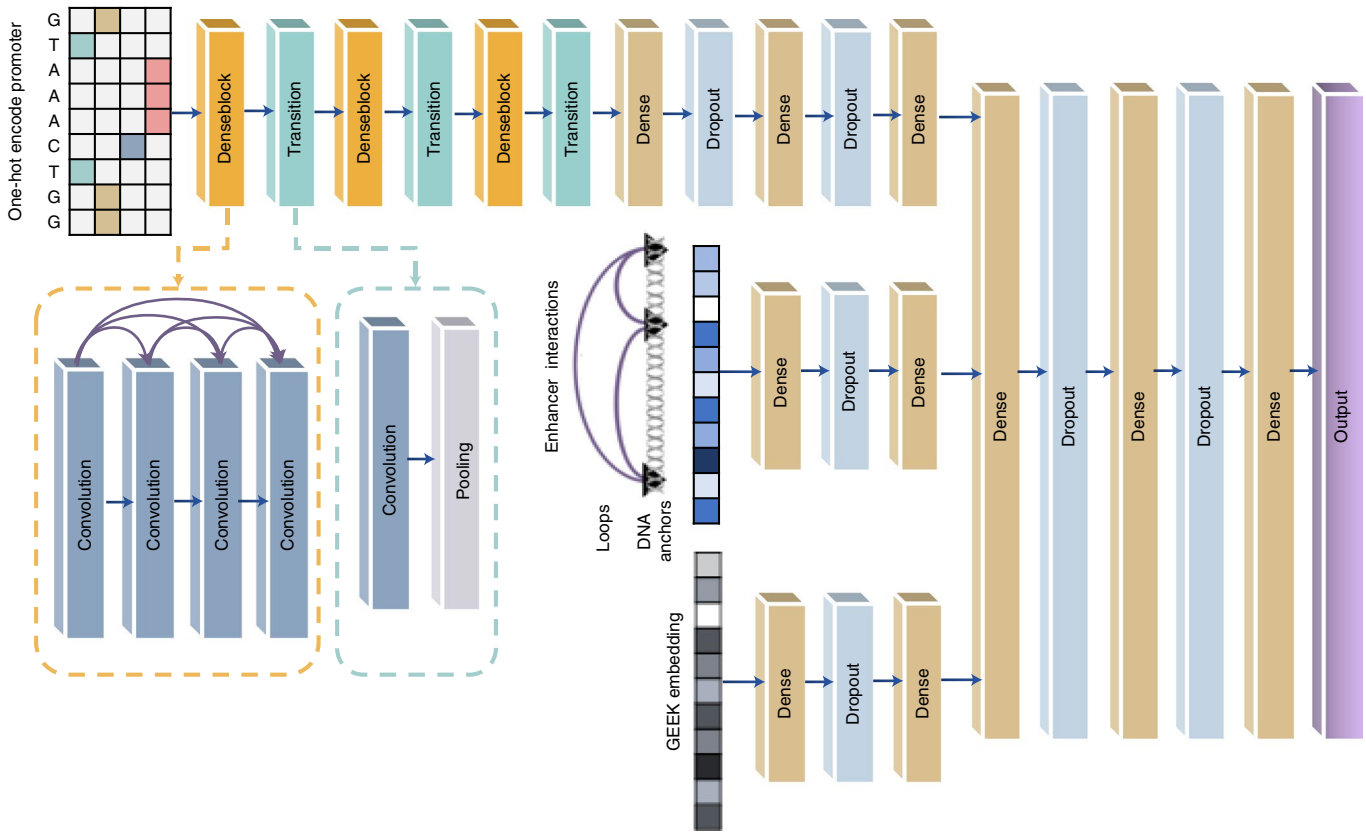

Fig. 1 | Modified DeepExpression framework with GEEK embedding vectors. The modified DeepExpression framework consists of three modules: (1) sequence-based embedding vectors, which take the one-hot promoter sequence as input, (2) 3D HiChIP interactions, which take the HiChIP signal as input and (3) GEEK embedding vectors, which take the GEEK vectors as input. The network structures of the Sequence and HiChIP modules remain the same as the original DeepExpression, while the GEEK module is the same as the HiChIP interaction module.

Table 1 | Regression performance of different combinations of inputs measured as PCC in 10-fold cross-validation experiments

Input	Mean PCC score
Sequence + Embedding	0.6837
Sequence + HiChIP	0.6552
HiChIP	0.4396
Embedding	0.5142
Sequence	0.5982

PCC is calculated between gene expression levels from RNA-seq and predicted values from different types of feature.

with the same cross-validation process). Among these features, sequence-based embedding vectors achieved higher performance than GEEK embedding vectors and 3D interactions. The combination of GEEK embedding vectors and the other two types of feature further improved the regression performance, demonstrating the importance of DNA sequence and 3D interactions. We performed additional experiments to embed the vector by only using network information (N) and only using DNase I hypersensitivity attributes (A), and compared these two strategies with our current implementation (N + A). We substituted the GEEK Embedding module from N + A separately into N and A. We found that the combination of the Sequence module (S) and GEEK Embedding module (N + A) yielded the best performance, with a PCC score of 0.652. This confirms that the combination of these two types of feature (N + A) improves the performance. With the Sequence module, the network information (S + N) and DNase I hypersensitivity (S + A) attributes performed similarly in predicting the gene expression level, with PCC scores of 0.612 and 0.613. This is slightly different from the ablation results in supplementary fig. 14 of ref. [2] after combination with the Sequence module. Simply dropping gene expression data and chromatin accessibility data from the GEEK learning leads to reasonable performance by leveraging the power of the Sequence module. Indeed, with the Sequence module, incorporation of the context-specific data of GEEK mildly outperformed the foundational metapath2vec model[1].

Prioritizing genetic variants for high-altitude adaptation

We next utilized the regulatory information embedded in GEEK vectors to prioritize genetic variants. The idea is to leverage the regulatory impact on gene regulation embedded in the omics data of relevant cell types to correlate large-scale phenotype-associated noncoding variants. We used the attention score[5] to summarize an embedded vector as a regulatory score, and then associated this score with variants. In detail, the embedded vectors were obtained from GEEK pre-trained vectors in normal HUVEC cells. Let X be a $p \times N$ embedding matrix, where N denotes the number of genes and p the length of the embedding vector ($p = 128$). Let $\mathbf{x}_i$ be the ith column of X, representing the embedding vector of gene i. Similarly, let W be the embedding matrix with $p \times M$ dimension, where M denotes the total number of 10-kb windows. Let $\mathbf{w}_j$ be the jth column of W, denoting the embedding vector of window j. The attention score of the ith gene and the jth window, α_{ij}, is calculated as

$$\alpha_{ij} = \frac{\exp\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_j\right)}{\sum_{k=1}^{M}\exp\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_k\right)} \quad (2)$$

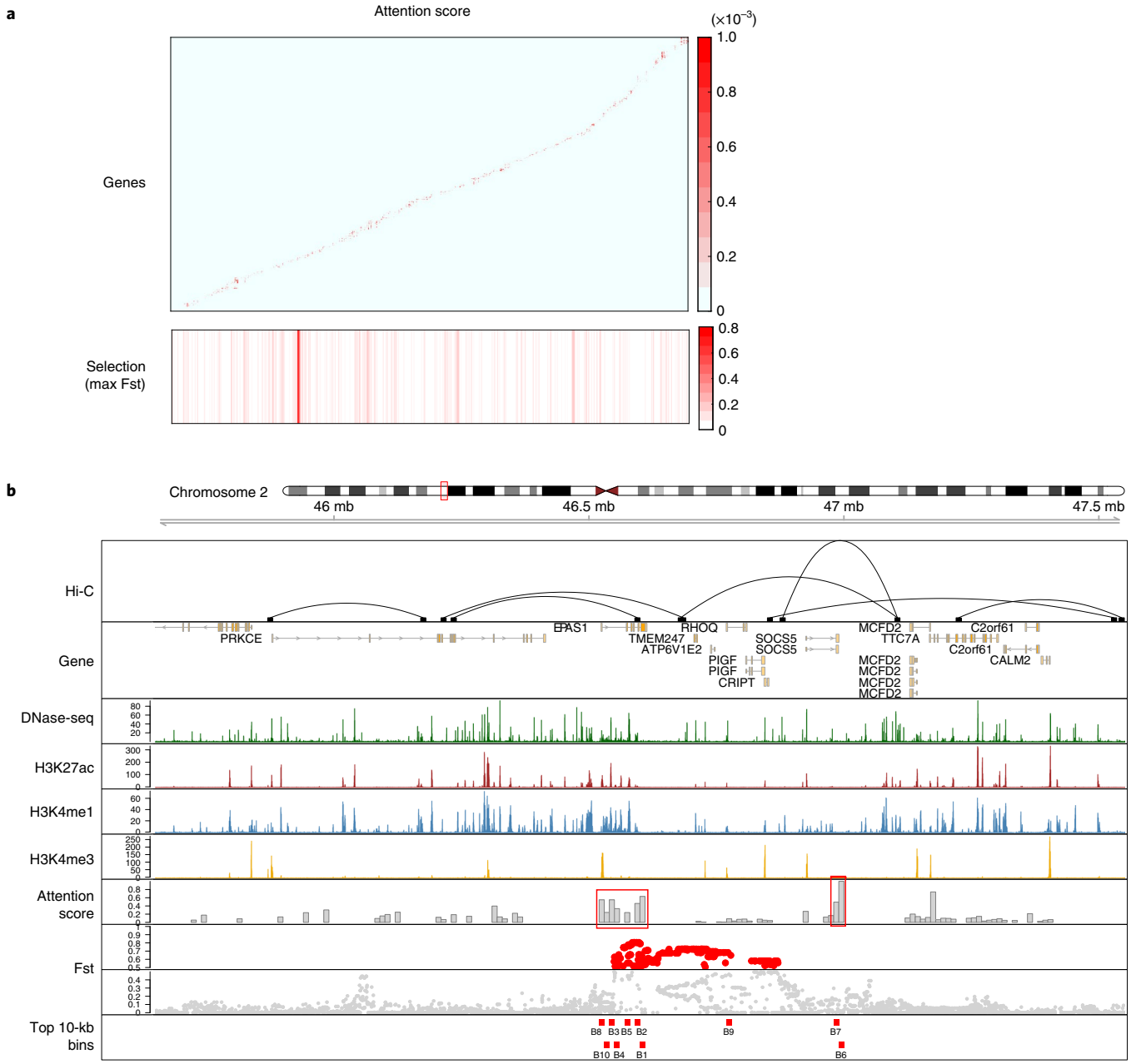

Fig. 2 | Identifying key regions around *EPAS1* in HUVEC cells for high-altitude adaptation by attention score. a, Attention score for all genes and their neighbouring 10-kb windows on chromosome 2. Top: attention scores (as a heatmap) for each gene and all 10-kb windows within [−1M, +1M] of the TSS. Bottom: the selection score for each 10-kb window, calculated as the maximum Fst score of all SNPs within the window. All genes and windows are sorted by genomic coordinate. **b**, Multi-omics data around *EPAS1* in HUVEC cells. Hi-C loops, epigenomic marks from ENCODE, attention score, selection score (Fst) for each SNP and the top 10 selected 10-kb windows are visualized in the [−1M, +1M] region from the TSS of *EPAS1*. The *y* axes of traces for DNase-seq, H3K27ac, H3K4me1 and H3K4me3 represent $-\log_{10} P$ values. The attention scores (ASs) are calculated using equation (2) and are normalized by the maximum value, that is, normalized AS = AS/max(AS). The top 10 bins were selected as the geometric mean of the attention score and the selection score (Table 2).

We then associated the attention score with positively selected variants underlying the high-altitude adaptation of Tibetan individuals[6] in HUVEC cells, which are oxygen-sensitive and serves as a classic model to study oxidative stress and cellular responses to hypoxia. We obtained a total of 4,627,029 variants after variant calling from the whole-genome sequencing data of 38 Tibetan highlanders and 39 Han Chinese lowlanders[6]. For each single nucleotide polymorphism (SNP), we computed the fixation index (Fst)[7], a widely used statistic in population genetics to detect potential positive selection among different populations. *EPAS1* showed the strongest signal in selective sweeps[8,9], so we focused on variants with high Fst scores around *EPAS1*.

We calculated the attention matrix of all genes on chromosome 2 (hg19) and each 10-kb window (in the range [−1M, +1M] from TSSs) in HUVEC cells with equation (2), then applied quantile normalization over the columns of the attention matrix to correct the bias between genomic windows (Fig. 2a). The selection status of each 10-kb window was calculated as the maximum Fst score of all SNPs within the window (Fig. 2a). For the [−1M, +1M] region around *EPAS1*, several windows achieve a high attention score, for

Table 2 | Top 10 regions identified by attention and selection scores

10-kb window	Attention score	Normalized attention score	Selection score	Geomean square	Name
chr2.46600000.46610000	2.00×10^{-5}	0.63	0.7902	1.58×10^{-5}	B1
chr2.46590000.46600000	1.46×10^{-5}	0.46	0.8044	1.18×10^{-5}	B2
chr2.46540000.46550000	1.74×10^{-5}	0.55	0.5732	1.00×10^{-5}	B3
chr2.46550000.46560000	1.06×10^{-5}	0.33	0.6569	6.96×10^{-6}	B4
chr2.46570000.46580000	7.56×10^{-6}	0.24	0.776	5.87×10^{-6}	B5
chr2.46990000.47000000	3.17×10^{-5}	1.00	0.1479	4.69×10^{-6}	B6
chr2.46980000.46990000	1.56×10^{-5}	0.49	0.2963	4.63×10^{-6}	B7
chr2.46520000.46530000	1.74×10^{-5}	0.55	0.2545	4.43×10^{-6}	B8
chr2.46770000.46780000	2.74×10^{-6}	0.09	0.6822	1.87×10^{-6}	B9
chr2.46530000.46540000	7.67×10^{-6}	0.24	0.2385	1.83×10^{-6}	B10

The attention scores are calculated using equation (2) and normalized by the maximum value: normalized AS = AS/max(AS). The selection score for each window is calculated from the maximum Fst score of all SNPs for the window. Geometry mean square (Geomean) is the attention score multiplied by the selection score for each window. The top 10 windows were selected by their Geomean value. Bin names are shown in Fig. 2.

example B6 and B7 downstream of *EPAS1* and B1–B5 and B8 in the gene body (Fig. 2b and Table 1). These positions with high attention scores locate in functional genomic regions such as open chromatin and histone modifications. The DNase-seq, H3K27ac, H3K4me1 and H3K4me3 data were downloaded from the ROADMAP epigenomics project (https://egg2.wustl.edu/roadmap/web_portal/imputed.html). The Hi-C loops in HUVEC cells were obtained from the GEO database (accession no. GSE63525)[10]. To identify functional regions under strong selection, we incorporated the attention score and the Fst score as a geometric mean. Top 10 regions were identified from both regulation and selection evidence (Table 2 and Fig. 2b). The top five regions, that is, B1–B5 in Table 2, were also identified as potential causal regulatory regions in our recent study[4].

We note that regulatory information embedded in GEEK vectors can provide local information for the effects of genetic variants. There is no global concordance between Fst score and attention score. The Spearman correlation between the attention score from HUVEC cells and Fst across chromosome 2 is −0.0086. This is to be expected, because the Fst score encodes the information in DNA by natural selection, which has effects in many cell types and tissues. For example, we counted the number of overlaps of a given set of 31 SNPs with high Fst scores near *EPAS1* with predicted enhancers in 128 cell types (H3K27ac gapped peaks) from the ROADMAP database. Eighty-one cell types overlapped at least one high-Fst SNP, with a maximum of 23 overlapped, while HUVEC cells covered 11 SNPs and missed the other 20 SNPs. This is consistent with the fact that attention score is derived from HUVEC cell types and encodes the condition-specific regulatory information. Thus, the utility of our prioritization approach is not guaranteed globally and the GEEK embedding vector should be interpreted carefully as a mixture of context-specific and non-specific information.

Discussion

The above results indicate that regulatory information in GEEK embedding vectors is useful for modelling gene expression and potentially helpful in prioritizing genetic variants in some applications. Additionally, the minimal demonstrative GEEK pipeline at Code Ocean (https://codeocean.com/capsule/3404879/tree/v1)[11] can reproduce the original results in ref. [2] and further facilitate the development of downstream applications. Overall, we have demonstrated the robustness and re-usability of the GEEK embedding framework and shown its convenience for follow-up studies by representing the network as a vector. After systematic evaluation of GEEK and other applications, we expect that the incorporation of sequence-based features and 3D chromatin interaction-based features into a unified framework may provide a holistic perspective to understand gene transcriptional control and potentially provide insights to identify genome-wide association study risk variants. In addition, multidimensional summaries of omics data can be further integrated into sophisticated statistical models. For example, the STAAR model introduces 'annotation principal components' to effectively summarize multiple qualitative and quantitative variant functional annotations to boost the power of variant set tests for continuous and binary traits in whole-genome sequencing rare variants association studies[12].

Data availability

The data used in our K562 and HUVEC studies with the retrained GEEK model are available at https://zenodo.org/record/4797001#.YK3HLS21FN0[11]. All the GEEK data[2] are available at http://yiplab.cse.cuhk.edu.hk/geek/, https://zenodo.org/record/3040059, http://www.ncbi.nlm.nih.gov/geo/ (accession no. GSE145774) and the Genome Sequence Archive (project no. CRA002025).

Code availability

GEEK is freely available at https://codeocean.com/capsule/3404879/tree/v1[13]. Modified DeepExpression for reproduction is freely available at https://github.com/wanwenzeng/DeepExpression[14]. vPECA is freely available at https://github.com/jxxin22/vPECA[15]. Details of the methods are available in refs. [2–4].

Received: 22 October 2020; Accepted: 11 June 2021;
Published online: 8 July 2021

References

1. Dong, Y., Chawla, N. V. & Swami, A. metapath2vec: scalable representation learning for heterogeneous networks. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 135–144 (ACM, 2017).
2. Cao, Q. et al. A unified framework for integrative study of heterogeneous gene regulatory mechanisms. *Nat. Mach. Intell* **2**, 447–456 (2020).
3. Zeng, W., Wang, Y. & Jiang, R. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics* **36**, 496–503 (2020).
4. Xin, J. et al. Chromatin accessibility landscape and regulatory network of high-altitude hypoxia adaptation. *Nat. Commun.* **11**, 4928 (2020).
5. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30, 5998–6008 (NIPS, 2017).
6. Lu, D. et al. Ancestral origins and genetic history of Tibetan Highlanders. *Am. J. Hum. Genet.* **99**, 580–594 (2016).
7. Weir, B. S. & Cockerham, C. C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

8. Peng, Y. et al. Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol. Biol. Evol.* **28**, 1075–1081 (2011).
9. Simonson, T. S. et al. Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72–75 (2010).
10. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
11. Zeng, W., Xin, J., Jiang, R. & Wang, Y. Compressing regulatory networks to vectors for interpreting gene expression and genetic variants (Zenodo, 2021); https://doi.org/10.5281/zenodo.4797001
12. Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
13. Cao, Q. et al. GEEK (Gene Expression Embedding frameworK) demo (GM12878, chromosome 1) (Code Ocean, 2020); https://doi.org/10.24433/CO.1518993.V1
14. Zeng, W. wanwenzeng/DeepExpression: DeepExpression (Zenodo, 2021); https://doi.org/10.5281/zenodo.4798333
15. Xin, J. vPECA (Zenodo, 2021); https://doi.org/10.5281/zenodo.4797172

Acknowledgements

We acknowledge funding from the National Key Research and Development Program of China (grants 2018YFC0910404 and 2020YFA0712402), the National Natural Science Foundation of China (grants 11688101,12025107, 11871463, 61621003, 61873141, 61721003, 61573207 and 62003178), and a grant from the Guoqiang Institute, Tsinghua University.

Author contributions

Y.W. and R.J. conceived and supervised the project. W.Z. and J.X. designed the experimental/analytical approach and performed numerical experiments and data analysis. All authors wrote, revised and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.J. or Y.W.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.