Virtual Gene Concept and a Corresponding Pragmatic Research Program in Genetical Data Science

Łukasz Huminiecki

Evolutionary, Computational, and Statistical Genetics, Department of Molecular Biology, Institute of Genetics and Animal Biotechnology, Polish Academy of Sciences, Postępu 36A, Jastrzębiec, 05-552 Warsaw, Poland; l.huminiecki@igbzan.pl

Abstract: Mendel proposed an experimentally verifiable paradigm of particle-based heredity that has been influential for over 150 years. The historical arguments have been reflected in the near past as Mendel's concept has been diversified by new types of omics data. As an effect of the accumulation of omics data, a virtual gene concept forms, giving rise to genetical data science. The concept integrates genetical, functional, and molecular features of the Mendelian paradigm. I argue that the virtual gene concept should be deployed pragmatically. Indeed, the concept has already inspired a practical research program related to systems genetics. The program includes questions about functionality of structural and categorical gene variants, about regulation of gene expression, and about roles of epigenetic modifications. The methodology of the program includes bioinformatics, machine learning, and deep learning. Education, funding, careers, standards, benchmarks, and tools to monitor research progress should be provided to support the research program.

Keywords: gene concept; technology; scientific method; experimentalism; molecular biology; genomics; bioinformatics; computational biology; data science; virtualization

1. Introduction

I start by underlining the long-term significance of the Mendelian gene concept. Then, I argue that genetics is transformed into an interdisciplinary data science. Molecular genes are computationally modeled with various levels of biochemical detail and integrate various functional data. This computational modeling gives rise to a virtual gene concept in which genes are not physically existent, but appear to be so when being modeled at various levels of abstraction [1] as database records, programming objects, or elements of computer graphics. Finally, I argue that the virtual gene concept can shape a vibrant research program if applied pragmatically.

In his seminal 1866 paper [2], Gregor Mendel proposed a paradigm of intra-cellular, discrete, and particulate differentiating element mediating heredity (differirenden Zelelemente in German). Due to the limited experimental methodology available to him, the postulated particle of heredity was thought of in purely abstract terms. It was only later given the name we know today: gene. Mendel provided strong empirical evidence in support of his hypothesis. The evidence was in the form of brilliantly designed experiments on plant hybridizations. Moreover, Mendel analyzed the implications of the paradigm very well. His insights were profound, mathematical, and reached into the contemporary genomics era [1,3–10].

At the beginning of the 20th century, the gene continued to be seen as an abstract concept. However, later in the century, the reductionism [11] of molecular biology promoted a material—physicochemical—view of the gene. The material gene focused on the bio-organic chemistry of DNA, the genetic code, and the protein sequence defined in the order of exonic base pairs [7–10,12,13].

There is a danger that genetical data science could be an observational field drowning in data, particularly when practiced by the inexperienced. As a remedy, there is a need for...
research programs; that is, the clearly defined and well-reviewed sequences of theories. According to Lakatos [14], a scientific research program has a hard core of central theories that are regarded as certain and a soft shell of satellite falsifiable theories. There is also a group of practitioners who are interested in such theories or can apply them for practical purposes. The aim of this opinionated review is to argue for the support, funding, formalization of goals and achievements, benchmarking, and methodological development of the research program that is focused on the virtual gene concept.

2. Integrating Genomics Data Gives Rise to the Virtual Gene Concept

In the 21st century, data from the Human Genome Project (HGP) were integrated with data from functional genomics consortia. The functional genomics projects were designed to functionally characterize genomes of model species, humans, and other vertebrates. For example, a database of single-nucleotide polymorphisms (dbSNP) was developed. There were hundreds of genome-wide association studies (GWAS), or a database of expressed sequence tags (dbEST) [15], and a project for functional annotation of mammalian genomes (FANTOM). The first FANTOM catalogue [16] was based on sequencing full-length transcripts and revealed the existence of widespread antisense transcription [17] and many complex loci [18]. Moreover, FANTOM and RNA-seq [19] confirmed the existence of ubiquitous alternative splicing.

Presently, most functional genomic data are generated using next generation sequencing (NGS). For example, the FANTOM consortium developed a new technology for single-base resolution expression profiling called cap analysis of gene expression—CAGE [20]. Similar to dbEST, CAGE was applied to survey genome-wide transcriptional activity in human and mouse. The upgraded expression catalogue showed that transcriptional activity is spread widely throughout the genome, and that most human genes have distant enhancers [21], as well as multiple transcriptional start sites (TSSes). Other studies suggested that the TSSes likely evolved in correlation with splicing [22,23].

Additionally using NGS, a consortium developed that was dedicated to generating an Encyclopedia of DNA Elements—ENCODE [24]—and parallel encyclopedias for model species—modENCODE [25]. These encyclopedias characterized the patterns of DNA binding for dozens of transcription factors in humans, the fly, and the worm.

3. The Virtual Gene Concept Is Helpful in Genetics Education, in Computational Biology Research, and as a Focus of Data Integration

Students of genetics have learnt from undergraduate textbooks about Mendel’s paradigm using classic examples of the genes responsible for monogenic diseases, such as cystic fibrosis, or multiple genes controlling eye color or continuous traits such as height. Pedigree charts are shown to demonstrate different patterns of heredity, and basic information is provided about genetic mapping and molecular gene structure.

However, students today are also likely to have an early interactive encounter with the virtual gene concept using genome browsers. The three most popular genome browsers are: (1) European Bioinformatics Institute’s Ensembl; (2) a browser provided by the National Center for Biotechnology Information—NCBI; and (3) a browser provided by the University of California in Santa Cruz—UCSC.

Similarly, in computational genetical research, we increasingly interact with the virtual gene concept using bioinformatics tools. Both anti-reductionist and reductionist approaches can be used [1], but interactivity makes this classic methodological distinction less clear. For example, genome browsers pragmatically integrate different levels of abstraction, hyperlinking phenotypes or physical and genetic maps with sequence data. Note that a pragmatic tradition in philosophy suggests that bioinformaticians ought to focus on how well the variants of the gene concept work for practical problem solving, prediction, and action.

The pragmatism is evident in resources provided by the leading data integrator at the National Institutes for Health: NCBI [26]. NCBI’s gene database hyperlinks nucleotide
sequences with a wide range of annotation databases, though mapping remains a challenge, being hard to standardize or describe statistically. A desire to pragmatically integrate as much data as possible appears as a priority, rather than ensuring that the mapping is completely non-redundant, always reproducible, or is statistically fully described.

In a second example, the UCSC Genome Browser [27–31] represents genes as pragmatic graphical objects that can be interactively repositioned or zoomed-in on. Each locus is annotated in a practical manner with genetic and functional data in parallel tracks. The genes are also hyperlinked to “Gene Views” containing nucleotide/protein sequences, to biomedical and biotechnological literature, gene expression data, or experimental protein structures.

Note that the trend towards data integration is accelerating [32–39]. Integrated, gene-related data may now include single-cell/spatial expression profiles [40–42], digital histopathology [43–45], or clinical imaging [46]. Indeed, data integration is now possible for omics datasets even at a single-cell level [47,48].

4. The Virtual Gene Concept Can Define a Practical Research Program
4.1. Methods of the Research Program

Traditionally, computational biology relied on purpose-developed bioinformatics methods [49–51]. Such bioinformatics tools are now well-described in well-edited textbooks and monographs; they are described either in a general context [49] or in more specific contexts, such as NGS informatics [51] or molecular evolution [52]. There are also numerous monographs devoted to specific technologies of genomics, such as the analysis of variability [53] or microarrays [50].

However, genetical data science now requires not only tools of bioinformatics but also of applied statistics, such as computer-age and large-scale statistical inference [54,55], techniques of statistical learning [56], or probabilistic modeling [57]. Moreover, algorithms of machine learning [58] and deep learning [59–61] have advantages over statistical models if associations between variables go beyond linearity and additive. For example, support vector machines, random forests, or dense neural networks detect non-linear or non-additive interactions between inputs. Convolutional neural networks—CNN—can recognize spatial associations between variables and non-additive effects. Recurrent neural networks can also be used to model spatial effects, for instance in the case of sequence data.

Generally speaking, if one can compile a set of functionally related nucleotide sequences, then they can be learnt as a computational model. The model can then be used for motif interpretation or prediction. This procedure is easy to apply for some of the in silico hypotheses generated by the virtual gene concept (in particular, those outlined below as (a), (b), and (c)). Deep learning offers new opportunities in this area as complex sequence motifs can be detected more flexibly than with statistical or machine learning. That is to say, deep learning robustly learns functional motifs in the whole sequence range covered by model inputs regardless of the positions of motifs in relation to each other. Moreover, sequence motifs can be flexibly combined with functional experimental data such as protein/DNA interactions, splicing events, post-transcriptional modifications, etc.

4.2. Goals of the Research Program

There are a number of research questions in which statistical data modeling, machine learning, and deep learning intersect non-trivially with the virtual gene concept. Such intersections suggest goals for the research program proposed here. The investigation of pseudo-genes was a successful example of such research [13,62–74]. In the future, a number of further in silico-testable hypotheses about the nature of genes could be included. Representative examples of such hypotheses are provided below.

(a) Alternative TSSes can drive contrasting patterns of gene expression [75–77]. This is not only because of the use of alternative promoters, but also because alternative 5’-UTRs might confer different mRNA stability, translational efficiency, or affect polymerase II (Pol II) pausing (Pol II pausing is increasingly recognized as having a regulatory
(b) Of course, alternative TSSes must produce alternatively structured transcripts. While many such transcripts have been previously sequenced and deposited in nucleotide sequence databases, they were typically attributed to alternative splicing rather than to alternative promoter usage. This fairly trivial deduction prompts a number of non-trivial questions. Crucially, do alternative TSSes result in neo-functionalized protein isoforms if the first exon is skipped? For example, truncated dominant negative members of protein complexes could bind a ligand or co-receptor but do not have enzymatic activity to propagate a signal. Alternatively, protein isoforms could differ in subcellular localizations if a signal peptide is affected; examples include Arabidopsis glutathione S-transferase F8 [77]. Note that fusion transcripts [83], scrambled neighboring transcripts [84], or prematurely terminated transcripts [85] can give rise to functional protein variants, particularly in cancer. Pathological neo-functionalization is well-known for fusion transcripts resulting from chromosomal re-arrangements in cancer, e.g., constitutively active BCR-ABL1 fusion tyrosine kinase in chronic myeloid leukemia, which is successfully targeted using tyrosine kinase inhibitors [86].

c) Are different varieties of non-coding transcripts generally functional genes? Or are they more commonly a type of biological noise that results from open chromatin in transcriptionally active chromosomal domains? There are already many known examples of functional antisense [87–89], functional expressed pseudo-genes, and functional long non-coding RNAs [67,69,71,73]. The quantification of genome-wide trends is needed, now.

d) Are enhancers and insulators typically associated with individual genes or rather with large but linear transcriptional domains? Are the transcriptional domains of universal significance, or do they only play a role in specific cell types? To what extent do the transcriptional domains correlate with the 3D organization of the genome? Early studies using a technique called optical reconstruction of chromatin architecture—ORCA—to image chromatin in embryos [90] suggest that transcriptionally active domains are sharply defined by borders of Polycomb-repressed DNA, but change with cell identity [91]. A deep learning model trained with ORCA data proved that 3D chromatin architecture strongly correlates with gene expression, and the effect is complex and diffuse, extending beyond direct contact of sequence-defined motifs such as promoters or enhancers [92]. (In fact, the contact of promoters with enhancers was not a good predictor of transcription [92]).

e) Can DNA sequences be usefully modeled on their own, or should DNA methylation, nucleosome occupancy, and histone modifications also typically be taken into account? For example, it is well known that several distinct types of histone modifications correlate positively with transcription [93,94]. However, there is no certainty if any of the histone modifications are early causative events. Rather, current knowledge suggests that pioneering transcription factors are primary causative agents for active transcription while the histone modifications are merely a later/downstream consequence of transcriptionally active chromatin [95–98]. On the other hand, DNA hyper-methylation and dense nucleosome occupancy in promoter regions appear to be early events of transcriptional silencing [99,100].

For each of the above hypotheses, there are intersections with further questions about genetic variability in populations, the impact on sequence alignment, or the calculation of distances/evolutionary rates between sequences. For example, one might ask about different kinds of polymorphisms, especially SNPs and small indels. Do the polymorphisms...
affect promoter usage, splicing, interactions between coding and non-coding transcripts, functions of enhancers/insulators, or the establishment of epigenetic marks? Moreover, should alternative TSSes be taken into account in sequence alignment or for the calculation of distances?

Determining which datasets will need integration in preparation for data mining to tackle such questions will, of course, depend on context. For some applications, alternative splice variants or promoters may not matter. For other projects, the focus will be precisely on the nuances of gene definition, or structure, or variability, or epigenetics, or on functional annotations.

4.3. Recent Examples

Deep learning has already been applied to predict alternative poly-adenylation [101,102], noncoding variants that interfere with splicing [103,104], gene regulatory networks [105–108], the expression of copy number variants [109], and in single-cells [110,111] or the targets of non-coding RNAs.

Tables 1 and 2 list a few published examples of applications of deep learning related to the virtual gene concept. Table 1 is focused on simple model gene expression. Table 2 lists three examples focused on gene structure: promoter prediction or prediction of alternative splicing coordinated with polymerase II pausing. For each of the examples, the tables list references, the main result of the study, their biological interpretations, and the data inputs/outputs of the deep learning model.

Of course, Tables 1 and 2 are intended to be illustrative rather than comprehensive. Indeed, it would be difficult to be comprehensive in such a dynamic field; important papers are published every month and any review will be outdated by the time of publication.

4.4. Potential Weaknesses of the Research Program

An obvious area for improvement in deep learning is model interpretability. Note that deep learning was developed as an enabling technology for industrial applications such as artificial vision and natural language processing. In such industrial applications, the predictive performance of models is prized over their interpretability. In scientific applications, however, the priorities are frequently exactly the opposite. Academic reviewers are likely to be as interested in the mechanisms of genetical phenomena as in prediction. Already, predictive black box models are less valued than transparent models from which functional sequence motifs—such as splice sites or transcription factor (TF) binding sites—can be extracted. Although several computational methods for enhancing interpretability in deep learning were proposed, successes are still mostly limited to the recovery of sequence motifs [112,113].

Moreover, interpretability itself may require a better theoretical framework [114] so that gradual improvements in interpretability can be quantified or benchmarked.

Note that if interpretability is of much higher value than prediction, traditional statistical data modeling using some flavor of multivariate statistics may be preferred in practice. This is because inference and learning theory is more developed in statistics than in computer science [115,116]. Moreover, visualizations are well developed to help in the diagnosis and interpretation of statistical models [117].

Of further note, both statistical and machine learning models may need adapting to specific genetic applications. For example, Teschendorff and Relton discussed adapting feature selection to the context of the analysis of DNA methylation data [118].

Hopefully, some universally applicable guidelines for variable selection and choice of algorithm will be developed with time. For example, in statistical data modeling, a useful rule of thumb may be to use only those variables that add to the predictive power of the model. Furthermore, generally speaking, automatic variable selection outperforms manual/human variable selection [119]. Common approaches to reduce the number of trainable parameters involve regularization steps (e.g., dropout, L1/L2 regularization, etc.) Note, however, that some gene-related variables may add predictive power to gene
models [95] but could take away from interpretability. For example, I noticed that a chromatin-associated gene variable such as DNASE1/methylation signal or GC- and CpG-content had great predictive value for classifying genes as either housekeeping or tissue-specific [95]. However, the chromatin-associated variables had no explanatory value as the casual association for the breadth of expression was with the number of transcription factors binding a proximal promoter [96].

4.5. Related Research Programs

One can identify a number of successful research programs in genetical data science that are related but not focused on the concept of the gene itself. For example, systems genetics [120,121] focuses on the interpretation of phenotypes and has already yielded profound insights into widespread genetic pleiotropy [122,123]. Genomic prediction has revolutionized breeding and animal science over the last two decades. Network medicine aims to explain diseases [124,125].

Deep learning is currently making a great impact across all these related fields. Its applications have already been reviewed for general omics [126–129], gene function prediction [130,131], disease prediction [132], predicting the impact of genetic variation in genomics [129,133], predicting gene regulatory networks [133,134], regulatory genomics [135], sequence motifs of transcription factors and enhancers [133,134,136–142], variant calling and pathogenicity scores [143], precision medicine [144,145], pharmacogenomics [128], and even the prediction of CRISPR targets [146].

In this opinionated and forward-looking review, I argue for the formalization of a new research program in genetical data science. The program should be focused on the concept of the gene. Although related to fields mentioned in the two previous paragraphs, the proposed research program is markedly different. Crucially, the research program is a part of biology that is understood as a basic science. Its central question—gene concept—has long been a focus of theoreticians of biology and its philosophers [7,9,10,147].

| Reference          | Main Result                                                                 | Biological Interpretation of the Model                                      | Data Inputs                                                                 | Model Output                                                                 | Other Points                                                                 |
|--------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Rajpurkar et al.   | Convolutional neural networks predict gene expression better than density neural networks or a random forest. Blanking could reveal important motifs for example, enhancers and silencers. | Chromatin architecture predicts gene expression. However, the effect is diffuse, extending beyond sequence-identifiable motifs such as promoters or enhancers. | Optical reconstruction of chromatin architecture (ORCA) of the Bithorax gene cluster in Drosophila. ORCA images were pre-processed to preserve structure but not viewing angle. | ON or OFF binary prediction of expression. | This was a remarkably innovative approach building on the strength of a remarkably novel dataset. |
| Zrimec et al. [148]| Up to 82% of the variation of transcript levels could be predicted from DNA sequences. | Both coding and cis-regulatory regions contribute to prediction of gene expression. | DNA sequences of proximal promoters *, plus 64 codon frequencies from coding regions, and eight mRNA stability variables. | Expression levels recoded as transcripts per million. | Motif interactions were key for the control of the dynamic range of gene expression. |
| Singh et al. [149] | A model derived from histone marks predicts expression better than traditional machine learning. | Histone marks correlate with expression, although it is unclear which marks are causative. | Histone marks from 56 different cell types [150] around TSSes in consecutive intervals. | Binary * or low gene expression level. | Complex interactions of chromatin features could be detected and visualized for intuitive interpretation. |
| Cuperus et al. [142]| A CNN trained on random 50 5′-UTRs can predict the expression of a reporter gene from both artificial and native UTRs. | Alternative 5′-UTRs confer different mRNA stability or translational efficiency. | Nucleotide sequence of the 5′-UTR. | Scalar score for each UTR (proportional to protein expression). | Shorter UTRs did not work as well. |

* One-hot encoding is a standard approach to sequence representation in deep learning.
Table 2. Examples of deep learning applied to structural annotation of virtual genes.

| Reference                | Main Result                                                                 | Biological Interpretation of the Model                                                                 | Data Inputs                                                                 | Model Output                                                                 | Other Points                                                                 |
|--------------------------|------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Oubounyt et al. [151]    | The prediction method improves performance over comparable approaches (fewer false positives). The improvement is attributed to a novel negative learning set. | Short eukaryote promoter sequences are sufficient to predict both TATA and non-TATA promoters in both human and mouse. | Genomic sequence from −249 to +50 bps relative to the TSS *.                    | A real-valued promoter score.                                                  | Impact of mutations on output scores was also studied (150 substitutions on the interval from −40 to +10 bps relative to the TSS). |
| Kelley et al. [152]      | The method uses chromatin accessibility to predict gene promoters with high accuracy. | Promoters and transcription factor binding motifs could be predicted, but the method was developed to annotate point mutations. | DNase-seq mapping accessible genomic sites in 164 cell types (encoded as a binary vector of length 164). Plus, a DNA sequence of 600 bps *. | Probability value for chromatin accessibility.                                | Every mutation in the genome could be annotated with respect to its impact on chromatin accessibility. |
| Feng et al. [153]        | A deep learning model can predict Pol II pausing events. (An attention layer and data integration provide good interpretability.) | The pausing events also provide insights into alternative splicing, TF binding sites, and epigenetic modifications. | 200-50-bp DNA sequence * integrated with ChIP-seq and epigenetic data via an attention layer. | Probability value for Pol II pausing.                                          | Strongest sequence determinants were typically −14 to 12 bp around the pausing sites. The model was relatively interpretable due to an attention mechanism analogous to DeepHINT [154]. |

* One-hot encoding is a standard approach to sequence representation in deep learning.

4.6. The Need for Training, Funding, Benchmarking, and Monitoring of Progress

My goal was to define a research program named after the object of study: the virtual gene concept. Less experienced or younger researchers could benefit the most from such “manuals for doing research”. There can be no doubt that rich opportunities are available. However, success will always depend on the skills and drive of individual researchers. Data mining is a difficult and time-consuming art, which calls for a wide range of skills, patience, good judgment, and a perfectionist attitude. An understanding of the theoretical and philosophical aspects of gene concept will also be necessary. Ultimately, principal investigators—especially those who work individually rather than in consortia—will develop their own preferred way of practicing the data science. Postgraduate education, research funding, and a career structure in academia must be provided for this to be realistic. Methods developed in the field of virtual gene concept should be, of course, standardized and benchmarked. Ideally, methods could additionally be developed to quantitatively monitor the progress of research in the field.

Funding: L.H. was funded to perform this work by the National Science Centre, Poland, grant POLONEZ 2 (grant number 2016/21/P/NZ2/03926). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 665778.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data is contained within the article.

Acknowledgments: L.H. was funded to perform this work by the National Science Centre, Poland.

Conflicts of Interest: The author declare no conflict of interest.

References
1. Huminiecki, L. Models of the Gene Must Inform Data-Mining Strategies in Genomics. *Entropy* **2020**, *22*, 942. [CrossRef] [PubMed]
2. Mendel, J.G. Versuche über pflanzenhybriden. In *Verhandlungen des Naturforschenden Vereines in Brünn*; Naturforschender Verein in Brünn: Brno, Czech Republic, 1866; Volume iv, pp. 3–47.
3. Huminiecki, L. A Contemporary Message from Mendel’s Logical Empiricism. *Bioessays* **2020**, *42*, e2000120. [CrossRef] [PubMed]
4. Abbott, S.; Fairbanks, D.J. Experiments on Plant Hybrids by Gregor Mendel. *Genetics* **2016**, *204*, 407–422. [CrossRef]
5. Bateson, W. Mendel's Principles of Heredity: A Defence, with a Translation of Mendel's Original Papers on Hybridisation; Cambridge University Press: Cambridge, UK, 2009; p. 236.

6. Miko, I. Gregor Mendel and the principles of inheritance. Nat. Educ. 2008, 1, 134.

7. Portin, P.; Wilkins, A. The Evolving Definition of the Term “Gene”. Genetics 2017, 205, 1353–1364. [CrossRef]

8. Portin, P. The Development of Genetics in the Light of Thomas Kuhn’s Theory of Scientific Revolutions. Recent Adv. DNA Gene Seq. 2015, 9, 14–25. [CrossRef] [PubMed]

9. Griffiths, P.; Stotz, K. Gene. In The Cambridge Companion to the Philosophy of Biology; Hull, D.L., Ruse, M., Eds.; Cambridge University Press: Cambridge, UK, 2007; Volume xxvii, p. 513.

10. Griffiths, P.E.; Stotz, K. Genes in the postgenomic era. Med. Bioeth. 2006, 27, 499–521. [CrossRef]

11. Rosenberg, A. Reductionism (and antireductionism) in biology. In The Cambridge Companion to the Philosophy of Biology; Hull, D.L., Ruse, M., Eds.; Cambridge University Press: Cambridge, UK, 2007; Volume XXVII, p. 513.

12. Falk, R. The gene in search of an identity. Hum. Genet. 1984, 68, 195–204. [CrossRef]

13. Portin, P. Historical development of the concept of the gene. J. Med. Philos 2002, 27, 257–286. [CrossRef]

14. Lakatos, I. The Methodology of Scientific Research Programmes: Philosophical Papers; Cambridge University Press: Cambridge, UK, 1978.

15. Skrabanek, L.; Campagne, F. TissueInfo: High-throughput identification of tissue expression profiles and specificity. Nucleic Acids Res. 2001, 29, e102. [CrossRef]

16. Carinci, P.; Kasukawa, T.; Katayama, S.; Gough, J.; Frith, M.C.; Maeda, N.; Oyama, R.; Ravasi, T.; Lenhard, B.; Wells, C.; et al. The transcriptional landscape of the mammalian genome. Science 2005, 309, 1559–1563. [CrossRef]

17. Katayama, S.; Tomaru, Y.; Kasukawa, T.; Waki, K.; Nakanishi, M.; Nakamura, M.; Nishida, H.; Yap, C.C.; Suzuki, M.; Kawai, J.; et al. Antisense transcription in the mammalian transcriptome. Science 2005, 309, 1564–1566. [CrossRef] [PubMed]

18. Engstrom, P.G.; Suzuki, H.; Ninomiya, N.; Akalin, A.; Sessa, L.; Lavorgna, G.; Brozzi, A.; Luzi, L.; Tan, S.L.; Yang, L.; et al. Complex Loci in human and mouse genomes. PloS Genet. 2006, 2, e47. [CrossRef]

19. Merkin, J.; Russell, C.; Chen, P.; Burge, C.B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 2012, 336, 1593–1599. [CrossRef] [PubMed]

20. Kanamori-Katayama, M.; Itoh, M.; Kawaji, H.; Lassmann, T.; Katayama, S.; Koijima, M.; Bertin, N.; Kaiho, A.; Ninomiya, N.; Daub, C.O.; et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res. 2011, 21, 1150–1159. [CrossRef]

21. Andersson, R.; Gebhard, C.; Miguel-Escalada, I.; Hoof, I.; Bornholdt, J.; Boyd, M.; Chen, Y.; Zhao, X.; Schmidl, C.; Suzuki, T.; et al. An atlas of active enhancers across human cell types and tissues. Nature 2014, 507, 455–461. [CrossRef]

22. Fiszbein, A.; Krick, K.S.; Begg, B.E.; Burge, C.B. Exon-Mediated Activation of Transcription Starts. Cell 2019, 179, 1551–1565.e17. [CrossRef] [PubMed]

23. Willson, J. Exons as enhancers. Nat. Rev. Genet. 2020, 21, 68–69. [CrossRef]

24. Dunham, I.; Kundaje, A.; Aldred, S.F.; Collins, P.J.; Davis, C.A.; Doyle, F.; Epstein, C.B.; Fritze, S.; Harrow, J.; Kaul, R.; et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012, 489, 57–74. [CrossRef]

25. Gerstein, M.B.; Lu, Z.J.; Van Nostrand, E.L.; Heinig, M.; Uhlenhaut, N.H.; Kawai, J.; et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 2012, 336, 1559–1563. [CrossRef]

26. Wheeler, D.L.; Barrett, T.; Benson, D.A.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; DiCuccio, M.; Edgar, R.; Federhen, S.; et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2007, 35, D5–D12. [CrossRef]

27. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. Genome Res. 2002, 12, 996–1006. [CrossRef]

28. Karolchik, D.; Hinrichs, A.S.; Kent, W.J. The UCSC Genome Browser. Curr. Protoc. Bioinform. 2012, 40. [CrossRef]

29. Zweig, A.S.; Karolchik, D.; Kuhn, R.M.; Haussler, D.; Kent, W.J. UCSC genome browser tutorial. Genomics 2008, 92, 75–84. [CrossRef]

30. Mangan, M.E.; Williams, J.M.; Mathe, S.M.; Karolchik, D.; Mathe, W.C. III. UCSC genome browser: Deep support for molecular biomedical research. Biotechnol. Annu. Rev. 2008, 14, 63–108. [CrossRef]

31. Kent, W.J.; Hsu, F.; Karolchik, D.; Kuhn, R.M.; Clawson, H.; Trumbower, H.; Haussler, D. Exploring relationships and mining data with the UCSC Gene Sorter. Genome Res. 2005, 15, 737–741. [CrossRef] [PubMed]

32. Manzoni, C.; Kia, D.A.; Vandrovcova, J.; Hardy, J.; Wood, N.W.; Lewis, P.A.; Ferrari, R. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Brief. Bioinform. 2018, 19, 286–302. [CrossRef] [PubMed]

33. Cavill, R.; Jennen, D.; Kleinjans, J.; Briédé, J.J. Transcriptomic and metabolomic data integration. Brief. Bioinform. 2016, 17, 891–901. [CrossRef] [PubMed]

34. Das, T.; Andrieux, G.; Ahmed, M.; Chakraborty, S. Integration of Online Omics-Data Resources for Cancer Research. Front. Genet. 2020, 11, 578345. [CrossRef]

35. Höllbacher, B.; Balázs, K.; Heinig, M.; Uhlenhaut, N.H. Seq-ing answers: Current data integration approaches to uncover mechanisms of transcriptional regulation. Comput. Struct. Biotechnol. J. 2020, 18, 1330–1341. [CrossRef]

36. Yugi, K.; Kubota, H.; Hatano, A.; Kuroda, S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple ‘Omic’ Layers. Trends Biotechnol. 2016, 34, 276–290. [CrossRef]
70. Cheetham, S.W.; Faulkner, G.J.; Dinger, M.E. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat. Rev. Genet.* 2020, 21, 191–201. [CrossRef] [PubMed]

71. Singh, R.K.; Singh, D.; Yadava, A.; Srivastava, A.K. Molecular fossils "pseudogenes" as functional signature in biological system. *Genes Genom.* 2020, 42, 619–630. [CrossRef] [PubMed]

72. Bok, I.; Karreth, F.A. Strategies to Study the Functions of Pseudogenes in Mouse Models of Cancer. *Methods Mol. Biol.* 2021, 2324, 287–304. [PubMed]

73. Salmena, L. Pseudogenes: Four Decades of Discovery. *Methods Mol. Biol.* 2021, 2324, 3–18.

74. Troskie, R.L.; Jafrani, Y.; Mercer, T.R.; Ewing, A.D.; Faulkner, G.J.; Cheetham, S.W. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome Biol.* 2021, 22, 146. [CrossRef]

75. FANTOM5-Consortium. A promoter-level mammalian expression atlas. *Nature* 2014, 507, 462–470. [CrossRef]

76. Humniecki, L. Magic roundabout is an endothelial-specific ohnolog of ROBO1 which neo-functionalized to an essential new role in angiogenesis. *PLoS ONE* 2019, 14, e0208952.

77. Thatcher, L.F.; Carrie, C.; Andersson, C.R.; Sivasithamparam, K.; Whelan, J.; Singh, K.B. Differential gene expression and subcellular targeting of Arabidopsis glutathione S-transferase F8 is achieved through alternative transcription start sites. *J. Biol. Chem.* 2007, 282, 28915–28928. [CrossRef]

78. Rose, S.; Biers, D.; Svejstrup, J.Q. Causes and consequences of RNA polymerase II stalling during transcription elongation. *Nat. Rev. Mol. Cell Biol.* 2021, 22, 3–21. [CrossRef] [PubMed]

79. Zhang, P.; Dimont, E.; Ha, T.; Swanson, D.J.; Hide, W.; Goldowitz, D. Relatively frequent switching of transcription start sites during cerebellar development. *BMC Genom.* 2017, 18, 461. [CrossRef]

80. Koenigsberger, C.; Chicca, J.J., II; Amoureux, M.C.; Edelman, G.M.; Jones, F.S. Differential regulation by multiple promoters of the gene encoding the neuron-restrictive silencer factor. *Proc. Natl. Acad. Sci. USA* 2000, 97, 2291–2296. [CrossRef]

81. Thorsen, K.; Spechler, T.; Øster, B.; Rasmussen, M.H.; Vang, S.; Wang, K.; Hansen, K.Q.; Lamy, P.; Pedersen, J.S.; Eller, A.; et al. Tumor-specific usage of alternative transcription start sites in colorectal cancer identified by genome-wide exon array analysis. *BMC Genom.* 2011, 12, 505. [CrossRef]

82. Karlsson, K.; Lonnerberg, P.; Linnarsson, S. Alternative TSSs are co-regulated in single cells in the mouse brain. *Mol. Syst. Biol.* 2017, 13, 930. [CrossRef]

83. Luo, J.H.; Liu, S.; Zuo, Z.H.; Chen, R.; Tseng, G.C.; Yu, Y.P. Discovery and Classification of Fusion Transcripts in Prostate Cancer and Normal Prostate Tissue. *Am. J. Pathol.* 2015, 185, 1834–1845. [CrossRef]

84. Qin, F.; Song, Z.; Babiceau, M.; Song, Y.; Facemire, L.; Singh, R.; Adli, M.; Li, H. Discovery of CTCF-sensitive Cis-spliced fusion RNAs between adjacent genes in human prostate cells. *PLoS Genet.* 2015, 11, e1005001.

85. Karlic, R.; Chung, H.R.; Lasserre, J.; Vlahovicek, K.; Vingron, M. Histone modification levels are predictive for gene expression. *Nat. Rev. Mol. Cell Biol.* 2016, 17, 447–452. [CrossRef] [PubMed]

86. Rosikiewicz, W.; Makalowska, I. Biological functions of natural antisense transcripts. *Acta Biochim. Pol.* 2016, 63, 665–673. [CrossRef] [PubMed]

87. Pugh, C.W. Modulation of the Hypoxic Response. *Adv. Exp. Med. Biol.* 2016, 903, 259–271.

88. Li, K.; Ramchandran, R. Natural antisense transcript: A concomitant engagement with protein-coding transcript. *Oncotarget* 2010, 1, 447–452. [CrossRef]

89. Vavouri, T.; Lehner, B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Proc. Natl. Acad. Sci. USA* 2010, 107, 2926–2931. [CrossRef] [PubMed]

90. Huminiecki, L. Modeling of the breadth of expression from promoter architectures identifies pro-housekeeping transcription factors. *PLoS ONE* 2018, 13, e0198961. [CrossRef]

91. Hesson, L.B.; Sloane, M.A.; Wong, J.W.; Nunez, A.C.; Srivastava, S.; Ng, B.; Hawkins, N.J.; Bourke, M.J.; Ward, R.L. Altered promoter nucleosome positioning is an early event in gene silencing. *Epigenetics* 2014, 9, 1422–1430. [CrossRef] [PubMed]
100. Han, H.; Cortez, C.C.; Yang, X.; Nichols, P.W.; Jones, P.A.; Liang, G. DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum. Mol. Genet.* **2011**, *20*, 4299–4310. [CrossRef] [PubMed]

101. Vainberg Stutskin, I.; Weinberger, A.; Segal, E. Sequence determinants of polyadenylation-mediated regulation. *Genome Res.* **2019**, *29*, 1635–1647. [CrossRef]

102. Bogard, N.; Linder, J.; Rosenberg, A.B.; Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **2019**, *178*, 91–106.e23. [CrossRef]

103. Jaganathan, K.; Kyriazopoulou Panagiotopoulou, S.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.L.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **2019**, *176*, 535–548.e24. [CrossRef]

104. Bao, S.; Moakley, D.F.; Zhang, C. The Splicing Code Goes Deep. *Cell* **2019**, *176*, 414–416. [CrossRef]

105. Yuan, Y.; Bar-Joseph, Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 27151–27158. [CrossRef]

106. Kang, M.; Lee, S.; Lee, D.; Kim, S. Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning With Regulatory Latent Space. *Front. Genet.* **2020**, *11*, 869. [CrossRef]

107. Yang, Y.; Fang, Q.; Shen, H.B. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. *PloS Comput. Biol.* **2019**, *15*, e1007324. [CrossRef] [PubMed]

108. Muzio, G.; O’Bray, L.; Borgwardt, K. Biological network analysis with deep learning. *Brief. Bioinform.* **2021**, *22*, 1515–1530. [CrossRef]

109. Seal, D.B.; Das, V.; Goswami, S.; De, R.K. Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration. *Genomics* **2020**, *112*, 2833–2841. [CrossRef]

110. He, Y.; Yuan, H.; Wu, C.; Xie, Z. DISC: A highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. *Genome Biol.* **2020**, *21*, 170. [CrossRef]

111. Fortelny, N.; Bock, C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol.* **2020**, *21*, 190. [CrossRef] [PubMed]

112. Talukder, A.; Barham, C.; Li, X.; Hu, H. Interpretation of deep learning in genomics and epigenomics. *Brief. Bioinform.* **2021**, *22*, bbaa177. [CrossRef] [PubMed]

113. Koo, P.K.; Ploenzke, M. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nat. Mach. Intell.* **2021**, *3*, 258–266. [CrossRef]

114. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]

115. Breiman, L. Statistical modeling: The two cultures. *Stat. Sci.* **2001**, *16*, 199–215. [CrossRef]

116. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]

117. Wickham, H.; Cook, D.; Hofmann, H. Visualizing statistical models: Removing the blindfold. *Stat. Anal. Data Min.* **2015**, *8*, 203–225. [CrossRef]

118. Teschendorff, A.E.; Relton, C.L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **2018**, *19*, 129–147. [CrossRef]

119. Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from Genomic, Clinical, and Textual Data. *Stat. Anal. Data Min.* **2019**, *12*, 219–232. [CrossRef] [PubMed]

120. Nadeau, J.H.; Dudley, A.M. Genetics. Systems genetics. *Am. J. Hum. Genet.* **2011**, *89*, 767–773. [CrossRef] [PubMed]

121. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [CrossRef] [PubMed]

122. Stearns, F.W. One hundred years of pleiotropy: A retrospective. *Genetics* **2010**, *186*, 767–773. [CrossRef]

123. Sivakumaran, S.; Agakov, F.; Theodoratou, E.; Prendergast, J.G.; Zgaga, L.; Manolio, T.; Rudan, I.; McKeigue, P.; Wilson, J.F.; Campbell, H. Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **2011**, *89*, 607–618. [CrossRef]

124. Barabasi, A.L. Network medicine—From obesity to the “diseasome”. *N. Engl. J. Med.* **2007**, *357*, 404–407. [CrossRef]

125. Barabasi, A.L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef] [PubMed]

126. Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform.* **2017**, *18*, 851–869. [CrossRef] [PubMed]

127. Zhang, Z.; Zhao, Y.; Liao, X.; Shi, W.; Li, K.; Zou, Q.; Peng, S. Deep learning in omics: A survey and guideline. *Brief. Funct. Genom.* **2019**, *18*, 41–57. [CrossRef] [PubMed]

128. Kalinin, A.A.; Higgins, G.A.; Reamaroon, N.; Soroushmehr, S.; Allyn-Feuer, A.; Dinov, I.D.; Najarian, K.; Athey, B.D. Deep learning in pharmacogenomics: From gene regulation to patient stratification. *Pharmacogenomics* **2018**, *19*, 629–650. [CrossRef] [PubMed]

129. Eraslan, G.; Avsec, Z.; Gagneur, J.; Theis, F.J. Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **2019**, *20*, 389–403. [CrossRef] [PubMed]

130. Zou, Q.; Sangahia, A.K.; Mrozek, D. Editorial: Machine Learning Techniques on Gene Function Prediction. *Front. Genet.* **2019**, *10*, 938. [CrossRef]

131. Mahood, E.H.; Kruse, L.H.; Moghe, G.D. Machine learning: A powerful tool for gene function prediction in plants. *Appl. Plant Sci.* **2020**, *8*, e11376. [CrossRef]
132. Wong, A.K.; Sealfon, R.S.G.; Theesfeld, C.L.; Troyanskaya, O.G. Decoding disease: From genomes to networks to phenotypes. Nat. Rev. Genet. 2021, 22, 774–790. [CrossRef] [PubMed]

133. Telenti, A.; Lippert, C.; Chang, P.C.; DelPrisco, M. Deep learning of genomic variation and regulatory network data. Hum. Mol. Genet. 2018, 27, R63–R71. [CrossRef]

134. Min, X.; Lu, F.; Li, C. Sequence-Based Deep Learning Frameworks on Enhancer-Promoter Interactions Prediction. Curr. Pharm. Des. 2021, 27, 1847–1855. [CrossRef]

135. Zrimec, J.; Buric, F.; Kokina, M.; Garcia, V.; Zeleznia, A. Learning the Regulatory Code of Gene Expression. Front. Mol. Biosci. 2021, 8, 673363. [CrossRef]

136. Miraldi, E.R.; Chen, X.; Weirauch, M.T. Deciphering cis-regulatory grammar with deep learning. Nat. Genet. 2021, 53, 266–268. [CrossRef]

137. King, D.M.; Hong, C.K.Y.; Shepherdson, J.L.; Granas, D.M.; Maricic, B.B.; Cohen, B.A. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. Elife 2020, 9, e41279. [CrossRef] [PubMed]

138. Chen, L.; Capra, J.A. Learning and interpreting the gene regulatory grammar in a deep learning framework. Genome Res. 2021, 31, 1082–1096. [CrossRef]

139. Bravo Gonzalez-Blas, C.; Minnoye, L.; Papasokrati, D.; Aibar, S.; Hulselmans, G.; Christiaens, V.; Davie, K.; Wouters, J.; Aerts, S.; et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. 2021, 53, 354–366. [CrossRef] [PubMed]

140. Atak, Z.K.; Taskiran, I.I.; Demeulemeester, J.; Flerin, C.; Mauduit, D.; Minnoye, L.; Hulselmans, G.; Christiaens, V.; Ghanem, G.E.; Wouters, J.; et al. Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. Genome Res. 2021, 31, 1185–1196. [PubMed]

141. Koumakis, L. Deep learning models in genomics; are we there yet? Comput. Struct. Biotechnol. J. 2020, 18, 1466–1473. [CrossRef]

142. Wang, J.; Zhang, X.; Cheng, L.; Luo, Y. An overview and metaanalysis of machine and deep learning-based CRISPR gRNA design tools. RNA Biol. 2020, 17, 13–22. [CrossRef]

143. Griffiths, P.; Stotz, K. Genetics and Philosophy: An Introduction; Cambridge University Press: Cambridge, UK, 2013; p. 280.

144. Zrimec, J.; Börlin, C.S.; Buric, F.; Muhammad, A.S.; Chen, R.; Siewers, V.; Verendel, V.; Nielsen, J.; Töpel, M.; Zeleznia, A. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat. Commun. 2020, 11, 6141. [CrossRef] [PubMed]

145. Singh, R.; Lanchantin, J.; Robins, G.; Qi, Y. DeepChrome: Deep-learning for predicting gene expression from histone modifications. Bioinformatics 2016, 32, i639–i648. [CrossRef] [PubMed]

146. Kelley, D.R.; Snoek, J.; Rinn, J.L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016, 26, 990–999. [CrossRef] [PubMed]

147. Kelley, D.R.; Snek, J.; Rinn, J.L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genomic Res. 2016, 26, 990–999. [CrossRef] [PubMed]

148. Zrimec, J.; Buric, F.; Muhammad, A.S.; Chen, R.; Siewers, V.; Verendel, V.; Nielsen, J.; Töpel, M.; Zeleznia, A. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat. Commun. 2020, 11, 6141. [CrossRef] [PubMed]

149. Singh, R.; Lanchantin, J.; Robins, G.; Qi, Y. DeepChrome: Deep-learning for predicting gene expression from histone modifications. Bioinformatics 2016, 32, i639–i648. [CrossRef] [PubMed]

150. Koumakis, L. Deep learning models in genomics; are we there yet? Comput. Struct. Biotechnol. J. 2020, 18, 1466–1473. [CrossRef] [PubMed]

151. Wang, J.; Zhang, X.; Cheng, L.; Luo, Y. An overview and metaanalysis of machine and deep learning-based CRISPR gRNA design tools. RNA Biol. 2020, 17, 13–22. [CrossRef]