Human Molecular Genetics, 2018, Vol. 27, No. R1 R29–R34

doi: 10.1093/hmg/ddy088
Advance Access Publication Date: 16 March 2018
Invited Review

INVITED REVIEW

Biomedical informatics and machine learning for clinical genomics

James A. Diao[1,2], Isaac S. Kohane[1] and Arjun K. Manrai[1,*]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA and [2]Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

*To whom correspondence should be addressed. Tel: +1 6174322144; Fax: +1 6174320693; Email: arjun_manrai@hms.harvard.edu

Abstract

While tens of thousands of pathogenic variants are used to inform the many clinical applications of genomics, there remains limited information on quantitative disease risk for the majority of variants used in clinical practice. At the same time, rising demand for genetic counselling has prompted a growing need for computational approaches that can help interpret genetic variation. Such tasks include predicting variant pathogenicity and identifying variants that are too common to be penetrant. To address these challenges, researchers are increasingly turning to integrative informatics approaches. These approaches often leverage vast sources of data, including electronic health records and population-level allele frequency databases (e.g. gnomAD), as well as machine learning techniques such as support vector machines and deep learning. In this review, we highlight recent informatics and machine learning approaches that are improving our understanding of pathogenic variation and discuss obstacles that may limit their emerging role in clinical genomics.

Introduction

Over the past two decades, advances in the acquisition and analysis of genetic sequence data have enabled the association of many thousands of genetic variants with myriad diseases and traits (1–3). These efforts have improved our understanding of basic disease mechanisms and transformed their treatment in clinical practice. While the catalogue of variants associated with human disease is extensive, it consists of efforts from different communities where evidentiary criteria for disease association can vary profoundly (4). Study designs like genome-wide association studies (GWASs) are relatively agnostic to assumptions about candidate genes, and they consistently address population stratification and multiplicity, leading to highly reproducible associations (1,5); it is likely that the full clinical impact of risk factors distributed across the genome (6) has yet to be realized. By contrast, new sequencing technologies are increasingly penetrating the clinic for many cancers and Mendelian diseases, but such applications often lack a precise quantitative understanding of disease risk. In fact, it is now well-recognized that many of these applications may be of questionable utility (7,8).

The scale of clinical genomics is already substantial: as of February 2018, the NIH Genetic Testing Registry (9) contained over 53 000 genetic tests for over 11 000 inherited conditions (10). However, many of these tests may involve variants of uncertain or conflicting significance (11,12), which have the potential to mislead or potentially harm patients (13). In response, researchers are increasingly turning to new analytical approaches and data sources to better evaluate new variants and re-evaluate previously implicated variants.

Many data sources that might improve clinical genomics are complex, high-dimensional and scattered across institutions (14). Collecting, sharing and harmonizing these diverse data streams remain important challenges. Nonetheless, several studies have already leveraged the increasing accessibility of existing data modalities such as electronic health records (EHRs). EHRs allow investigators to infer computationally

Received: February 21, 2018. **Revised:** March 8, 2018. **Accepted:** March 8, 2018

derived phenotypes and conduct genotype-phenotype studies. Growing evidence suggests that EHR-based methods can be powerful predictors of mortality, readmission, prolonged stays and final diagnoses (15), if models and data representations are sufficiently flexible (16,17). Research efforts involving EHR data have already contributed to a variety of applications including pharmacogenomics (18) and community health (19), and are becoming increasingly feasible following increased adoption in the United States (20) and worldwide (21).

As clinical data have become more accessible, attention has turned to joining clinical measures with other data modalities. A series of initiatives around the world have assembled biobanks of longitudinal cohorts that integrate a range of molecular, clinical and environmental measures. Such programs include the United States All of Us Research Program (22), the UK Biobank (23), the China Kadoorie Biobank (24) and the Estonian Biobank (25). These efforts aim to build a rich collection of patient data to accelerate medical research efforts, both broadly and within specific foci (e.g. cancer genomics). At the same time, cohorts of ancestrally diverse populations have recently been harmonized and aggregated across various large-scale sequencing projects (26,27). These databases play a central role in clinical genomics by providing precise estimates of allele frequency across ancestrally diverse populations.

In this review, we highlight recent efforts from biomedical informatics and machine learning that leverage these data sources to improve the evidence base for clinical genomics. We argue that the major challenge in genomics has shifted from exploration (i.e. collection and interpretation) to exploitation (i.e. translation to clinical knowledge and tools). Finally, we discuss potential solutions for bridging clinical knowledge gaps, including methods for integrating data, interpreting models and improving accessibility by clinicians.

Clinical Genomics

A fundamental concept in clinical genomics is 'pathogenicity', which refers to the likelihood that a genetic variant is disease-causing. Pathogenicity is assigned on a categorical scale; the American College of Medical Genetics and Genomics recommends describing variants using the terms 'pathogenic', 'likely pathogenic', 'uncertain significance', 'likely benign' and 'benign', with the modifier 'likely' indicating a >90% certainty (28). The concept of pathogenicity is separate from that of 'penetrance', which refers to the probability of disease among patients with an associated variant or set of variants (4). Pathogenicity is classified based on an assortment of evidentiary standards, weighted from 'supporting' (e.g. co-segregation with disease in multiple family members) to 'very strong' (e.g. predicted null variant where loss of function is a known disease mechanism), with opportunities for expert judgment to evaluate the full body of evidence (28).

As specific as these guidelines can be, laboratories will often disagree on a variant's pathogenicity status. The Clinical Sequencing Exploratory Research program (CSER) piloted a set of 99 variants across nine molecular diagnostic laboratories and found acceptable within-laboratory concordance—79%, but poor between-laboratory concordance—34% (12). Even after discussions between laboratories, concordance improved to just 71%. In practice, variants in shared databases like ClinVar (29) often include different pathogenicity assertions from different laboratories, with limited means of resolving differences.

Similarly, penetrance is poorly understood for the majority of variants in clinical use, even for variants with consensus on pathogenicity. For some rare diseases, this is due to persistently small sample sizes. When estimates are available, they are often influenced by the 'winner's curse', where initial estimates from discovery studies are inflated by ascertainment bias (30). An individual's genomic background may also influence the expression of genes implicated in disease via modifier effects; these effects remain incompletely understood and contribute additional uncertainty to estimates of penetrance. While the number of pathogenicity assertions continues to grow, the number of variants with validated penetrance estimates is orders of magnitude smaller, with some notable exceptions (31–33). Overall, it is likely that many pathogenic variants are perceived to be more penetrant than they actually are, as observed with diseases like breast cancer (31) and hemochromatosis (34).

Retrospective clinical audits present one feasible and cost-effective method for ensuring that current clinical applications are ultimately improving patient care (35). However, there is likely no evidentiary substitute for randomized control trials (RCTs) to validate existing variant interpretations and the interventions they inform. The first randomized study of precision-guided treatment in oncology was the SHIVA trial, published in 2015 (36). This trial compared targeted molecular agents against treatment at physician's choice in 195 patients with solid tumors and found that progression-free survival was not significantly longer with targeted therapy. Although the study was preliminary and does not preclude the possibility that targeted agents can be effective in patient subgroups, it raises concerns about the overall effect of precision-guided therapies in clinical practice. Other trials have evaluated the effects of whole genome sequencing (WGS) in clinical practice, with mixed results. The MedSeq Project, for example, has sought to test how WGS will impact patients and physicians (7). A recent pilot RCT found unclear clinical value of WGS, suggesting that the benefits of sequencing otherwise healthy patients may be outweighed by the costs and risks (8). Current studies still lack sufficient sample sizes for definitive conclusions about the costs and benefits of many modern genetic testing practices. Expanding this evidence base will become increasingly important for justifying or refining their continued use.

Biomedical Informatics for Clinical Genomics

Technological advances have dramatically lowered the cost of sequencing, and the subsequent wealth of data has since prompted rapid increases in the number of disease-associated variants. Many of the usual informatics challenges of genome-scale analysis, such as data storage, analysis and security, are present in clinical genomics, but clinical genomics in particular has seen important new innovations. One major contribution has been the creation and release of the public NIH database ClinVar (29), which allows different testing laboratories and companies to share pathogenicity assertions at the variant-phenotype level. Additionally, these documented assertions can be critically evaluated using tools like the Clinical Genome Resource (ClinGen) Pathogenicity Calculator, which automates and standardizes the application of ACMG/AMP guidelines (37).

Another major success has been the widespread dissemination of large-scale allele frequency databases across ancestrally diverse populations. For example, as of its February 2017 release, the Genome Aggregation Database (gnomAD) contained data from sequenced genomes and exomes from 138 632 individuals (38). This is more than twice as many as its precursor, the Exome Aggregation Consortium (ExAC) dataset, which itself

was 10 times larger than any previously available population database. By providing refined estimates of allele frequency across ancestrally diverse populations, these resources allow researchers to use ancestry and disease-specific allele frequency thresholds to reclassify variants interpreted as pathogenic or genes listed in recommended reporting guidelines (39). Allele frequency based approaches have proven especially useful when combined with large disease-specific cohorts, as demonstrated for cardiomyopathy (40) and prion disease (33). Nonetheless, even with gnomAD and large case cohorts, such analyses are often only able to make claims at the gene level (as opposed to the individual variant level) given the rarity of many disease-associated variants.

Clinical genomics has also benefitted from increased data availability across clinical modalities like EHRs and insurance claims data. New integrated datasets have enabled an approach to personalized medicine based on a broad picture of health (19,41). At the same time, many of these clinical datasets are not representative of the general population, and care must be taken to avoid selection bias. For example, genetic studies using EHR data from one geographic location may not generalize to new locations with different demographics. Other local biases stem from unique usage patterns across clinicians, departments and hospitals, making it difficult to conduct rigorous cross-system studies. Common to almost all EHRs, however, is the selection bias of hospital entry. Patient records tend to reflect a sicker and less ancestrally diverse subset of the general population. It is important to remember that EHR systems are designed primarily for clinical, administrative and financial purposes, including documentation, billing and public health surveillance; research functions such as data mining and clinical studies are largely secondary. Care must be taken when making conclusions with such data (42).

One example of the potential utility from integrative informatics approaches is the Integrated Personal Omics Profile (iPOP) (43). A study of an individual over a 14-month period combined longitudinal data on diet, stress and activity levels with broad omics data to uncover dynamic changes across molecular and physiological factors for both healthy and diseased states. Another effort to combine data types is the eMERGE Network, a consortium of biorepositories linked to EHRs (44). Like the iPOP profiles, the eMERGE network consolidates clinical data with omics data, albeit with a larger cohort at lower resolution. eMERGE represents a novel first step towards combining heterogeneous data sources at scale and has enabled several genome-wide association studies across a variety of phenotypes. However, further studies are needed to assess whether such large-scale integrative efforts will improve clinical utility when applied at a population scale.

Machine Learning for Clinical Genomics

Machine learning algorithms in clinical genomics generally take three main forms: supervised, unsupervised and semi-supervised. Supervised methods require data with observed labels (e.g. positive or negative disease status of a patient; pathogenic or benign status of a variant) that can be used to predict unobserved labels for new data. Unsupervised methods extract patterns from data features and do not require labels. Semi-supervised methods use the structure of unlabeled data to improve label predictions; this approach is especially useful for prediction tasks when data are plentiful but labels are not. All of these methods aim to optimize a performance measure related to the quality of predictions.

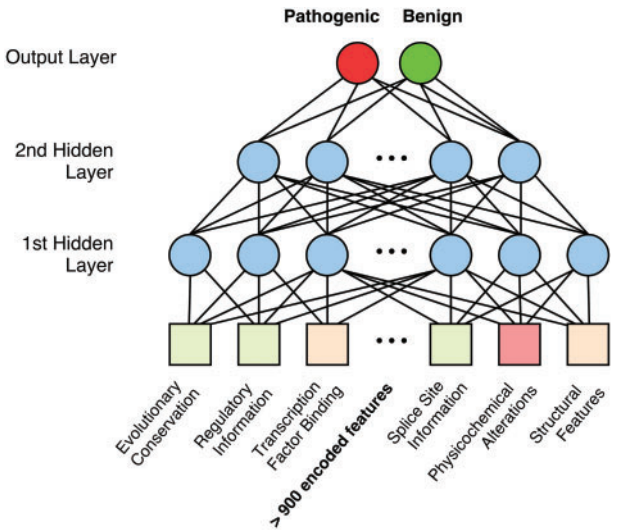

Figure 1. Predicting variant pathogenicity status using a neural network. Schematic representation of a neural network that predicts the pathogenicity status (pathogenic versus benign) of a genetic variant using a large number of input features, including sequence conservation, regulatory information and protein-level annotations. Feature scores are passed serially through successive interconnected layers and trained using a large set of labeled variants. Two hidden layers are shown in the schematic diagram above, but modern networks often consist of many more layers.

Recent algorithmic and hardware improvements, combined with large-scale data, have made it possible for machine learning methods to achieve state-of-the-art results on a wide range of tasks. Over the last decade, 'deep learning' methods in particular have outstripped traditional methods on many prediction and classification tasks (45), revolutionizing fields like image classification (46) and speech recognition (47). Deep learning utilizes a model known as an artificial neural network, consisting of many interconnected processing units. These units, also known as nodes or neurons, are arranged into layers that successively accept input from previous layers. Modern deep learning models are often parameterized by millions of different values that are iteratively adjusted using backpropagation and gradient descent methods. Deep learning enables rich and hierarchical representations of diverse and heterogeneous data types. Best outfitted for tasks that are complex and data-rich, deep learning appears to be well suited for many biological and clinical problems (16), including improving pathogenicity calls for variants and reframing the pathogenicity concept (Fig. 1).

Within the past decade, deep learning has made significant contributions to our ability to understand and interpret genomic data, which is generally characterized by very high dimensionality and sparsity. In principle, measurements may be collected across billions of genomic coordinates and hundreds of known sequences, cell types and tissues. Methods such as deep convolutional neural networks have been particularly effective at predicting sequence function and activity in different cell types. For example, the open-source software Basset (48) leverages deep convolutional neural networks to predict tissue-specific functional activity (e.g. DNase 1 hypersensitivity) from genomic sequence, trained on *in silico* and *in vitro* data from the ENCODE Consortium (49). Future experimental work will be important to clarify the validity of predictions from such approaches.

To many researchers, deep learning is seen as a way to improve performance on prediction tasks currently tackled by other models (e.g. linear models and support vector machines). In clinical genomics, supervised machine learning approaches have leveraged support vector machines for classifying deleterious variants (50) and scoring variant pathogenicity in hypertrophic cardiomyopathy (51). Similar methods have also successfully predicted the function, interactions and activity of variants and DNA sequences. More recent tools like DANN (52) have used deep learning to better capture complex relationships between input features and pathogenicity, but predictive performance guarantees for such approaches remain needed. Other potential uses of deep learning involve assessment of drug bioactivity and interactions, prediction of patient trajectories, and assignment to cohorts in clinical trials. While it remains unclear how frequently or consistently these tools are used across testing laboratories in clinical practice, they are central to a rapidly growing area in clinical genomics research.

More broadly, some of the major recent successes of machine learning in biomedical research have been achieved in image analysis, including segmentation, classification and diagnosis. This includes identifying bodily structures and landmarks in medical scans, predicting prognosis for patients with non-small cell lung carcinoma from stained histopathology slides (53), and diagnosing diabetic retinopathy from retinal fundus images (54). In many of these applications, data augmentation and 'transfer learning' from unrelated images have proven instrumental in overcoming small sample sizes. Future research will undoubtedly test the utility of integrating imaging techniques with genomic data. For example, researchers could evaluate the relationship between variants believed to be pathogenic for inherited heart disease and features extracted from automated analyses of cardiac imaging data (e.g. cardiac MRI and echocardiography).

Despite the successes of machine learning and deep learning in particular, applications in medicine face several unique challenges. One of the most significant problems is the dearth of reliably labelled examples. Data labels often come from clinicians or genetic counselors, who may be uncertain about their classifications or disagree with other experts. Moreover, because assembling such datasets may require time from specialist physicians, labelling large amounts of data may be prohibitively expensive. Interpretability presents another issue. Due to the stakes involved, clinical care requires a higher standard of justification than most applications of machine learning. Doctors, patients and lawyers may all want to know how an algorithm arrived at a certain decision or finding. Although researchers are investigating new methods to visualize and understand the inner workings of neural networks (55), such approaches remain underexplored in clinical genomics. Many such methods aim to show the importance of certain nodes, or 'average' representations of predicted classes, and not the decision-tree-like workflows that characterize differential diagnosis.

Some of these challenges may be addressed using other informatics approaches. For example, advances from natural language processing may allow weak phenotyping of images from radiological reports without the need for complete expert labelling. At the same time, researchers are developing algorithms that improve on current methods for interpreting neural networks. Although the best performing algorithms may simply be too complex to be meaningfully summarized, new methods and improvements to old methods may strike a better balance between accuracy and interpretability. Many solutions to healthcare-specific problems may also leverage the knowledge of human experts, including clinicians and providers. For example, the 'anchor and learn' framework uses expert knowledge to derive relationships between high-confidence observations and expected phenotypes that may be used to reliably infer labels (56). As argued several decades ago for medical reasoning more broadly (57), strategies that carefully blend categorical and probabilistic forms of reasoning about variants will likely prove most effective in clinical genomics.

Conclusion

Massive data repositories, modern algorithms and increased attention to actionability continue to inform and improve the clinical applications of genomics. These efforts have both enriched and complicated our understanding of the clinical utility of genetic testing. Recent efforts in informatics and machine learning have introduced improved models for representing pathogenicity, estimating penetrance and identifying incorrect or weakly supported variant classifications. However, the interpretability, generalizability and clinical validity of computational models still present significant limitations. Models that directly guide diagnoses and prognoses must demonstrate their reliability and accessibility to patients, clinicians and other stakeholders. It is critical that new methods continue to be developed and rigorous testing be performed in order for genomic information to be used effectively in the clinic.

Acknowledgements

The authors were supported by grants NIH BD2K U54HG007963, NIH OT3 OD025466-01 and NHLBI OT3 HL142480-01.

Conflict of Interest statement. None declared.

References

1. Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.
2. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. and Shendure, J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.
3. Rehm, H.L., Bale, S.J., Bayrak-Toydemir, P., Berg, J.S., Brown, K.K., Deignan, J.L., Friez, M.J., Funke, B.H., Hegde, M.R. and Lyon, E. (2013) ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.*, **15**, 733–747.
4. Manrai, A.K., Ioannidis, J.P.A. and Kohane, I.S. (2016) Clinical genomics: from pathogenicity claims to quantitative risk estimates. *JAMA*, **315**, 1233–1234.
5. Panagiotou, O.A., Willer, C.J., Hirschhorn, J.N. and Ioannidis, J.P.A. (2013) The power of meta-analysis in genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.*, **14**, 441–465.
6. Loh, P.R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., De Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S. *et al.* (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.*, **47**, 1385–1392.
7. Vassy, J.L., Lautenbach, D.M., McLaughlin, H.M., Kong, S.W., Christensen, K.D., Krier, J., Kohane, I.S., Feuerman, L.Z., Blumenthal-Barby, J., Roberts, J.S. *et al.* (2014) The MedSeq

Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials*, **15**, 85.
8. Vassy, J.L., Christensen, K.D., Schonman, E.F., Blout, C.L., Robinson, J.O., Krier, J.B., Diamond, P.M., Lebo, M., Machini, K., Azzariti, D.R. *et al.* (2017) The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: a pilot randomized trial. *Ann. Intern. Med.*, **167**, 159–169.
9. Rubinstein, W., Maglott, D., Lee, J.M., Kattman, B.L., Malheiro, A.J., Ovetsky, M., Hem, V., Gorelenkov, V., Song, G., Wallin, C. (2013) The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res.*, **41**, D925–D935.
10. Genetic Testing Registry (GTR) - NCBI. https://www.ncbi.nlm.nih.gov/gtr/; date last accessed February 11, 2018.
11. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L. *et al.* (2015) ClinGen - The Clinical Genome Resource. *N. Engl. J. Med.*, **372**, 2235–2242.
12. Amendola, L.M., Jarvik, G.P., Leo, M.C., McLaughlin, H.M., Akkari, Y., Amaral, M.D., Berg, J.S., Biswas, S., Bowling, K.M., Conlin, L.K. *et al.* (2016) Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. *Am. J. Hum. Genet.*, **98**, 1067–1076.
13. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J. and Kohane, I.S. (2016) Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.*, **375**, 655.
14. Weber, G.M., Mandl, K.D. and Kohane, I.S. (2014) Finding the missing link for big biomedical data. *JAMA*, **311**, 2479–2480.
15. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Liu, P.J., Liu, X., Sun, M., Sundberg, P., Yee, H. *et al.* (2018) Scalable and accurate deep learning for electronic health records. arXiv: 1801.07860.
16. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Xie, W., Rosen, G.L. *et al.* (2018) Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv, J. R. Soc. Interface*, **15**(141), doi: 10.1098/rsif.2017.0387.
17. Beaulieu-Jones, B.K. and Greene, C.S. (2016) Semi-supervised learning of the electronic health record for phenotype stratification. *J. Biomed. Inform.*, **64**, 168–178.
18. Kohane, I.S. (2011) Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.*, **12**, 417–428.
19. Carey, D.J., Fetterolf, S.N., Davis, F.D., Faucett, W.A., Kirchner, H.L., Mirshahi, U., Murray, M.F., Smelser, D.T., Gerhard, G.S. and Ledbetter, D.H. (2016) The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet. Med.*, **18**, 906–913.
20. Charles, D., Gabriel, M. and Searcy, T. (2015) Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008–2014. ONC Data Brief, no.35. Office of the National Coordinator for Health Information Technology: Washington DC.
21. Jha, A.K., Doolan, D., Grandt, D., Scott, T. and Bates, D.W. (2008) The use of health information technology in seven nations. *Int. J. Med. Inform.*, **77**, 848–854.
22. Collins, F.S. and Varmus, H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.
23. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.*, **12**, e1001779.
24. Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L., Lancaster, G., Yang, X., Williams, A. *et al.* (2011) China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.*, **40**, 1652–1666.
25. Leitsalu, L., Alavere, H., Tammesoo, M.-L., Leego, E. and Metspalu, A. (2015) Linking a population biobank with national health registries: the Estonian experience. *J. Pers. Med*, **5**, 96–106.
26. Lek, M., Karczewski, K.J., Samocha, K.E., Banks, E., Fennell, T., O, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., Birnbaum, D.P. *et al.* (2016) Analysis of protein-coding genetic variation in 60 706 humans. *bioRxiv*, **536**, 285.
27. NHLBI GO Exome Sequencing Project (ESP) Exome Variant Server. http://evs.gs.washington.edu/EVS/.
28. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–423.
29. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M. and Maglott, D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
30. Zöllner, S. and Pritchard, J.K. (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.*, **80**, 605–615.
31. Struewing, J.P., Hartge, P., Wacholder, S., Baker, S.M., Berlin, M., McAdams, M., Timmerman, M.M., Brody, L.C. and Tucker, M.A. (1997) The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N. Engl. J. Med.*, **336**, 1401–1408.
32. Chen, S. and Parmigiani, G. (2007) Meta-analysis of BRCA1 and BRCA2 penetrance. *J. Clin. Oncol.*, **25**, 1329–1333.
33. Minikel, E.V., Vallabh, S.M., Lek, M., Estrada, K., Samocha, K.E., Sathirapongsasuti, J.F., McLean, C.Y., Tung, J.Y., Yu, L.P.C., Gambetti, P. *et al.* (2016) Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.*, **8**, 322ra9.
34. Beutler, E., Felitti, V.J., Koziol, J.A., Ho, N.J. and Gelbart, T. (2002) Penetrance of 845G–> A (C282Y) HFE hereditary haemochromatosis mutation in the USA. *Lancet*, **359**, 211–218.
35. Hamblin, A., Wordsworth, S., Fermont, J.M., Page, S., Kaur, K., Camps, C., Kaisaki, P., Gupta, A., Talbot, D., Middleton, M. *et al.* (2017) Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: retrospective validation and prospective audit in the UK National Health Service. *PLoS Med.*, **14**, e1002230.
36. Le Tourneau, C., Delord, J.P., Gonçalves, A., Gavoille, C., Dubot, C., Isambert, N., Campone, M., Trédan, O., Massiani, M.A., Mauborgne, C. *et al.* (2015) Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol.*, **16**, 1324–1334.
37. Patel, R.Y., Shah, N., Jackson, A.R., Ghosh, R., Pawliczek, P., Paithankar, S., Baker, A., Riehle, K., Chen, H., Milosavljevic, S. *et al.* (2017) ClinGen pathogenicity calculator: a configurable system for assessing pathogenicity of genetic variants. *Genome Med.*, **9**, 3.

38. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60 706 humans. *Nature*, **536**, 285–291.
39. Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., Macarthur, D. *et al.* (2017) Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.*, **19**, 1151–1158.
40. Walsh, R., Thomson, K.L., Ware, J.S., Funke, B.H., Woodley, J., McGuire, K.J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J.C. *et al.* (2017) Reassessment of Mendelian gene pathogenicity using 7855 cardiomyopathy cases and 60 706 reference samples. *Genet. Med.*, **19**, 192–203.
41. Jensen, P.B., Jensen, L.J. and Brunak, S. (2012) Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.*, **13**, 395–405.
42. Ioannidis, J.P.A. (2013) Are mortality differences detected by administrative data reliable and actionable? *JAMA*, **309**, 1410–1411.
43. Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y.K., Chen, R., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E. *et al.* (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, **148**, 1293–1307.
44. McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R., Masys, D.R., Ritchie, M.D., Roden, D.M. *et al.* (2011) The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics*, **4**, 13.
45. LeCun, Y.A., Bengio, Y. and Hinton, G.E. (2015) Deep learning. *Nature*, **521**, 436–444.
46. Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) {ImageNet} classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, doi: 10.1016/j.protcy.2014.09.007.
47. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. *et al.* (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.*, **29**, 82–97.
48. Kelley, D.R., Snoek, J. and Rinn, J.L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.
49. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
50. Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
51. Jordan, D.M., Kiezun, A., Baxter, S.M., Agarwala, V., Green, R.C., Murray, M.F., Pugh, T., Lebo, M.S., Rehm, H.L., Funke, B.H. *et al.* (2011) Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. *Am. J. Hum. Genet.*, **88**, 183–192.
52. Quang, D., Chen, Y. and Xie, X. (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.
53. Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L. and Snyder, M. (2016) Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.*, **7**, 12474.
54. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. *et al.* (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J. Am. Med. Assoc.*, **316**, 2402–2410.
55. Shrikumar, A., Greenside, P. and Kundaje, A. (2017) Learning important features through propagating activation differences. arXiv:1704.02685.
56. Halpern, Y., Horng, S., Choi, Y. and Sontag, D. (2016) Electronic medical record phenotyping using the anchor and learn framework. *J. Am. Med. Informatics Assoc.*, **23**, 731–740.
57. Szolovits, P. and Pauker, S.G. (1978) Categorical and probabilistic reasoning in medical diagnosis. *Artif. Intell.*, **11**, 115–144.