Human Genetics (2021) 140:231–240
https://doi.org/10.1007/s00439-020-02208-5

REVIEW

Population genetics: past, present, and future

Atsuko Okazaki[1,2] · Satoru Yamazaki[3] · Ituro Inoue[4] · Jurg Ott[2]

Received: 12 May 2020 / Accepted: 14 July 2020 / Published online: 18 July 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
We present selected topics of population genetics and molecular phylogeny. As several excellent review articles have been published and generally focus on European and American scientists, here, we emphasize contributions by Japanese researchers. Our review may also be seen as a belated 50-year celebration of Motoo Kimura's early seminal paper on the molecular clock, published in 1968.

Introduction

In recent years, large amounts of DNA sequencing data have been generated in various projects such as 1000 Genomes (Genomes Project et al. 2010, 2012, 2015), the ALSPAC database (Fraser et al. 2013; Hameed et al. 2017), and Icelandic (Gudbjartsson et al. 2015), and Japanese populations (Nagasaki et al. 2015). Major achievements of these efforts have been as follows: (1) Larger genetic variation is observed within populations than between populations, and (2) each individual harbors large numbers of variants with low allele frequencies. These findings have long ago been predicted by population genetics and evolutionary studies. Therefore, it is instructive to look back at historic achievements in population genetics.

Excellent reviews of population genetics have been written (Chakraborty 2006; Charlesworth and Charlesworth 2017; Crow 1987; Crow and Kimura 1970) documenting the development of population genetics from early achievements by Mendel (1866), Hardy (1908), and Weinberg (1908) up to highly sophisticated theoretical developments, mostly by American, British, and Japanese scientists. Here, we review selected aspects of population genetics, genome evolution, and molecular phylogeny with an emphasis on contributions by Japanese researchers.

Historical aspects of population genetics and road to the neutral theory

Darwin's theory of evolution through selection very well explains changes in time of heritable phenotypes. In the early 1900s, focusing on the evolution of genetic variants in the population, R. A. Fisher, S. Wright, and J. B. S. Haldane made fundamental theoretical contributions to population genetics (Provine 1971), Fisher in his 1922 paper (Fisher 1922), which was the first to introduce diffusion equations into population genetics, and Haldane in developing in 1927 (Haldane 1927) the approximation of change of numbers of copies of very rare mutants by branching processes. Wright (1938) developed the theory on the effects of genetic drift, that is, random changes in small populations. While his theory was supported only by a minority of scientists in an era when the molecular basis of genes had yet to be proven and the effects of genetic drift were underestimated, Wright's theory made a great contribution to connecting Mendelian Genetics with the Darwinian theory of evolution.

More recently, it has become apparent that many molecular changes have no effects on phenotypes. Based on Wright's drift hypothesis and Haldane's approximation model of an advantageous mutation (Haldane 1927), Motoo Kimura (1964) then developed his neutral theory based on backward diffusion models, which showed the probability of fixation to zero of a variant in the population to be equal to 2 $s(N_e/N)$, where s is the selection

✉ Jurg Ott
ott@rockefeller.edu

[1] Intractable Disease Research Center, Juntendo University, Tokyo, Japan

[2] Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, NY 10065, USA

[3] Department of Molecular Pharmacology, National Cerebral and Cardiovascular Center, Osaka, Japan

[4] Division of the Human Genetics, National Institute of Genetics, Shizuoka, Japan

coefficient, N the size of the breeding population, and N_e the effective population size.

Mutations and selection are driving forces for evolution. Basically, mutations occur at random DNA bases. Harmful mutations tend to be eliminated within a short period of time and do not contribute to long-term evolution. This process is called negative or purifying selection as opposed to positive selection. Before Kimura (1964) proposed his neutral theory, there was little notion of neutral variation, although, at about the same time, Lewontin and Hubby (1966) considered the possibility of neutral mutation as a possible reason for a large amount of variation which they found in electrophoretic mobility. Still, natural selection was the mainstream hypothesis with the idea that advantageous variations in populations are the driving forces for evolution, and deleterious variations are removed in a rapid manner.

At the time, population genetics usually considered two alleles at each gene locus based on the assumption of genes being base pairs. On the other hand, Kimura and Crow (1964) assumed an infinite allele model ("neutral isoalleles") and proposed that genetic variation in populations arises as to the balance between mutations and genetic drift. Comparing hemoglobin molecules between different organisms, Kimura (1968) postulated that amino-acid substitution rates are so high that they can only be explained by neutral mutations. In other words, mutation and random changes in a finite population can maintain considerable variation through random fixation of selectively neutral or nearly neutral mutants. In the light of current knowledge, however, Kimura's reasoning appears somewhat flawed. For example, he argued that the "cost of natural selection" would be too high otherwise—more consideration has shown that no cost is imposed by beneficial mutations in the absence of environmental deterioration. He also used the total amount of DNA without distinguishing protein-coding regions and non-coding regions. Nonetheless, Kimura's contributions to population genetics have been tremendous.

Together with the Darwinian selection hypothesis, the neutral theory is one of the two pillars of genome evolution. Thus, 'survival of the luckiest, and not necessarily of the fittest' may be a good explanation for the evolution of a great majority of genetic changes (Chakraborty 2006). Interestingly, Kimura (1969) also proposed the "infinite sites model". In this model, if the mutation rate is low and the effective population size is small ($\theta = 4N_e\mu \ll 1$), a mutant variant will always appear at a different site in the genome. If so, identity by state at the variant can be regarded as identity by descent, and in this respect, the infinite sites model represents one of the bases for genome-wide association studies using SNPs as genetic markers in unrelated individuals (Sella and Barton 2019).

The nearly neutral theory

The evolutionary rate, $\lambda = f\mu$, in the neutral theory (f is the proportion of neutral mutations among all mutations in a gene, μ is the mutation rate) disregards mutations favorable to survival and simply classifies other mutations into neutral (f) and deleterious ($1 - f$) mutations. However, the extent of harmfulness measured by the selection coefficient, s, is a continuous quantity. Based on these ideas, Tomoko Ohta (Ohta 1973, 1992, 2002), who had built the foundation of the neutral theory with Motoo Kimura, proposed the "nearly neutral" theory, where slightly disadvantageous mutations (attenuated mutations) could persist in the population by chance if the population is small. Thus, according to her publications (Ohta 1973, 1992, 2002), a substantial fraction of changes is caused by random fixation of nearly neutral changes, a class that includes intermediates between neutral and advantageous, as well as between neutral and deleterious classes, although other population geneticists may disagree with this view (Kondrashov 1995; Nei 2005).

A difference from the neutral theory is that the nearly neutral theory allows for interactions between (1) genes having occurred through weak natural selection (or weak deleterious selection) and (2) genes without weak natural selections, and for the two types of genes to jointly contribute to evolution by opposing the action of genetic drift (Hurst 2009). In the nearly neutral theory, the effect of genetic drift is weakened, and slightly disadvantageous mutations are excluded from a population if the population is extremely large; if a population is small, then slightly disadvantageous mutations are kept (some are even fixed) by the effects of genetic drift. It seems that the structure of very large datasets such as 1000 Genomes or the Exome Sequencing Project 6500 can be explained by the nearly neutral theory, because there is increasing evidence that selection pressure in small populations such as mammals including humans is weaker compared to that in ancestral species, and slightly disadvantageous mutations have been accumulating in populations (Kosiol et al. 2008; Nelson et al. 2012; Nielsen et al. 2009; Tennessen et al. 2012).

Evolutionary rate of pseudogenes

In the second half of 1970, accumulated sequencing data confirmed the prediction by King and Jukes (1969) that mutation rates of synonymous variants are higher than those of non-synonymous variants, which supports the neutral theory. Kimura (1977) asserted that according to the neutral mutation-random drift hypothesis, most mutant substitutions detected among organisms should be the results of random fixation of selectively neutral or nearly neutral mutations. This conjecture was verified by the analysis of mutation rates of pseudogenes, that is, of genes with sequences similar to normal genes having lost their functions as they were duplicated to another location in the genome, and in the process, their transcription sequences were not preserved. Based on the neutral theory, Takashi Miyata calculated the replacement rates of non-synonymous variants and synonymous variants in nucleotide sequences of several pseudogenes, α and β globin, and compared them with those in their functional counterparts (Miyata and Hayashida 1981). Results showed that replacement rates were uniformly the same in different pseudogenes and almost equal to the mutation rate, with no other gene evolving at a faster rate. This observation clearly supported the neutral theory.

Junk DNA, a term publicized by Susumu Ohno (1972) but rarely used today (see below), contains inter-genic regions, most of which are SINEs (*S*hort *IN*terspersed *E*lements) and LINEs (*L*ong *IN*terspersed *E*lements). The term ‘junk DNA’ was mentioned by a few other authors in 1972 and even 9 years earlier in a paper little known to human geneticists (Ehret and De Haller 1963), but Ohno’s name tends to be most closely associated with this term.

Evolutionary rates of junk DNA are expected to be similar to those of synonymous mutations and pseudogenes. In mammals, most of the genome regions, likely well more than 90%, are predicted to be junk DNA. Therefore, evolutionary rates of whole genomes can be approximated as being those of junk DNA.

In 2012, the Encyclopedia of DNA elements (ENCODE) project (Consortium 2012) proved biochemical functions of 80% of the genome, especially outside of protein-coding regions, which was once considered junk DNA. The findings from the ENCODE project enable us to further explore the function of the human genome.

Genes and genomic duplication

In higher organisms, genomic duplication is known to be extremely important for evolution. Early on, Susumu Ohno proposed that evolution is caused by genomic duplication, which was a visionary idea at a time when large sequencing data were not yet available (Ohno 1970). It has been shown empirically and by theoretical considerations that the advantage of creating new copies of genomes (or individual genes) can result in higher fitness. An alternative model explaining genomic duplication is DDC (*D*uplication *D*egeneration *C*omplementation) (Lynch and Conery 2000). In the DDC model, regulatory elements each controlling independent functions are duplicated and random null mutations in the regulatory elements through degeneration lead to sub-functionalization, where the regulatory elements complement each other to achieve the full ancestral repertoires. What is important in the process is that it does not require the help of positive selection, that is, functional diversification. In practice, it has been proposed that the selection of slightly disadvantageous mutations works with the expression level of each gene changing. Therefore, genetic duplication is predicted to proceed in a nearly neutral manner based on mutation pressure and genetic drift. In addition, “concerted evolution” in minisatellites used as markers for hyper-polymorphisms, and in other sequences such as rRNA genes can be explained well by Ohno’s theory (Hillis et al. 1991; Jeffreys et al. 1985).

Molecular phylogeny

Through evolution, currently, living organisms have descended from common ancestors. Systematic biology seeks to unravel relationships among organisms and to establish evolutionary trees. As every biology student knows, the classical approach to such discoveries is through painstaking analysis of morphological details. Depending on which of these phenotypes are considered most important, different relationships among organisms emerge.

Rather than relying on phenotypes that may or may not be heritable, molecular phylogeny relies on DNA sequences and their comparisons among organisms. Researchers with various backgrounds have made significant contributions to methods of creating phylogenetic trees and the evaluation of phylogenetic relationships. In this field, Joseph Felsenstein almost single-handedly established this field as a special branch of population genetics (Felsenstein 2004). For example, he introduced the maximum-likelihood method of establishing phylogenetic trees (Felsenstein 1978) (see below). One of his other contributions is the “Felsenstein Zone” (Huelsenbeck and Hillis 1993), which involves the phenomenon of “long-branch attraction”; that is, long branches will appear similar to each other and appear as sister taxa on a tree even though they do not share a common ancestry. The Zone is the set of trees on which long-branch attraction occurs. Such phenomena have been observed in many datasets and simulation analyses, and have led to the discovery of long-branch attraction, which leads to wrongly assuming phylogeny where none exists (Huelsenbeck and Hillis 1993). Furthermore, Felsenstein contributed greatly to molecular phylogeny by developing a program package, PHYLIP, combining various phylogenic tree estimation methods including DNAML. Thanks to his contributions, molecular phylogeny has become increasingly popular for empirical molecular evolutionists.

The development of molecular phylogeny may not seem to be related to disease gene discovery. However, it greatly contributes to such discoveries through interpretation of huge sequencing datasets obtained from the 1000 Genomes project and other projects. Generating a molecular phylogenetic tree for phylogenetic relationships between species led to the discovery of gene families (orthologs and paralogs). The coalescent theory, which examines the gene tree in a species by reversing the time, was also applied to reconstruct the demographic history of species of interest. In particular, regarding the coalescent theory, Tajima (1983) estimated nucleotide diversity based on the limited DNA polymorphic data, calculated the time of coalescence of genes sampled from a single population, and their theory applies to a few genes at the time of population splitting. Takahata and Nei (1985) further developed a coalescent theory from DNA sequencing data and theoretically showed that alleles with deep coalescences are relatively rare.

The neighbor-joining method

Many methods for creating (estimating) phylogenic trees have been developed. Historically, these methods can roughly be classified into two groups, distance matrix methods and character state methods. The former uses a distance matrix and estimates evolutionary distance such as the number of amino-acid substitutions or base substitutions based on all possible pairs of OTUs (Operational Taxonomic Units). This method was first applied to create phylogenic trees in the form of the UPGMA (Unweighted Pair Group Method with Arithmetic mean) method, where clusters of neighboring OTUs are created and connected in a stepwise fashion. The method is used not only for amino-acid or base-pair sequences but also in numerical taxonomy, which deals with expression analysis using microarray (Eisen et al. 1998) or trait-encoded information (Sokal and Michener 1958). However, since this method assumes constant evolutionary speed, it is problematic to apply to amino-acid or base-pair sequence data. To overcome this problem, distance methods were developed that did not assume a molecular clock (Fitch and Margoliash 1967). Masatoshi Nei and Naruya Saitou greatly improved upon this method and developed a much faster procedure (Saitou and Nei 1987). This method is one of the "star decomposition" methods that determine which, of a given pair of sequences, reduces length of the total tree most and combine neighboring nodes until all OTUs are included. In the neighbor-joining method, "neighbors" keep track of nodes on a tree rather than taxa or clusters of taxa. A modified distance matrix is obtained in which the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes. The tree is constructed by joining the least-distant pair of nodes in this modified matrix. When two nodes are joined, their common ancestral node is added to the tree and the terminal nodes with their respective branches are removed from the tree. At each stage in the process, two terminal nodes are replaced by one new node. This iterative operation finds "neighbors" one after another, which creates the final phylogenetic tree. The neighbor-joining method is the most commonly used distance matrix method. Starting in 1971, Nei proposed that Nei's distance be used for phylogenetic tree estimation, which was later incorporated into the neighbor-joining program package MEGA (Kumar et al. 1994; Saitou and Nei 1987).

The second group, character state methods, do not use a distance matrix and define characters (phenotypes) and use them for exploring tree topology. One of the examples of character state methods is the maximum-likelihood method discussed in the next section.

The maximum-likelihood method

Maximum likelihood (ML) was developed by Fisher (1922) as a method to estimate parameters in statistical models. It has several advantages over other methods, but tends to be more complicated to apply than simpler methods. In population genetics, Luigi Luca Cavalli-Sforza first applied the ML method to an approach for creating phylogenic trees based on allele frequencies (Cavalli-Sforza and Edwards 1967). The first use of maximum-likelihood inference of trees from molecular sequences was by Jerzy Neyman (Felsenstein 2001; Neyman 1971). Felsenstein proposed ML for creating phylogenic trees based on allele frequencies as continuous quantities (Felsenstein 1973a), thus improving on the method previously proposed by Cavalli-Sforza, and introduced ML for estimating trees based on discrete datasets and the maximum parsimony criterion (Felsenstein 1973b). Masami Hasegawa incorporated this approach into the MOLPHY program package and pioneered in the use of model selection methods such as AIC in comparing phylogenies (he was a member of Akaike's institute) (Adachi and Hasegawa 1992, 1996).

The ML method is the most efficient approach among all tree construction methods. For example, false-positive evidence of relationships of long branches ("long-branch attraction") will not occur when trees are estimated by ML and the model of evolution is correct, although it can occur when the model is not correct. However, the ML method tends to be time-consuming and, for some large trees, may be impossible to apply.

Impact of variants on multifactorial disorders and missing heritability

Based on the material mentioned so far, we will now cover some topics on how progress in population genetics, genome evolution, and phylogenic studies can be applied to medical research.

Multifactorial disorders are assumed to occur through interactions between multiple genetic and environmental factors. Therefore, identifying disease susceptibility genes has been considered difficult, and detecting interactions with environmental factors even more so. Especially in the 1990s, such considerations were widespread, quite in contrast to the relative ease with which increased numbers of gene identifications for monogenic disorders have been achieved. However, there was a researcher to struggle with the solution for genetic causes of multifactorial disorders at that time. Ituro Inoue succeeded in narrowing down disease loci using linkage analysis with affected sib-pairs and constructing haplotypes of the angiotensinogen (AGT) gene using limited data (Inoue et al. 1997). Inoue assessed linkage disequilibrium (LD) at each site in the AGT gene and further demonstrated by in vitro functional assay that the combination between A (− 6) and T235 alleles affects the expression of the AGT gene. This study was visionary, since LD block structures had yet to be proved at that time.

After that, genome-wide association studies with large SNP data over the whole genome became available thanks to the HAPMAP project, SNP collections by Perlegen Science, LD block measurements, and construction of haplotype maps (HapMap 2005; Hinds et al. 2005). Although such genome-wide studies contributed to narrowing down locations of disease susceptibility genes, results are still insufficient for identifying many specific disease susceptibility genes, for example Moyamoya disease (Liu et al. 2011). A remaining challenge has been that identified susceptibility loci show only small odds ratios, and all susceptibility loci combined only explain up to 30% of most of the disease causes. These numbers are generally smaller than the heritability calculated in the previous twin studies, which is known as "missing heritability" (Manolio et al. 2009). Nowadays, however, methods for calculating SNP-based heritability have been developed (Yang et al. 2017) that come up with heritability estimates close to those obtained by classical segregation analysis, and part of the problem seems to be resolved.

Out-of-Africa hypothesis

Recent advances in sequencing technology have enabled the identification of whole genome structures at population levels. These successes have made it possible to compare current human genome sequences with ancient genomes such as *Homo neanderthalensis* or *Denisova hominin*, which greatly contributed to the understanding of the origin of *Homo sapiens* (Nielsen et al. 2017). Allan Wilson, along with Rebecca Cann and Mark Stoneking, first proposed the "out-of-Africa" hypothesis (Cann et al. 1987), which claims that *Homo sapiens* originated in Africa and then spread all over the world. They based their results on the analysis of mitochondrial DNA of various populations, which represented the first phylogenic tree of *Homo sapiens*. Work by Masatoshi Nei contributed to the out-of-Africa hypothesis: In the 1970s, Nei calculated heterozygosity for various protein isozymes and created phylogenic trees of *Homo sapiens* (Nei and Roychoudhury 1972, 1974; Nielsen et al. 2017). An interesting finding based on this work is that genetic variation estimated by Nei's distance or Wright's F_{st} is larger within populations than between populations (Lewontin 1972), which was later confirmed by the 1000 Genomes project. In other words, there are greater differences among individuals in a given population than between populations. However, this notion has also been challenged (Edwards 2003).

Relationship between recent explosive population growth and origin of deleterious variants

Numerous human genome sequence projects such as 1000 Genomes revealed that each individual harbors considerable numbers of private mutations. This fact had been proposed by Haldane in his "genetic load" theory, which predicted an association between the numbers of variants possessed over populations and survival rate (Haldane 1937). In his theory, he claimed that if we consider genetic load for the whole genome rather than a given locus, the fitness decrease by mutations is equal to the mutation rate, v, irrespective of the extent of selection. He also claimed that pathogenic mutations accumulate in the form of heterozygous variants unless such mutations are excluded as lethal homozygous mutations (Haldane 1937) (this theory is also known as the Haldane–Muller principle). The theory of genetic load was further elaborated upon by Kimura (1960); for neutral mutations, there is no load. Based on this background, for variants whose distributions differ among populations, estimating the age of each variant becomes possible, which is important for understanding the history of human evolution, as well as for developing novel methods for disease gene discovery. The mathematical theory of coalescence allowing haplotype and allele ages to be calculated was developed by John Kingman (2000), and Kimura and Ohta (1973) proposed a formula for determining allele age, $-2x(1-x)/\log(x)$. This formula represents the expected age of a neutral mutation of frequency x in a stationary population based on a diffusion process used in classical population genetics. Although there was a discussion regarding the restrictive assumption that the age distribution of a mutant allele with population frequency x should be the same as the distribution of the time to extinction of the allele, conditional on extinction, it made a great contribution to later calculations of allele age (Fu et al. 2013). Calculating allele age assuming the infinite many sites of model of mutation developed Kimura and Ohta formula, it showed that about three-quarters of all protein-coding SNV predicted to be deleterious across in the past 5000 years (Fu et al. 2013). This attempt provides important practical information that can be prioritized variants in disease gene discovery.

Inbreeding (mating between relatives) has so far not been discussed here as it does not lead to changes in allele frequencies. It does, however, lead to a decrease in heterozygotes and a corresponding increase in homozygotes. As is well known, at a bi-allelic locus with allele frequency p, the proportion of heterozygotes is given by $2p(1-p)(1-F)$, where F is the inbreeding coefficient. In many human populations, F tends to be rather small; for example, $F = 0.00038$ in the UK (Pattison 2016). An exception is offspring of first cousins ($F = 1/16$). For rare deleterious recessive traits with disease allele frequency p, recessive offspring of first-cousin marriages occur with probability $p^2 + p(1-p)F$ (Haldane and Moshinsky 1939). Through genetic linkage of such a trait with SNPs surrounding it, rare recessive traits tend to be located in long runs of homozygous SNPs (homozygosity mapping (Lander and Botstein 1987)). More modern approaches have been developed, for example, based on the Hamming distance between chromosomes in affected and control individuals (Imai et al. 2015). This approach revealed a mutation, p.H96R in the BOLA3 gene, possibly having originated in a single Japanese founder individual (Imai et al. 2016).

Darwinian (evolutionary) medicine

From the viewpoint of Darwinian medicine (or evolutionary medicine), which is medicine based on evolution (Williams and Nesse 1991), we discuss a few aspects of how discovering variants can translate into medical care.

In the 1960s, Richard Lewontin discovered in Drosophila populations that heterozygosity is more often observed than expected (Lewontin and Hubby 1966). He interpreted this finding as advantageous fitness of heterozygosity compared to the homozygous state of the wild type or mutant (so-called over-dominance, or balancing selection) and emphasized its importance for survival. After the establishment of the neutral theory, as described below, the importance of balancing selection for some types of variants with high allele frequencies was rediscovered. Theoretical studies on natural selection also greatly progressed and “Tajima’s D”, developed by Fumio Tajima, is computed as the difference between two measures of genetic diversity: the mean number of pairwise differences and the number of segregating sites, each scaled so that they are expected to be the same in a neutrally evolving population of constant size. This is a unique contribution to statistical genetics by Japanese researchers in that this method can assess whether a given variant scattered over the whole genome is neutral or under selection pressure (Tajima 1989).

Analyzing genome sequences in several populations using the techniques of next-generation sequencing reveals some signals with positive selection pressure. One such example is infection-related diseases. Regarding the natural selection for resistance of a pathogen, this was revealed by next-generation sequencing to represent the strongest positive selection pressure in human evolution; that is, the well-known balancing signals on glycoproteins and positive selection signals on TLRs (Ferrer-Admetlla et al. 2008). Applying the history of evolution for various pathogens to disease susceptibility research will likely identify functional variants as well as intra-cellular mechanisms and treatment for various diseases. We believe that selection pressure for ancient pathogens will affect not only infectious and auto-immune diseases but also other traits. Recently, the association between life-style diseases and natural selection has become an attractive topic. Using 40 traits from the UK Biobank, functional low-frequency variants have been revealed to be under negative selection (Gazal et al. 2018). An alternative suggestion has been that positive selection acts on susceptibility loci for life-style diseases. An example is the thrifty gene hypothesis. At the dawn of the era of genomic medicine, the ancient history of human evolution is a powerful tool for understanding human biology leading to improving human health.

Discussion

In this outline, we deliberately emphasized contributions to population genetics by Japanese researchers—in this field, Japanese scientists have arguably carried out comprehensive fundamental work. Thus, we feel justified in presenting this short review of population genetics from a Japanese point of view.

In terms of future developments in population genetics, we expect DNA sequencing to play an ever-increasing role. In an era where human genome sequence projects are underway around the world, established population genetics principles will be applied to reveal more detailed migration history, population history, and mechanisms of selection pressure, particularly in small ethnic populations (Antonio et al. 2019; Lipson et al. 2020).

Technological advances have changed the landscape of genetic screening (Ceyhan-Birsoy et al. 2019). Together with epidemiological and molecular genetics studies, population genetics approaches have demonstrated the association between disease mechanisms and mutations in populations. Cystic fibrosis is one such successful example (Bell et al. 2020). By identifying the relationship between specific mutations and a cystic fibrosis transmembrane conductance regulator (CFTR) defect, we can improve patient care including disease monitoring and treatment decisions. In the future, improvement of patient care in more diseases can be achieved by the combination of population genetics, epidemiological studies, and molecular genetics studies.

With the huge amount of genomic information currently available, it is challenging to link genotypes to phenotypes, predict regulatory functions, and classify mutant types. Therefore, new and innovative approaches are needed for further understanding of medical biology and connections to genetic disease. One approach is to collect previously reported SNV information and create a suitable mathematical model. As an example, a study by Davis et al. (2016) describes a biophysical metric of cardiomyocyte function, which accurately predicts human cardiac phenotypes.

Another approach is based on neural networks to automatically extract relevant features from input data (Zou et al. 2019). Since advances in sequencing technologies provide large amounts of data, it is realistic to utilize machine learning as a tool for analysis in the field of clinical healthcare and population genetics. Although deep learning has great potential, attempts to apply it to genomics have only just begun. For example, SpliceAI, a 32-layer deep neural network (DNN) was developed for predicting de novo mutations with predicted splice-altering consequences in patients with neurodevelopmental disorders, which paves the way for the application of deep learning on complex genetic variant prediction (Jaganathan et al. 2019). To identify pathogenic mutations in patients with rare diseases, a DNN model was developed combining common variants derived from human and six non-human primate species. The proposed model achieved an 88% accuracy and found 14 unreported candidate genes associated with intellectual disability (Sundaram et al. 2018).

Finally, epidemics and pandemics of viruses and their sequences provide rich sources of information. For example, population genetic analyses of 103 SARS-CoV-2 genomes indicated the presence of two major lineages, although the implications of these evolutionary changes remained unclear (Tang et al. 2020).

Acknowledgements Helpful comments by Prof. Joseph Felsenstein on an earlier version of this manuscript are gratefully acknowledged. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant numbers JP20K08497 and JP18K15863 (A. O.), and Grant number 19K09408 (S. Y.).

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

Adachi J, Hasegawa M (1992) MOLPHY, programs for molecular phylogenetics. I, PROTML, maximum likelihood inference of protein phylogeny. Computer science monographs, no. 27. Institute of Statistical Mathematics, Tokyo, pp 1–14

Adachi J, Hasegawa M (1996) MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Computer science monographs, no. 28. Institute of Statistical Mathematics, Tokyo, pp 1–150

Antonio ML, Gao Z, Moots HM, Lucci M, Candilio F, Sawyer S, Oberreiter V, Calderon D, Devitofranceschi K, Aikens RC, Aneli S, Bartoli F, Bedini A, Cheronet O, Cotter DJ, Fernandes DM, Gasperetti G, Grifoni R, Guidi A, La Pastina F, Loreti E, Manacorda D, Matullo G, Morretta S, Nava A, Fiocchi Nicolai V, Nomi F, Pavolini C, Pentiricci M, Pergola P, Piranomonte M, Schmidt R, Spinola G, Sperduti A, Rubini M, Bondioli L, Coppa A, Pinhasi R, Pritchard JK (2019) Ancient Rome: a genetic crossroads of Europe and the Mediterranean. Science 366:708–714. https://doi.org/10.1126/science.aay6826

Bell SC, Mall MA, Gutierrez H, Macek M, Madge S, Davies JC, Burgel PR, Tullis E, Castanos C, Castellani C, Byrnes CA, Cathcart F, Chotirmall SH, Cosgriff R, Eichler I, Fajac I, Goss CH, Drevinek P, Farrell PM, Gravelle AM, Havermans T, Mayer-Hamblett N, Kashirskaya N, Kerem E, Mathew JL, McKone EF, Naehrlich L, Nasr SZ, Oates GR, O'Neill C, Pypops U, Raraigh KS, Rowe SM, Southern KW, Sivam S, Stephenson AL, Zampoli M, Ratjen F (2020) The future of cystic fibrosis care: a global perspective. Lancet Respir Med 8:65–124. https://doi.org/10.1016/S2213-2600(19)30337-6

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet 19:233–257

Ceyhan-Birsoy O, Murry JB, Machini K, Lebo MS, Yu TW, Fayer S, Genetti CA, Schwartz TS, Agrawal PB, Parad RB, Holm IA, McGuire AL, Green RC, Rehm HL, Beggs AH, BabySeq Project T (2019) Interpretation of Genomic Sequencing Results in Healthy and Ill Newborns: Results from the BabySeq Project. Am J Hum Genet 104:76–93. https://doi.org/10.1016/j.ajhg.2018.11.016

Chakraborty R (2006) Population Genetics: Historical Aspects. eLS. Wiley, Chichester, pp 1–3

Charlesworth B, Charlesworth D (2017) Population genetics from 1966 to 2016. Heredity (Edinb) 118:2–9. https://doi.org/10.1038/hdy.2016.55

Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247

Crow JF (1987) Population genetics history: a personal view. Annu Rev Genet 21:1–22. https://doi.org/10.1146/annurev.ge.21.120187.000245

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York

Davis J, Davis LC, Correll RN, Makarewich CA, Schwanekamp JA, Moussavi-Harami F, Wang D, York AJ, Wu H, Houser SR, Seidman CE, Seidman JG, Regnier M, Metzger JM, Wu JC, Molkentin JD (2016) A tension-based model distinguishes hypertrophic versus dilated cardiomyopathy. Cell 165:1147–1159. https://doi.org/10.1016/j.cell.2016.04.002

Edwards AW (2003) Human genetic diversity: Lewontin's fallacy. BioEssays 25:798–801. https://doi.org/10.1002/bies.10315

Ehret CF, De Haller G (1963) Origin, development and maturation of organelles and organelle systems of the cell surface in Paramecium. J Ultrastruct Res 23:1–42

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Felsenstein J (1973a) Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet 25:471–492

Felsenstein J (1973b) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Biol 22:240–249

Felsenstein J (1978) The number of evolutionary trees. Syst Biol 27:27–33. https://doi.org/10.2307/2412810

Felsenstein J (2001) Taking variation of evolutionary rates between sites into account in inferring phylogenies. J Mol Evol 53:447–455. https://doi.org/10.1007/s002390010234

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland

Ferrer-Admetlla A, Bosch E, Sikora M, Marques-Bonet T, Ramirez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F (2008) Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol 181:1315–1322. https://doi.org/10.4049/jimmunol.181.2.1315

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil Trans Roy Soc A202:309–368

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284. https://doi.org/10.1126/science.155.3760.279

Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA (2013) Cohort profile: the avon longitudinal study of parents and children: ALSPAC mothers cohort. Int J Epidemiol 42:97–110. https://doi.org/10.1093/ije/dys066

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, Project NES, Akey JM (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220. https://doi.org/10.1038/nature11690

Gazal S, Loh P-R, Finucane HK, Ganna A, Schoech A, Sunyaev S, Price AL (2018) Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat Genet 50:1600–1607

Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadottir HT, Johannsdottir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K (2015) Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 47:435–444. https://doi.org/10.1038/ng.3247

Haldane J (1927) A mathematical theory of natural and artificial selection, Part V: selection and mutation. Math Proc Cambridge Philos Soc 23:838–844. https://doi.org/10.1017/S0305004100015644

Haldane JBS (1937) The effect of variation on fitness. Am Nat 71:337–349

Haldane JBS, Moshinsky P (1939) Inbreeding in mendelian populations with special reference to human cousin marriage. Ann Eugen 9:321–340

Hameed MA, Lingam R, Zammit S, Salvi G, Sullivan S, Lewis AJ (2017) Trajectories of early childhood developmental skills and early adolescent psychotic experiences: findings from the ALSPAC UK birth cohort. Front Psychol 8:2314. https://doi.org/10.3389/fpsyg.2017.02314

HapMap (2005) A haplotype map of the human genome. Nature 437:1299–1320

Hardy GH (1908) Mendelian proportions in a mixed population. Science 28:49–50

Hillis DM, Moritz C, Porter CA, Baker RJ (1991) Evidence for biased gene conversion in concerted evolution of ribosomal DNA. Science 251:308–310. https://doi.org/10.1126/science.1987647

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079. https://doi.org/10.1126/science.1105436

Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the four-taxon case. Syst Biol 42:247–264. https://doi.org/10.1093/sysbio/42.3.247

Hurst LD (2009) Evolutionary genomics and the reach of selection. J Biol 8:12. https://doi.org/10.1186/jbiol113

Imai A, Nakaya A, Fahiminiya S, Tetreault M, Majewski J, Sakata Y, Takashima S, Lathrop M, Ott J (2015) Beyond homozygosity mapping: family-control analysis based on hamming distance for prioritizing variants in exome sequencing. Sci Rep 5:12028. https://doi.org/10.1038/srep12028

Imai A, Kohda M, Nakaya A, Sakata Y, Murayama K, Ohtake A, Lathrop M, Okazaki Y, Ott J (2016) HDR: a statistical two-step approach successfully identifies disease genes in autosomal recessive families. J Hum Genet 61:959–963. https://doi.org/10.1038/jhg.2016.85

Inoue I, Nakajima T, Williams CS, Quackenbush J, Puryear R, Powers M, Cheng T, Ludwig EH, Sharma AM, Hata A, Jeunemaitre X, Lalouel JM (1997) A nucleotide substitution in the promoter of human angiotensinogen is associated with essential hypertension and affects basal transcription in vitro. J Clin Invest 99:1786–1797. https://doi.org/10.1172/JCI119343

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ, Farh KK (2019) Predicting splicing from primary sequence with deep learning. Cell 176(535–548):e24. https://doi.org/10.1016/j.cell.2018.12.015

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73. https://doi.org/10.1038/314067a0

Kimura M (1960) Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. J Genet 57:21–34

Kimura M (1964) Diffusion models in population genetics. J Appl Probab 1:177–232

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Kimura M, Ohta T (1973) The age of a neutral mutant persisting in a finite population. Genetics 75:199–212

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Kingman JF (2000) Origins of the coalescent. 1974–1982. Genetics 156:1461–1463

Kondrashov AS (1995) Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? J Theor Biol 175:583–594. https://doi.org/10.1006/jtbi.1995.0167

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144. https://doi.org/10.1371/journal.pgen.1000144

Kumar S, Tamura K, Nei M (1994) MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput Appl Biosci 10:189–191

Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570

Lewontin RC (1972) The apportionment of human diversity. In: Dobzhansky T, Hecht MK, Steere WC (eds) Evolutionary biology, vol 6. Appleton-Century-Crofts, New York, pp 381–398

Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54:595–609

Lipson M, Ribot I, Mallick S, Rohland N, Olalde I, Adamski N, Broomandkhoshbacht N, Lawson AM, Lopez S, Oppenheimer J, Stewardson K, Asombang RN, Bocherens H, Bradman N, Culleton BJ, Cornelissen E, Crevecoeur I, de Maret P, Fomine FLM, Lavachery P, Mindzie CM, Orban R, Sawchuk E, Semal P, Thomas MG, Van Neer W, Veeramah KR, Kennett DJ, Patterson N, Hellenthal G, Lalueza-Fox C, MacEachern S, Prendergast ME, Reich D (2020) Ancient West African foragers in the context of African population history. Nature 577:665–670. https://doi.org/10.1038/s41586-020-1929-1

Liu W, Morito D, Takashima S, Mineharu Y, Kobayashi H, Hitomi T, Hashikata H, Matsuura N, Yamazaki S, Toyoda A, Kikuta K, Takagi Y, Harada KH, Fujiyama A, Herzig R, Krischek B, Zou L, Kim JE, Kitakaze M, Miyamoto S, Nagata K, Hashimoto N, Koizumi A (2011) Identification of RNF213 as a susceptibility gene for moyamoya disease and its possible role in vascular development. PLoS ONE 6:e22542. https://doi.org/10.1371/journal.pone.0022542

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Mendel GJ (1866) Versuche über Pflanzen-Hybriden. Verh Naturforsch Ver Brünn 4:3–47

Miyata T, Hayashida H (1981) Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. Proc Natl Acad Sci USA 78:5739–5743. https://doi.org/10.1073/pnas.78.9.5739

Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, Sato Y, Mimori T, Tsuda K, Saito R, Pan X, Nishikawa S, Ito S, Kuroki Y, Tanabe O, Fuse N, Kuriyama S, Kiyomoto H, Hozawa A, Minegishi N, Douglas Engel J, Kinoshita K, Kure S, Yaegashi N, To MJRPP, Yamamoto M (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 6:8018. https://doi.org/10.1038/ncomms9018

Nei M (2005) Selectionism and neutralism in molecular evolution. Mol Biol Evol 22:2318–2342. https://doi.org/10.1093/molbev/msi242

Nei M, Roychoudhury AK (1972) Gene differences between Caucasian, Negro, and Japanese populations. Science 177:434–436

Nei M, Roychoudhury AK (1974) Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. Am J Hum Genet 26:421–443

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zollner S, Whittaker JC, Chissoe SL, Novembre J, Mooser V (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337:100–104. https://doi.org/10.1126/science.1217876

Neyman J (1971) Molecular studies of evolution: a source of novel statistical problems. In: Gupta SS, Yackel J (eds) Statistical decision theory and related topics. Academic Press, New York, pp 1–27

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG (2009) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19:838–849. https://doi.org/10.1101/gr.088336.108

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (2017) Tracing the peopling of the world through genomics. Nature 541:302–310. https://doi.org/10.1038/nature21347

Ohno S (1970) Evolution by gene duplication. Springer, New York

Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23:366–370

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96–98

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286

Ohta T (2002) Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci USA 99:16134–16137. https://doi.org/10.1073/pnas.252626899

Pattison JE (2016) An attempt to integrate previous localized estimates of human inbreeding for the whole of Britain. Hum Biol 88:264–274

Provine WB (1971) The origins of theoretical population genetics. University of Chicago Press, Chicago

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sella G, Barton NH (2019) Thinking about the evolution of complex traits in the era of genome-wide association studies. Annu Rev Genomics Hum Genet 20:461–493. https://doi.org/10.1146/annurev-genom-083115-022316

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–1438

Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, Xu J, Batzoglou S, Li X, Farh KK (2018) Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 50:1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J (2020) On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 7:1012–1023. https://doi.org/10.1093/nsr/nwaa036

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, Broad GO, Seattle GO, Project NES (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69. https://doi.org/10.1126/science.1219240

Weinberg W (1908) Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64:369–382

Williams GC, Nesse RM (1991) The dawn of Darwinian medicine. Q Rev Biol 66:1–22. https://doi.org/10.1086/417048

Wright S (1938) Size of population and breeding structure in relation to evolution. Science 87:430–431

Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017) Concepts, estimation and interpretation of SNP-based heritability. Nat Genet 49:1304–1310. https://doi.org/10.1038/ng.3941

Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A (2019) A primer on deep learning in genomics. Nat Genet 51:12–18. https://doi.org/10.1038/s41588-018-0295-5

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