model of particle physics, the accepted model for describing the subatomic Universe. The discovery therefore marked a pivotal moment in the search for physics beyond this model.

Neutrino oscillations are studied using an approach known as a long-baseline experiment[4], in which neutrinos produced by a particle-accelerator complex are first detected a short distance away, and then at a detector far away (Fig. 1). These experiments are carried out at large facilities such as the forthcoming Deep Underground Neutrino Experiment (DUNE)[4], which will consist of two detectors: one near the source of the neutrino beam at Fermilab in Illinois, and the other 1,300 kilometres away in South Dakota. Long-baseline experiments are also under way at the Hyper-Kamiokande detector in Japan[5].

Current efforts at these facilities include Fermilab's NOvA experiment (https://novaexperiment.fnal.gov) and the T2K experiment (http://t2k-experiment.org) in Japan. Both experiments measure the properties of neutrino oscillations by monitoring how the distribution of neutrino energies changes as the particles propagate (Fig. 1). But these energies cannot be measured directly — instead, they must be reconstructed using information about the charged particles that are produced when the neutrinos interact with a target. In the case of NOvA and T2K, the targets are made of atomic nuclei, because this maximizes the number of interactions.

Khachatryan *et al.* tested the accuracy of the models used in such experiments to reconstruct neutrino energies. To do so, they exploited the observation that the probability of interactions between electrons and nuclei takes a similar mathematical form to the probability of interactions between neutrinos and nuclei, although the strength of the interactions is different. This promising strategy implies that models describing the probabilities of neutrino–nucleus interactions should also apply to those of electron–nucleus interactions, the properties of which are well known. So, using an electron beam with a known energy, Khachatryan *et al.* tested how accurately these models could reconstruct this energy using information about the electron–nucleus interactions.

Most of the events that occur when an electron hits a nuclear target involve the electron interacting with the target's protons and neutrons, collectively known as nucleons, after which a single nucleon is emitted. But other events are possible, including processes in which the electron couples to interacting nucleons, leading to two nucleon emissions. In this case, one of the nucleons hit by the electron is subsequently reabsorbed by the nucleus. The incoming electron can also excite the nucleon it hits, producing pions, which are the particles responsible for holding the nucleus together.

Khachatryan *et al.* used reconstruction models to simulate all these possibilities, and focused only on those with one electron, one proton and zero pions in the final state. Their crucial finding was that most of the energies obtained in this way failed to reconstruct the correct incident energy.

A key novelty of Khachatryan and colleagues' work is the use of semi-exclusive data, which means that one or more of the products of interactions were measured directly rather than inferred — in this case, the scattered electrons and charged particles knocked out from the nucleus. The value of the electron energy used in the authors' experiments is commensurate with that used for the neutrino beams in current and planned oscillation experiments. The results therefore indicate the need for substantial improvement in the accuracy of the way in which neutrino interactions are modelled. They also highlight the power of electron-scattering data in ensuring that these models achieve the level of accuracy needed to fully exploit the future discovery potential of these high-precision experiments.

Noemi Rocco is in the Theory Department, Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA.
e-mail: nrocco@fnal.gov

1. Khachatryan, M. *et al. Nature* **599**, 566–570 (2021).
2. Fukuda, Y. *et al. Phys. Rev. Lett.* **81**, 1562–1567 (1998).
3. Ahmad, Q. R. *et al. Phys. Rev. Lett.* **89**, 011301 (2002).
4. DUNE Collaboration. Preprint at https://arxiv.org/abs/1807.10334 (2018).
5. Hyper-Kamiokande Proto-Collaboration. Preprint at https://arxiv.org/abs/1805.04163 (2018).

The author declares no competing interests.

Genetics

From genes to health

Yukinori Okada & Qingbo S. Wang

The protein-coding portions of more than 450,000 individuals' genomes have been sequenced, and analysed together with the individuals' health data, revealing rare and common gene variants linked to various health-related traits. **See p.628**

To link genetic variations to human health, we need to collect genetic data and health-related information from as many individuals, and in as much detail, as possible. The principles of such association testing are long-standing[1], and the field of human genomics has expanded in scale and quality in the past few decades. On page 628, Backman *et al.*[2] report a new milestone. They have analysed the protein-coding parts of the genome — called the exome — of more than 450,000 individuals whose health and genetic data have been collected in the UK Biobank, identifying 12.3 million variants that lead to changes in the encoded protein, which are known as coding variants. The authors then tested these variants for associations with 3,994 health-related traits (health phenotypes) and found 8,865 such associations (Fig. 1). This work is on an unprecedented scale in terms of the number of participants and the quantity of genetic and clinical data gathered.

Initial analysis of data from the UK Biobank[3] included genetic information that was limited to common variants already known to occur in a relatively large proportion of individuals (for example, more than 1% of the tested population). However, such variants are probably only the tip of the iceberg of the total variation in the human exome. The use of whole-exome sequencing (WES) technology was the key to obtaining detailed information about exome variation.

In 2016, the Exome Aggregation Consortium[4] presented aggregated WES data for a group of more than 60,000 people, which was later expanded to include exome and genome data from more than 140,000 individuals[5]. These studies provided a systematic quantification of the biological impact of human genetic variants. However, the impact was not linked to specific phenotypes, because the data set lacked detailed clinical information, highlighting the value of the health records of the UK Biobank.

Previous analyses have also combined exome (or even whole-genome) data with detailed phenotypic data. For example, the DiscovEHR collaboration sequenced about 50,000 exomes[6], the Trans-Omics for Precision Medicine Program sequenced almost 54,000 exomes[7], and previous studies using the UK Biobank data analysed nearly 50,000 exomes[8] and more than 280,000 exomes[9]. However, the results of these previous analyses consistently indicated that it would be valuable to expand the sample sizes further. Indeed, Backman and co-workers demonstrate that, by increasing the sample size by less than 10-fold, about 20 times

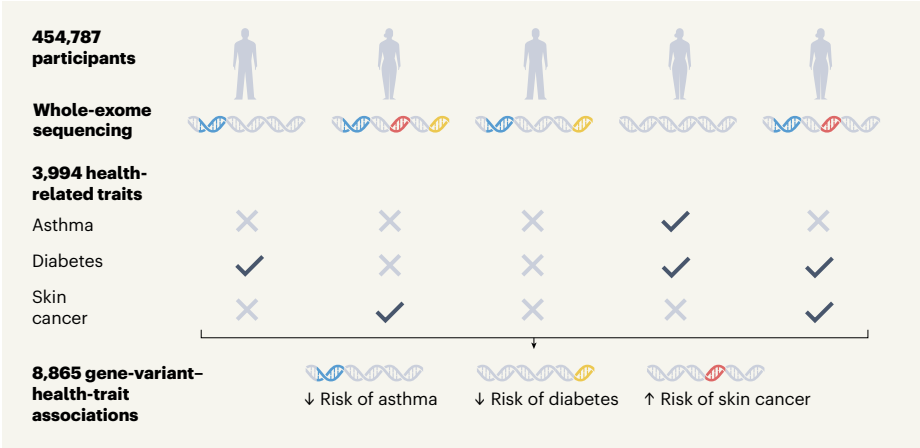

Figure 1 | A large-scale initiative for analysing associations between genes and human health. Backman *et al.*[2] sequenced the protein-coding portions of the genomes of more than 450,000 participants in the UK Biobank. In doing so, they identified 12.3 million genetic variants that lead to changes in the sequence of amino-acid residues in the encoded protein — called coding variants. Almost all were relatively under-studied, rare variants (occurring in less than 1% of the population), whereas only 0.4% were well-studied, common variants (occurring in more than 1% of the population). The authors screened the variants for associations with thousands of health-related traits, such as risk of having diabetes, childhood asthma or skin cancer, identifying 8,865 such associations. Predicting which variants might be associated with increased or decreased risk of various disorders should aid future drug-discovery efforts by indicating which encoded proteins could make good drug targets.

more gene–phenotype associations[8] (from 26 to 564) could be identified.

The 12.3 million coding variants identified by Backman and colleagues' WES analysis represent a 30% increase compared with the number of coding variants identified in previous WES analyses combined[5,7]. Almost all of the 12.3 million were rare variants — occurring in less than 1% of people across all ancestries — and roughly half of the variants were observed in only one person in the data set (called singleton variants), highlighting the value of WES at such a large scale.

Describing the biological effects of the identified variants on encoded proteins and their functions helps researchers to prioritize variants and genes that are probably clinically relevant. For example, putative loss-of-function (pLOF) variants are those that are expected to disrupt the biological function of the normally encoded protein[5]. Backman *et al.* discovered more than 900,000 pLOF variants, corresponding to more than 200 pLOF variants per individual, and about 50 pLOF variants per gene. Missense variants lead to changes in the sequence of amino-acid residues in the encoded protein that might affect the protein's biochemical and biological properties. The authors identified almost 7.9 million missense variants in total, corresponding to more than 600 variants per gene, on average. Furthermore, 23% of these missense variants were predicted with high confidence to have deleterious effects for the individual carrying them.

Notably, the authors constructed a comprehensive genotype–phenotype catalogue of associations between the identified deleterious variants and 3,994 health phenotypes, including 3,706 'binary' health phenotypes, such as the presence (or absence) of specific types of cancer. The health data set covered a wide range of phenotypes, including diseases, measurements and proportions of the body, and levels of molecular markers of health and disease. The authors discovered 8,865 statistically significant associations (involving 564 genes and 492 traits).

Using human genotype–phenotype associations to inform and accelerate drug-discovery efforts is a topic of interest[10]. The authors found that genes encoding proteins that are targeted by drugs approved in the United States showed 3.6-fold more associations with health-related phenotypes than did the remaining genes, enforcing the potential of genome-based drug discovery. The authors specifically highlighted pLOF variants that were associated with a reduced risk of disease; for example, pLOF variants of the *SLC27A3* gene were associated with a reduced risk of childhood asthma. Such associations could inform possible therapeutic strategies — for example, by motivating the design of compounds that inhibit the protein products of such genes.

"Almost all of the 12.3 million variants were rare — occurring in less than 1% of people."

Interpreting the effects of well-studied common variants together with those of relatively under-studied rare variants in a single framework is often challenging. Backman and co-workers demonstrate that the associations of rare variants with health phenotypes mostly remained statistically significant even when accounting for the effects of common variants. This observation paves the way for the interpretation of the integrated effects of gene variants that occur as a spectrum of frequencies in the population, spanning from common to extremely rare ones, including singleton variants. Furthermore, by examining pairs of rare (coding) and common (potentially non-coding) variants associated with the same health trait that are nearby in the genome, Backman and colleagues suggested that the gene closest to a (non-coding) genomic variant associated with a particular phenotype is extremely likely to be a gene that affects that phenotype.

What should be expected for the next milestone? Several future strategies are warranted. First is the continued accumulation of samples. As the authors show, even in their exceptionally large WES study, their data still have insufficient statistical power to find most of the rare protective variants. The second strategy is to delve much more deeply into human phenotypes, for example by assessing clinical records over extended periods of time, and looking at behavioural and socio-economic data, as well as levels of gene expression, proteins and metabolites. Third is to shift from WES to whole-genome sequencing, as the authors plan to do in the near future. Rare non-coding variations associated with certain phenotypes might be waiting for us, as yet another chunk of the iceberg. Finally, the involvement of diverse populations is crucial. Although people with non-European ancestry are included in the UK Biobank and in Backman and colleagues' analysis, the proportion of people with non-European ancestry in the sample is still relatively small. Greater ancestral diversity, most effectively achieved by multi-population and multi-biobank initiatives[11,12], is necessary not only to maximize our efforts to discover variations, but also to achieve equity in human genetics.

Importantly, Backman and colleagues are making the WES resources publicly available. Although we live in a time when we have perhaps more data than we can analyse, the whole scientific community is ready to tackle massive data sets with combinations of statistical, computational and biological knowledge.

Yukinori Okada and **Qingbo S. Wang** are in the Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan. **Y.O.** is also at the Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.
e-mails: yokada@sg.med.osaka-u.ac.jp; qingbow@sg.med.osaka-u.ac.jp

1. Risch, N. & Merikangas, K. *Science* **273**, 1516–1517 (1996).
2. Backman, J. D. *et al. Nature* **599**, 628–634 (2021).
3. Bycroft, C. *et al. Nature* **562**, 203–209 (2018).
4. Lek, M. *et al. Nature* **536**, 285–291 (2016).
5. Karczewski, K. J. *et al. Nature* **581**, 434–443 (2020).
6. Dewey, F. E. *et al. Science* **354**, aaf6814 (2016).
7. Taliun, D. *et al. Nature* **590**, 290–299 (2021).
8. Van Hout, C. V. *et al. Nature* **586**, 749–756 (2020).
9. Wang, Q. *et al. Nature* **597**, 527–532 (2021).
10. Reay, W. R. & Cairns, M. J. *Nature Rev. Genet.* **22**, 658–671 (2021).
11. Choudhury, A. *et al. Nature* **586**, 741–748 (2020).
12. Sakaue, S. *et al. Nature Genet.* **53**, 1415–1424 (2021).

The authors declare no competing interests.
This article was published online on 25 October 2021.

Materials science

How to make macroscale non-crystalline diamonds

Alfonso San-Miguel

A diamond shatters easily, despite it being the hardest natural material. Atomically disordered forms of diamond made from buckyballs might not only overcome this problem, but also allow other properties to be optimized. **See p.599 & p.605**

The brilliant facets of diamonds have entranced people throughout history and are a result of the ordered atomic structure of these gemstones. But this order comes at a cost: it makes diamonds fragile. In contrast to quartz and many other crystalline materials that produce atomically disordered forms, a disordered — and potentially less fragile — form of diamond has not been available. In this issue, Shang *et al.*[1] (page 599) and Tang *et al.*[2] (page 605) report how to produce atomically disordered diamond-like materials with millimetre-scale dimensions, constituting a breakthrough for materials science.

Atomic-scale order can be a problem for materials scientists, as illustrated by the mineral quartz. Quartz is a bad choice of material for a car windscreen, because it is difficult to shape and easily broken by objects hitting it. Glass is a much better choice, because it can be conveniently engineered, is easier to shape and does not disintegrate so readily[3]. Yet quartz and glass are made of the same atoms, have the same chemical formula, and their architectures are built from the same tetrahedral SiO_4 subunits.

The crucial difference between glass and quartz is the atomic order: quartz is crystalline, which means that its atomic bonds follow a regular pattern throughout the material; whereas glass lacks any such order and is said to be amorphous. This difference contributes to the divergence of the mechanical properties of the two materials — the lack of long-range atomic-scale order in glass means that there are no planes of atoms that provide directions in which the material breaks easily. These 'cleavage planes' are the reason why crystalline gemstones can be cut by applying a blow in just the right direction. They also explain why diamond is fragile, despite it being the hardest naturally occurring material.

Diamond consists of tetrahedrally arranged carbon atoms, linked by covalent bonds. The physical properties of diamond strongly differ from those of graphite, which is another crystalline form of carbon. These differences derive from the organization of electrons — called electron hybridization — in the carbon bonds. In graphite, electrons organize in a way that allows each atom to form three bonds in a plane, thereby producing a 2D hexagonal lattice; this is called sp^2 hybridization. Graphite is made by stacking planes of sp^2-hybridized carbon atoms, and it has many disordered forms[4].

By contrast, the electrons in diamond exhibit sp^3 hybridization, which enables each carbon atom to form four bonds pointing in different directions from the atom's centre to the corners of a tetrahedron. Diamond therefore has a 3D architecture generated by the sp^3-hybridized carbon bonds, and no form of it was known that had long-range atomic disorder and could be produced as 3D macroscale samples. A type of sp^3-hybridized carbon known as diamond-like carbon comes close to having such disorder, but only films of this material have been produced, which incorporate varying amounts of hydrogen atoms[5]. Samples of materials proposed to be amorphous diamond have also been reported, but only in quantities similar to the size of dust particles[6,7].

The methods now reported by Shang *et al.* and Tang *et al.* for making millimetre-scale forms of disordered diamond are broadly similar to the original method used to make crystalline diamond: a carbon-based material, such as graphite[8] or carbon nanotubes[9], is subjected to high pressure and temperature in a large press. However, in contrast to previous work, both research groups used fullerite as a starting material, which consists of a crystalline arrangement of fullerenes — soccer-ball-shaped C_{60} molecules that are also known as buckyballs.

Moreover, the syntheses were carried out

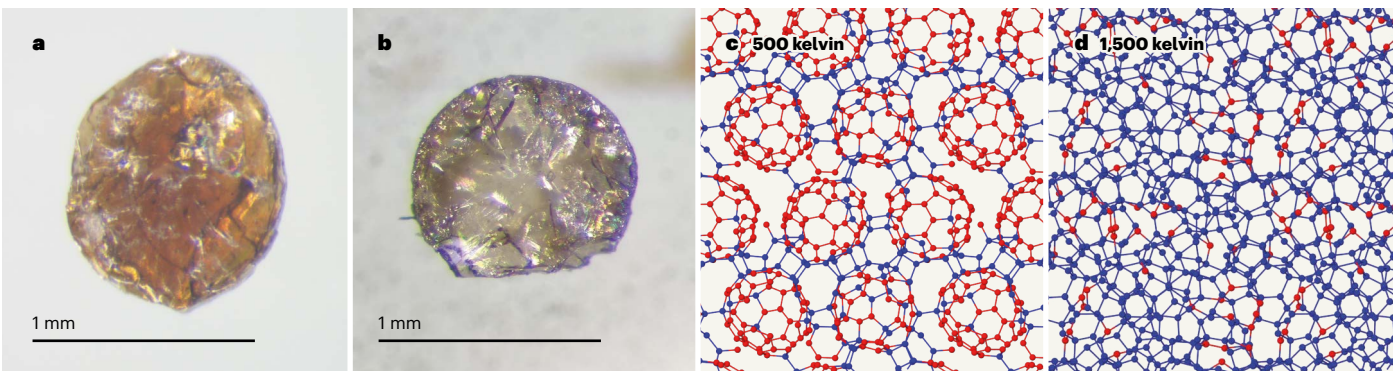

Figure 1 | Atomically disordered diamond forms from buckyballs. a, Shang *et al.*[1] subjected buckyballs to high temperatures and pressures, and produced amorphous diamond — a form of carbon that has the same type of chemical bonds (sp^3 bonds) as crystalline diamond, but that lacks atomic order. **b**, Tang *et al.*[2] used the same approach to make paracrystalline diamond, which can be thought of as a composite that consists of an amorphous carbon matrix containing nanometre-scale, severely distorted diamond crystals. **c**, **d**, Tang and colleagues' computer simulations show how the pressurized buckyballs (which contain sp^2 bonds, red) polymerize and then collapse to form a disordered structure that comprises mostly sp^3 bonds (blue) as the temperature increases.