the detection of illegal fishing operations and measurement of the informal economy[2] (economic activities that are neither taxed nor monitored by governments). Indeed, the findings reveal a high concentration of illicit fishing to the west of the Korean Peninsula and on the North African coast. Vessel tracking could also transform environmental conservation efforts by revealing encroachment on protected areas — Paolo *et al.* report that two such areas, the Galapagos Marine Reserve and the Great Barrier Reef Marine Park, were visited by an average of more than 5 and more than 20 fishing vessels per week, respectively.

Second, a lot of global satellite imagery is freely available to the public, and machine-learning tools for processing these data are continuously being developed, often as open-source software (Paolo and colleagues' data set and software are also freely available). Such efforts democratize access to data and tools and allow researchers, analysts and policymakers in low-income countries to leverage tracking technologies at low cost, for example to monitor exclusive economic zones — areas of the ocean for which sovereign states have exclusive exploration and usage rights.

The study also highlights some general limitations of Earth observation by satellites. The spatial resolution of publicly available satellite imagery (such as the Sentinel-1 data used by Paolo *et al.*, which have a ground resolution of about 20 metres per pixel) prevents the detection of objects such as small fishing vessels. Crucially, this leaves many small-scale fishers in coastal communities off the map. And although satellites cover every corner of the planet, some areas are more difficult to map than others — for example, because of persistent cloud coverage.

Emerging data sources and increasingly powerful machine-learning systems will overcome some of these limitations, and thereby enable the quantification of previously unmeasured aspects of human development and change. High-resolution radar instruments can see through clouds, for instance; and advances in 'unsupervised' deep learning will make it possible to gain insight from satellite images with less laborious hand-labelling by humans than is currently needed[3].

The complexity and sheer volume of Earth-observation data remain challenges that are driving research among machine-learning and Earth scientists[4]. But this should not be the only research focus. Earth-observation technologies should also be designed to help capture human activity in a way that is equitable and community-centred. For example, unlike Western large-scale farming operations, many sustenance farmers in low-income countries operate on small plots of land. Earth-observation models can fail to delineate those fields if the imagery available is not of sufficiently high resolution or because models trained on Western data have not been exposed to such small-scale patterns.

Moreover, cross-disciplinary expertise and the involvement of multiple stakeholders are often essential to interpret satellite-based observations in specific regions and put them into context. As an illustration of this, Indigenous peoples often have a fundamentally different, and frequently more complete, understanding of their local ecosystems than can be captured on a satellite image. Failure to incorporate those insights might lead to the development of methods that fail to meet real-world needs, ignore important equity and transparency considerations, and ultimately limit the utility of satellite-based observations[5]. The development of approaches that prioritize computational accessibility — publicly available data, open-source software, and efficient algorithms that don't require high-performance computers — will also be crucial[6].

Paolo and colleagues' study adds to a growing body of work highlighting how deep learning can facilitate ocean monitoring, with applications ranging from the detection of marine debris[7] to tracking algal blooms[8]. The task now is for interdisciplinary collaborations to build on these prototype technologies to set up large-scale observation systems that focus on stakeholder engagement and local community efforts. This will ensure that advances in deep learning for Earth observation achieve their potential to address pressing local and global challenges.

Konstantin Klemmer is at Microsoft Research New England in Cambridge, Massachusetts 02142, USA. **Esther Rolf** is at the Harvard Data Science Initiative and the Center for Research on Computation and Society, Harvard University, Boston, Massachusetts 02134, USA.
e-mails: kklemmer@microsoft.com; erolf@g.harvard.edu

1. Paolo, F. *et al. Nature* **625**, 85–91 (2024).
2. Angrist, N., Goldberg, P. K. & Jolliffe, D. *J. Econ. Persp.* **35**, 215–242 (2021).
3. Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L. & Zhou, X. X. *IEEE Geosci. Remote Sensing Mag.* **10**, 213–247 (2022).
4. Reichstein, M. *et al. Nature* **566**, 195–204 (2019).
5. De-Arteaga, M., Herlands, W., Neill, D. B. & Dubrawski, A. *ACM Trans. Mgmt Inf. Syst.* **9**, 9 (2018).
6. Rolf, E. *et al. Nature Commun.* **12**, 4392 (2021).
7. Rußwurm, M., Venkatesa, S. J. & Tuia, D. *iScience* **26**, 108402 (2023).
8. Wang, M. & Hu, C. *Remote Sens. Environ.* **264**, 112631 (2021).

The authors declare no competing interests.

Human genetics

Linking the non-coding genome to human health

Ryan S. Dhindsa & Slavé Petrovski

An expanded version of a human-genome database called gnomAD, containing 76,156 whole-genome sequences, has enabled investigation of how variants in non-protein-coding regions of the genome affect health. **See p.92**

Scientists have long suspected that many disease-causing genetic mutations reside in the 98% of the genome that does not encode proteins, especially in regions that have roles in regulating gene expression. However, it has been challenging to differentiate systematically between harmful and neutral mutations, partly because researchers lack a clear picture of which stretches of the non-coding genome are essential for human health. On page 92, Chen *et al.*[1] address this challenge, introducing a tool that can analyse large collections of human genomes to identify non-coding regions that have the greatest potential to cause disease when mutated.

This work represents the most recent iteration of the Genome Aggregation Database (gnomAD), a publicly available catalogue of human genetic variation. The first version[2], released in 2020, included sequence data from the protein-coding DNA of 125,748 people and the whole genomes of 15,708 people. Since then, the consortium has greatly expanded the database; the resource now includes whole-genome sequences from 76,156 individuals of diverse ancestries, providing a much deeper picture of human genetic variation.

GnomAD has transformed human genetics, especially in terms of diagnosing rare diseases. The genome of any individual differs from those of other people at millions of sites. Most of these genetic variants are clinically insignificant, particularly those that are common in the general population. When clinical

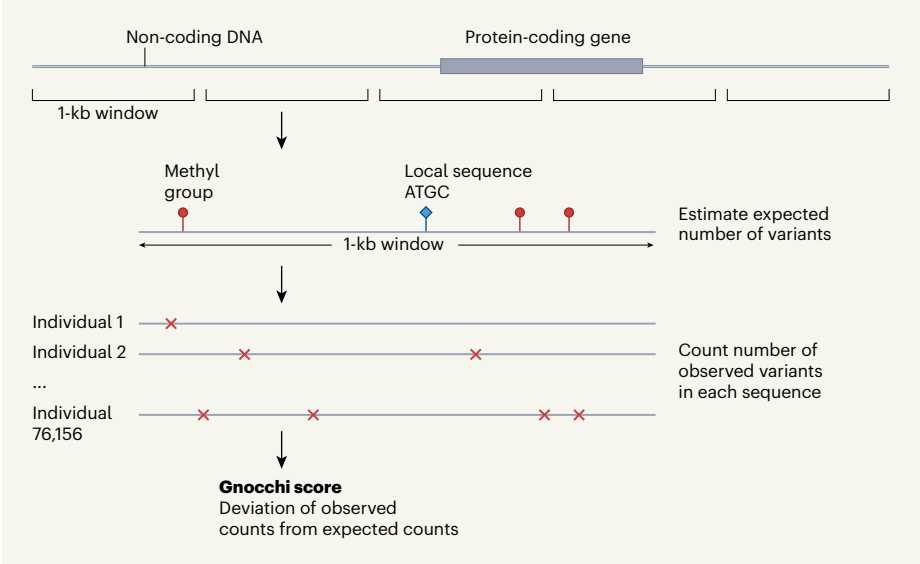

Figure 1 | A metric to measure variation in non-protein-coding DNA. The Genome Aggregation Database (gnomAD) contains 76,156 human whole-genome sequences — a sample size large enough to examine genetic variation in the non-protein-coding portions of the genome. Chen *et al.*[1] developed a metric called Gnocchi to measure this variation. They divided the genome into tiled 1,000-base-pair-long (1 kb) windows. They calculated the expected number of variants in each window using an algorithm that took into account several genomic features, such as local DNA sequences (denoted by bases dubbed A, T, G and C) and the way in which the DNA is modified by methyl groups. They then compared this expected number with the actual number of variants present in each sequence in the gnomAD cohort to calculate the Gnocchi score. Regions that had many fewer mutations than expected received a larger Gnocchi score. This indicated that they are more intolerant of mutations, and so more likely to be relevant to disease.

geneticists analyse the DNA of a person who is suspected of having a rare genetic disease, they must sift through all of the variants, eliminating common ones to find those most likely to cause disease. GnomAD made it possible for a clinician to easily look up a person's variants and rule out those that are common either worldwide or in a certain genetic ancestry. The larger and more ancestrally diverse number of whole genomes in this newest version of gnomAD will allow scientists to more easily identify which variants are rare — and thus more likely to be disease-related — in the non-coding genome.

Large data sets such as gnomAD have also enabled researchers to develop 'intolerance metrics' to examine how many protein-altering variants are observed in a gene in a large sample of the human population, compared with how many are expected to arise at random during evolution[3–6]. These measures help to determine which genes are intolerant of genetic variation. Genes with less variation than expected are more intolerant — and more likely to be disease-associated — than are genes that harbour as much or more variation than is expected by chance.

There has been a growing effort to extend intolerance metrics to include the non-coding sections of the human genome[7–9]. Chen *et al.* add to this list of tools with an approach that they call Gnocchi (Fig. 1). Whereas protein-coding genes have well-defined boundaries, non-coding regions are not as conveniently divided into functional units. To circumvent this issue, the authors divided the genome into 1,000-base-pair windows, and calculated the intolerance of each window.

Although conceptually similar to other non-coding intolerance scores, the major advance of Gnocchi is in how it calculates the theoretical expected number of variants in each window. Mutation rate varies considerably across the genome, being affected by factors such as local-sequence context and the way in which the DNA is modified by the addition of methyl groups. Chen and colleagues introduce a statistical model that includes these different features to better estimate the mutation rate in each window.

"The gnomAD resource now includes whole-genome sequences from 76,156 individuals."

The authors validated the ability of Gnocchi to identify relevant regions of the genome in several ways. First, they showed that protein-coding regions were, on average, more intolerant of variation than were non-coding regions, which was consistent with expectations. Second, they found that the most intolerant regions of the non-coding genome are enriched for gene-regulatory elements such as promoters and enhancers. Third, they demonstrated that Gnocchi can distinguish between putatively benign variants and curated lists of disease-causing mutations in the non-coding genome. Fourth, they showed that individuals diagnosed with developmental disorders are more likely than healthy people to have copy-number variants (large variants that result in duplications or deletions of DNA) in intolerant regions of the genome.

Chen and colleagues also demonstrated that Gnocchi can be used to bolster conventional, gene-level intolerance scores. They compared the intolerance of a gene's non-coding enhancer to variation (measured using Gnocchi) with the intolerance of coding regions of the gene to variants that disrupt their normal function (measured using a separate metric, called LOEUF)[2]. The two metrics generally concurred, but there were instances in which a gene that seemed tolerant of loss of function had an intolerant enhancer. These occurrences arose mostly for small genes, for which the accuracy of gene-level intolerance scores is currently limited by sample size. Combining a gene's LOEUF score with the Gnocchi score of its enhancer improved overall intolerance estimates for small protein-coding genes.

We have previously shown similar performance gains when combining gene-level intolerance scores with scores that measure the intolerance of a gene's untranslated regions[10] (regions that are transcribed into messenger RNA but not translated into proteins). Such approaches could be further refined in future work by combining intolerance metrics for each gene and all of its nearby and distant regulatory sequences.

It is noteworthy that Gnocchi seems to outperform existing metrics in its ability to identify non-coding, disease-associated variants — including one metric that was developed last year using nearly twice as many genome sequences, which were available through a publicly available repository called the UK Biobank[11]. The better performance of Gnocchi might be explained by the differences in how its scores are formulated and in its modelling of how mutations arise. However, the fact that the collection of genome sequences in gnomAD includes many more individuals of non-European ancestries than does the UK Biobank could also explain some of the performance differences. It is commendable that nearly half the gnomAD genome samples are from individuals of non-European ancestries, but researchers must continue to strive for larger, more-diverse human reference sets, both to increase the accuracy of intolerance metrics and to improve health equity.

The gnomAD consortium set a gold standard for data aggregation and sharing with its first iteration, and continues to be exemplary in this regard. This resource will continue to grow

under the strong leadership of the gnomAD consortium, which has made it clear that the priority is to continually expand the database to be more representative of the global population. In doing so, it will equip scientists with ever-more tools with which to reveal the hidden secrets of our genome.

Ryan S. Dhindsa is in the Department of Pathology and Immunology, Baylor College of Medicine, Houston, Texas 77030, USA, and the Centre for Genomics Research, Discovery Sciences, AstraZeneca, Waltham, Massachusetts. **Slavé Petrovski** is in the Centre for Genomics Research, Discovery Sciences, R&D, AstraZeneca, Cambridge CB2 0AA, UK.
e-mails: ryan.dhindsa@bcm.edu; slav.petrovski@astrazeneca.com

1. Chen, S *et al. Nature* **625**, 92–100 (2024).
2. Karczewski, K. J. *et al. Nature* **581**, 434–443 (2020).
3. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. *PLoS Genet.* **9**, e1003709 (2013).
4. Traynelis, J. *et al. Genome Res.* **27**, 1715–1729 (2017).
5. Samocha, K. E. *et al. Nature Genet.* **46**, 944–950 (2014).
6. Dhindsa, R. S., Copeland, B. R., Mustoe, A. M. & Goldstein, D. B. *Am. J. Hum. Genet.* **107**, 83–95 (2020).
7. Gussow, A. B. *et al. PLoS ONE* **12**, e0181604 (2017).
8. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. *Nature Commun.* **12**, 1504 (2021).
9. di Iulio, J. *et al. Nature Genet.* **50**, 333–337 (2018).
10. Petrovski, S. *et al. PLoS Genet.* **11**, e1005492 (2015).
11. Halldorsson, B. V. *et al. Nature* **607**, 732–740 (2022).

The authors declare competing interests. See go.nature.com/3bgd7vn for details.
This article was published online on 6 December 2023.

Condensed-matter physics

Tailoring graphene for electronics beyond silicon

Francesca Iacopi & Andrea C. Ferrari

The integration of non-silicon semiconductors into systems on chips is needed for advanced power and sensing technologies. A semiconducting graphene 'buffer' layer grown on silicon carbide is a step on this path. **See p.60**

Electronics relies on the use of switching devices made from semiconducting materials, in which an electric current can be controlled, ideally down to the movement of single charges. Semiconductors achieve this because the allowed energies of their electrons leave a gap between a low-energy and a high-energy band, and electrons can be excited to cross this 'bandgap'. Materials with different-sized bandgaps can have complementary functions, such as performing logic operations, supplying power or acting as a sensor. But integrating these materials into a single device is challenging. On page 60, Zhao *et al.*[1] report a way of growing a graphene-like layer with a narrow bandgap on a material with a wide bandgap.

Graphene is a material that is gapless because its electronic bands touch at one point, called the Dirac point, and then diverge. This peculiar behaviour makes it ideal for applications in photonics and optoelectronics[2,3]. Although its gapless nature means that it is not the material of choice for electronic devices, it can be used in components that operate in the terahertz portion of the electromagnetic spectrum (1 THz is 10^{12} Hz), where it fills a role that very few other materials can[4].

There has nevertheless been a constant effort over the past 20 years to 'open a bandgap' in graphene, to convert this versatile material into a semiconductor. One way of doing so involves cutting or shaping graphene into nanoribbons[5,6], which can now be achieved with atomic precision[7]. In this case, the bandgap opens because the electrons are further confined to a single dimension. However, nanoribbons are subject to sample-to-sample variations, and it is currently difficult to produce them at the scale required for consumer electronics. Another way to create a bandgap involves leveraging how graphene interacts with the substrate on which it is grown[8]. This is the route that Zhao *et al.* took.

Graphene is a single layer of carbon atoms arranged in a honeycomb lattice. It can be grown by heating the semiconducting material silicon carbide (SiC) until the silicon atoms on its surface sublimate, leaving a carbon-rich layer that can recrystallize[9]. The resulting layer has a hexagonal structure, similar to that of SiC, with some carbon atoms covalently bonded to the substrate. Subsequent layers form as normal graphene, but the partial bonding makes this first 'buffer' layer a semiconductor[10]. A bandgap opens not through dimensional confinement, as for nanoribbons, but as a result of the bonding constraints that SiC imposes[11].

Graphite (of which graphene is a single layer) has been produced from SiC since at least 1896 (ref. 12), and the growth mechanism has been investigated since the 1960s (ref. 13). Over the past 20 years, this approach has been refined and used systematically to obtain graphene at the scale of a wafer (that is, the typical size used for mass production of electronic devices)[9].

It was known as early as 2008 that the graphene buffer layer that forms on SiC could be a semiconductor[14], but achieving wafer-scale samples has been a challenge. Zhao *et al.* succeeded in creating a controlled environment by sandwiching two SiC chips so that the silicon surface of the top chip was opposite the carbon surface of the bottom one. When the system was heated at ambient pressure, carbon atoms were transported from the carbon surface to the silicon surface to form the buffer layer. This differs from other routes, in which atoms are depleted from the silicon surface and lost to the system's surroundings[14].

The authors' technique allowed them to

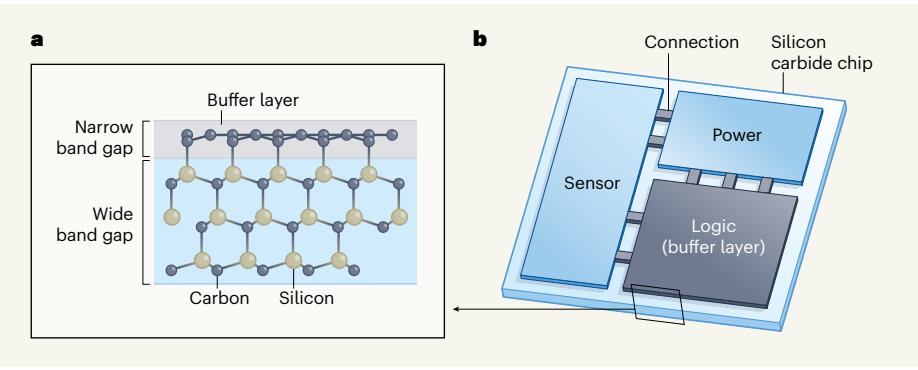

Figure 1 | Combining 'beyond-silicon' materials to make an integrated system on a chip. Graphene (a single layer of carbon atoms arranged in a honeycomb pattern) can be grown on silicon carbide (SiC), which has a wide 'bandgap' between the allowed energies of its electrons. **a**, Graphene is 'gapless', but Zhao *et al.*[1] devised a way of tailoring a graphene 'buffer' layer on SiC, such that the layer has a bandgap (albeit a narrow one) and is therefore semiconducting. **b**, This approach could be used to integrate materials with wide and narrow bandgaps into the same chip, where they could serve as sensing, logic and power components, to create integrated systems on chips. The conducting graphene layers (not shown in **a**) that grow on top of the buffer layer could be used as connections between chip components.