spectrum of pLoF tolerance. The larger sample size means that gene length is less of a problem in the gnomAD analysis, but even so, the authors could not definitively assess pLoF frequency in the 30% of genes that were expected to have few pLoF variants.

Despite this limitation, the group use their approach to gain fresh insights into the genetics of disease. For example, they found rare variants in genes that do not tolerate LoF more often in people who have an intellectual disability or autism spectrum disorder than in people who do not. These data might help researchers to understand the complex genetic structure that underlies these traits.

In the second paper of the collection, Cummings *et al.*[4] investigated why genes that seem intolerant to pLoF can sometimes carry these variants with apparently little consequence. Genes can be transcribed in different ways, with some protein-coding regions (exons) expressed only in a limited fashion. Cummings and colleagues demonstrated that, when an individual carries a pLoF variant in an 'intolerant' gene, the variant is often in an exon that shows this restricted expression, thus limiting its effect.

In the third paper, Minikel *et al.*[5] assessed how the pLoF database might improve our ability to identify genetic targets for drugs. The identification of individuals who carry two pLoF variants in a given gene is desirable in drug discovery — if these individuals also exhibit a change in a particular trait, it provides evidence that the gene could be a good drug target[8]. The group showed that there are still many errors when identifying pLoF variants; that quality control is needed when identifying these variants; and that instances of an individual carrying two pLoF variants in the same gene are sufficiently rare that we will need cohorts roughly 1,000 times bigger than gnomAD to gather definitive evidence of their existence in most genes.

One of the most exciting aspects of the gnomAD project is the production of a catalogue of structural variants, described in the final paper by Collins and co-workers[6]. There have been excellent efforts at cataloguing structural variants using long-read sequencing technology[9]. However, sample sizes have been small, owing to the expense and lack of standardized analysis pipelines for this approach — although I expect this situation to improve in the near future. By contrast, identifying structural variants in short-read sequences is technically challenging, because the variants are often larger than a typical short sequence read, and they can arise through a variety of mutational mechanisms, resulting in many variant types (duplication, deletion or inversion of DNA, for instance) that each leave different footprints in the genome. This has led to the development of many tools for identifying structural variants from short reads, but no 'standard' pipeline.

Collins *et al.* sought to remedy this problem by creating a pipeline that allows for harmonized analysis over thousands of genomes; this could become the industry standard for structural-variant detection from short-read sequences on a population scale. The authors generated a catalogue of more than 300,000 high-quality structural variants — more than twice as many as previous analyses. They then began to assess the contribution of structural variants to physiological traits. This analysis revealed some evidence for natural selection against structural variants in non-coding sequences that control gene expression. Unsurprisingly, selection against structural variants was stronger in protein-coding regions. This suggests that more variation is tolerated in non-coding than in coding regions, and that even-larger cohorts (or other approaches) will be needed to begin to robustly dissect non-coding variation. The authors also found that structural variants account for roughly one-quarter of protein-truncating events.

The routine analysis of structural variants, integrated with analysis of SNVs and gene expression, will be crucial for interpreting individual genomes. Collins *et al.* have taken an important step in this direction, and the gnomAD resource provides tools for others to continue on this path.

An interesting, recurring theme in these papers is that — despite the size of the cohort — we still lack the numbers required for many analyses. The sequencing of ever-larger cohorts should no doubt continue. However, this approach alone will not enable us to fully understand the relationships between human genetics and traits at both cellular and organismal levels. We need scalable approaches to program genetic variation into human cells, and well-characterized cellular traits that can be monitored to allow us to directly interrogate the physiological impact of this variation. Such interventional biology will substantially augment population genetics and accelerate our understanding of human biology.

The gnomAD consortium has already made its data publicly available. The impact that the project will have on science goes well beyond the current collection, which includes not only the papers in this issue, but several published in *Nature*'s sister journals (go.nature.com/2zgfxr2). The gnomAD resource, like ExAC before it, will change how we interpret individual genomes. The consortium's work has revealed how much information about human variation we had been missing and has provided tools that help us to better understand the genome at both the population and individual level. I can't wait to see what comes next.

Deanna M. Church is at Inscripta, Boulder, Colorado 80301, USA.
e-mail: deanna.church@inscripta.com

1. Auton, A. *et al. Nature* **526**, 68–74 (2015).
2. Chiang, C. *et al. Nature Genet.* **49**, 692–699 (2017).
3. Karczewski, K. J. *et al. Nature* **581**, 434–443 (2020).
4. Cummings, B. B. *et al. Nature* **581**, 452–458 (2020).
5. Minikel, E. V. *et al. Nature* **581**, 459–464 (2020).
6. Collins, R. L. *et al. Nature* **581**, 444–451 (2020).
7. Lek, M. *et al. Nature* **536**, 285–291 (2016).
8. Plenge, R. M., Scolnick, E. M. & Altshuler, D. *Nature Rev. Drug Discov.* **12**, 581–594 (2013).
9. Chaisson, M. J. P. *et al. Nature Commun.* **10**, 1784 (2019).

The author declares competing financial interests: see go.nature.com/3eklyoi for details.

Materials science

A model of perfection for light-activated catalysts

Simone Pokrant

Efforts to make hydrogen from water directly using sunlight have been hampered by the inefficiency of the catalysts that promote the process. A model system demonstrates that almost perfectly efficient catalysts can be made. **See p.411**

Since the emergence of Greta Thunberg's 'Fridays For Future' movement in August 2018, the need to prevent climate change and to find 'green' alternatives to fossil fuels have become topics of broad public interest. But although public awareness has advanced rapidly, progress in the search for cost-effective technological solutions has not. One promising sustainable energy carrier is hydrogen, if it can be produced using renewable energy sources — hydrogen is a green fuel, because its combustion produces only pure water. On page 411, Takata *et al.*[1] report a breakthrough in catalyst design that might accelerate the development of large-scale processes for making hydrogen from water using sunlight.

The largest potential source of renewable energy is the Sun[2]: about 0.02% of the solar

energy absorbed by Earth's surface annually would be enough to cover current global energy consumption. Many approaches for converting solar energy into the chemical energy stored in hydrogen are therefore being investigated, using 'water-splitting' reactions in which water is broken down into hydrogen and oxygen. Some of these approaches are already being tested in pilot facilities – such as solar-power-to-gas units, in which electricity produced by solar cells is used to split water through electrolysis[3]. Hydrogen produced in this way could be used for the long-term storage of solar energy, or as fuel for vehicles. However, solar-to-gas conversion processes are not yet economically viable.

A study of the technical and economic feasibility of solar-energy production[4] has shown that systems based on light-activated catalysts (photocatalysts) are attractive alternative options for water splitting. In these systems, photocatalytic semiconductor particles are suspended in a bed filled with an aqueous electrolyte; when sunlight shines on the suspension, hydrogen and oxygen gases are produced. The technical simplicity of this approach should enable economically competitive hydrogen production, if the photocatalyst can convert solar energy to hydrogen with a minimum efficiency of about 10%.

However, the conversion efficiencies of photocatalytic semiconductors are typically much lower than 10%. This is because the photocatalytic process is highly complex and requires the semiconductor particles to have a combination of several properties. They must: absorb light; generate and separate electron–hole pairs (holes are positively charged quasiparticles produced when photons knock electrons out of an atomic lattice); enable holes and electrons to travel to the particle–water interface; and catalyse the production of hydrogen and oxygen from water (reactions that require electrons and holes, respectively). Side processes that can occur at each step can lower the overall conversion efficiency. Materials scientists are therefore trying to design photocatalysts that minimize such efficiency losses.

A key measure of the effectiveness of a photocatalyst is the fraction of absorbed photons that it can use to produce hydrogen, a quantity called the internal quantum efficiency (IQE). A perfect photocatalyst that converts all of the absorbed photons to hydrogen would have an IQE of 1 (or 100%). However, IQE cannot be determined from experiments.

A related quantity that can be experimentally determined for a reaction is the external quantum efficiency (EQE): the fraction of photons illuminating the reaction vessel that the photocatalyst can use to produce hydrogen. This value is always lower than the IQE, because an unknown portion of the illuminating photons will not be absorbed by the photocatalyst, but will instead be lost to other processes, such as scattering. If similar photocatalyst-particle suspensions are investigated using the same experimental set-up, ensuring that the same fraction of light is absorbed, then EQE can be used as an indirect measure of IQE. But EQEs determined using different set-ups cannot be used as a way of comparing IQEs of photocatalytic systems, because the relationship between EQE and IQE is different for each set-up – therefore making it difficult for different research groups to compare results.

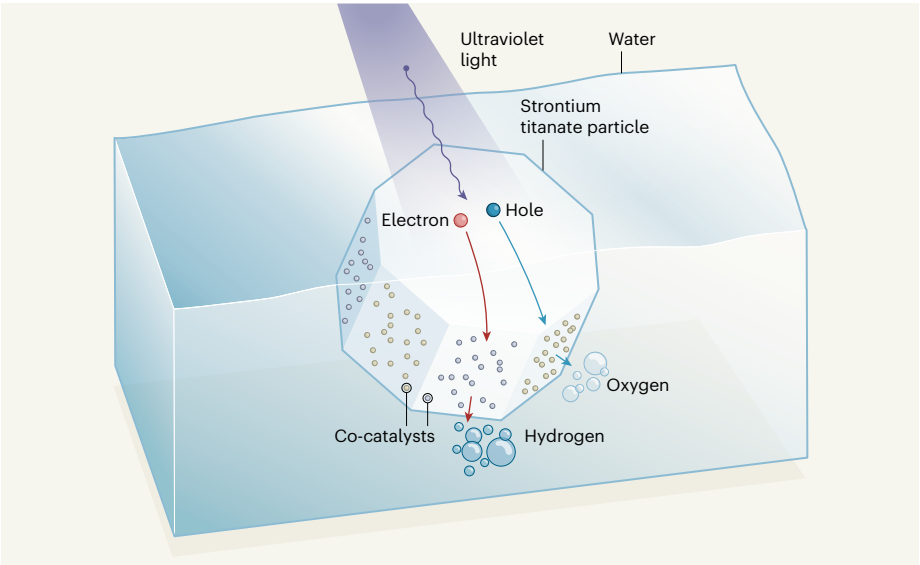

Figure 1 | A photocatalytic particle engineered for water splitting. Particles of light-activated catalysts (photocatalysts) can be used to drive water splitting – the reaction in which water is broken down into hydrogen and oxygen gases. Takata *et al.*[1] prepared photocatalysts made of highly crystalline strontium titanate that contains a small number of aluminium atoms. When the particles are suspended in water and irradiated with ultraviolet light, electrons and holes (positively charged quasiparticles) are produced, and travel to different facets of the crystalline particle. The authors selectively deposited appropriate co-catalysts on the facets to promote hydrogen production (using the electrons) at the electron-collecting facets, and oxygen production (using the holes) at the hole-collecting facets. The selective migration of electrons and holes to different reaction sites contributed to the almost perfectly efficient conversion of light to hydrogen molecules by the photocatalyst.

"This combination of complex mitigation strategies proved highly successful."

Takata *et al.* focus on strontium titanate ($SrTiO_3$) – one of the first materials found to split water photocatalytically, as reported[5] in 1977. Strontium titanate produces electron–hole pairs by absorbing ultraviolet light. Because the Sun's intensity is highest in the visible-light range, it is unlikely that UV-driven catalysts will enable sustainable hydrogen production on a large scale. However, strontium titanate is an excellent model system for studying the influence of photocatalyst parameters on quantum efficiency (both EQE and IQE), because the mechanisms that cause efficiency losses in this material are well understood, and strategies for mitigating the losses have been proposed.

The authors used a combination of approaches to address specific loss mechanisms. One such mechanism is charge recombination, a process in which electrons and holes recombine before they can take part in water splitting. Takata and colleagues suppressed recombination in several ways. The first approach was to improve the crystallinity of the photocatalyst particles, thereby reducing the number of lattice defects. Another method was to reduce the number of chemical defects in the crystal lattice using aluminium doping – a process in which small quantities of aluminium atoms are incorporated into the lattice. These two approaches work because any defect (a lattice defect or a chemical defect) can act as a potential centre for recombination[6].

Takata and colleagues also took advantage of the fact that electrons and holes in their strontium titanate crystals collect at different crystal facets – a feature that further suppresses charge recombination. The authors selectively deposited appropriate co-catalysts on the facets to promote hydrogen production at the electron-collecting facets, and oxygen production at the hole-collecting facets (Fig. 1); this approach was previously proposed[7] and developed[8] by other research groups. Finally, the authors prevented an unwanted side reaction (the oxygen-reduction reaction) by encasing the rhodium co-catalyst

for the hydrogen-producing reaction in a protective shell of a chromium compound.

This combination of complex mitigation strategies proved highly successful: the authors reported EQEs of up to 96% when their photocatalysts were irradiated with light in the wavelength range of 350–360 nanometres. This is excellent news, because it means they have designed an almost perfect photocatalyst – the IQE must be between 96% and 100%.

This is a spectacular result for several reasons, even though strontium titanate is 'just' a model system for visible-light photocatalysts. First, it demonstrates that experiments can be designed in which EQEs come close to IQEs within an acceptable error margin of less than 4%. Improved experimental set-ups in which measured EQEs are very near to IQEs should facilitate the comparison of photocatalysts and therefore accelerate progress in this field.

Second, it proves that the combination of design strategies used by the authors can indeed eliminate efficiency losses associated with recombination. It is to be expected that the strategies used to improve the efficiency of strontium titanate will also apply to photocatalysts driven by visible light – and could therefore enable the conversion of solar energy to hydrogen with efficiencies of about 10%.

Finally, and most importantly, Takata and colleagues' findings will inspire and encourage other researchers to continue their work on photocatalysts. One of the authors of the work, Kazunari Domen, published his first paper[9] on the use of strontium titanate as a photocatalyst in 1980. This shows the timescale needed for success in this area. Although we do not yet have a route for the sustainable and economically viable production of hydrogen, we stand a good chance of finding one in the next few decades. This paper vouches for it.

Simone Pokrant is in the Department of Chemistry and Physics of Materials, University of Salzburg, Salzburg A-5020, Austria.
e-mail: simone.pokrant@sbg.ac.at

1. Takata, T. *et al. Nature* **581**, 411–414 (2020).
2. Abbott, D. *Proc. IEEE* **98**, 42–66 (2010).
3. Ochoa Robles, J., De-León Almaraz, S. & Azzaro-Pantel, C. in *Hydrogen Supply Chains* 8–11 (Academic, 2018).
4. Pinaud, B. A. *et al. Energy Environ. Sci.* **6**, 1983–2002 (2013).
5. Wrighton, M. S., Wolczanski, P. T. & Ellis, A. B. *J. Solid State Chem.* **22**, 17–29 (1977).
6. Ham, Y. *et al. J. Mater. Chem. A* **4**, 3027–3033 (2016).
7. Giocondi, J. L. & Rohrer, G. S. *J. Am. Ceram. Soc.* **86**, 1182–1189 (2003).
8. Mu, L. *et al. Energy Environ. Sci.* **9**, 2463–2469 (2016).
9. Domen, K., Naito, S., Soma, M., Onishi, T. & Tamaru, K. *J. Chem. Soc. Chem. Commun.* 543–544 (1980).

Molecular biology

Evolution of a molecular machine

Michael Berenbrink

The multi-subunit protein haemoglobin relies on complex interactions between its components to function properly. Analysis of ancient precursors suggests that its evolution from a simple monomer involved only a few steps. **See p.480**

The oxygen-transporting protein haemoglobin has undergone repeated adaptations as animals evolved to conquer new environments – from the depths of the oceans[1] to high mountain ranges[2]. These adaptations relied on changes in the long-range interactions between oxygen-binding sites buried in the protein's subunits, and between these regions and binding sites for a multitude of small effector molecules on the protein's surface[3]. How did this complex molecular machine, which can respond so exquisitely to available levels of both oxygen and several other effector molecules, come into being? On page 480, Pillai *et al.*[4] reconstruct the stepwise evolution of haemoglobin from precursors that existed more than 400 million years ago.

Almost nothing was previously known about how the four-subunit (tetrameric) form of haemoglobin that is found in modern-day jawed vertebrates evolved from ancient monomers. Tetrameric haemoglobin consists of two α- and two β-subunits. Pillai *et al.* computationally reconstructed an evolutionary tree to chart the protein's ancient history, using the amino-acid sequences of a large collection of the closely related vertebrate globin proteins, which exist as either monomers or tetramers.

"Pillai and colleagues' work serves as one of the clearest examples so far of how such complexity can arise."

The authors' tree was constructed taking into account that amino-acid substitutions a given protein shares with close relatives tend to have originated in more-recent common ancestors than have those it shares with more-distant relatives. The reconstructed evolutionary tree indicates that multiple rounds of gene duplication and subsequent divergence gave rise to the globin family and, by way of several ancestral proteins, to tetrameric haemoglobin (Fig. 1).

What is special about the study is that Pillai and colleagues went on to resurrect several of these extinct ancestral proteins, generating them from the amino-acid sequences predicted by the tree. The group then tested these proteins' functions.

First, Pillai and colleagues analysed whether each ancestral protein could form dimers and tetramers of like or unlike subunits. The earliest protein – a common ancestor of haemoglobin and the monomeric globin protein myoglobin, named AncMH by the authors – exists only as a monomer. A later protein, named Ancα/β, which is the ancestor of all existing haemoglobin subunits, forms homodimers when expressed at high levels. The authors' tree indicates that Ancα/β underwent gene duplication to produce two proteins: the ancestors of all existing α- or β-subunits, which the group respectively named Ancα and Ancβ. These proteins also form homodimers, or even homotetramers, when expressed alone. However, when the two are expressed together in equal proportions, they can form heterodimers, which then further align to yield haemoglobin tetramers.

The group next investigated the oxygen-binding affinity of the ancestral proteins, along with their oxygen cooperativity (the ability of oxygen-binding subunits to interact with one another) and their 'allosteric' regulation by a potent, artificial effector molecule, inositol hexaphosphate (IHP). They found that only Ancα and Ancβ – when expressed together at high concentrations – show similar oxygen-binding affinity, cooperativity and allosteric regulation to today's haemoglobin protein. These features are shared by all living jawed vertebrates, but are absent or achieved in a different way in jawless vertebrates, whose haemoglobin proteins are of more ancient origin. This indicates that the basic functions of jawed-vertebrate haemoglobin had already evolved in a common ancestor of these animals but at some time after the split with jawless vertebrates.

Next, Pillai *et al.* modelled the stepwise changes in α- and β-subunit interfaces that might have allowed Ancα and Ancβ first to form heterodimers with one another, and later heterotetramers from pairs of such dimers. The modelling indicated that strikingly few amino-acid substitutions might have been needed to transform a simple monomeric