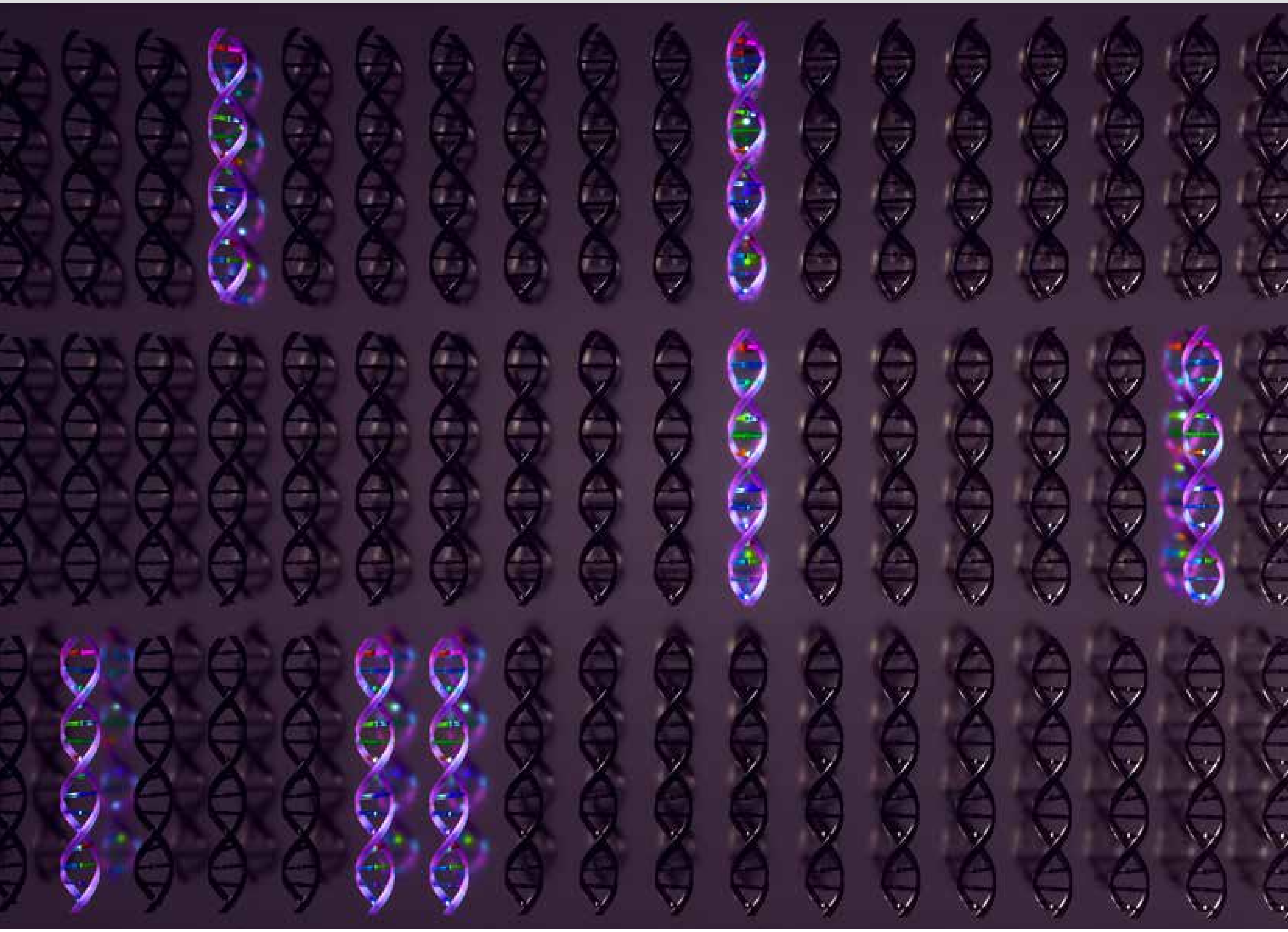
HOW TO BRING MORE DIVERSITY INTO POLYGENIC RISK SCORES



illumina®

nature research custom media

HOW TO BRING MORE DIVERSITY INTO POLYGENIC RISK SCORES

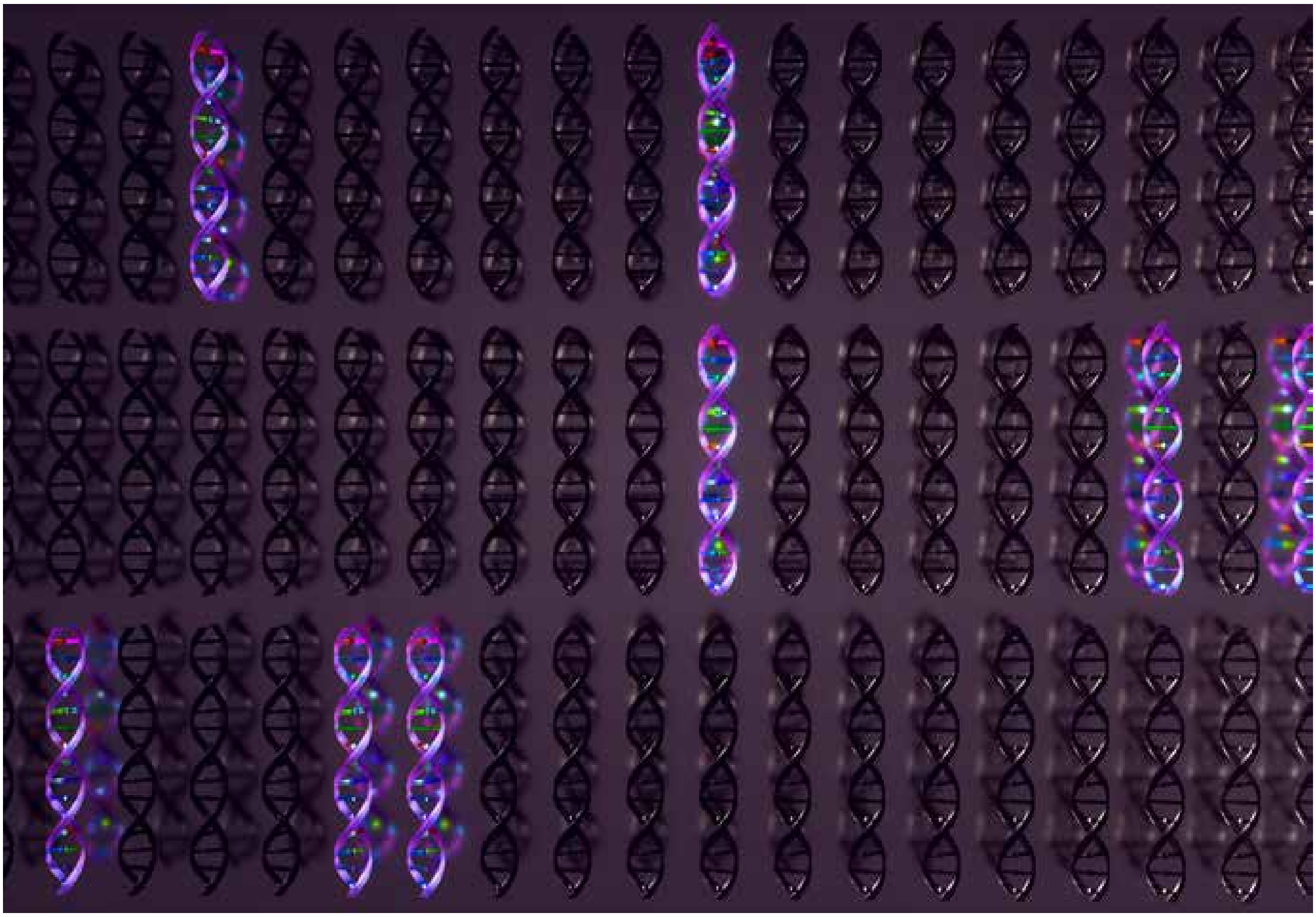
EURO-CENTRIC GENOMIC DATA skew attempts to calculate polygenic risk. Statistical adjustments allow research to move on, even if they don't solve the underlying problem.

A polygenic risk score (PRS) is used to predict how likely it is that someone will develop a particular disease, based on the presence of a vast number of tiny variant regions in their genome. The PRS concept is steadily moving from the margins of research towards the mainstream. But the genomic data that PRS is based on has a diversity problem. The largest databases most commonly used to generate the scores, such as the UK Biobank, are hugely biased towards populations of limited ancestry, making PRS much less accurate for some populations.

"I am concerned that any polygenic risk score I use is more applicable to someone of European ancestry, with less relevance to African-Americans or people of Asian ancestry," says cardiologist and professor of medicine Iftikhar Kullo, in the Department of Cardiovascular Medicine at the Mayo Clinic, Rochester, Minnesota.

Ying Wang, a statistical and population geneticist at Massachusetts General Hospital, says: "At present, the prediction performance of PRS in African populations may achieve only about 20% of the level of accuracy achieved in European populations."

A person's PRS is generated



Ella Maru Studios

▲ **Lack of diversity in the underlying genetic data hampers broad applicability of polygenic risk scores.**

by searching for the presence of variants in their DNA that have been previously linked to a specific disease by genome-wide association studies (GWAS) involving vast numbers of individuals. Most of the variants are single-nucleotide polymorphisms (SNPs, pronounced 'snips'), which are individually rare — but collectively common — changes to one DNA base.

PRS researchers all agree that the only real fix is to generate more representative primary genomic data and conduct more diverse GWAS. However, while these time-consuming tasks advance, there are measures that can be used in the meantime. Wang and Kullo are part of a movement among researchers to develop ways to at least partially correct the bias in PRS, using statistical methods among other innovations. Such approaches indicate that PRS progress is possible, even in this imperfect scenario.

SOLUTIONS IN STATISTICS

In order to improve the applicability of PRS, researchers make statistical modifications when analysing the existing GWAS data. Many different methods are being tried; the most common strategy is to adjust the weightings given to the genetic variants deemed to be most significant for both the non-European populations, and for the population that each individual belongs to. This is assisted by taking account of the biological functions of genetic variants, a process known as functional annotation, rather than just using variants without considering their likely effects. Other key factors involve looking at data across multiple genetic ancestries and exploring more than one specific disease or genetic trait.

Wang and her co-authors have appraised the wide variety of existing PRS methods, many of which make some progress towards at least partially overcoming the diversity problem[1]. She adds that more approaches are under development. The reported improvements in performance, compared to raw un-manipulated data, can be substantial — although still far from perfect. "I do think they are making a significant difference," she says, "although there is no one-size-fits-all method, and their success depends on trait-specific genetic architecture."

Alkes Price, a geneticist at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, is also working on the PRS problem. "Including family history improves the accuracy of polygenic risk scores, especially in diverse populations," he says.

Price and his colleagues have demonstrated the potential of this method by combining raw GWAS data from the UK Biobank with family history of disease[2]. They found that average prediction success of PRS alone in non-British Europeans, South Asians and Africans was only 5.8%, 4.0% and 0.53% respectively. But when knowledge of family history of the target disease was added, to create what they call PRS-FH scores, the success rates rose to 13%, 12% and 10%. The study looked at major diseases such as stroke, lung disease, and several common cancers. They confirmed that PRS predictions, while initially weak, were improved by the additional information to be superior to looking at family histories alone.

"Additional statistical improvements are certainly possible, and are being investigated by many research groups, including ours," Price says.

Wang is a member of the research group at Massachusetts General Hospital, led by Alicia Martin, which has recently published two methods to improve PRS predictions across diverse populations. One, which they call PRS-CSx, integrates the data in the commonly used large Euro-centric GWAS databases with the limited data already available from GWAS studies in Japan and Taiwan[3]. This approach almost doubled the accuracy of predicting schizophrenia in some non-European populations.

> **"WE LOOK AT THE TRANSCRIPTS OF AROUND 20,000 OR SO GENES, RATHER THAN JUST SNPS; I LIKE TO DESCRIBE OUR METHOD AS GETTING CLOSER TO THE BIOLOGICAL TRUTH."**

The second method, developed in collaboration with Price's group, is called PolyPred[4]. This is a hybrid of two other modified PRS predictors – PolyFun-pred, which takes account of specific causative effects of genetic variants to make scores more relevant to target populations, and BOLT-LMM, which is claimed to capture genotype signals more effectively in any target population. The team applied PolyPred to 23 diseases or traits and found up to 24% improvement in accuracy in non-European populations. But they do acknowledge that "prediction accuracy in non-Europeans is still substantially lower compared with Europeans".

EXPRESSIONS OF INTEREST

A new and significantly different approach to predicting disease risk has been developed by a team led by Hae Kyung Im, at the University of Chicago. Called Polygenic Transcriptome Risk Scoring (PTRS), it uses data from the same biobanks used for conventional PRS to predict the level of disease-associated messenger RNA, known as gene transcripts[5].

"Our scores are based on genes rather than genomic locations," says Im. "We have already found evidence that they can be applied more accurately to different populations." She cites work showing that PTRS significantly improves the reliability of disease trait prediction in populations of African ancestry and, when combined with conventional PRS, the accuracy is even further improved. "We look at the transcripts of around 20,000 or so genes, rather than just SNPs; I like to describe our method as getting closer to the biological truth," Im says. She emphasizes that information at the level of the specific gene activity should be more useful and more clinically relevant than statistical associations considering millions of SNPs.

The PTRS team also has a preprint manuscript[6] suggesting that the data in human biobanks could have relevance to disease even in other species. "We have trained PTRS predictions for height using data from humans and found that they can be transferred to rats, which as far as I know is the first example of polygenic risk analysis being transferrable from humans to another species," Im says. She suggests that her team's plans to investigate the transferability of human GWAS data to diseases in other species can only help efforts to apply GWAS data more widely across different populations of humans.

COMING TO THE CLINIC

While much of the research in PRS is still at the level of exploration and development, for some diseases it is already being put to clinical use. Kullo, at the Mayo Clinic, focuses on applying it specifically to coronary heart disease. He says that this is probably the condition for which PRS is closest to being routinely applied. "In certain situations, it can reveal why some patients had a heart attack, or if a patient is at a higher risk of an attack in the future." Poor transferability remains a problem, he adds, "but the most recent polygenic risk scores for coronary heart disease have improved to the point that these may be useful even in non-European ancestry populations."

Kullo acknowledges that statistical techniques to improve the portability of the scores to diverse ancestries are getting better, although the scores are the weakest for people of African ancestry.

Statistical manipulations, and other methods, strive to reduce the limitations of the existing genetic data, but true equity in PRS will only come from more globally diverse, large genomic datasets, says Kullo. "There's probably a limit to what can be done with statistical techniques."

The necessary efforts to address the data deficiencies are underway. For example, the Global Biobank Meta-Analysis Initiative, Human Heredity and Health in Africa, the Million Veterans Program, AllofUs and TOPMED are all actively recruiting individuals of diverse ancestry to widen their representation.

Martin notes that PRS is already being marketed by some companies and made available by researchers to doctors, patients and curious consumers for some high priority diseases like heart disease and breast cancer. "Despite the lack of PRS generalizability across different ancestry populations, many major efforts are underway to significantly improve the situation. There is a methods deluge."

While PRS is still not ready for the clinic, research into its application is nonetheless having an important effect. Martin says all these efforts are laying the groundwork for future translational uses, such as prophylactic statin prescriptions or altering mammography screening recommendations based on baseline risk. "PRS is already teaching us about how heritable disease predispositions associate with other aspects of people's lives — and how we should alter preventative risk models to be more comprehensive." ■

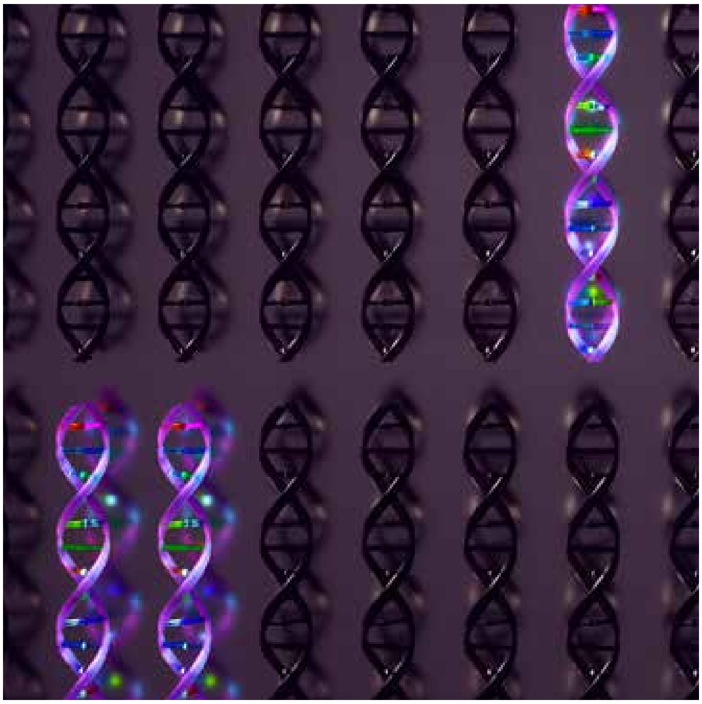
▲ **Most polygenic risk scores are calculated using data from European populations, which do not translate to other ethnicities.**

> **"THE MOST RECENT POLYGENIC RISK SCORES FOR CORONARY HEART DISEASE HAVE IMPROVED TO THE POINT THAT THESE MAY BE USEFUL EVEN IN NON-EUROPEAN ANCESTRY POPULATIONS."**

REFERENCES

1. Wang, Y. *et al. Annu. Rev. Biomed. Data Sci*. **5**, 293-320 (2022).
2. Hujoel, M. L. A. *et al. Cell Genomics*, **2**, 100152 (2022).
3. Ruan, Y. *et al. Nature Gen*. **54**, 573-580 (2022).
4. Weissbrod, O. *et al. Nature Gen*. **54**, 450-458 (2022).
5. Liang, Y. *et al. Genome Biol*. **23**, 23 (2022).
6. Santhanam, N., *et al*. Preprint at bioRxiv https://doi.org/10.1101/2022.06.03.494719 (2022).

illumina®



Ella Maru Studios

This is a reprint of an advertisement feature originally published on nature.com.
It can be accessed at www.nature.com/articles/d42473-022-00315-7