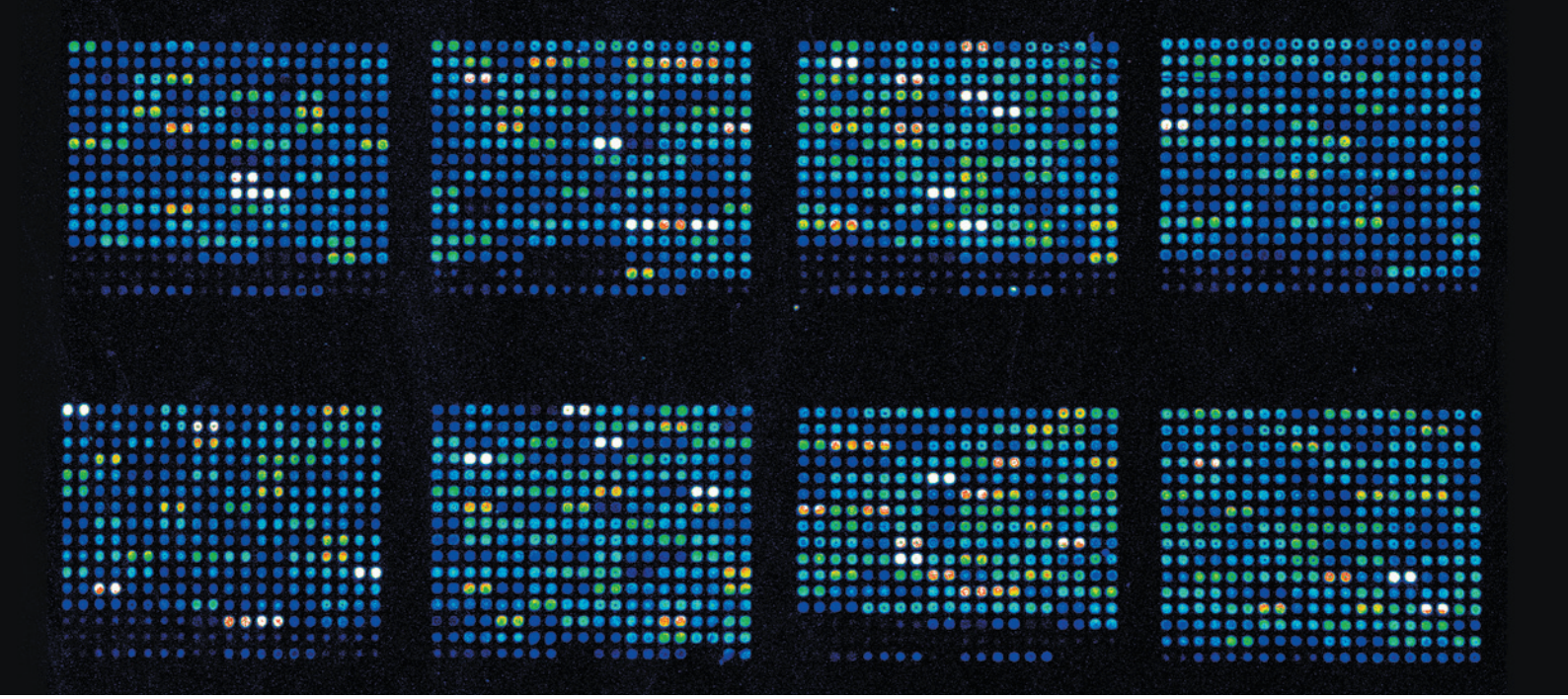


DNA microarrays allow researchers to analyse the expression of a huge number of genes simultaneously.

GENOMICS

Gene data to hit milestone

With close to one million gene-expression data sets now in publicly accessible repositories, researchers can identify disease trends without ever having to enter a laboratory.

BY MONYA BAKER

Purvesh Khatri sits in front of an oversized computer screen, trawling for treasure in a sea of genetic data. Entering the search term 'breast cancer' into a public repository called the Gene Expression Omnibus (GEO), the postdoctoral researcher retrieves a list of 1,170 experiments, representing nearly 33,000 samples and a hoard of gene-expression data that could reveal previously unseen patterns.

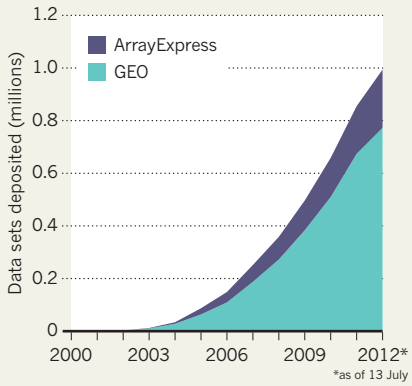
That is exactly the kind of search that led Khatri's boss, Atul Butte, a bioinformatician at the Stanford School of Medicine in California, to identify a new drug target for diabetes. After downloading data from 130 gene-expression studies in mice, rats and humans, Butte looked for genes that were expressed at higher levels in disease samples than in controls. One gene was strikingly consistent: *CD44*, which encodes a protein found on the surface of white blood cells, was differentially expressed in 60% of the studies (K. Kodama *et al. Proc. Natl Acad. Sci. USA* **109,** 7049–7054; 2012). The CD44 protein is not widely investigated as a drug target for diabetes, but Butte's team found that treating obese mice with an antibody against it caused their blood glucose levels to drop.

Butte and his team are now using publicly available data to answer a diverse range of questions — Khatri, for instance, hopes to discover secrets behind kidney-transplant rejection. "We don't do wet lab experiments for discovery," he says. Those are for validating hypotheses. The beauty of analysing data from multiple experiments is that biases and artefacts should cancel out between data sets, helping true relationships to stand out, Butte says. "There is safety in numbers."

And those numbers are rising rapidly. Since 2002, many scientific journals have required that data from gene-expression studies be deposited in public databases such as GEO, which is maintained by the National Center for Biotechnology Information in Bethesda, Maryland, and ArrayExpress, a large gene-expression repository at the European Bioinformatics Institute (EBI) in Hinxton, UK. Some time in the next few weeks, the number of deposited data sets will top one million (see 'Data dump').

DATA DUMP

The number of gene-expression data sets in publicly available databases has climbed to nearly one million over the past decade.



SOURCES: NIH, EBI

The result is an unprecedented resource that promises to drive down costs and speed up progress in understanding disease. Gene-sequence data are already shared extensively, but expression data are more complex and can reveal which genes are the most active in, say, liver versus brain cells, or in diseased versus healthy tissue. And because studies often look at many genes, researchers can repurpose the data sets, asking questions other than those posed by the original researchers.

It is easy to track how many data sets are being deposited — much harder is working out how they are being used. Heather Piwowar, who studies data reuse with the National Evolutionary Synthesis Center from the University of British Columbia in Vancouver, Canada, found that 20% of data sets deposited in GEO in 2005 and 17% of those in 2007 had been cited by the end of 2010. But those rates are certainly underestimates, she says. The PubMed Central repository, which her study relied on, holds only about one-third of the relevant papers, and her algorithms identify reuse only when researchers cite database accession numbers, which many don't do. More studies are reusing data every year, she says. "We have every reason to believe it is game-changing."

Having access to such data is "immensely

A. NANTEL/SHUTTERSTOCK

valuable," agrees Enrico Petretto, a genomicist at Imperial College London. "We would never be in a position to look across multiple tissues and species with the money we have." But he cautions that using other people's data can be tricky. If data sets give contradictory outcomes, it is unclear whether that is because the underlying data contradict each other or because something went wrong with the analysis. "That's why people sometimes don't trust this," he says.

CHANGE OF PRACTICE

Still, few researchers are using the data to their greatest potential, says Alvis Brazma, a bioinformatician at the EBI. "Being able to reuse functional genomics data is a really new thing," he says. Researchers rarely download more than half a dozen data sets, and most use the data only to compare with their own results. Studies that use only other scientists' data to come up with new findings are still unusual.

That makes Butte and Khatri trailblazers. Another pioneer is Gustavo Stolovitzky, a computational biologist at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York, who has used publicly available data to train algorithms to recognize gene signatures for diseases such as lung cancer, chronic obstructive pulmonary disease (COPD) and psoriasis. Not only can the algorithms distinguish lung cancer from COPD, they can also tell squamous-cell carcinoma from adenocarcinoma. "There is enough info in existing databases to predict disease in samples that algorithms have never seen before," Stolovitzky says.

Other efforts promise to unleash even more power from the growing repositories. In 2009, for instance, curators of ArrayExpress used their database to create the Gene Expression Atlas, which allows researchers to look at how the expression of a gene might vary across tissues, disease states and species without having to download any data.

Curators will have to adjust to the ways that data are changing, says Tanya Barrett, coordinator at GEO. A growing proportion of the data finding their way into repositories are derived from RNA sequences, which poses challenges: the files are larger, methods are still in flux and integration with conventional microarray data is difficult. But the biggest factor to limit data reuse could be cultural. Many researchers are reluctant to use data that are in different formats, or from other experimental designs or materials, says Ann Zimmerman, who studies data reuse at the University of Michigan in Ann Arbor. Familiarity could help to solve the problem, says Barrett. The more examples of data reuse that scientists see, the more ways they will find to reuse data. ■

➲ NATURE.COM
To read a *Nature* supplement on genomics, see:
go.nature.com/ftkwlr

HEALTH

Wary approval for drug to prevent HIV

US regulators seek to mitigate risks of combined pill.

BY AMY MAXMEN

US regulators took a step into the unknown this week when they approved the first drug to prevent HIV infection. US Food and Drug Administration (FDA) commissioner Margaret Hamburg hailed the pill, Truvada, as a tool for reducing the rate of infection in the United States, where 50,000 people are diagnosed each year. But the drug combines low doses of two antiretroviral agents normally used to treat infection, and some researchers fear that its use in healthy people could have unacceptable side effects and spark the emergence of resistant viruses.

US insurers must now decide whether they will pay for Truvada, which costs roughly US$10,000 for a year's supply. Moreover, health-policy experts must script guidelines on how to prescribe it, and how to monitor side effects and HIV infections in people using the drug. "There are a lot of questions about how to implement it," says Connie Celum, an HIV researcher at the University of Washington in Seattle, who led a large trial[1] of the drug in East Africa and has begun studies to answer practical delivery questions, such as which subsets of people are at highest risk.

Developed by Gilead Sciences in Foster City, California, Truvada proved particularly effective in the East African trial[1], published last week: it reduced the incidence of HIV by 75% in people with partners who had been infected. In an earlier trial[2] in the United States, HIV incidence dropped by 44% in men who have sex with men.

But concerns emerged on 10 May at a public meeting of a panel that advised the FDA on its decision. Most members voted in favour of approval, but the researchers, doctors and patient advocates in attendance wrestled with the issue of drug resistance. The two drugs in Truvada, emtricitabine and tenofovir, are effective antiretroviral treatments, but trials have shown that viruses exposed to lower doses in the acute phase of infection can become resistant, said meeting attendees. In six people who tested negative on enrolment but turned out to be HIV-positive, the drugs were no longer effective. Another fear, unconfirmed in trials, was that people might not take the pill consistently, and might contract a strain of HIV that became drug-resistant as a result of exposure to low levels of antiretrovirals.

To mitigate these risks, the FDA requires that Truvada be prescribed only once an individual has tested negative for HIV. The agency also advises that people use the drug in combination with safe sex practices, and get tested for the virus every three months while taking it. Some experts at the advisory meeting proposed stricter policies, such as making the tests mandatory, but these were dismissed as impractical. Another idea was to limit the drug to specific populations who are at the very highest risk, such as homosexual people who use intravenous drugs, but the FDA adopted a vaguer category encapsulating anyone at high risk of contracting HIV. "We want to reach marginalized populations," says Celum, "and restricting access would mean that Truvada would be less likely to have a public-health impact."

Wayne Chen, acting chief of medicine at the AIDS Health Foundation in Los Angeles, California, regrets the decision to approve the drug, saying that condoms are cheaper and can be a more effective preventative. "The best thing would be to have this drug withdrawn from the market, and if it's not, there should at least be mandatory testing because we know that people don't take this as prescribed," he says, citing a Truvada clinical trial[3] in Africa that was ended prematurely because the drug was not preventing infection. Blood tests later confirmed that fewer than 40% of the study participants on Truvada had been taking the pills daily.

"Truvada is now the only technology we have that empowers women."

To proponents, however, the promise of the drug is bright. Salim Abdool Karim, director of the Center for the AIDS Programme of Research in South Africa in Durban, hopes that Truvada might soon be available in his country, where up to one-quarter of women have HIV by the age of 20. "Truvada is now the only technology we have that empowers women," he says. "I don't think we'll be able to slow the HIV epidemic in South Africa without something to protect them." ■

1. Baeten, J. M. *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1108524 (2012).
2. Grant, R. M. *et al. N. Engl. J. Med.* **363,** 2587–2599 (2010).
3. Van Damme, L. *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1202614 (2012).