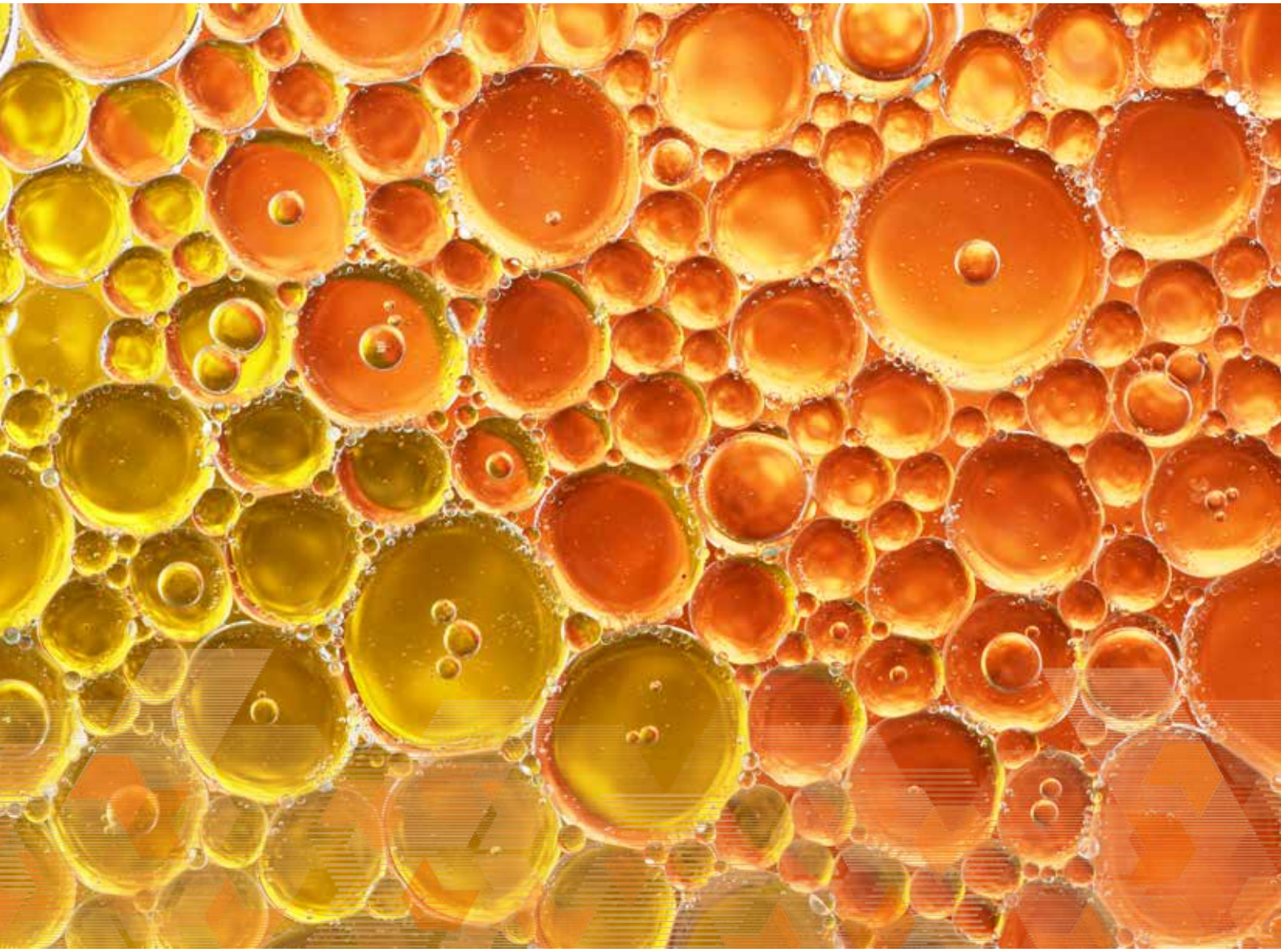
WHITEPAPER

Interpreting Next Generation Sequence Data



SUMMARY
The development of personalized medicines and the study of rare diseases are two areas of life sciences research that are being revolutionized by stunning advances in DNA sequencing technology. In the 13 years since the Sanger sequencing method was employed to sequence the first whole human genome, an array of exponentially faster and far less costly technologies has emerged.

ELSEVIER

Next generation sequencing

INTRODUCTION

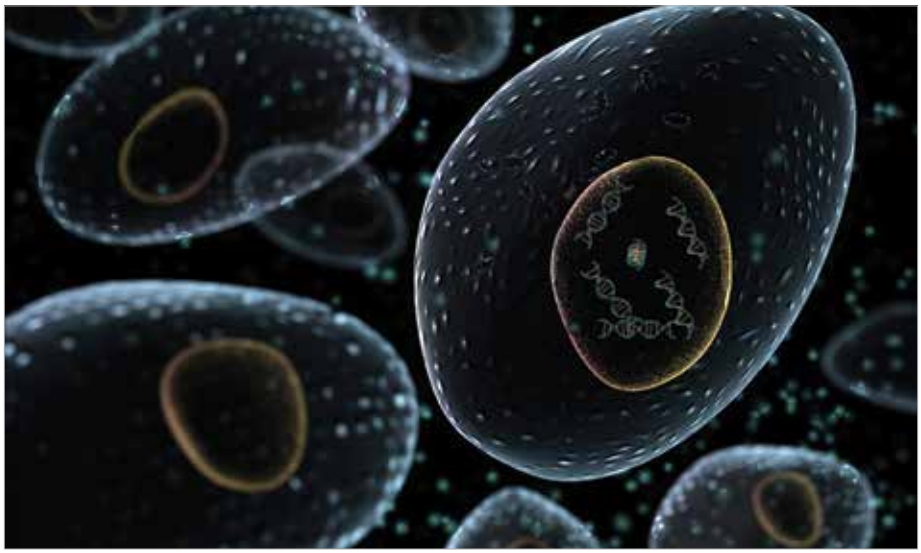
Known collectively as next generation sequencing, or NGS, technologies, a variety of state-of-the-art platforms on the market today offer a massively parallel approach. By processing millions of DNA fragments simultaneously, users of NGS platforms can sequence entire human genomes in just a few days for 1/50,000th the price of former methods. Consequently, genome sequencing is finding more widespread applications in medicine.

In clinical drug trials, for instance, collecting and sequencing participants' DNA samples is becoming routine. With the ability to make genetic comparisons between groups of patients who respond well to an experimental drug and those who do not, researchers can now drill down to the precise genetic variation that is best served by a drug's mechanism of action. Pharmaceutical companies can thereby execute more successful clinical trials and bring to market faster products that successfully treat specific genetic segments of the population.

Particularly in studies of cancer therapies, genomic data collection has already become standard. Because cancer is known to be a result of varied mutations and combinations of mutations, building a large repository of genomic data collected from a critical mass of patients is crucial to developing an understanding of cancer types and their response to treatments.

NGS is also transforming the study of Mendelian disorders, or diseases inherited via single-gene defects. Family-wide genomic studies would have been cost-prohibitive and too time consuming to be of value a decade ago. Now, to get to the bottom of what is causing a mysterious pediatric disease, for instance, researchers can sequence the sick child's DNA, as well as that of his healthy siblings and both his parents, to identify disease-associated markers and define the genetic mechanism of the disease. In fact, NGS has already enabled researchers to discover the mutations that cause more than 40 rare diseases.

In the not too distant future, NGS will increasingly be used in clinical care, too. The day is coming when an individual's genomic sequence will be considered in evaluating disease risk, making dietary recommendations and offering lifestyle guidance.



ANALYSIS IS THE LIMITING FACTOR

Cost and time were once impediments to sequencing DNA, but today the limiting factor to deriving medically useful information from DNA is effective data analysis. With massive amounts of data generated by every genomic study, culling for specific clues becomes an overwhelming undertaking.

In cancer research, for instance, zeroing in on the causal mutations that might be targeted by therapies involves examining huge numbers of mutations that appear to be associated with cancer phenotypes. In drug studies, hundreds, thousands or even millions of genetic variations can be detected among participants in just one clinical trial. Elsevier Life Science Solutions Research Director Nikolai Daraselia explains: "If you have 100 people who responded well to an experimental drug and 100 who did not, and 100,000 genetic variations are detected in each genome sequenced, it is an enormous challenge to compare those genomes and find the mutations that are responsible for the difference in drug response."

Because most variations have no relevance to drug response, NGS data analysis relies on a series of filters to discard mutations that are extremely common, that have low-quality data, that do not pertain to the disease being investigated or that do not modify protein function or binding.

This narrowing-down process relies on a variety of sources of information. Consulting accumulated public knowledge about the frequency of certain variations in the human population or the functions of certain genes is a useful first step in such a workflow: common variations with no relevance to disease can be filtered out and genetic variations widely known to be implicated in cancer cell growth can be retained for further examination.

But to tap the full scope of the most current and potentially useful knowledge on gene function, protein function, cellular pathways or the interactions of multiple regulatory networks, researchers need to comb through dozens if not hundreds or even thousands of experimental results reported in the scientific literature to find facts relevant to their specific research.

FINDING MEANING FROM THE SCIENTIFIC LITERATURE

For that work, many labs still rely on human curators. People, whether scientists with 20 years' experience, or graduate students paid on a per-article basis, are still considered the most accurate arbiters when it comes to finding relevant clues in scientific publications. Indeed, for some types of information, particularly where conclusions, assumptions or summaries are needed, human curation is still the gold standard for finding what researchers need.

Instead of combing through hundreds of scientific papers to find precisely relevant research, users select specific data points or types of relationships of interest and let this solution show them the relevant biological processes, along with the handful of supporting papers that deserve deeper attention.

For genomic data analysis, finding mention of specific genes or proteins and their interactions across scientific publications is a more straightforward task. Elsevier's Pathway Studio, utilizing a proprietary natural-language-processing text-mining tool running on a standard PC, can perform those tasks as much as 1,000 times faster than a human curator—at a rate of between 80,000 and 100,000 full-text articles in an hour—and with comparable accuracy to a highly-trained human annotator. This automated, rapid and accurate accumulation of specific information gives Pathway Studio the largest knowledgebase of molecular interaction data.

Elsevier's Pathway Studio employs expertly-curated terminologies and ontologies specifically constructed for biomedical research to search more than 3 million full-text scientific publications and more than 23 million PubMed abstracts, extracting sentences that contain key concepts, such as "gene", "protein," or "drug," and then reviews the sentences looking for relationships like "binds to," or "activates" or "inhibits." Those sentences, as well as biological pathways developed and validated by Elsevier experts, are loaded into the Pathway Studio knowledgebase for researchers to search, view and apply to analyze their experimental data.

Investigators use the process of pathway analysis to filter out reams of data that are not related to the disease or phenotype of interest, as well as to zero in on the proteins, genes, diseases, drugs and other entities that the data-mining process identifies as relevant.

Instead of combing through hundreds of scientific papers to find precisely relevant research, users select specific data points or types of relationships of interest and let this solution show them the relevant biological processes, along with the handful of supporting papers that deserve deeper attention. Researchers are thereby able to avoid information overload, gain understanding of the biological processes, and more rapidly develop and confirm hypotheses.

Scientists can also use Pathway Studio to build and visualize complex disease models—representing the biology of a disease or a drug response—and uncover its cause at the level of molecular interactions. Modeling the complex, redundant, overlapping interaction networks that regulate the disease-driving expression of genes and proteins can advance basic and applied research, and can help researchers communicate complex biological concepts with others, sharing their discoveries.

Whether for unraveling the genetic pathways that drive patient responses in drug trials, or pinpointing the precise genetic variation or mutations responsible for a rare disease, relying on the largest knowledgebase of literature-based relationships provides Pathway Studio users with greater confidence in their findings. And to researchers tasked with finding meaning in the mountains of next generation sequencing data it gives them a sharp competitive edge by pointing them to the most current and most relevant data faster than any other analysis method available.

Discover more at
www.elsevier.com/pathway-studio