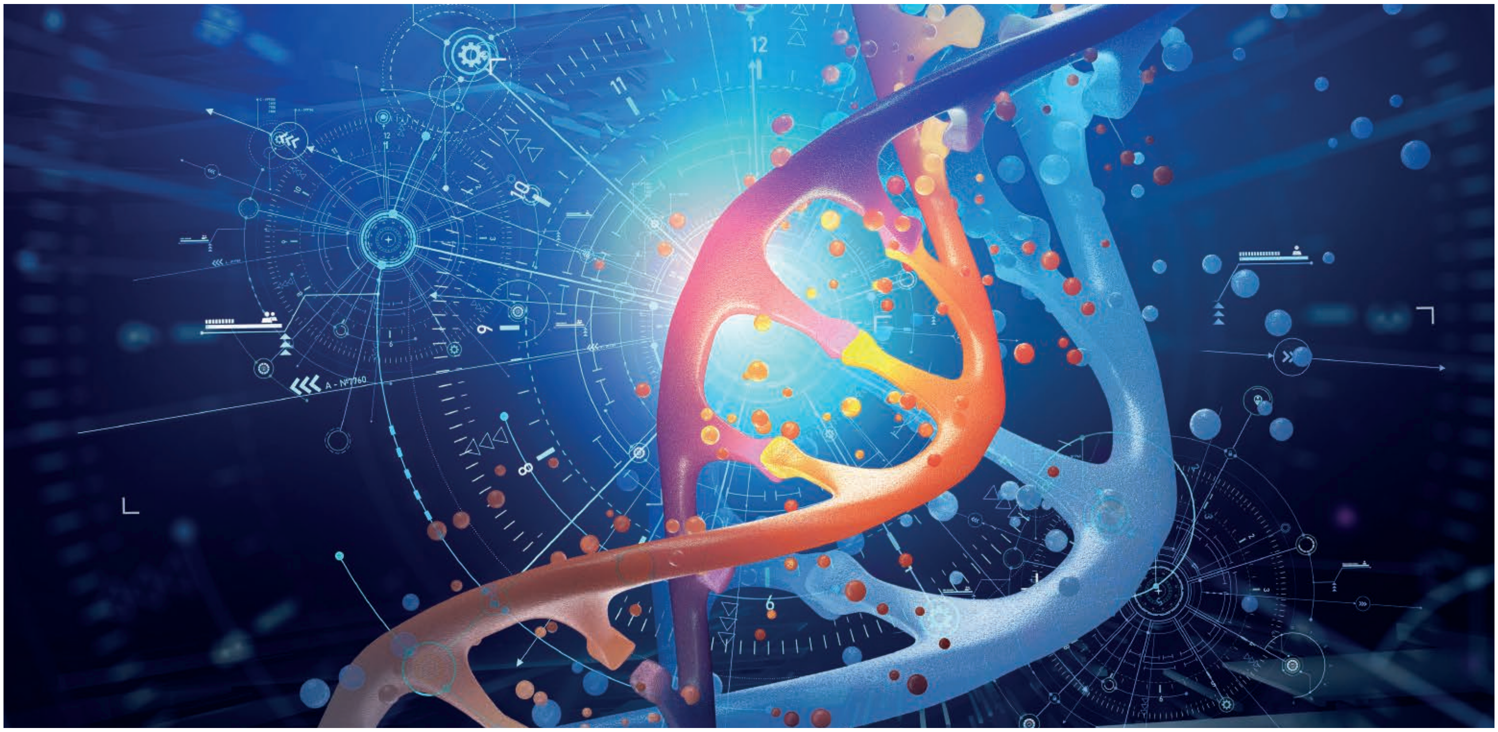

Imputation and critical variants: Fast becoming essential components for predictive genomics

By Jeanette Schmidt, Head of Bioinformatics at Thermo Fisher Scientific

Realizing the full potential of predictive genomics to transform healthcare

Bolstered by foundational research and powerful analytical tools, predictive genomics carries the potential to transform healthcare and improve health outcomes. An example of the potential of predictive genomics is its ability to leverage population data to reveal an individual's risk for developing diseases and help guide preventive lifestyle choices and medical decisions such as earlier screening.

Predictive genomics can also help when a healthcare provider advises a patient as to how the patient may respond to particular drugs. Informed with such advice, the patient can avoid drugs that may cause an adverse reaction and choose options that may elicit a better response and is consistent with the patient's lifestyle. In addition, predictive genomics can help people determine potential health risks for future children through carrier screening to determine if parents carry any genetic mutations linked to inherited health conditions. In addition to provider-patient relationships, public health advisors can use genomic data from these studies to gain insight into population-level trends that can inform policies and help improve health outcomes at scale – ultimately helping to reduce healthcare spend.

Predictive genomics can also deepen our understanding of the genetics of large populations. Today's genomic tools are reliable and scalable to enable extensive global studies, such as genome-wide association studies (GWAS) with sample sizes up to and beyond a million participants.[1] One challenge that is increasingly coming to focus is the lack of representation of various global populations in genomic databases; more than 95% of available genomic research is on people of European descent.[2] Therefore, for predictive genomics to realize its full potential to transform healthcare, genetic datasets must be more inclusive and diverse.

We need accessible technology to make this possible. Having these technologies and large, diverse data sets will also allow imputing data through valid inferences in the populations under study, thereby further extending the reach of predictive genomics. Such imputed connections could not have been possible without the confluence of these technologies to create large-scale datasets validated through predictive genomics that result in a virtuous cycle of data generation, predictions testing, and technology extensions. In this paper, we tie these three technologies together in a coherent, consistent story.

Enabling technology platforms make their mark

Affymetrix, which was acquired by Thermo Fisher Scientific in 2016, was the first company to introduce a commercial microarray system in the early 1990s and the first genotyping arrays a few years later. As a result of their pioneering efforts, profiling technologies have become a valuable and broadly used gateway for genomic research. Today's robust versions of microarray technology can be customized to expand the reach of genetic research to more populations, making genomic databases more inclusive, diverse, and accessible.

As robust microarray and informatics technology platforms have evolved to deliver trustworthy data, they have grown in utility as a powerful tool for making large population studies more affordable and accessible. In turn, researchers around the world are adopting this technology to expand the reach of large-scale genomic studies to populations that have been historically underrepresented. We have gained greater insights as the technology continued to advance, on one hand enabling deep analysis of critical and rare variants and on the other hand improving imputation, the process of inferring unobserved genotypes in a sample of individuals. Imputation has become a key step prior to a GWAS for genomic prediction.

Imputation and critical variants

As head of bioinformatics at Thermo Fisher, I lead a team that specializes in working with genomic scientists and informaticists to design custom arrays to power their research for novel and better-annotated arrays. Our work includes discovering new associations, confirming and refining previous discoveries, or investigating known important variants.

To start, we work with the biology research team to determine the critical variants to include on the array, e.g., SNPs, insertions, deletions, and duplications of DNA segments of interest. We then collaborate with independent, principal investigators to further optimize the array for their specific research goals. We tile the targeted variants on the array in order of importance so we can prioritize the most critical variants to help ensure they are well represented with strategic placement and backup probes.

Once the critical variants of interest are fully covered, we work with the research team to select strategic variants that enable the team to later impute to many millions of variants *not specifically designed* on the array for their populations of interest. Imputation helps researchers get a more complete picture by statistically inferring, or predicting, additional genotypes in the genome.

Imputation works because the genome is not a random collection of genotypes. Our common ancestries share haplotypes (groups of genomic variants) that tend to be passed down from generation to generation. Using existing reference panels from relevant populations of common ancestry we can infer the genotypes that were not directly assayed. While a microarray may directly assay up to a million markers, imputation enables researchers to expand their view to include tens of millions more variants. This expansion through valid inferences has thereby extended even further the potential of predictive genomics.

Genomic profiling: The evolution of array and bioinformatics technology platforms

The bioinformatics community has made incredible progress in developing and improving the tools for genomic research. We started with simple algorithms to select the markers tiled onto the array for the purpose of imputation, but more than a decade ago, my team worked with one of our academic research partners and a large healthcare system in California to pioneer a new approach to optimize array coverage [see **Inset 1: Pioneering A New Approach to Optimize Array Coverage**]. We have continued to refine the method since then, and today our team uses a proprietary algorithm optimized for choosing markers to design an imputation grid for the best outcome.

We are continuously investing in our array capabilities, quality, and fidelity. When we collaborated on this (then) new method in 2011, we used one reference panel for each population, and selected markers for each population separately. Since then, reference panels have evolved significantly and so have our algorithms. Now that we have optimized the methods, we can look at multiple populations together and see better results. Plus, arrays and imputation grids can be customized so researchers can provide their own sequencing data to serve as the reference panel and optimize it for their specific niche.

Deciding when to use arrays for genetic research

Several benefits justify the use of an array; perhaps most importantly, a genotyping panel designed to study variants of interest is more cost-effective than whole genome sequencing when researchers do not need access to the whole genome.

Inset 1

Pioneering A New Approach to Optimize Array Coverage

More than a decade ago, Thermo Fisher worked with a large healthcare system in California and an academic research hospital to collaborate on a genome-wide association study (GWAS) with 100,000 participants. They planned to include large multiracial and multiethnic cohorts with greater genetic diversity than populations studied in previous research.

While earlier GWAS started from any fixed set of SNPs and first used conventional SNP tagging and later imputation to capture and map common variations, to unleash the full power of imputation the team needed to design a more efficient way to select SNPs during the array design process.

The team did this by developing an algorithm to select high-quality SNPs that maximize imputation coverage. The approach would also allow more room on the array to tile ultra-rare SNPs that cannot be assayed by imputation. With this novel approach the team was able to impute more information from fewer SNPs to provide complete and redundant coverage of regions known to be associated with diseases, traits and outcomes.

The imputation-aware array designs yielded a significant overall increase in statistical power.[7] Using this new method, the team designed four high-density Axiom arrays for studies of European, African American, Latino and East Asian populations.

Read more: "Biobanks and Beyond: Genotyping for the Future."

https://www.thermofisher.com/blog/behindthebench/biobanks-and-beyond-genotyping-for-the-future/

Inset 2

Taiwan Precision Medicine Initiative

The Taiwan Precision Medicine Initiative (TPMI) partnered with Thermo Fisher Scientific to design a custom genotyping research array that contains more than 700,000 genetic markers to capture variants associated with disease risk or drug reactions.

TPMI is a joint effort across 16 partner hospital systems across Taiwan. The main goal is to collect genetic profiles and clinical data from a large Taiwanese cohort to establish a sustainable precision health ecosystem in Taiwan, including future guidelines for managing patients with early disease screening, effective treatment, and prevention.

In the project's first phase, TPMI collaborated with Thermo Fisher to design a custom genotyping research array for identifying genetic factors associated with the risk of certain diseases. In the next phase, TPMI will work with Thermo Fisher to optimize the Axiom genotyping platform with the aim of translating the results to clinical use and expanding the development of disease risk prediction algorithms for at least 20 common diseases relevant to Han Chinese people, approximately 18% of the global population.

More than 500,000 people have enrolled in the study, which is halfway to TPMI's goal of one million participants. It is the largest predictive genomics project outside the United States and Europe.

To put it in context: one could genotype 1,000 or even 10,000 people for the same budget that would be needed to sequence 100 individual whole genomes. With the lower price per sample, researchers can investigate larger populations typically needed when researchers are looking for predictive genetic markers. Making genetic research cost-effective and scalable helps break down a critical barrier to expanding research to underrepresented populations.

In addition, focusing on data of interest can be a potential benefit by saving the need to sort through additional information that comes with sequencing the entire genome. The issue is not necessarily parsing through the superfluous data, as such advanced algorithms can handle the analysis; rather, the concern is having to report on information that may not be pertinent to the study. Arrays can be customized and designed to arm researchers with a more manageable dataset that tells them precisely the information they need in a timely manner at a manageable cost.

Why predictive genomics now?

A decade ago, genomics leaders knew the potential for predictive genomics, but the healthcare community was not ready to put this knowledge into practice. Researchers understood how individual genotypes played a role in pharmacogenomics to some extent, but not in large numbers. The UK Biobank made huge strides toward helping the clinical and scientific communities appreciate the potential of predictive genomics more fully.[4] Suddenly, researchers had access to genetic and medical information for hundreds of thousands of people, enabling large-scale studies to understand the impact of genomics with statistical significance.

Fortunately, we are seeing more biobanks now. For instance, the Taiwan Precision Medicine Initiative has enrolled more than 500,000 participants and is the largest predictive genomics project outside of the United States and Europe[5] [see **Inset 2: Taiwan Precision Medicine Initiative**].

Today's research will improve health tomorrow

Array technology has been available for over two decades, such that arrays have become a robust and trusted technology and a cost-effective, scalable solution for genomic testing, including large-scale studies. Furthermore, advances in bioinformatics and related algorithms are consistently getting better at connecting the dots between biomarkers and diseases to predict health outcomes.

As research matures in the coming years, insights gained from today's studies are expected to have a great impact on future health outcomes. By arming patients and clinicians with information that could be used to prevent health issues, predictive genomics could help keep people healthy longer.

Already, countries and health systems around the world are starting to look at how insights from predictive genomics studies can be used to improve how we take medications and guide when we get screened for disease. While early research primarily focused on Caucasian populations, customizable, scalable microarray technology is now enabling more inclusive and diverse studies that aim to expand the reach of predictive genomics globally.[6]

Dr. Jeanette Schmidt

Jeanette is head of bioinformatics at Thermo Fisher Scientific. She earned her Ph.D. in applied mathematics and computer science from the Weizmann Institute of Science and has worked in bioinformatics for more than 30 years. Her background was initially computational, and she dove into biological applications while in a faculty role at Polytechnic University in New York (now part of NYU). She spent six years at Stanford, joined Affymetrix in 2010, and has led Thermo Fisher's bioinformatics team since then. Dr. Schmidt's work focuses on how to provide the best research tools and most accurate genotypes to Thermo Fisher customers to assist them in their research questions, which range from understanding disease to understanding and managing choice and dosage of medication based on genetic profile.

Summary Points

- Genomic data from genome-wide association studies (GWAS) can provide insights into population-level trends to inform policies and help improve health outcomes at scale – ultimately helping to reduce healthcare spending.
- For predictive genomics to transform healthcare fully, genetic datasets must be more inclusive and diverse. We need accessible technology like array and bioinformatics platforms side-by-side with whole genome sequencing to make this possible.
- Since robust microarray technology was first introduced in the 1990s, it has become a valuable and broadly used tool for genomic research. During that time, Thermo Fisher Scientific has continued to invest in and advance the technology by coupling deep analysis of critical variants with imputation to provide greater insights.
- As microarray technology has evolved, it has gained utility as a powerful tool for making large population studies more affordable and accessible. In turn, researchers around the world are adopting this technology to expand the reach of large-scale genomic studies to populations that have been historically underrepresented.
- While early research primarily focused on Caucasian populations, customizable, scalable microarray technology is now enabling more inclusive and diverse studies that aim to expand the reach of predictive genomics globally.

References

1. Uffelmann, E., Huang, Q.Q., Munung, N.S. et al. 2021. Nat Rev Methods Primers. Genome-wide association studies.
2. GWAS Diversity Monitor.
3. Hoffman, T.J., Kvale, M.N., Hesselson, S.E., et al. 2011. Genomics. Next generation genome-wide association tool: Design and coverage of a high-throughput European optimized SNP array.
4. Conroy, M., Sellors, J., Effingham, M. et al. 2019. J Intern Med. The advantages of UK Biobank's open-access strategy for health research.
5. Formosa TV English News. Taiwan Precision Medicine Initiative exceeds 500,000 participants [Video file]. (2022, August 18). Retrieved from http://youtube.com/
6. Fatumo, S., Chikowore, T., Choudhury, A. et al. 2022. Nature Medicine. A roadmap to increase diversity in genomic studies.
7. Hoffman, T.J., Kvale, M.N., Hesselson, S.E., et al. 2011. Genomics. Next generation genome-wide association tool: Design and coverage of a high-throughput European optimized SNP array.