Human phenotyping on a population scale

Vivien Marx

Large-scale phenotyping is generating much data that geneticists can harness. Amid the excitement about the possibilities, there are some points of caution.
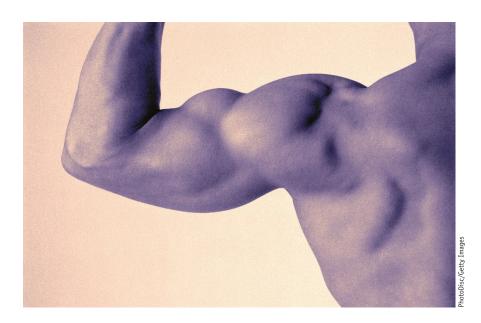
The whirlwind of data collection in human genomics has a dynamic new partner: the collection of human phenotypic data on a massive scale. These data—on eating behavior, anatomical differences or biochemical markers, for example—will be important for interpreting the troves of genetic information. Population-scale phenotyping can power genetic association studies and contribute to deciphering the genetic basis of human biology and disease. It can also help define what is healthy and ferret out cryptic or latent disease.

The data flood will have many sources. Some projects leverage mobile phones to amass data from large numbers—potentially millions—of people. The hope is to vastly increase participation in research and to expand data collection beyond what is collected during physician visits. Project initiators seek an interactive relationship with participants, who choose which data they will share when, and who directly receive information they can discuss with their physicians.

Critics assert that large-scale phenotyping projects raise privacy concerns, have a problematic lack of control groups and have the potential to overtreat asymptomatic people[1]. Proponents say that participant data and records can be protected and that these initiatives will support preventive healthcare and deeply inform genetics research[2].

New apps, new data
The Google Baseline Study is a phenotyping venture to define a health baseline in people. Along with scientists at Duke and Stanford Universities' Schools of Medicine, the organizers will analyze DNA from up to 10,000 people, perform biochemical tests and track other health-related aspects such as eating habits.

Exercise regimes, eating habits and other lifestyle and health-related data are being captured in large-scale initiatives. Geneticists can look forward to a new data tsunami.

The project's pilot has begun with 175 paid participants who undergo a physical exam as well as blood, urine and saliva-based tests and who are asked about their medical history and exercise habits. After assessing these first results, organizers plan to scale up the project to include genome sequencing and cell-based assays, such as immune cell profiling.

The Apple ResearchKit software helps researchers tap into the iPhone user base, letting them develop apps to collect data from the healthy and from people with disease. Once an app is downloaded, the scientists can build their groups of participants.

One such app-based study is the Mobile Parkinson Observatory for Worldwide, Evidence-based Research, mPower, sponsored by the nonprofit organization Sage

Bionetworks, directed by Stephen Friend, and which also co-developed the app. Scientists and clinicians at five universities and at Sage Bionetworks, will be analyzing the data from people with various stages of Parkinson's disease.

After they consent to join the mPower study, participants receive questionnaires on their iPhones about their health, diet and exercise habits. They are asked to tap on their phone a certain way as part of manual dexterity exercises. Participants make choices about if and how they want personal data to be shared; their data will be stored in Sage Bionetworks' platform, Synapse.

There are also Android-based apps. And other phenotype-oriented data-capture projects are operated through foundations; the Michael J. Fox Foundation, for

instance, has launched a virtual, web-based clinical study called Fox Insight, which will be linked to mPower later this year.

Other large-scale projects include both phenotypic data collection and genome sequencing from the start. The 100K Wellness Project—launched by Leroy Hood, president of the Institute for Systems Biology (ISB), and his team—just completed a ten-month study of 100 individuals. Besides sequencing these participants' whole genomes, the team conducted blood, saliva and stool analysis. Information about sleep patterns and heart rate was captured through activity trackers such as the Fitbit bracelet. The data analysis takes place at ISB, and project organizers are planning to scale up the project to profile 100,000 individuals. ISB has also spun out a company called Arivale that will be involved in this project.

The US National Institutes of Health (NIH) is creating a cohort of ultimately a million people as part of the Precision Medicine Initiative announced by US President Barack Obama earlier this year. The cohort will be composed of volunteers willing to share medical and lifestyle information; their DNA, RNA and proteins will be profiled from tissue samples, all with a view to participant privacy. The details of this audacious effort, NIH director Francis Collins said in a statement, are still unfolding, but the results from studying such as large group "will build the scientific evidence necessary for moving precision medicine from concept to reality."

The Harvard Personal Genome Project (PGP) has recruited over 4,000 volunteers willing to openly share genomic and health data. Harvard University researcher and PGP founder George Church says that PGP has been working with Google since 2007 and that the teams are discussing how to combine PGP and Baseline cohorts.

PGP lets participants capture data around the clock, and new data types can be added to the various 'omics and imaging data already in the database, says Church. As the world's only cohort with fully open-access data sharing, the PGP cohort is especially attractive for research projects, he says, and, "it would make great sense if the NIH 1 million cohort included a major subset from PGP."

One of the most important benefits of these phenotyping projects comes from the expanded genomic data sets on healthy individuals, says Erin Cox, deputy director of the Institute for Genomic Medicine at Columbia University Medical Center. The data allow for better inferences when scientists study

disease genomes or genomes from individuals with unique phenotypes. Analyses with sample sizes this large and with carefully collected phenotypic data in healthy people profiled so comprehensively have never before been possible, she says. It will be more within researchers' grasp to do multidimensional analyses that look for associations between multiple coexisting polymorphisms and constellations of phenotypes, says Cox.

At the same time, it is unclear whether these sample sizes will be enough in all cases. As Cox explains, "we do not have sufficient information yet about the genetic architectures of most traits to know for sure what the outcomes will be, but these studies are certainly a first step towards resolving those questions."

Small surprises
By connecting genotypic and phenotypic information, scientists can find clusters of unexpected manifestations of certain genomic variants, says Teri Manolio, who directs the Division of Genomic Medicine at the NIH National Human Genome Research Institute (NHGRI). One study, with which Manolio was not involved, delivered this kind of surprise in the form of subtle disease-related phenotypes that are not readily detected without genetic information.

In June, Leslie Biesecker, who directs NHGRI's Medical Genomics and Metabolic Genetics Branch, and his team published results relating to the sequenced genomes of nearly 1,000 people[3]. The team sifted through more than 100,000 genetic variants. They focused on 100 people with rare variants associated with disease and were able to reach 79 of them for a thorough clinical checkup. Of these individuals, 34 had what turned out to be conditions—detectable phenotypes including heart and lung conditions as well as skin and hearing disorders—that had not been noted previously. This follow-up to the genomic information is what the scientists are calling iterative phenotyping, which involves studying people with disease-causing variants more closely for subtle bodily changes indicative of early disease.

According to the study authors, the results show how such data can play a role in preventive medicine and help to capture the full spectrum of genotype-phenotype correlation. The results also are relevant for population genetics: previous estimates indicated that around 0.2% of the US population may have a genetic condition, but this

study indicates that number might actually be closer to 3%.

One individual taking part in ISB's 100K Wellness Project, who has identified himself in articles, turned out to have elevated homocysteine levels in his blood, which is associated with heart disease[4]. Genome analysis helped to explain this phenotype. His genome contains a variant of the methylenetetrahydrofolate reductase (NAD(P)H) (MTHFR) gene, which can interfere with the body's ability to absorb the B vitamin folate and can lead to a buildup of homocysteine. He was able to address this issue by adjusting his vitamin intake.

For iterative phenotyping, it is useful to also include family members in such ventures to address potential medical issues, says NIH's Manolio. Their inclusion can also help advance research projects.

But as scientists dive through genomic and phenotypic information, they will want to proceed with caution. "Make sure you have participants' consent to look for and report a wide range of phenotypes, not just the phenotypes you originally proposed to study," says Manolio. And, she points out, identifying an association between a genotype and a phenotype does not automatically mean it is causal. "Amazing how often people leap to causal inferences on the slimmest of evidence," she says.

The payoff of large-scale human phenotyping for genetics research will likely be large, says Manolio. These results can expand the understanding of the spectrum of disease-causing mutations. And, she says, they can help to interpret the thousands, if not millions, of variants of unknown significance that are being identified by exome and genome sequencing.

Speaking the same language
Privacy rules govern the use of electronic health records in research. But even when consent is granted, making genotype-phenotype associations with these data can be difficult. Patients' phenotypic traits tend to be described in free text, making them not readily

Phenotyping helps scientists study complex disease, says Helen Parkinson. They can then compare phenotypes shared between common and rare diseases.

EMBL-EB

In medical records, patients' phenotypic traits tend to be described in free text, says Michael Brudno. There are ways to make these descriptions computable.

University of Toronto

computable, says Michael Brudno, a computer scientist at the University of Toronto; he also directs projects involving phenotypic analysis of individuals with rare disorders at The Hospital for Sick Children. Ontologies can help, such as the Human Phenotype Ontology, which is a standardized vocabulary with over 11,000 entries that describes all defects related to human disease conditions. But, says Brudno, health records often do not use ontologies to describe a patient's condition.

For example, if clinician-scientists want to study neurodevelopmental disorders, they might encounter records with the abbreviation MR, indicating the diagnosis of mental retardation. That term has been replaced with the expression 'intellectual disability'. Or, a physician might write that a child spoke his or her first words at age five instead of writing that the child has expressive language delay.

To help make such phenotypic descriptions computable, Brudno and his group have built PhenoTips, software that helps to standardize data about pediatric patients. The main goal, he says, is to help researchers find and group patients. Then researchers can search for genomic data on these individuals and explore the genes shared by patients with similar phenotypes.

PhenoTips is used in labs and hospitals and by research ventures around the world, says Brudno. One public-private project using PhenoTips is Neuromics, which includes efforts to connect phenotypic and genotypic data about individuals with neuromuscular and neurodegenerative diseases.

The Human Phenotype Ontology is built into PhenoTips, and the software draws on resources of the Monarch Initiative, an organization that supports computational approaches for cross-species phenotype analysis. One such approach is Exomiser, software with which to compare mouse and human phenotypic data[5]. Tools are one part of the equation; collaboration between researchers is another.

As these large-scale phenotyping projects unfold, Helen Parkinson at the European Bioinformatics Institute (EBI) sees an important role for the Global Alliance for Genomics and Health (GA4GH). This organization draws together scientists in academia and companies to develop standards to enable sharing of genomic and clinical data in a reproducible and robust way. In the alliance's working groups, researchers clarify and compare the particulars of data capture in different fields such as cancer, rare diseases or nutrition. This conversation is needed because it is not straightforward to standardize questions about, for example, the eating habits of people or to make this information computable, says Parkinson. The data are more comparable if the same ontology terms are associated with a given question and the results.

Parkinson works on a number of phenotyping and genomics projects with human data, such as the NHGRI-EBI genome-wide association study (GWAS) catalog, and does tool development and data analysis also with data on mice, stem cells and plants[6]. The larger the phenotypic data collection, the greater the need to make these data computable and the more necessary ontologies become, says Parkinson.

'Data wranglers' are research staff who help to facilitate computational genotype-phenotype analysis as part of the Mouse Phenotyping Informatics Infrastructure consortium, which is an NIH-funded effort to manage the large data volumes of the International Mouse Phenotyping Consortium (IPMC) and which includes the EBI, Medical Research Council Harwell and the Wellcome Trust Sanger Institute. Data wranglers understand a given research domain and spot errors or inconsistencies in data collections. In a longitudinal study, an assay or an instrument might be modified, and the data description associated with the measurement must reflect that shift. People are better than computers at detecting such changes, says Parkinson, but she and her team are working on how to teach computers to do so.

Lessons from the mouse

Parkinson sees plenty of lessons for the new human phenotyping ventures from mouse phenotyping projects—for example, from the large-scale IMPC. IMPC scientists at 18 research institutions around the world produce knockout mice for 20,000 mouse genes and also profile the mice phenotypes. The consortium has five national funders and also two corporate sponsors, Charles River Laboratories and Taconic Biosciences.

One of Parkinson's projects involves linking human GWAS data to IMPC data to help find new mouse models of disease. One goal is to facilitate the search for mouse phenotypes that might be relevant for an experiment based on a human phenotype of interest.

There are a few caveats to address with such cross-species data comparisons, such as finding comparable measurement approaches. Mice can be studied in ways that humans cannot, such as with calorimetry. "You can put them in a jar and measure everything that they eat, everything that they pee out, all the oxygen they consume," says Parkinson, referring to a metabolic cage. It would be unethical to study humans in this fashion. But there are controlled tests to capture human metabolic data, such as a gestational diabetes test that measures how fast a person's blood sugar level drops after imbibing a sugary drink.

Parkinson and her colleagues have found terms in mouse data annotations that do not have a corresponding term in the Human Phenotype Ontology. To address this conundrum she, along with scientists at Medical Research Council Harwell and the Sanger Institute, has been applying the PhenoDigm (phenotype comparisons for disease genes and models) software[7]. In the absence of an exact or lexical match, the software can predict orthogonal matches. For example, she says, mice do not speak, so one cannot readily find parallels for human speech disorders. Scientists can use PhenoDigm in these instances. For example, it will map the human difficulty of articulating speech to abnormalities of the larynx in mice.

Images/HemeraThinkstock

The new human phenotyping ventures can draw on lessons learned from mouse phenotyping projects, such as the International Mouse Phenotyping Consortium (IMPC).

Geneticists know to be cautious when using such predicted matches, but having it on hand helps to narrow a large pool of genetic variants down to a smaller group as scientists hunt for clinically relevant phenotype-genotype connections.

Phenotyping helps scientists work on complex disease. They can start with rare-disease phenotypic data. When looking at phenotypes shared between common and rare diseases, says Parkinson, they can use what they learn about rare disease to explore the mechanism of action in the common disease.

Some researchers use the shorthand 'knockout human' to refer to a person with a completely inactive, and often rare, form of a gene. Using genotype and phenotype data about these individuals, researchers can devise a model organism–based gene-editing experiment. And then, says Parkinson, they might be able to find phenotypes that correspond to these rare genetic conditions and perhaps also choose a humanized mouse as the basis for additional research.

Statistics caveats

Big phenotypic data sets have great prospects, but they raise big challenges, too, says Yoav Benjamini, a statistician at Tel Aviv University. Benjamini co-developed the concept of false discovery rate, which showed how to address the pitfalls of drawing conclusions when many variables are measured[8].

Even in well-controlled experiments involving mouse behavioral phenotyping with two strains of mice in one lab, the results might not be readily replicated in another lab even when measuring the

Big phenotypic data sets are promising, and they raise big challenges, says Yoav Benjamini.

Tel Aviv University

same phenotype with the same strains, says Benjamini. This experience highlights issues to be expected with human phenotype databases, with measurements that are vaguely defined or that are based on different measurement techniques.

In well-designed human trials, says Benjamini, there are documented measurement errors when recording eating habits, for example. In the new phenotyping projects, there will be plenty more variation due to factors such as age or living conditions. "We have learned that standardization of phenotype definitions is not enough," he says. Methods are needed to address these statistical challenges. Although the knowledge for such methods exists, he is concerned there may be a lack of awareness in the research community, particularly among casual users of phenotyping data. He is working with IPMC biostatisticians and says the consortium is aware of the issues.

Another challenge with population-scale phenotyping is connected to the implications of searching across many phenotypes and selecting only the few promising ones, says Benjamini. In that case, the burden of proof is higher than when looking at many phenotypes connected to one particular genomic location. "The increased burden is essential in order to limit the chance for the discoveries to be false," he says.

The proposed data sets will have many more layers of multiplicity: there will be multiple genomic aspects such as gene expression, single-nucleotide polymorphisms and methylation. And there will be multiple phenotype subsets—for example: male, 18–20 years old, student, with a certain grade point average range, who grew up in a small town, whose parents both work from home, whose body mass index is above 20, and other traits.

Unlike typical genome-wide association scans, a database open to scientists and that allows repeated querying, possibly even automatically by a 'bot', adds a unique dimension of multiplicity, says Benjamini. "In view of that, how can we take care that the 'extremely interesting and highly significant' results mined from the database stand the scrutiny of replication and avoid flooding the scientific literature, and the public in general, with false discoveries?" He and colleagues at the Technion, Tel Aviv University and Stanford University have been looking at statistical methods to address two layers of multiplicity, in which a genome-wide search for associations is conducted over multiple phenotypes. "Addressing more layers of complexity is the next challenge statisticians should tackle," he says, adding that the basic framework for this selective inference exists, an approach he and others are working on.

For the conclusions derived from these large-scale phenotyping projects to be valid, "the open database, in spite of its name, has to be monitored and even actively managed," says Benjamini. "No bots-generated discoveries, please."

1. Diamandis, E.P. *BMC Med.* **13**, 5 (2015).
2. Hood, L., Lovejoy, J.C. & Price, N.D. *BMC Med.* **13**, 4 (2015).
3. Johnston, J.J. *et al. Am. J. Hum. Genet.* **96**, 913–925 (2015).
4. Bleicher, A. 100,000 people, 250 biomarkers, and the quest for good health. *IEEE Spectr.* (28 May 2015).
5. Robinson, P.N. *et al. Genome Res.* **24**, 340–348 (2014).
6. Deans, A.R. *et al. PLoS Biol.* **13**, e1002033 (2015).
7. Smedley, D. *et al. Database* 2013, bat025 (2013).
8. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

Vivien Marx is technology editor for *Nature* and *Nature Methods* (v.marx@us.nature.com).