Deep phenotyping predicts Huntington's genotype

Douglas M Ruderfer & Joel T Dudley

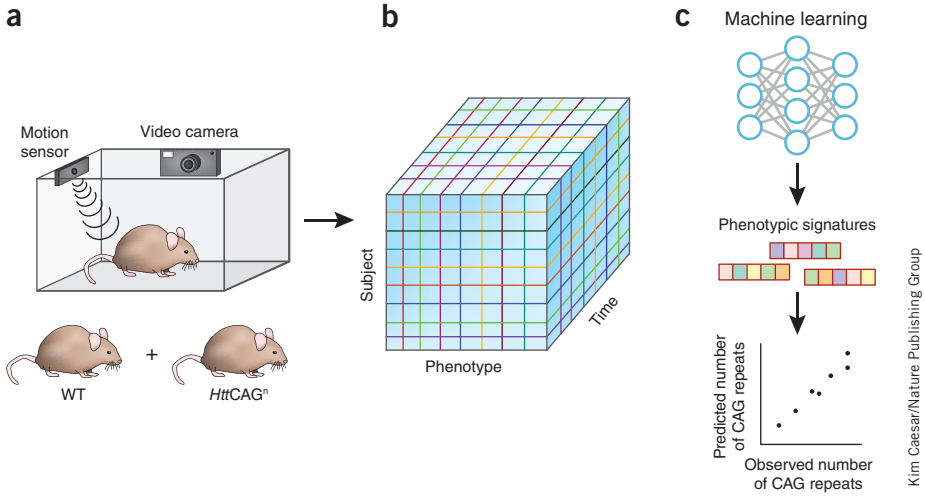
Comprehensive phenotyping of Huntington's disease in mice yields phenotypic signatures of disease-causing genetic variants.

Most genomics studies are designed to scan across the entire genome for genetic variation that underlies a single, narrowly defined phenotype. Such analyses rely on defining phenotypes a priori in a hypothesis-driven manner, which precludes systematic discovery of more complex relationships between phenotype and genotype. In contrast, the inverse approach—efficient phenome-wide discovery of complex phenotypic traits and their relationships to genetic variation—has rarely been achieved, owing to a lack of studies that measure both high-quality, high-dimensional phenotypic data and comprehensive genetic data from large numbers of individuals. In this issue, Alexandrov *et al.*[1] show how 'deep' phenotyping of mice with different Huntington's genetic variants allows the disease genotype of individual animals to be predicted from phenotypic measurements alone. Their results lend support to efforts to mine high-dimensional phenotypic data from humans, obtained from sources such as wearable health sensors, medical records and patient-reported information, with the aim of defining complex phenotypes and their genetic origins.

Throughout the history of medicine, physicians have observed disease-associated human phenotypes in the form of symptoms. Recurrent patterns of severe symptoms were recorded, reported in the literature and eventually labeled as syndromes. With the advent of technologies for genetic analysis, many syndromes were shown to have a genetic basis. For example, the development of karyotyping in the 1950s enabled the discovery that Down syndrome, first characterized as a severe and distinctive form of mental disability in the nineteenth century, is caused by the presence of an extra copy of part or all of chromosome 21. More recently, genomic methods, such as DNA sequencing, have been used to link variants in particular genes to an array of severe syndromes[2] and to subtypes of heterogeneous diseases such as autism[3].

As the amount of genetic information increases, our understanding of the genetic basis of diseases will surely continue to improve. Yet disorders that are characterized by phenotypic heterogeneity and variable expressivity (the extent to which a trait is expressed) will remain difficult to reduce to specific genetic pathologies. Alexandrov *et al.*[1] set out to study this problem using a relatively well-understood genetic disorder. Huntington's disease is caused by a CAG repeat expansion in the huntingtin (*Htt*) gene, and the number of CAG repeats determines the severity and course of the illness[4]. The authors wanted to discover whether increasing both the number of dimensions and the resolution of phenotypic data collected in mice would allow the number of repeats in individual mice to be inferred from fine-grained phenotypic signatures.

First, they generated six heterozygous knock-in mouse lines expressing different numbers of CAG repeats ranging from 20 to 175 (**Fig. 1**). Using a set of high-throughput, proprietary phenotyping devices that objectively quantify a series of behaviors, including motor, social, anxiety-like and gait, through computer vision software, they were able to capture a total of 3,086 distinct behavioral



Kim Caesar/Nature Publishing Group

Figure 1 Strategy to predict complex genotypes from high-dimensional phenotypic data on mouse models of Huntingon's disease. (**a**) Mice bred to have between 20 and 175 CAG repeats in the huntingtin (*Htt*) gene and wild-type mice are monitored in an 'intellicage', which enables multiple phenotypes to be recorded from freely moving animals, including cognitive, motor, circadian, social, anxiety-like and gait phenotypes. Recording is through motion and vision sensors and processing is through dedicated algorithms. (**b**) Deep phenotyping data is obtained and multiple features are analyzed. (**c**) Machine learning approach identifies phenotypic signatures that allow accurate prediction of the number of CAG repeats in an independent mouse sample.

Douglas M. Ruderfer is at the Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA, & Joel T. Dudley is at the Department of Genetics and Genomics Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
e-mail: joel.dudley@mssm.edu

phenotypes. Data analysis using custom-built software and machine-learning algorithms based on support vector machines identified a set of ~200 features that accurately predicted the length of CAG repeats in individual mice. Importantly, no smaller set of phenotypes could account for the predictive accuracy. These findings highlight both the importance of comprehensive (or deep) phenotyping as well as the profound phenotypic complexity that emerges from aberrant *Htt* expression.

The approach taken by Alexandrov *et al.*[1] is impressive for the extent of the phenotyping and the difficulty of working with subtle effects of the particular genetic variants studied. A previous paper analyzed autistic phenotypes in humans with a machine-learning method to predict one of six possible causative genetic disorders[5]. The prediction accuracy obtained was 63% overall, and between 10% to 77% for individual disorders. In that study, however, the genetic causes were entirely different (22q11.2 deletion, Down syndrome, Prader-Willi, supernumerary marker chromosome 15, tuberous sclerosis complex and Klinefelter syndrome), whereas Alexandrov *et al.*[1] distinguished subtle phenotypic effects arising from different numbers of a CAG repeat in a single gene.

There have been several other efforts to identify subgroups of human patients with shared pathology within heterogeneous, complex disorders using broad collections of phenotypic measures (e.g., hundreds of laboratory test measurements and thousands of diagnosis codes from electronic medical records). One recent study[6], in which the authors examined the electronic medical records of nearly 11,000 patients, including 2,251 with type 2 diabetes, enabled better patient stratification. Specifically, Li *et al.*[6] identified three distinct subtypes of type 2 diabetes patients based on similarities and differences among hundreds of biometric and laboratory test values present in the electronic medical records. Importantly, these variables included some not known to be relevant to diabetes. The three identified subtypes were associated with different risks for kidney, eye and cardiovascular complications—information that should enable more effective clinical management of the disease. Furthermore, the subtypes were associated with unique sets of genetic variants, which point toward putative biomarkers or mechanisms for each subtype.

A different strategy, the phenome-wide association study (PheWAS), maps phenotypes, usually derived from electronic medical records, to genetic variants. Unlike the machine-learning approaches of Alexandrov *et al.*[1] and Bruining *et al.*[5], a PheWAS is carried out for each phenotype, one at a time. PheWAS provides another means to infer relationships between targeted genotypes (e.g., loci previously identified by genomic association studies) and clinical phenotypic traits in an unbiased manner[7]. However, it is limited by the relatively sparse representation and biased collection of phenotypes in electronic medical records.

The findings of Alexandrov *et al.*[1] may inspire similar methods to exploit human phenotypic complexity in medicine. The utility of such efforts will be confirmed if they lead to improved understanding of disease biology or more effective treatments. But the difficulties of translating the approach to human populations should not be underestimated. Humans have a much greater diversity of genetic backgrounds and phenotypes than mice, and they live in more variable environments, so capturing the full breadth and depth of the phenome is far more challenging. In addition, developing appropriate high-dimensional phenotyping technologies for data acquisition may be tricky. Although, in principle, comprehensive phenotypic data could be gathered through a combination of wearable health devices (e.g., physiology sensors, activity trackers, sleep trackers), medical records, photos and social media, this assumes both that suitable technologies are available and that study participants are willing and able to use them.

Large-scale digital health studies are already being facilitated through the Apple ResearchKit framework, and they will contribute the main set of data collected from patient cohorts in the 1-million-patient Precision Medicine Initiative being sponsored by the National Institutes of Health. It will not be long before analyses of these data will begin to reveal whether comprehensive phenotypic data collection in humans is feasible, and, more important, whether such data are clinically useful.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nbt.3648).

1. Alexandrov, V. *et al. Nat. Biotechnol.* **34**, 838–844 (2016).
2. Ng, S.B. *et al. Nat. Genet.* **42**, 30–35 (2010).
3. Bernier, R. *et al. Cell* **158**, 263–276 (2014).
4. Langbehn, D.R., Hayden, M.R. & Paulsen, J.S *Am. J. Med. Genet.* **153B**, 397–408 (2010).
5. Bruining, H. *et al. Mol. Autism* **5**, 11 (2014).
6. Li, L. *et al. Sci. Transl. Med.* **7**, 311ra174 (2015).
7. Denny, J.C. *et al. Nat. Biotechnol.* **31**, 1102–1110 (2013).

An edible switch for gene therapy

Xavier M Anguela & Katherine A High

The expression of therapeutic transgenes in mice is made responsive to the amino acid content of the diet.

A central challenge in the field of gene therapy is how to control the expression of a transgene over time. Temporal regulation of therapeutic transgenes has not yet been described in the clinical literature, and the solutions that have been developed in animal models may not have the requisite safety and efficacy for use in patients. In the July issue of *Nature Biotechnology*, Chaveroux *et al.*[1] demonstrate an alternative strategy in mice—a natural signaling pathway, activated by deficiencies in essential amino acids, that is harnessed to regulate transgene expression simply by modifying the amino acid content of the diet. Because the approach does not rely on drugs or transcription-factor ligands, it should avoid some of the concerns that have hindered clinical translation of other regulatable systems.

Technology development in gene therapy has shown success in controlling the spatial expression of transgenes, using cell-type-specific promoters or microRNA targets to restrict expression to specific target cells, such as hepatocytes or myocytes. But regulating the timing and level of transgene expression has proved more difficult. Achieving this is especially important for the treatment of diseases with a narrow therapeutic window—the dose range of a drug that is effective without being toxic. In type 1 diabetes, for example, normal blood sugar control requires exquisite glucose-induced regulation of insulin synthesis and secretion, and attempts at extrapancreatic regulation of insulin expression have so far been inadequate[2].

At the other extreme are diseases with a wide therapeutic window, such as hemophilia. Even for individuals with severe forms of hemophilia, a modest increase in clotting-factor activity markedly improves the clinical phenotype.

Xavier M. Anguela and Katherine A. High are at Spark Therapeutics, Philadelphia, Pennsylvania, USA.
e-mail: Kathy.High@sparktx.com