Molecular Systems Biology 5; Article number 273; doi:10.1038/msb.2009.32
Citation: ***Molecular Systems Biology* 5:273**
© 2009 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/09
www.molecularsystemsbiology.com

molecular
systems
biology

EDITORIAL

Personal phenotypes to go with personal genomes

Molecular Systems Biology **5**: 273; published online 19 May 2009; doi:10.1038/msb.2009.32

With the cost of DNA sequencing decreasing rapidly, it is likely that the genome sequences of many individuals will be determined. In fact, if half of the individuals in industrialized countries choose to have their genomes sequenced, then well over 500 million personal genome sequences will be determined. Currently, such genetic information is likely to be of limited value to the individual, as the number of loci that provide useful predictive information is quite small (probably less than 200). Indeed, recent analyses of common complex traits such as diabetes, body mass and height show that in each case the genetically identifiable contribution from multiple candidate loci (18 in the case of diabetes) is only a small percentage (less than 7%) of the total identifiable genetic load (Gaulton *et al*, 2008; Willer *et al*, 2009); thus, the interpretable genetic contributions that can be identified are quite minor. Presumably, either many low-frequency alleles at different loci contribute to the genetic load or perhaps the many phenotypes are because of other phenomena such as synergistic effects between variants at more than one locus or between different loci and factors in the environment, recurrent spontaneous mutations, or epigenetic defects.

Regardless of which proves to be correct (likely a differing mixture of effects for different diseases), the ability to accurately correlate all bases with precise phenotypes is likely to be powerful only if a common set of phenotypes are scored. The power of 500 million sequences correlated with 500 million phenotypes can show both small contributions as well as help identify potential causative mutations. Indeed, a data set of this size would greatly exceed that of even the large genome-wide association studies that typically analyze thousands of individuals to tens of thousands of individuals (Willer *et al*, 2009). Although in the short term this information is not likely to be helpful for prediction of common diseases, it may provide a generic tool for interpretation and prevention of a large number of individually uncommon or rare recessive disorders. In the long term, it is likely to be of enormous value to scientists for understanding which types of genes and pathways are involved in a particular biological process and for determining the underlying nature of complex diseases. Furthermore, the entire community is expected to ultimately benefit from this information, which would help in the diagnosis and treatment of disease.

The need for phenotypes to go with genotypes

Although the prospects of a large number of genome sequences might seem daunting, the biggest stumbling block for a genotype–phenotype correlation is not likely to be the acquisition of the DNA sequence, but rather the phenotypic information. Indeed, the phenotyping of large numbers of individuals might well prove to be more expensive, complex and difficult to implement than the genomic sequencing. However, without common and accurate phenotypes the power of the genome sequences will be extremely limited.

Table I Examples of data types to consider for collection

General	Anatomical, height, body mass Blood pressure Morphometric Medical History (disease conditions, medical treatment, medication, etc. asthma, infections, cancer, other diseases)
Behavioural & Cognitive	Anxiety, depression, hyperactivity, sleep Cognitive attributes (learning and memory, 'intelligence')
Molecular[a]	RNA expression Proteomics (mass spectrometry; antibody profiling) Metabolomics Microbiome metagenomics

[a] Types of samples to analyze: saliva, plasma, serum, urine, breath, skin (stem cells), feces (microbiology).

Deciding exactly what to phenotype is not trivial. Some types of information, such as height, body mass, blood pressure, and many aspects of medical history (infectious and other diseases, etc) are quite common and obvious. Furthermore, much of this information is already available (albeit not always consistently obtained or available in a useful manner) (Table I). Other types of information, such as anatomical features and skeletal information, could be digitized and converted into useful format for morphometrical analysis. Phenotypes that would be particularly powerful to analyze using large data sets are behavioral (e.g. anxiety, depression) and cognitive attributes (e.g. 'intelligence tests'). Some of these data are likely to be controversial and raise issues regarding safeguarding the privacy of information. Nonetheless, the larger the collection of phenotypes, the more powerful the genetic information. In order to be useful, these phenotypes must be stored electronically and in a manner in which quantitative information can be obtained. In this respect, having all medical records and information stored in a digital format would be extremely valuable for sharing and analyzing data.

Perhaps, even more important than the phenotypes that should be measured are the implementation of common methods and standards for their collection. Phenotypic data are only likely to be useful if the same types of information are obtained, and only if the samples and measurements are obtained using the same methodology. Many parameters such as medical and psychiatric histories and physical examinations are not always collected under comparable conditions or with similar rigor. Although it may be difficult to have a common method used in all cases, ideally a prioritized set of standards could be prepared, and minimally it will be essential to record the types of methods used for each sample.

Molecular omic phenotypes

One way to provide a larger quantity of phenotypic information and potentially in a more standardized format is to shift from measuring macroscopic properties to analyzing molecules. In addition, the quantities of molecules are expected to be responsible for the observable bodily phenotypes, and molecules can be more directly related to the genomic sequence and its variations.

Traditionally, only a limited number of molecular markers are monitored, typically during blood tests. However, it is likely that large-scale and precise measurements of gene expression or protein abundance in specific types of cells are more consistent indicators of a given organism's phenotype. One can accurately quantify the RNA levels of all genes and/or exons using DNA microarrays or RNA sequencing (Wang *et al*, 2009), and the levels of many thousands of proteins and their modifications can be followed using mass spectrometry (Aebersold and Mann, 2003) and ultimately might be quantified using affinity reagents. Hundreds of metabolites can also be monitored using mass spectrometry (e.g. see Sreekumar *et al*, 2009). These components can easily be measured in blood samples, and proteins and metabolites can be measured in urine. Other samples such as saliva and breath could also be possibly measured. In the future, one could even consider the analysis of patient-derived stem cells and microbiome samples from oral and fecal samples. The analysis of microbiome using metagenomic sequencing could prove to be a useful indicator of both environment and phenotype.

Although RNA and metabolites might be relatively straightforward, analysis of proteins in complex samples such as plasma, sera and urine may be particularly susceptible to how the samples are prepared, which can have a significant influence on the outcome. For example, a recent proteome analysis of human sera showed significant differences in outcome depending upon the buffers and inhibitors present in the collection samples (Omenn *et al*, 2005). Nonetheless, robust analytical procedures need to be established to ensure that the results can be reproduced in different laboratories. Furthermore, even if comprehensive monitoring of molecular markers is difficult, accurately quantifying even a large subset is likely to be extremely valuable.

The collection of molecular phenotypes is expected to be extremely valuable for helping us understand the basic mechanisms involved in human disease. For example, activation of signaling pathways can be readily deduced from RNA and protein expression and post-translational modification data. This in turn can be related to the genome sequence. In addition, molecular information will greatly facilitate medical diagnostics. Currently, diagnostic tests are carried out on small numbers of proteins whose functions are usually, although not always, known. One can readily envision a future in which simple blood or urine tests involving profiling of thousands of protein and/or metabolic components will be much more valuable for both early and accurate diagnostics. Control experiments will obviously have to be carried out to account for parameters such as diet and the time of day at which the samples are collected. Nonetheless, such an information is expected to be extremely useful in conjunction with genomic and epigenomic analyses.

Moving forward

Several large consortia have formed around global genome sequencing projects such as the 1000 Genomes Project and The Cancer Atlas Project. Although smaller advisory committees have discussed the collection of common phenotypes (see Church, 2005), a large consortium is needed to decide what common phenotypes and samples should be collected and how would they be of equal impact. Arguably, the best way to accomplish this is in conjunction with organization of the large genome sequencing projects. The effort involved in obtaining a standard set of phenotypes should be no less than that expended in developing a standard set of gene functions through the Gene Ontology consortium.

There is no doubt that a large number of human genome sequences will be a valuable resource. However, it will only be valuable in the context of a large number of accurate phenotypes. With the first sequences now being determined, we need to aggressively develop guidelines for deciding what phenotypes should be collected and establish common standards for collecting those phenotypes.

Michael Snyder[1,2]**, Sherman Weissman**[3] **and Mark Gerstein**[2,4,5]
[1]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA;
[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA;
[3]Department of Genetics, Yale University, New Haven, CT, USA;
[4]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA and
[5]Department of Computer Science, Yale University, New Haven, CT, USA

References

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* **422** (6928)**:** 198–207

Church G (2005) The personal genome project. *Mol Syst Biol* **1:** 2005.0030

Gaulton KJ, Willer CJ, Li Y, Scott LJ, Conneely KN, Jackson AU, Duren WL, Chines PS, Narisu N, Bonnycastle LL, Luo J, Tong M, Sprau AG, Pugh EW, Doheny KF, Valle TT, Abecasis GR, Tuomilehto J, Bergman RN, Collins FS *et al* (2008) Comprehensive association study of type 2 diabetes and related quantitative traits with 222 candidate genes. *Diabetes* **57:** 3136–3144

Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, Kapp EA, Moritz RL, Chan DW, Rai AJ, Admon A, Aebersold R, Eng J, Hancock WS, Hefta SA, Meyer H *et al* (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5:** 3226–3245

Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, Laxman B, Mehra R, Lonigro RJ, Li Y, Nyati MK, Ahsan A, Kalyana-Sundaram S, Han B, Cao X, Byun J, Omenn GS, Ghosh D, Pennathur S, Alexander DC *et al* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457:** 910–914

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63

Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C, Lettre G, Lim N, Lyon HN, McCarroll SA, Papadakis K, Qi L, Randall JC, Roccasecca RM, Sanna S, Scheet P *et al* (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41:** 25–34