

## From Phenotype to Genotype: Experiences in Navigating the Available Information Resources

Joyce A. Mitchell, Ph.D.<sup>a,b</sup>, Alexa T. McCray, Ph.D.<sup>a</sup>, Olivier Bodenreider, M.D., Ph.D.<sup>a</sup>  
<sup>a</sup>National Library of Medicine; <sup>b</sup>University of Missouri

*As part of an investigation of connecting health professionals and the lay public to both the disease and genomic information, we assessed the availability and nature of the data from the Human Genome Project relating to human genetic diseases. We focused on a set of single gene diseases selected from main topics in MEDLINEplus. We used publicly available websites to investigate specific questions about the genes and gene products associated with the diseases. We also investigated questions of knowledge and data representation for the information resources and navigational issues. We assessed the major challenges encountered when navigating from phenotype to genotype.*

Many people want to know, “What data is coming from the Human Genome Project that relates to my disease and my risks for this disease?” This project was undertaken to assess what data are available on common single gene diseases and how accessible are the data to health professionals and the lay public. The work focused on the major information resources containing genome and proteome knowledge and the pathways to navigate them.

An analysis of MEDLINEplus<sup>1</sup>, the NLM’s principal resource focused on consumers, revealed that several specific diseases were main topics in MEDLINEplus and also fulfilled four other criteria: (1) entries in the Online Mendelian Inheritance in Man (OMIM)<sup>2</sup> for a specific disease, (2) entries in Locus-Link<sup>3</sup> for specific gene products for the OMIM disease, (3) disease summary in GeneClinics<sup>4</sup>; and (4) at least three commercial laboratories doing DNA tests as listed in GeneTests<sup>4</sup>. Eleven of these diseases were examined in detail to investigate the potential links between phenotype (disease) and genotype and the current systems that contained the data of interest. The eleven diseases studied are achondroplasia, cystic fibrosis, Duchenne muscular dystrophy, Gaucher disease, Huntington disease, Marfan syndrome, dystrophia myotonica, neurofibromatosis, phenylketonuria, polycystic kidney disease, and tuberous sclerosis.

The specific questions investigated included:

- What gene(s) causes the disease?
- On which chromosomes are the genes located?

- What are the normal function of the gene product(s) and the function of the mutated gene product(s)?
- Are there clinical trials for the disease?
- Where are the laboratories that do DNA tests?
- What mutations have been found in the genes?
- Do the genes cause any other diseases besides the target disease?

All of the data sought for the set of eleven diseases could be found in the systems examined but several resources need to be traversed to find all of the answers. From these eleven target diseases, there were 16 genes, 117 gene names (including synonyms), 39 gene products (including isoforms), and 182 disease names (including synonyms). The list of synonyms for the gene and gene product names is undoubtedly incomplete because there is no single source to collect all of them.

Four major challenges were encountered when navigating from phenotype to genotype. (1) the complexity of data, including the gene names, makes it difficult for naïve users to find information ; (2) the dynamic nature of the data from advances in science contributes to constant changes in both data and systems; (3) the diversity of information resources makes searching difficult and time consuming and requires training, and (4) the lack of routine use of standard data and knowledge representation methods makes information retrieval more complex than if such methods were used.

There is a tremendous amount of data arising from the results of the Human Genome Project. Most of the current information resources related to the HGP are focused on the scientific research audience, but those information resources contain information that could be useful to non-geneticist clinicians and the lay public.

1. <http://medlineplus.gov>
2. <http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=OMIM>
3. <http://www.ncbi.nlm.nih.gov/>
4. <http://www.genetests.org/>