Human disease genes

Gerardo Jimenez-Sanchez*, Barton Childs* & David Valle*†

** Department of Pediatrics, McKusick-Nathans Institute of Genetic Medicine, and † Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA*

The complete human genome sequence will facilitate the identification of all genes that contribute to disease. We propose that the functional classification of disease genes and their products will reveal general principles of human disease. We have determined functional categories for nearly 1,000 documented disease genes, and found striking correlations between the function of the gene product and features of disease, such as age of onset and mode of inheritance. As knowledge of disease genes grows, including those contributing to complex traits, more sophisticated analyses will be possible; their results will yield a deeper understanding of disease and an enhanced integration of medicine with biology.

To test the proposal that classifying disease genes and their products according to function will provide general insight into disease processes[1,2], we have compiled and classified a list of disease genes. To assemble the list, we began with 269 genes identified in a survey of the 7th edition of *Metabolic and Molecular Bases of Inherited Disease*[2]. We then searched the 'morbid map' and allelic variants listed in the Online *Mendelian Inheritance in Man*[3] (OMIM), an online resource documenting human diseases and their associated genes (www.ncbi.nlm.nih.gov), and increased the total disease gene set to 923. This sample included genes that cause monogenic disease (97% of the sample) and genes that increase susceptibility for complex traits. We excluded genes associated only with somatic genetic disease (such as non-inherited forms of cancer) or the mitochondrial genome.

Functional classification

We categorized each disease gene according to the function of its

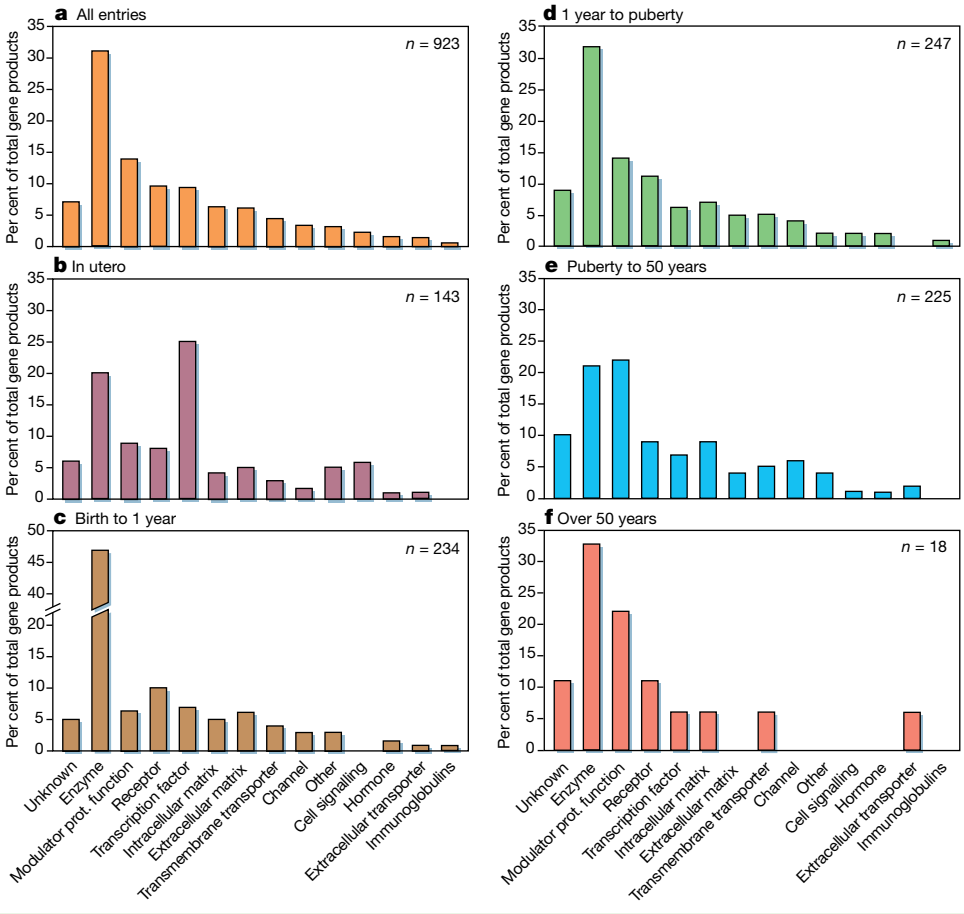

Figure 1 The functions of the protein products of disease genes. **a**, The entire disease gene set. **b–f**, Disease genes stratified according to the typical age of onset of the disease phenotype. The fraction of disease genes encoding transcription factors in the *in utero* onset disorders (25%) differs from the fraction encoding transcription factors for disorders with onset after birth (6%; $\chi^2 = 49.4$, $P < 0.001$). Similarly, the fraction of disease genes encoding enzymes causing a disorder with onset in the first year of life (47%) is different from the fraction encoding enzymes causing disorders with other ages of onset (25.8%; $\chi^2 = 35.8$, $P < 0.001$).

protein product (see Supplementary Information). Our approach differed in two ways from that used by the International Human Genome Sequencing Consortium (IHGSC) to annotate the working draft human sequence[4]. First, we focused on the function of the protein itself without consideration of its biological context, whereas the IHGSC used the classification employed by the Gene Ontology project[5] which integrates three aspects of function: biochemical activity, biological process and subcellular location. Second, our functional designations were largely informed by features of pathology, whereas those of the IHGSC were almost entirely based on sequence homology to proteins of known function in model organisms. We also scored each disease gene for features related to clinical presentation, including age of onset, mode of inheritance, frequency, severity, extent of tissue involvement and association with malformations.

The results of our functional classification of the proteins encoded by 923 disease genes are shown in Fig. 1a. The largest functional category, comprising genes encoding enzymes, accounts for 31.2% of the total. This represents about twice as many as the next highest category, designated modulators of protein function (13.6 %), which includes proteins that stabilize, activate, fold or otherwise influence the function of a second protein. Each of the remaining 12 categories accounts for less than 10% of the total sample. The abundance of enzymes in the disease gene set may reflect some historical bias towards metabolic disorders in the study of human inherited disease[2]. In contrast, only 15% of 114 positionally cloned genes (updated from http://genome.nhgri.nih.gov/clone/) encode enzymes, but this set may have its own biases (see Supplementary Information). Indeed, protein domains associated with enzymes were identified in 27% of 8,360 *Drosophila* proteins scored for these motifs[6]. This observation suggests that in higher eukaryotes the fraction of genes encoding enzymes may be 25–30%, or close to the fraction identified in our disease gene set.

Gene function and disease characteristics

We analysed the disease gene set for evidence of correlations

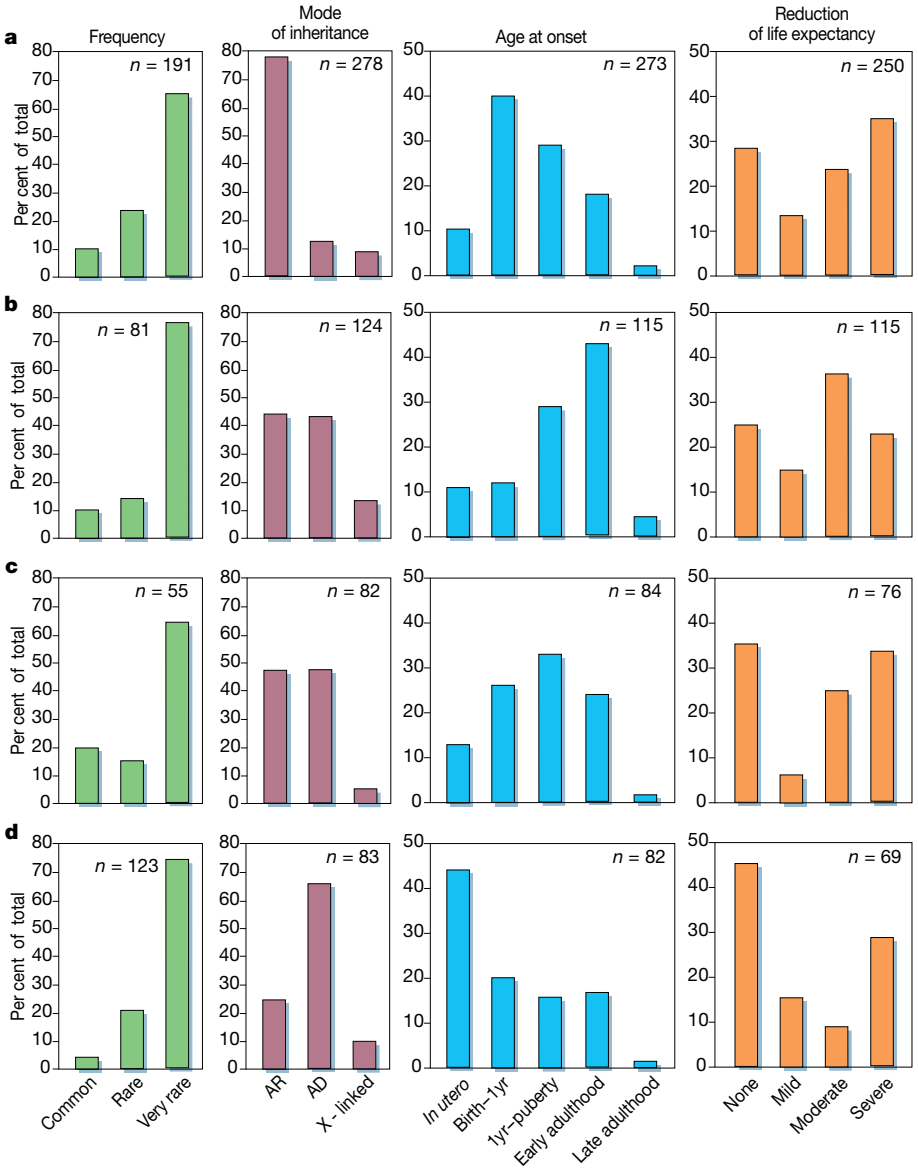

Figure 2 Characteristics of disease arranged by function of the protein encoded by the disease gene. **a**, Disease genes encoding enzymes; **b**, disease genes encoding modifiers of protein function; **c**, disease genes encoding receptors; **d**, disease genes encoding transcription factors. The columns of disease features are labelled at the top. AR, autosomal recessive; AD, autosomal dominant; early adulthood, puberty to <50 years; late adulthood, >50 years.

between the function of a gene product and the age of onset of its associated disease (Fig. 1). Several aspects of this analysis are of interest. First, diseases associated with genes encoding proteins in all the functional categories can appear at any stage of life. The only apparent exception is for diseases presenting after 50 years of age (Fig. 1f) but the sample of genes in this category is small and a more general distribution of protein function may emerge as the number increases. Second, genes encoding transcription factors are over-represented among genes causing genetic disease with onset *in utero* (Fig. 1b). This concentration of diseases resulting from abnormalities of transcription factors probably reflects the important role of these proteins in orchestrating development. It is therefore not surprising that genes encoding transcription factors account for more than 30% of the genes associated with malformation phenotypes (see Supplementary Information).

An extraordinarily high fraction of diseases with onset in the first year of life are caused by defects in genes encoding enzymes (47%; Fig. 1c). This too fits with biological expectations and clinical evidence. The developing fetus has access to its mother's metabolic homeostatic systems through the placenta. Thus, infants with inborn errors caused by enzyme deficiencies are typically normal at birth and develop symptoms only after the defect in their homeostatic system is exposed by demands on their own metabolism[7]. The fraction of disease genes encoding enzymes falls with later disease onset (Fig. 1d–f). Disorders with onset after age 50 are an apparent exception, with the fraction of genes encoding enzymes increasing to more than 33%. But the number of disorders in this category (18) is small, and our understanding of the genes that contribute to complex traits with onset in this age range is limited. Three of the six genes encoding enzymes in this age of onset category are variants identified as susceptibility alleles rather than true disease-producing alleles.

We divided the disease gene list by function and compared disease characteristics including frequency, mode of inheritance, age at onset and reduction in life expectancy. Figure 2 shows the results for the four largest functional categories. Regardless of category, diseases caused by most genes in this analysis are rare or very rare in frequency. In part, this reflects our preliminary knowledge of the genes that contribute to common complex traits. Comparison of the inheritance patterns shows that disorders caused by genes encoding enzymes are primarily recessive, whereas those caused by genes encoding modifiers of protein function and receptors are split more-or-less evenly between recessives and dominants. Disorders caused by genes encoding transcription factors, by contrast, are more likely to be dominant. These results fit well with our understanding of how proteins of various functions contribute to development and homeostasis.

Interestingly, each of the four functional categories has a different peak age at onset. For transcription factors the peak is *in utero*; for enzymes it is in year 1; for receptors it is between year 1 and puberty; and for modifiers of protein function it is in early adulthood. These correlations provide biological support for the validity of the functional characterization and they hint at additional principles of disease. Perhaps disorders of receptors are most likely to present in childhood because this is a time of rapid growth and, especially during puberty, of intense signalling activity between various cells and tissues. Similarly, disorders involving modifiers of protein function may present later in life because the homeostatic systems are not completely disrupted by these defects; rather, they respond in ways that are less congruent with the demands placed on the organism and so become symptomatic more gradually. Finally, there is no apparent relationship between functional category and reduction in life expectancy. This may reflect a true lack of correlation or it may indicate that larger numbers and more sophisticated characterization of disease severity are required to discern such relationships.

Better functional annotation of the human genome and a comprehensive list of human disease genes should lead to much greater integration of medicine and biology. We believe that increasing knowledge of the genes associated with diseases will allow researchers to address more complicated issues, including the relative contributions to disease of genes in the core biological set shared by all species and those encoding proteins specific to humans[8]; how sequence features (such as conservation and polymorphism) relate to disease characteristics; and how protein function relates to the outcome of clinical treatment[9]. □

1. Childs, B. & Valle, D. Genetics, biology and disease. *Annu. Rev. Genomics Hum. Genet.* **1,** 1–19 (2000).
2. Jimenez-Sanchez, G., Childs, B. & Valle, D. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D.) 167–174 (McGraw-Hill, New York, 2001).
3. Antonarakis, S. E. & McKusick, V. A. OMIM passes the 1,000 disease gene mark. *Nature Genet.* **25,** 11 (2000).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
5. The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
6. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
7. Brusilow, S. W., Valle, D. & Arn, P. H. in *Current Therapy in Neonatal Perinatal Medicine* (ed. Nelson, N. M.) 164–169 (B. C. Decker, Philadelphia, 1989).
8. Chervitz, S. A. *et al.* Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282,** 2022–2028 (1998).
9. Treacy, E., Childs, B. & Scriver, C. R. Response to treatment in hereditary metabolic disease: 1993 survey and 10-year comparison. *Am. J. Hum. Genet.* **56,** 359–367 (1995).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank J. Amberger for assistance with OMIM, J. A. Escamilla for statistical advice and S. Muscelli for preparation of the manuscript. D.V. is an Investigator in the Howard Hughes Medical Institute.

Correspondence should be addressed to D.V. (e-mail: dvalle@jhmi.edu).