Data and text mining

Highly consistent patterns for inherited human diseases at the molecular level

Núria López-Bigas¹,*, Benjamin J. Blencowe² and Christos A. Ouzounis³

¹Genome Bioinformatics Laboratory, Center for Genomic Regulation, Universitat Pompeu Fabra, Pg. Martí i de la Barceloneta 37-49, E-08003, Barcelona, Spain, ²Banting and Best Department of Medical Research, C. H. Best Institute, University of Toronto, 112 College Street, Toronto, Ontario MG5 1L6, Canada and ³Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

Received on July 27, 2005; revised on October 27, 2005; accepted on November 11, 2005
Advance Access publication November 15, 2005
Associate Editor: John Quackenbush

ABSTRACT

Over 1600 mammalian genes are known to cause an inherited disorder, when subjected to one or more mutations. These disease genes represent a unique resource for the identification and quantification of relationships between phenotypic attributes of a disease and the molecular features of the associated disease genes, including their ascribed annotated functional classes and expression patterns. Such analyses can provide a more global perspective and a deeper understanding of the probable causes underlying human hereditary diseases. In this perspective and critical view of disease genomics, we present a comparative analysis of genes reported to cause inherited diseases in humans in terms of their causative effects on physiology, their genetics and inheritance modes, the functional processes they are involved in and their expression profiles across a wide spectrum of tissues. Our analysis reveals that there are more extensive correlations between these attributes of genetic disease genes than previously appreciated. For instance, the functional pattern of genes causing dominant and recessive diseases is markedly different. Also, the function of the genes and their expression correlate with the type of disease they cause when mutated. The results further indicate that a comparative genomics approach for the analysis of genes linked to human genetic diseases will facilitate the elucidation of the underlying molecular and cellular mechanisms.

Contact: nuria.lopez@crg.es; ouzounis@ebi.ac.uk
Supplementary information: http://genome.imim.es/~nlopez/supplements/supp_bioinformatics.pdf

1 INTRODUCTION

Elucidating the genetic basis of human-inherited disorders is a major goal of medical genetics. The application of this knowledge can lead to improvements in disease prevention and treatment. More than 1600 genes are now known to cause an inherited disease when they acquire a mutation. The information of genes and mutations causing disease is stored in several databases, such as the Online database of Mendelian Inheritance (OMIM) (Hamosh et al., 2000) LocusLink (Pruitt et al., 2000) and The Human Gene Mutation Database (Krawczak et al., 2000).

*To whom correspondence should be addressed.

The experimental analysis of genes and the consequences of mutations involved in particular diseases are important for the advancement of knowledge in human hereditary diseases. Moreover, now that the complete human genome sequence is known and the number of genes involved in disease is substantial, it is crucial to further develop other types of approaches in medical genomics that can provide us with a different point of view of hereditary diseases. Instead of focusing on one gene or one disease, the prospect of a genome-scale perspective for genes and diseases is emerging. This global viewpoint can help us to obtain a better understanding of disease mechanisms in a generalized manner.

A gene is involved in a hereditary disease when its sequence has been subjected to a mutation that impairs its function or expression strongly enough to produce a certain pathological phenotype that is classified as disease. The type of phenotype a particular mutation causes is typically related to the function and the biological process in which the gene is involved as well as the expression pattern of the gene. In fact, several methodologies have been developed to identify candidate disease genes for particular diseases that rely on the hypothesis that similar diseases will be caused by functionally similar genes (Perez-Iratxeta et al., 2002; Freudenberg and Propping, 2002; Turner et al., 2003; Van Driel et al., 2003; Silva et al., 2004) although this hypothesis has never been tested in detail. We propose that the classification of hereditary diseases and the classification of the genes involved in these diseases should reveal interesting correlations and tendencies of certain classes of genes to be involved in specific types of diseases. We have compiled a set of 1647 genes known to be involved in diseases from the OMIM database (Hamosh et al., 2000). In general, these are single-gene disorders, in which at least one mutation leading to the disease phenotype has been identified. These genes have been classified, where possible, according to the type of disease they cause, the mode of inheritance, their functional class and the biological process in which they are involved. The analysis of the properties of gene products and the characteristics of their associated phenotype reveal interesting correlations and allow us to obtain a global picture of the causes of inherited disease at multiple levels—from anatomical to physiological and from molecular to genetic.
2 CLASSIFICATION OF HEREDITARY DISEASE GENES

Genes involved in hereditary disease are catalogued in OMIM (Hamosh et al., 2000). We retrieved the list of genes from the ‘morbidity map’ table in OMIM. Using the NCBI LocusLink (Pruitt and Maglott, 2001) database (from tables mim2loc and loc2ref) and the Ensembl database (Hubbard et al., 2002) we located the corresponding gene sequence records. The result is a list of 1647 genes associated with human diseases.

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus maintained by the National Library of Medicine (http://www.nlm.nih.gov/mesh/). It consists of sets of descriptive terms arranged in a hierarchical structure that permits searching at various levels of specificity. We searched for MeSH terms of the ‘Diseases’ category (class C) for each disease gene using three complementary approaches: (1) text-mining automatic assignment; MeSH terms were mapped to the free text of the clinical synopsis section of OMIM, using keyword pattern matching and stemming—as described elsewhere (Iliopoulos et al., 2001), (2) Medline assignment; MeSH terms in Medline abstracts were retrieved through SRS (Zdobnov et al., 2002) and linked to the corresponding disease gene using the OMIM identification number, and (3) manual curation was used when no assignment was obtained using the two previous methods. In total, 1260 disease genes (77% of total) were thus assigned to (at least) one MeSH term related to the phenotype they cause when they have mutations. Note that some disease genes cause syndromes or diseases that affect several systems or organs and thus they are classified simultaneously in several disease categories. The distribution for the most general level of the hierarchical structure of MeSH disease terms is quite variable; in total there are 17 different categories. The two most frequent disease categories in human-inherited diseases are ‘nervous system’ diseases and ‘neonatal diseases and abnormalities’ (both at 14% of total), followed by ‘metabolic’, ‘musculoskeletal’ diseases and ‘neoplasms’ (Fig. 1a).

Disease genes were also classified according to the mode of inheritance of the diseases they cause using text-mining automatic extraction from the clinical synopsis section in OMIM and manual curation. Most of the disease genes (74% of total) cause either dominant or recessive diseases, or both, depending on the mutation (Fig. 1b). Other types of hereditary mutations include X-linked, chromosome rearrangements and association for complex traits (see Glossary in Supplementary data).

Each human gene was classified according to the molecular function of its protein product and the biological process it is involved in, according to Gene Ontology (GO) ‘slim’ terms (Camon et al., 2004). In total, 12,221 human genes and 1430 disease genes (87% of total) have at least one GO term assigned. The comparison of the functional classification or the biological process of all human genes and the disease genes give insights on which type of genes are known to be involved in human diseases and which are underrepresented in disease (Fig. 1c and d). This analysis reveals that although all GO Molecular Function classes are found to be involved in diseases, some classes, such as structural molecules, transporters or enzymes are overrepresented in the disease set, while others, such as nucleic acid binding, are markedly underrepresented (P-value for χ² analysis is 1.1e–13; Fig. 1c). The categorization of genes in GO Biological Process classes reveals that genes involved in developmental processes or stress response are clearly overrepresented in the disease set, while genes involved in cell communication are underrepresented (Fig. 1d). In order to assess the statistical significance of these differences we have performed two different tests: χ²-test (Fig. 1b and c) and 1000 random simulations for the distribution of GO terms (see Supplementary Table 1).

3 MODE OF INHERITANCE AND SEQUENCE CONSERVATION

We have analysed the disease gene set for evidence of correlations between the GO Molecular Function or Biological Process in which a gene product is associated and its mode of inheritance, namely dominant or recessive disorders (Fig. 2). On one hand, analysis reveals that diseases caused by mutations in genes coding for the GO Molecular Function classes of enzymes and transporters are mostly recessive (74 and 61.7%, respectively) (Fig. 2a). On the other hand, mutations in transcription regulators (72.7%), structural molecules (65.6%), nucleic acid binding genes (62.1%) and signal transducers (56.1%) are primarily dominant (Fig. 2a), consistent with previous reports (Jimenez-Sanchez et al., 2001) which used a slightly different functional classification. With respect to GO Biological Process classes, genes involved in metabolism or stress response cause mainly recessive diseases while genes involved in developmental processes, cell communication, cell cycle or cell motility are primarily involved in dominant disorders (Fig. 2b, χ²-test (Fig. 2) and 1000 random simulations of the GO distributions (see Supplementary Table 1) have been performed to assess the statistical significance of these results. These results are highly consistent with the general theory of dominance (Hurst and Randerson, 2000) in which information on the relevant mechanism(s) for different types of mutations can affect specific gene classes.

Mutations leading to disease can be classified into loss-of-function and gain-of-function mutations (see Glossary in Supplementary data). Loss-of-function mutations usually produce recessive phenotypes, because for most gene products a precise quantity is not crucial, and reduced molecular concentration is sufficient for normal function (Jimenez-Sanchez et al., 2001; Papp et al., 2003). For some gene products, however, reduction of the typical gene product concentration results in abnormal function, and haploinsufficiency generates a disease phenotype, which is therefore inherited in a dominant manner. Genes whose products act essentially in isolation, such as soluble metabolic enzymes, rarely show a dosage effect (Scrivener, 2002) thus their involvement is primarily observed in recessive diseases. On the other hand, mutations in proteins that function as dimers or multimers can produce dominant negative effects, i.e. when the mutant protein is able to interfere with the wild-type protein in a heterozygote cell. Proteins that form multimeric structures (such as structural proteins) are especially susceptible to dominant negative effects. An exemplary case is collagen, because the mutated protein can affect the polymerization of collagen fibres with wild-type molecules (Myllyharju and Kivirikko, 2001) We also find cases of dominant negative effects in non-structural proteins that dimerize or oligomerize. For example, many transcription factors function as dimers, thus their frequent involvement in dominant phenotypes. Mutant proteins unable to dimerize are usually involved in recessive phenotypes, while mutants that can interfere with the wild-type
protein producing inactive dimers cause dominant phenotypes (Strachan and Read, 2004). Signal transducers and nucleic acid binding proteins provide other examples of multimeric structures that are susceptible to dominant negative effects. Mutations that create a new allele associated with a new function (i.e. gain-of-function mutations) usually cause dominant phenotypes, because the presence of the normal allele does not necessarily suppress the new function acquired by the mutant protein. This is often the case of control or signalling proteins. Such typical examples are mutant proteins constitutively involved in non-regulated signalling in general, or process activation in particular.

The classification of gene functions into dominant or recessive mutation classes also reveals sequence conservation properties for genes in each of these categories. Previously, we have reported that genes involved in disease have less conserved paralogs than human genes in general (Lopez-Bigas and Ouzounis, 2004) presumably
because highly similar paralogs can potentially restore the function of a mutated protein—in which case a disease might not be observed. With the classification of disease genes by the mode of inheritance of the phenotype, this pattern is more pronounced in recessive compared with dominant disease genes. The average conservation score (cs) of paralogs (BLAST score versus the closest paralog over the maximum possible BLAST score, i.e. versus itslf) (Lopez-Bigas and Ouzounis, 2004) for all genes is 0.43, for genes causing dominant diseases is 0.40 and for those causing recessive diseases is 0.31, indicating the least conserved paralogs in the latter. Moreover, the number of genes with close paralogs (cs > 0.7) for genes causing dominant diseases is 72 out of 493 (15%) while this number for genes causing recessive diseases is only 61 out of 678 (9%) (the overall percentage of genes for paralogs in the human genome with cs > 0.7 being 24%). This is highly consistent with the current knowledge from genetics (Strachan and Read, 2004): recessive diseases are usually caused by loss-of-function mutations, in which case a close paralog can potentially restore the function of the mutated gene. Conversely, dominant diseases are often caused by dominant negative mutations or gain-of-function mutations, in which case the presence of a close paralog cannot restore the function of the gene.

4 GENE CLASSES AND DISEASE TYPE

In general, it would be expected that the GO Molecular Function and Biological Process in which the gene is associated should relate to the disease phenotype it causes when mutated. In fact, several methodologies to identify candidate disease genes use this principle (Perez-Iratxeta et al., 2002; Freudenberg and Propping, 2002; Turner et al., 2003; Van Driel et al., 2003; Silva et al., 2004) although a profound study on the magnitude of these relationships has never been performed. In order to investigate the extent to which this applies, we have classified the number of genes involved in each type of disease according to the GO Molecular Functional and Biological Process classes they belong to. We have assessed statistically the GO patterns observed for each type of disease by χ²-test (Fig. 3) and by performing 1000 random simulations of the GO distributions (see Supplementary Table 2). This analysis reveals that although genes encoding proteins in all functional categories or biological processes can be involved in different diseases, some classes of genes are preferentially involved in specific disease types (Fig. 3). Despite the fact that some of these patterns are highly consistent with current knowledge, this is the first time a quantitative correlation analysis of these properties has been performed.

With regard to molecular function, it is evident for instance that metabolic diseases are primarily caused by enzymes and also by transporters. It is also interesting to note that nucleic acid binding genes and transcription regulators are overrepresented in the group of genes that cause neoplasms and strongly underrepresented in metabolic diseases, or that skin diseases are mainly caused by structural molecules (Fig. 3a).

With regard to biological processes, it is apparent that genes involved in stress response and death are overrepresented in immunologic diseases, or that cardiovascular diseases are preferentially associated with cell motility genes. It is also interesting to note that genes causing inherited neoplasm diseases are mostly involved in cell cycle, stress response and cell growth and maintenance (Fig. 3b).

These correlations between classes of genes and type of disease mostly follow the expectations and reflect our current state of knowledge. For example, the fact that mutations in genes encoding enzymes are mostly associated with metabolic diseases and not in neoplasms can be easily appreciated. Another characteristic example is the class of structural molecules associated with phenotypes affecting the cardiovascular system or the skin, two tissues...
whose function crucially depends on properly formed structural proteins. This type of analysis also unravels the complexity of some diseases: for instance, metabolic diseases and neoplasms appear to have a more complex pattern of functional categories than, e.g. otorhinolaryngologic (ear/nose/throat), eye or skin diseases.

5 DISEASE TYPE AND TISSUE-SPECIFIC EXPRESSION PATTERNS

The type of phenotype caused by mutations in a particular gene must also be related to the tissue-specific expression pattern of that gene. We have thus analysed tissue-specific expression patterns of disease genes with regard to the type of disease they cause. This analysis has been performed using two different datasets, namely human (Su et al., 2004) (Fig. 4) and mouse tissue-specific expression data (Zhang et al., 2004) (Fig. 5). In total, 1545 human disease genes are represented in the human gene atlas (94% of total). For the mouse microarray dataset, genes were mapped to their corresponding human ortholog according to Supplementary Table 4 from the original publication (Zhang et al., 2004). In all, the gene expression pattern for 8553 human genes can be inferred from the mouse expression data, of which 740 are OMIM disease genes (45% of total). We only took into account those tissues in which every gene is prominently expressed. To ensure that values are comparable, we included genes expressed with a median-subtracted >4 for human, and a median-subtracted arcsinh value >1 for mouse. The number of genes notably expressed in each tissue and causing a particular disease type were compared with an expected value (following a \( \chi^2 \)-test) (Fig. 4 and 5). The matrices of \( \chi^2 \)-values were re-ordered for a better visual representation using the Sotarray clustering algorithm (Herrero et al., 2001) from the GEPAS server (Herrero et al., 2003).

It thus becomes manifestly evident that genes causing phenotypes affecting particular tissues are preferentially expressed in tissues in which the phenotype is noticeable. For example, the set of genes causing metabolic diseases are preferentially expressed in tissues involved in metabolism such as human kidney or liver (Fig. 4) and also heart, stomach or intestine in mouse, but not at the embryonic tissues (Fig. 5). Other striking examples include the positive association of lymph nodes with immune disease and bone marrow with hemic and lymphatic diseases in the human dataset (Fig. 4) or the negative association of skin diseases with kidney in the mouse dataset (Fig. 5), among multiple others. It is reassuring to observe that similar results are obtained with data from both human and mouse tissues. When only the 23 tissues shared by the two species datasets are considered, these patterns are remarkably comparable (Fig. 6).

6 DISCUSSION

Currently, there are more than 1600 human genes known to be associated with a particular disease phenotype. The analysis of...

**Fig. 3.** Distribution of types of disease caused by genes with specific (a) molecular function (the \( P \)-value for the \( \chi^2 \)-test is 3.4e-13) or (b) biological process (\( P \)-value = 5e-34). \( \chi^2 \)-values for each cell are represented with a colour code, as in Figure 1—except the dark grey cells that signify equal representation (\( \chi^2 \) value < 1). Light grey boxes signify absence of data (sample size < 10).
this group of genes from a global perspective has already revealed interesting insights about the nature of human disease (Lopez-Bigas and Ouzounis, 2004). Despite the importance of experimental analysis of human diseases at the molecular level, complementary computational approaches are able to investigate patterns of hereditary diseases in ways that medical genetics experiments cannot readily achieve. Importantly, new questions arise, assisting us to formulate accurate hypotheses to be further investigated.

Some important steps have already been made towards this direction. Previously, a report explored the functional classification of disease genes and described a correlation between the function of the gene product and general features of the disease, such as the age...
of onset, reduction of life expectancy and the mode of inheritance (Jimenez-Sanchez et al., 2001). Recently, we have shown that genes involved in human disease have specific sequence properties compared with the rest of genes in the human genome that make them more likely to undergo mutations leading to a diseases phenotype (Lopez-Bigas and Ouzounis, 2004). Sequence analysis for several eukaryotes revealed that human proteins with multiple long amino acid runs are often associated with diseases (Karlin et al., 2002). Others have identified 714 human disease genes matching 548 Drosophila sequences (Reiter et al., 2001).

Other recent work has focused on the global analysis of disease-associated mutations (Ferrer-Costa et al., 2002; Miller and Kumar,
These studies reveal that the mutated residues in disease proteins have common sequences and structural properties and are usually associated with radical changes in protein structure (Ferrer-Costa et al., 2002; Steward et al., 2003). Also it has been shown that most disease-mutated residues are evolutionarily conserved (Miller and Kumar, 2001; Mooney and Klein, 2002).

Genome-wide identification of disease-relevant human genes is a growing area of research. Recent studies describe different approaches based on computational methods to identify disease-associated genes. Automatic search using text mining in the biomedical literature has proven to be successful in the identification of known human disease genes and novel candidate genes (Lopez-Bigas and Ouzounis, 2004). Other computational methods for the identification of disease-related genes rely on the hypothesis that similar diseases will be caused by functionally similar genes (Freudenberg and Propping, 2002; Turner et al., 2003). Tools that integrate data from mapping, expression and phenotypic databases to retrieve a list of genes based on user-defined criteria have been devised (Van Driel et al., 2003). The term ‘patholog’ has been proposed to mean a homologue of a human disease-related gene and itself a candidate to be involved in disease. The identification of ‘pathologs’ from mouse cDNA datasets combined with information extraction methods in parallel with human expert analysis of MEDLINE abstracts has been recently proposed (Silva et al., 2004). Development of these methods is crucial to maximize the medically relevant knowledge that can be extracted from the Human Genome Project. Experimental cost can be largely reduced by the use of such methods to generate an accu-

Fig. 6. Plots of distribution of types of disease caused by genes with specific expression pattern for tissues comparable between human and mouse datasets. $\chi^2$-values for each cell are represented with a colour code, as in Figures 4 and 5.
rate and objective list of candidate genes for human genetic diseases.

7 CONCLUSION

This global analysis unravels correlations between modes of inheritance and sequence conservation, disease types, functional attributes such as molecular function or biological processes, and finally tissue-specific expression patterns of genes. The correlations detected provide a consistent picture that facilitates our understanding of the complex basis of hereditary diseases (Thornton-Wells, 2004), and depicts relationships between disease phenotypes and molecular properties of the genes involved that can be used for prediction of new candidate genes, as has been previously done for some of these correlations (Perez-Iratxeta et al., 2002; Freudenberg and Popping, 2002; Turner et al., 2003; Van Driel et al., 2003; Silva et al., 2004).

This work together with other recent reports on the global analysis of disease-related genes and their mutations emphasize the importance and potential application of comparative disease genomics. A genome-scale approach for the study of medical genomics. A genome-scale approach for the study of medical genomics.

ACKNOWLEDGEMENTS

We thank Leon Goldovsky for technical support, other members of the Computational Genomics Group for comments, Rodrigo Lopez for help with SRS and Barbara Brodsky for discussions. N.L.-B. is supported by a long-term post-doctoral fellowship from the Human Frontiers Science Program. C.A.O. acknowledges support from the UK Medical Research Council and IBM Research.

Conflict of Interest: none declared.

REFERENCES