BIOINFORMATICS DISCOVERY NOTE Vol. 25 no. 22 2009, pages 2891–2896
doi:10.1093/bioinformatics/btp538

Systems biology

Phenotypic categorization of genetic skin diseases reveals new relations between phenotypes, genes and pathways

Ruslan I. Sadreyev[1], Jamison D. Feramisco[2], Hensin Tsao[3,*] and Nick V. Grishin[1,4,*]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, [2]Department of Dermatology, University of California at San Francisco, San Francisco, CA 94115, [3]Wellman Center for Photomedicine, Department of Dermatology, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114, and [4]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received on May 11, 2009; revised on August 10, 2009; accepted on September 7, 2009

Advance Access publication September 10, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Systematic analysis of connection between proteins, their cellular function and phenotypic manifestations in disease is a central problem of biological and clinical research. The solution to this problem requires the development of new approaches to link the rapidly growing dataset of gene–disease associations with the many complex and overlapping phenotypes of human disease.

Results: We analyze genetic skin disorders and suggest a manually designed set of elementary phenotypes whose combinations define diseases as points in a multidimensional space, providing a basis for phenotypic disease clustering. Placing the known gene–disease associations in the context of this space reveals new patterns that suggest previously unknown functional links between proteins, signaling pathways and disease phenotypes. For example, analysis of telangiectasias (spider vein diseases) reveals a previously unrecognized interplay between the TGF-β signaling pathway and pentose phosphate pathway. This interaction may mediate glucose-dependent regulation of TGF-β signaling, providing a clue to the known association between angiopathies and diabetes and implying new gene candidates for mutational analysis and drug targeting.

Contact: grishin@chop.swmed.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Rigorous quantitative analysis of disease phenotypes is a key problem on our way to understanding systemic effects of human gene mutations. Such understanding would enable statistical prediction of clinical manifestations for genome abnormalities, inference of causative genes from complex disease phenotypes, as well as deeper insights into molecular mechanisms of pathophysiology. This tremendous task requires the development of new approaches to link the rapidly growing dataset of gene–disease associations with the many complex and overlapping phenotypes of human disease.

Previously reported approaches to this problem ranged from considering diseases as individual entities connected through shared causative genes (Goh *et al*., 2007) or co-occurrence in the same patient (Rzhetsky *et al*., 2007), to more detailed classifications involving the comparison of disease phenotypes, usually based on ontologies of phenotypic terms derived from natural-language phenotype descriptions through automated or semi-automated text analysis (Robinson *et al*., 2008; van Driel *et al*., 2006). These analyses may include additional high-throughput data on protein associations (Lage *et al*., 2007; Wu *et al*., 2008), improving prediction of new connections between diseases and proteins involved. Here we suggest a different approach to quantitative gene-phenotype analysis. By focusing on the set of genetic skin disorders, we are able to manually analyze the corresponding descriptions of phenotypic manifestations and design a set of elementary phenotypic features whose combinations define any given disease as a point in a multidimensional space. Placing the known gene–disease associations in the context of this space reveals new patterns that suggest previously unknown functional links between disease phenotypes, proteins and signaling pathways. In particular, analysis of telangiectasias (spider vein diseases), reveals a previously unrecognized interplay between the TGF-β signaling cascade and pentose phosphate pathway (PPP), which may mediate glucose-dependent regulation of TGF-β signaling in diabetes.

2 METHODS

The database of 560 genodermatoses, with their phenotypic representations, affected organ systems and associated genes (see Supplementary Material) was previously compiled by expert analysis of OMIM database (Feramisco *et al*. 2009). The elementary phenotypic features were manually selected as general clinical manifestations that can occur independently in different diseases and, when combined, can cover phenotype of any included genodermatosis (Feramisco *et al*., 2009).

Correlation matrix R of elementary phenotypes is calculated based on the set of genodermatoses represented as vectors in the phenotype space:

$$R_{ij} = \mathrm{corr}(P_i, P_j),$$

where $P = X^{\mathrm{T}}$ is the matrix composed of phenotype vectors P_i that corresponds to the transposed matrix X of disease vectors $X_i = \{x_k\}$, $k = 1,N$ (N is the dimension of phenotype space). $\mathrm{corr}(P_i, P_j)$ is Pearson's correlation coefficient of two vectors.

Principal components of the set of disease points in the phenotype space are derived as eigenvectors of covariance matrix $C = XX^{\mathrm{T}}$, with largest

*To whom correspondence should be addressed.

eigenvalues corresponding to the dimensions of strongest correlations within the data set X (Jolliffe, 2002).

Statistical significance of gene–phenotype associations is estimated from the observed counts of co-occurrence of a given phenotype with mutations in a given gene (Supplementary Tables S1–S5). Fisher's exact test is used as a more appropriate alternative to chi square association test in the cases where counts in some cells of contingency tables are low (<5). The significance level is adjusted for the multiple testing of all gene–phenotype associations using Bonferroni correction: $\alpha = \alpha_0/n = 1.2 \times 10^{-6}$, where $\alpha_0 = 0.05$ and $n = n(\text{genes})\ n(\text{phenotypes})$.

Phenotypic clustering is performed using UPGMA agglomerative method.

3 RESULTS

Using the Online Mendelian Inheritance in Man (OMIM) compendium of human mendelian inheritance (Amberger *et al*., 2009), we previously compiled a database of genetic skin diseases, their phenotypic representations, affected organ systems and associated genes (Feramisco *et al*., 2009). This database (see Supplementary Material) includes 560 diseases associated with 501 protein-coding genes, with ∼16% of diseases linked to two or more mutated genes and 18% of genes linked to two or more disease entities.

3.1 Phenotypic categorization of genodermatoses

Based on the manual analysis of disease phenotypes, we define a minimal set of elementary phenotypic features whose combinations cover all included genodermatoses, so that phenotypic manifestation of each skin disease can be represented as a combination of several elementary features. This set includes 42 elementary dermatologic features forming 18 groups: cornification phenotypes, pigmentation phenotypes, etc. and 29 elementary systemic phenotypes forming 17 groups (Feramisco *et al*., 2009) (see Supplementary Material). For example, the group of pigmentation features includes hyper- and hypo-pigmentation, café au lait, poikiloderma and nevi (birthmarks and moles). Most of the genodermatoses are characterized by more than one elementary phenotype, with the average number of assigned elementary phenotypes being ∼2.4 per disease (Feramisco *et al*., 2009). Distributions of the numbers of dermatologic and systemic phenotypes per disease and the numbers of diseases sharing an elementary phenotype are shown in Supplementary Figure S1.

As a combination of elementary features, each disease can be represented by a point in multi-dimensional space defined by 71 basis vectors corresponding to the features (Fig. 1). To check for the independence of these features in the phenotype space, we calculate their correlation matrix based on the set of all disease points (see 'Methods' section for details). This matrix (Fig. 2, see also Supplementary Material) does not show major correlation patterns in the incidences of different features in composite disease phenotypes. The highest non-diagonal correlation coefficient is 0.605 for poikiloderma versus alopecia/hypotrichosis phenotypes; the second highest is 0.505. When the distribution of disease points is analyzed in the plane (Supplementary Fig. S4A) or 3D-space (Supplementary Fig. S4B) of highest-correlated phenotypes, every possible combination of these phenotypes is found in a significant number of diseases.

In addition, we perform principal component analysis (PCA) of the disease set and visualize the distribution of points along principal components. Elongated or skewed distribution of points' projections onto the subspace of first two or three principal components, which correspond to the largest correlations within the dataset, would indicate a major correlation in the occurrence of phenotypic features. Neither the projection on the plane of top two components (Fig. 2B) nor the projection in the space of top three components (Supplementary Fig. S2) reveal such correlations. The corresponding projections for the separate dermatologic and systemic phenotype sets, along with phenotype correlation matrices are shown in Supplementary Figure S3. These results suggest that our set of phenotypic features is largely independent.

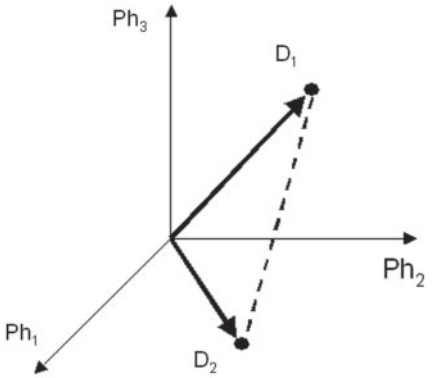

Fig. 1. Representation of diseases as points in the phenotype space. As an example, a 3D space of elementary phenotypic features Ph_1–Ph_3 is shown, with two diseases (D_1, D_2) defined by the combinations of these features. The similarity between these composite disease phenotypes is determined by the distance between points D_1 and D_2.

In phenotypic space, the similarity between two diseases can be determined by the distance between the corresponding points (Fig. 1). We use these distances to group diseases by phenotypes. Several tested distance metrics (Manhattan block, Euclidean distance, etc.) produce similar results, thus we further use Euclidean distances for simplicity. The distances range from zero for disorders with the same sets of phenotypic features to 3.9 for the most distant disorders that have 15 elementary phenotype differences.

Based on this disease representation, we analyze (i) phenotypes shared among diseases associated with a protein or a group of proteins (Fig. 3A); and (ii) functional links between proteins associated with phenotypically similar diseases (Fig. 3B). This analysis can provide significant insights into molecular mechanisms of pathogenesis. First, associations between proteins and disease phenotypes can be dissected in a statistically rigorous manner and previously unknown links can be revealed. Second, similarity in phenotypic manifestation may suggest common mechanisms of action for different proteins or signaling systems, and point to potential functional interactions. Finally, inferred phenotypic association of a protein group or pathway provides a set of new protein candidates that may cause similar diseases of yet unknown molecular mechanism (Fig. 1).

Our approach is different from previously reported approaches to the representation of disease phenotypes (Lage *et al*., 2007; Robinson *et al*., 2008; van Driel *et al*., 2006) in two essential points. First, we base our analysis on the data of high quality, with the phenotype descriptions being manually curated by an expert. Second, as opposed to building ontology of all phenotypic terms

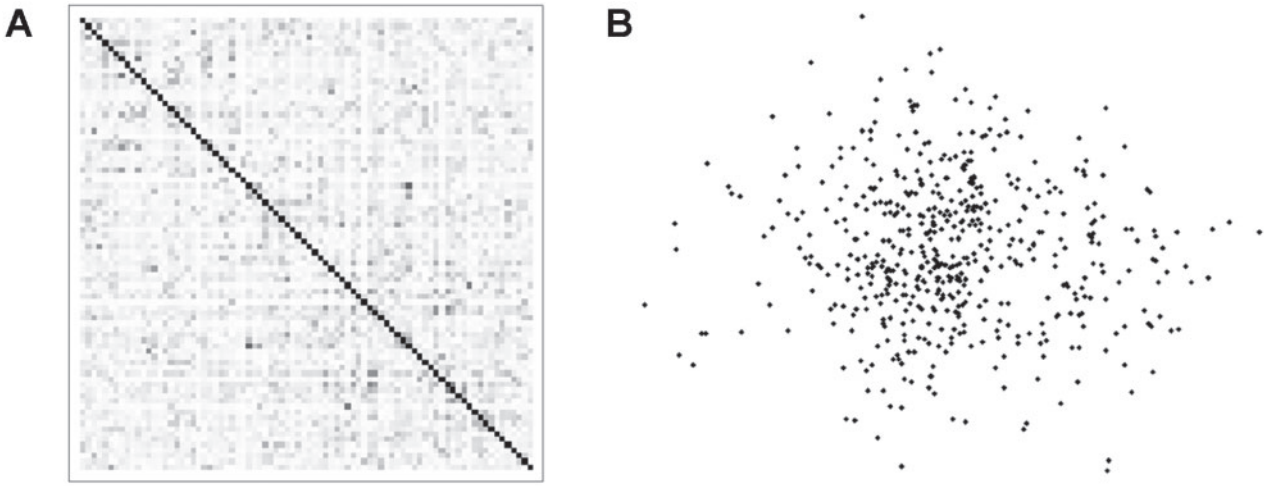

Fig. 2. Selected elementary phenotypic features are largely independent. (**A**) Correlation matrix of phenotypic features, based on the whole dataset of 560 disease phenotypes. The matrix is shown as a grayscale grid, with each square representing the absolute value of Pearson's correlation coefficient between two phenotypic features (see Methods). The values are in the range between 0.0 (white) and 1.0 (black). The matrix is symmetric, with ones on the diagonal. (**B**) Projection of the disease set on the plane of first two principal components in phenotypic space does not show any general correlations in the occurrence of elementary phenotypes among the diseases.

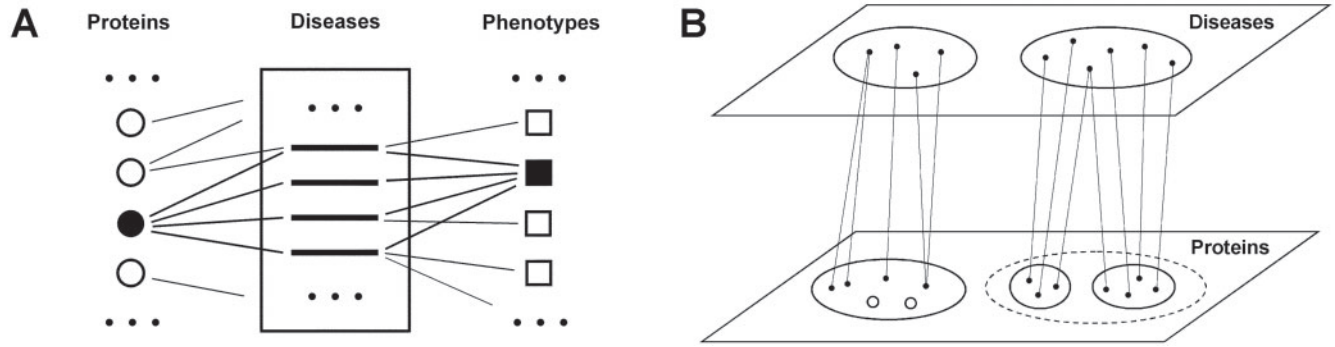

Fig. 3. (**A**) Inference of protein–phenotype associations. Decomposing disease phenotypes into elementary features allows for statistical analysis of co-occurrence between mutations in specific proteins and resulting elementary phenotypes. Significant correlations between defects of a protein (filled circle) and a manifested phenotypic feature (filled square) may suggest causation. (**B**) Inference of protein and pathway associations. Diseases are clustered by phenotypic presentation (upper plane) and corresponding groups of disease-associated proteins are considered (lower plane). Left, when a cluster of diseases corresponds to mutations in functionally related proteins (e.g. proteins from the same signaling cascade), other proteins of this functional group (open circles) may be suggested as new potential candidates for the association with the same class of diseases. Right, new relations between different protein groups may be inferred from their mapping to the same phenotypic cluster of related diseases.

and automatically tracing their relationships through shared parents, we are able to decompose complex phenotypes into elementary unrelated features that can be combined in an independent fashion.

3.2 Protein–phenotype associations

Decomposition of disease phenotypes into elementary features often reveals that an individual protein or a group of proteins is predominantly associated with a certain phenotypic feature (Fig. 3A). The statistical significance of such associations can be estimated from the frequencies of co-occurrence of mutations in a specific protein with an elementary phenotype. Our results confirm many previously known associations. For example, mutations in keratin I and collagens I and VII are significantly associated with disease phenotypes of hyperkeratosis (Supplementary Table S1, Fisher's exact test P-value: $P = 1.1 \times 10^{-6}$), atrophy/aplasia/fragile skin (Supplementary Table S2, $P = 9.2 \times 10^{-8}$) and bullous epidermal cohesion (blistering phenotype, Supplementary Table S3, $P = 1.5 \times 10^{-9}$), respectively. In addition, our results suggest previously unknown associations. For example, the link between mutations in the subunits of laminin 5(332) and mucosal phenotype group (Supplementary Table S4, $P = 7.2 \times 10^{-7}$), has been detected. As a major component of the basement membrane, laminin 5(332) is an essential structural component of the dermal–epidermal junction and is involved in cell adhesion and signal transduction. Our analysis suggests that various defects in laminin 5(332) consistently affect the integrity of mucous membranes. Extending our analysis to groups of functionally related proteins, we find statistically significant

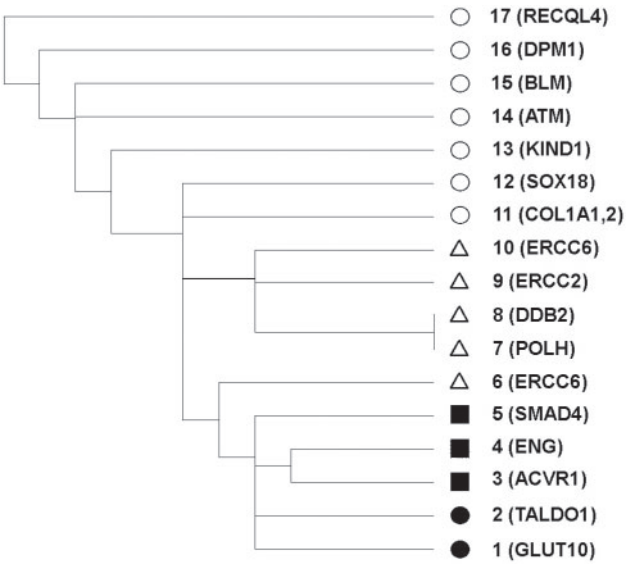

Fig. 4. Phenotypic clustering of diseases sharing the phenotype of telangiectasia ('spider veins') corresponds to a clear functional grouping of associated proteins and suggests the relation between TGF-β signaling cascade and pentose phosphate metabolic pathway. The cladogram of agglomerate hierarchical clustering is shown, the nodes marked by the numerical code of the disease (1–17) with the name of the associated protein in parentheses. Filled circles, proteins associated with PPP. Filled squares, proteins involved in TGF-β pathway. Open triangles, proteins involved in DNA excision repair. Open circles, other proteins. The following diseases are shown: 1: arterial tortuosity syndrome; 2: transaldolase deficiency; 3: Osler-Rendu-Weber syndrome 2; 4: telangiectasia, hereditary hemorrhagic, of Rendu, Osler and Weber; 5: juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome; 6: UV-sensitive syndrome; 7: xeroderma pigmentosum with normal DNA repair rates; 8: xeroderma pigmentosum, complementation group E; 9: xeroderma pigmentosum, complementation group D; 10: De Sanctis-Cacchione syndrome; 11: osteogenesis imperfecta, type IV; 12: hypotrichosis-lymphedema-telangiectasia syndrome; 13: Kindler syndrome; 14: ataxia-telangiectasia; 15: Bloom syndrome; 16: congenital disorder of glycosylation, type Ie; 17: Rothmund-Thomson syndrome.

associations for gap junction proteins, connexins, which show a highly significant connection to hyperkeratosis (Supplementary Table S5, $P = 1.7 \times 10^{-9}$).

Unsurprisingly, most of the elementary phenotypic features are associated with a wide range of protein functions. However, some features are associated with more functionally uniform groups of proteins. For example, bullous (blistering) phenotype is mainly caused by mutations in structural proteins (collagens, keratins), filament-associated proteins, or cell adhesion molecules.

3.3 Functional associations between proteins through their phenotypic representation

Similarity of disease phenotypes caused by mutations of different proteins may point to functional relationship between these proteins. In addition to many known protein relations, our dataset allows for the inference of previously unknown functional links.

As an example, Figure 4 shows phenotypic clustering of telangiectasias (spider vein diseases) that suggests an association between the TGF-β cascade and the PPP. The cladogram based on phenotypic composition reveals two major groups of similar, tightly clustered diseases. The first group is associated exclusively with the proteins of DNA excision repair (ERCC2, ERCC6, DDB2, POLH), whereas the second group is associated with proteins that are connected to two distinct systems: TGF-β cascade (SMAD4, ENG, ACVR1) and PPP (transaldolase TALDO1 and glucose transporter GLUT10). Involvement of TGF-β in vascular anomalies has been reported for several diseases (Coucke *et al*., 2006; Loeys *et al*., 2005; Tille and Pepper, 2004). The detected phenotypical similarity to the PPP-associated disorders suggests that PPP is implicated in angiogenesis through the same pathophysiologic mechanism. In particular, we hypothesize a cross-talk between PPP and TGF-β cascade, which may mediate the connection between glucose metabolism and abnormal vascular development in angiopathies. Among other implications, this hypothesis provides a potential explanation of the known link between diabetes and angiopathies (Miles *et al*., 2007; Simo *et al*., 2006), as well as suggests new potential angiopathy-linked proteins.

4 DISCUSSION

The role of proteins and their interactions in the living organism is a central focus of molecular and cellular biology, with both fundamental and clinical implications. In this respect, the catalogued links between protein mutations and their phenotypic manifestations in disease are, in a sense, the results of a grand mutagenesis 'experiment' that may prove useful for finding new associations and generating new hypotheses. Here, we present an analysis of phenotypes expressed in genetic skin disorders and the corresponding causative genes.

Although our method involved initial manual curation of elementary phenotypes, it can be readily generalized to other disease sets with available phenotype descriptions. The construction of the set of elementary phenotypes can start from all phenotypic terms derived from textual annotations of diseases of interest, further filtered by manual expert analysis and/or numerical selection of the maximal subset of independent phenotypic features, based, for example, on the analysis of correlation matrix (Fig. 2A) or on PCA. However, we believe that the set of phenotypes manually curated by an expert provides a more informative categorization, since it (i) involves additional knowledge not reflected in the brief disease annotations; and (ii) is based on well-defined clinical terms and thus is more accessible and relevant for clinical research community. As an example, analysis of such a dataset can link genes to specific clinical manifestations (Fig. 3A) that are easily recognized by medical practitioners, which facilitates data accumulation and hypothesis testing.

Involvement of a protein in the development of a specific disease phenotype is an important piece of functional information, which may lead to (i) better understanding of protein's function and molecular mechanism of disease; (ii) hypotheses about phenotypic effects of mutations in functionally related proteins; and (iii) phenotype-based prediction of potential proteins involved in similar diseases with unknown genetic causation.

The presented protein–phenotype associations, inferred from a relatively restricted statistical sample of OMIM, can be validated on larger datasets. This validation may be performed in at least two directions. First, phenotypic effects of mutations in a specific gene can be analyzed on a wider scale: most directly, in larger

clinical studies or in animal models. Second, for a considerable set of characterized skin diseases, causative genes are still unknown. With many of these diseases included in OMIM, an interesting further direction would be to test the patients for the mutations in the genes that have statistical associations with the manifested phenotypes.

Similarity in phenotypic effects of different proteins may lead to hypotheses about previously unknown functional associations (Fig. 3B). We present a potentially interesting example of such an association, the interaction between PPP and TGF-β pathway suggested by phenotypic clustering of telangiectasias (Fig. 4).

Utilizing glucose-6-phosphate as the substrate, PPP produces the major fraction of cell's NADPH, which plays a central role in maintaining intracellular redox potential by serving as a co-factor in the reduction of glutathione (Berg *et al*., 2001). As an example, PPP enzyme glucose-6-phosphate dehydrogenase is shown to play an important role in oxidative stress (Leopold *et al*., 2003; Park *et al*., 2006), which is thought to be the main mechanism of its involvement in cardiovascular disease (Matsui *et al*., 2005; Rajasekaran *et al*., 2007; Wiesenfeld *et al*., 1970). Changing the concentration of reduced glutathione alters redox state and affects, among other systems, TGF-β pathway (Maulik and Das, 2002; Shan *et al*., 1994). Altered redox state and the resulting oxidative stress are known to stimulate angiogenic response (Maulik and Das, 2002; Ushio-Fukai, 2006), and are shown to be involved in at least one disorder with telangiectasia phenotype (Nicotera *et al*., 1989). In addition, the effect of GLUT10 deficiency on TGF-β signaling in the arterial wall is shown in arterial tortuosity syndrome (Coucke *et al*., 2006).

We hypothesize that (i) defects in PPP may cause abnormal vascular development by altering intracellular redox state; and (ii) this effect may be mediated by TGF-β signaling cascade. This hypothesis has several important implications. First, the reported effects of glucose concentration on TGF-β cascade (Hua *et al*., 2003; Isono *et al*., 2000; Zhu *et al*., 2007) may be mediated by PPP. Furthermore, our hypothesis may explain the observed connection between microangiopathies and diabetes (Miles *et al*., 2007; Simo *et al*., 2006), suggesting that abnormal angiogenesis may be caused by changes in PPP activity due to the disruption of intracellular glucose homeostasis. Finally, our hypothesis suggests that other proteins of PPP, as well as of TGF-β pathway, may be associated with angiopathies, providing a new set of potential candidates for mutational analysis and drug targeting.

This hypothesis can be readily tested in various experiments, including (i) analysis of redox-state and TGF-beta responses to different levels of glucose concentration in the patients with telangiectasia phenotype carrying TALDO1 and GLUT10 mutations, or in corresponding animal models; (ii) analysis of these responses, as well as general phenotypic effects caused by mutations in other PPP proteins (in animal models or clinical studies); (iii) analysis of TGF-β response to expression changes or up- and down-regulation of PPP proteins *in vivo* and in culture; (iv) sequencing genes of PPP and TGF-β pathway in patients with telangiectasias of unknown genetic background; and (v) further experimental investigation of the role of PPP activity in the association between angiopathies and diabetes, in diabetes patients and animal models.

In conclusion, our manually curated decomposition of disease phenotypes is based on the set of elementary phenotypic features that serve as basis vectors in a multidimensional space, as opposed to previously reported automated (Lage *et al*., 2007; van Driel *et al*., 2006) or semi-automated (Robinson *et al*., 2008) ontologies of phenotypic terms. The potential value of this approach is shown by confirming known and revealing previously unknown gene–phenotype, gene–gene and pathway–pathway associations.

ACKNOWLEDGEMENTS

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources.

Funding: National Institutes of Health (GM67165 to N.V.G.); Welch Foundation (I1505 to N.V.G.); American Cancer Society (to H.T.).

Conflict of Interest: none declared.

REFERENCES

Amberger,J. *et al*. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.

Berg,J. *et al*. (2001) *Biochemistry*. W.H. Freeman & Co, New York.

Coucke,P.J. *et al*. (2006) Mutations in the facilitative glucose transporter GLUT10 alter angiogenesis and cause arterial tortuosity syndrome. *Nat. Genet.*, **38**, 452–457.

Feramisco,J.D. *et al*. (2009) Phenotypic and genotypic analyses of genetic skin disease through the online Mendelian inheritance in man (OMIM) database. *J. Invest. Dermatol.*, [Epub ahead of print, doi: 10.1038/jid.2009.108]

Goh,K.I. *et al*. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Hua,H. *et al*. (2003) High glucose-suppressed endothelin-1 Ca2+ signaling via NADPH oxidase and diacylglycerol-sensitive protein kinase C isozymes in mesangial cells. *J. Biol. Chem.*, **278**, 33951–33962.

Isono,M. *et al*. (2000) Stimulation of TGF-beta type II receptor by high glucose in mouse mesangial cells and in diabetic kidney. *Am. J. Physiol. Renal Physiol.*, **278**, F830–F838.

Jolliffe,I.T. (2002) *Principal Component Analysis.* Springer, New York.

Lage,K. *et al*. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Leopold,J.A. *et al*. (2003) Glucose-6-phosphate dehydrogenase overexpression decreases endothelial cell oxidant stress and increases bioavailable nitric oxide. *Arterioscler. Thromb. Vasc. Biol.*, **23**, 411–417.

Loeys,B.L. *et al*. (2005) A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2. *Nat. Genet.*, **37**, 275–281.

Matsui,R. *et al*. (2005) Glucose-6 phosphate dehydrogenase deficiency decreases the vascular response to angiotensin II. *Circulation*, **112**, 257–263.

Maulik,N. and Das,D.K. (2002) Redox signaling in vascular angiogenesis. *Free Radic. Biol. Med.*, **33**, 1047–1060.

Miles,P.D. *et al*. (2007) Impaired insulin secretion in a mouse model of ataxia telangiectasia. *Am. J. Physiol. Endocrinol. Metab.*, **293**, E70–E74.

Nicotera,T.M. *et al*. (1989) Elevated superoxide dismutase in Bloom's syndrome: a genetic condition of oxidative stress. *Cancer Res.*, **49**, 5239–5243.

Park,J. *et al*. (2006) Increase in glucose-6-phosphate dehydrogenase in adipocytes stimulates oxidative stress and inflammatory signals. *Diabetes*, **55**, 2939–2949.

Rajasekaran,N.S. *et al*. (2007) Human alpha B-crystallin mutation causes oxido-reductive stress and protein aggregation cardiomyopathy in mice. *Cell*, **130**, 427–439.

Robinson,P.N. *et al*. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Rzhetsky,A. *et al*. (2007) Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA*, **104**, 11694–11699.

Shan,Z. *et al*. (1994) Intracellular glutathione influences collagen generation by mesangial cells, *Kidney Int*, **46**, 388–395.

Simo,R. *et al*. (2006) Angiogenic and antiangiogenic factors in proliferative diabetic retinopathy. *Curr. Diabetes Rev.*, **2**, 71–98.

Tille,J.C. and Pepper,M.S. (2004) Hereditary vascular anomalies: new insights into their pathogenesis. *Arterioscler. Thromb. Vasc. Biol.*, **24**, 1578–1590.

Ushio-Fukai,M. (2006) Redox signaling in angiogenesis: role of NADPH oxidase. *Cardiovasc Res.*, **71**, 226–235.

van Driel,M.A. *et al*. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Wiesenfeld,S.L. *et al*. (1970) Elevated blood pressure, pulse rate and serum creatinine in Negro males deficient in glucose-6-phosphate dehydrogenase. *N. Engl. J. Med.*, **282**, 1001–1002.

Wu,X. *et al*. (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.

Zhu,Y. *et al*. (2007) Regulation of transforming growth factor beta in diabetic nephropathy: implications for treatment. *Semin. Nephrol.*, **27**, 153–160.