Genetic Association, Post-translational Modification, and Protein-Protein Interactions in Type 2 Diabetes Mellitus*[S]

Amitabh Sharma‡§¶, Sreenivas Chavali‡§¶, Anubha Mahajan‡¶, Rubina Tabassum‡¶, Vijaya Banerjee‡¶, Nikhil Tandon‖, and Dwaipayan Bharadwaj‡**

Type 2 diabetes mellitus is a complex disorder with a strong genetic component. Inherited complex disease susceptibility in humans is most commonly associated with single nucleotide polymorphisms. The mechanisms by which this occurs are still poorly understood. Here we focus on analyzing the effect of a set of disease-causing missense variations of the monogenetic form of Type 2 diabetes mellitus and a set of disease-associated nonsynonymous variations in comparison with that of nonsynonymous variations without any experimental evidence for association with any disease. Analysis of different properties such as evolutionary conservation status, solvent accessibility, secondary structure, etc. suggests that disease-causing variations are associated with extreme changes in the value of the parameters relating to evolutionary conservation and/or protein stability. Disease-associated variations are rather moderately conserved and have a milder effect on protein function and stability. The majority of the genes harboring these variations are clustered in or near the insulin signaling network. Most of these variations are identified as potential sites for post-translational modifications; certain predictions have already reported experimental evidence. Overall our results indicate that Type 2 diabetes mellitus may result from a large number of single nucleotide polymorphisms that impair modular domain function and post-translational modifications involved in signaling. Our emphasis is more on conserved corresponding residues than the variation alone. We believe that the approach of considering a stretch of peptide sequence involving a polymorphism would be a better method of defining the role of the polymorphism in the manifestation of this disease. Because most of the variations associated with the disease are rare, we hypothesize that this disease is a "mosaic model" of interaction between a large number of rare alleles and a small number of common alleles along with the environment, which is little contrary to the existing common disease common variant model. ***Molecular & Cellular Proteomics 4:1029–1037, 2005.***

Type 2 diabetes mellitus (T2DM)[1] is a genetically heterogeneous, polygenic disease with a complex inheritance pattern and is caused by genetic predisposition and environmental factors. The precise biochemical defects are unknown and almost certainly include impairments in insulin secretion and insulin action. T2DM is characterized by abnormal glucose homeostasis leading to hyperglycemia and is represented primarily by insulin resistance. The vast majority of insulin resistance in T2DM has been shown to arise due to defects at the postreceptor level (1). T2DM is also heterogeneous in the associated pathological and physiological symptoms leading to a variety of complications such as coronary heart disease, neuropathy, retinopathy, etc.

Genetic dissection of any complex trait is done based on two approaches: genome wide scan studies and association studies. The concept of association studies (2) is being widely applied as an experimental technique to identify single nucleotide polymorphisms (SNPs) underlying a complex phenotype, which represents the most common form (90%) of genetic variations in humans (3). Association is defined as a statistical statement about the co-occurrence of alleles or phenotypes. Because of the application of high throughput SNP detection techniques, the number of identified SNPs is growing rapidly, enabling detailed statistical studies. Over the past decade many laboratories have sought to clarify the etiology of T2DM by attempting to associate clear differences in metabolic phenotype with mutations or polymorphisms in the genes. As a result of this a large amount of data has accumulated, associating SNPs in a large number of candidate genes with the disease across different populations.

Unlike fully penetrant mutations that cause Mendelian diseases, SNPs involved in complex human phenotypes are not a necessary and sufficient condition defining the phenotype, but their effect depends on many other genetic and environmental components. In other words SNPs are shown to com-

From the ‡Functional Genomics Unit, Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research (CSIR), Delhi 110 007 and the ‖Department of Endocrinology, All India Institute of Medical Sciences, New Delhi 110 029, India

Received January 20, 2005, and in revised form, May 2, 2005

Published, MCP Papers in Press, May 10, 2005, DOI 10.1074/mcp.M500024-MCP200

[1] The abbreviations used are: T2DM, Type 2 diabetes mellitus; SNP, single nucleotide polymorphism; DCV, disease-causing variation; DAV, disease-associated nonsynonymous variation; PSIC, position-specific independent count; CNV, control nonsynonymous variation; IRS1, insulin receptor substrate 1; PPARγ, peroxisome proliferator-activated receptor γ; G3PD, glyceraldehyde-3-phosphate dehydrogenase; PH, pleckstrin homology; PTB, phosphotyrosine binding; SH, Src homology.

prise risk factors of having a specific phenotype more in a statistical sense. This raises the question as to whether the associated SNPs are only of statistical significance. If not, then what might be the reason for encountering differences in variation statistics across different populations as shown by Cargill *et al.* (4)? However, identifying SNPs responsible for specific phenotypes appears to be an enigma that is very difficult to solve. Several recent studies (5–10) have applied computational methods to predict the potential effects of the nonsynonymous coding SNPs in bringing about variations in humans.

A focus on the individual factors that highlight their maximum potential effect (whether positive or deleterious) is often optimistic because in practice they do not operate in isolation. Instead they work jointly to generate the disease gene architecture, and hence a study to determine the contribution of these interactions toward the disease is essential. Ideally the end point of disease gene identification should be functional analysis of the disease-associated allele and an understanding of the molecular mechanism of causation of the disease phenotype. The functional characterization can be facilitated by the computational analysis. Vitkup *et al.* (9) have shown that the probability of a nonsynonymous mutation causing a genetic disease increases monotonically with an increase in the degree of evolutionary conservation of the mutation site and a decrease in the solvent accessibility of the site; opposite trends are observed for non-disease polymorphisms.

In the current study we extensively analyzed the effect of nonsynonymous variations on the structure and function of proteins and attempted to determine their possible role in the disease phenotype.

EXPERIMENTAL PROCEDURES

Data Set Extraction—The data set considered for the study includes a set of 29 mutations shown to cause monogenetic T2DM in families or maturity onset of diabetes in young (disease causing variations (DCVs)), 113 polymorphisms associated with the disease in various populations in a total of 76 different candidate genes, and 92 random nonsynonymous variations in 32 genes that do not have any experimental evidence of association with any disease as a control data set (Supplemental Table 1). The selection of these random variations would help to distinguish specific behavior patterns of the disease-related variations from that of chance occurrence. Hence these random variations throughout the sequence in those genes that have been implicated with the T2DM were selected. The disease-associated polymorphisms fall into four major categories: nonsynonymous (45 polymorphisms), regulatory (42 polymorphisms), synonymous (11 polymorphisms), and intronic (15 polymorphisms). In this study we determine the effect of the disease-associated nonsynonymous variations (referred to hereafter as DAVs) in comparison to the control nonsynonymous variations (CNVs) on the phenotype. DCVs were obtained by querying Medline for "Type 2 Diabetes, Mutations"; DAVs were obtained by querying for "Type 2 Diabetes, SNPs" and "Type 2 Diabetes, Polymorphisms"; and CNVs were obtained from the Swiss-Prot Database (11). The extraction of protein sequences needed for the analysis of all these variations was done from Swiss-Prot.

The relationship between the genes harboring DAVs was determined using Pathway Assist (12). Pathway Assist is a software application for navigation and analysis of biological pathways, gene regulation networks, and protein interaction maps. It comes with the built-in natural language processing module MedScan and a comprehensive data base.

Evaluating Evolutionary Conservation Status of the Variations—The best method to evaluate the significance of a variation using evolutionary information is to consider the nature of the change with respect to the variability of the affected residue as estimated from the wild type sequences in different proteins of a protein family. A set of similar sequences can be characterized by a multiple sequence alignment within common sequence domains (in case of protein families) or just a small sequence region (motif). We carried out systematic examination of positions of the variations in motif region of proteins using the Pfam data base (13) of probabilistic models of protein domains and families derived using the HMM method and eMATRIX data base (14). eMATRIX (15) is a minimum risk method for estimating the frequencies of amino acids at conserved positions in a protein family. Minimum risk estimation finds the optimal weighting between a set of observed amino acid counts and a set of pseudofrequencies. This provides the information regarding the position of the variations in specific domains and functional motifs, respectively.

The prediction of residue conservation among the homologous proteins was performed by Scorecons (16). Scorecons algorithm scores each residue position with multiple sequence alignment in terms of conservation. Multiple sequence alignment of homologous protein was done by using ClustalW (17) algorithm and was formatted in ClustalX (version 1.81). The mutation matrix of Jones *et al.* (18) was used to determine the likelihood of a particular residue being replaced by another and to calculate a score based on the variability of each position. Normalized Shannon entropy scores for each amino acid position were calculated using the following general formulas (16).

$$C_{\text{ent}} = -\sum_a^k \mathrm{p}_a \log_2 p_a / \log_2 [\min(N, K)] \qquad \text{(Eq. 1)}$$

and

$$p_a = n_a / N \qquad \text{(Eq. 2)}$$

where n_a is the number of amino acid residues of type A, N is the number of residues in the sequence data base, and K is the number of residue types. The program Scorecons (www.biochem.ucl.ac.uk/cgi-bin/valder/Scorecons_server.pl) was used for all calculations. A score of zero indicates a lack of conservation at that position, whereas a score of 1 indicates very high sequence conservation.

Determining the Involvement in Formation of Specific Patterns—Non-conserved residues adjacent to the conserved residues in the primary sequence are generally less substitutable than other non-conserved residues, reflecting their involvement in a functionally important region (19). The peptide sequence containing the variant along with 10 neighboring residues on either side was selected from the protein sequence, and a pattern search was done using PROSITE (20) data base to determine the involvement of the variants in formation of specific patterns. PROSITE consists of biologically significant sites; patterns like phosphorylation, glycosylation, etc.; and profiles that help to reliably identify specific motifs within a peptide sequence.

Sequences involving variants showing potential phosphorylation sites were evaluated for the effect on phosphorylation using NetPhos 2.0. NetPhos 2.0 is an artificial neural network method that predicts phosphorylation sites in independent sequences with sensitivity in the range from 69–96% (21).

Assessing the Effect of Variation on Structural Parameters of Proteins—It is apparent that amino acid allelic variants have an impact on the protein structure and function, and this has been shown to be predicted by the analysis of multiple sequence alignments and protein three-dimensional structures (8). To assess the effect of the variations

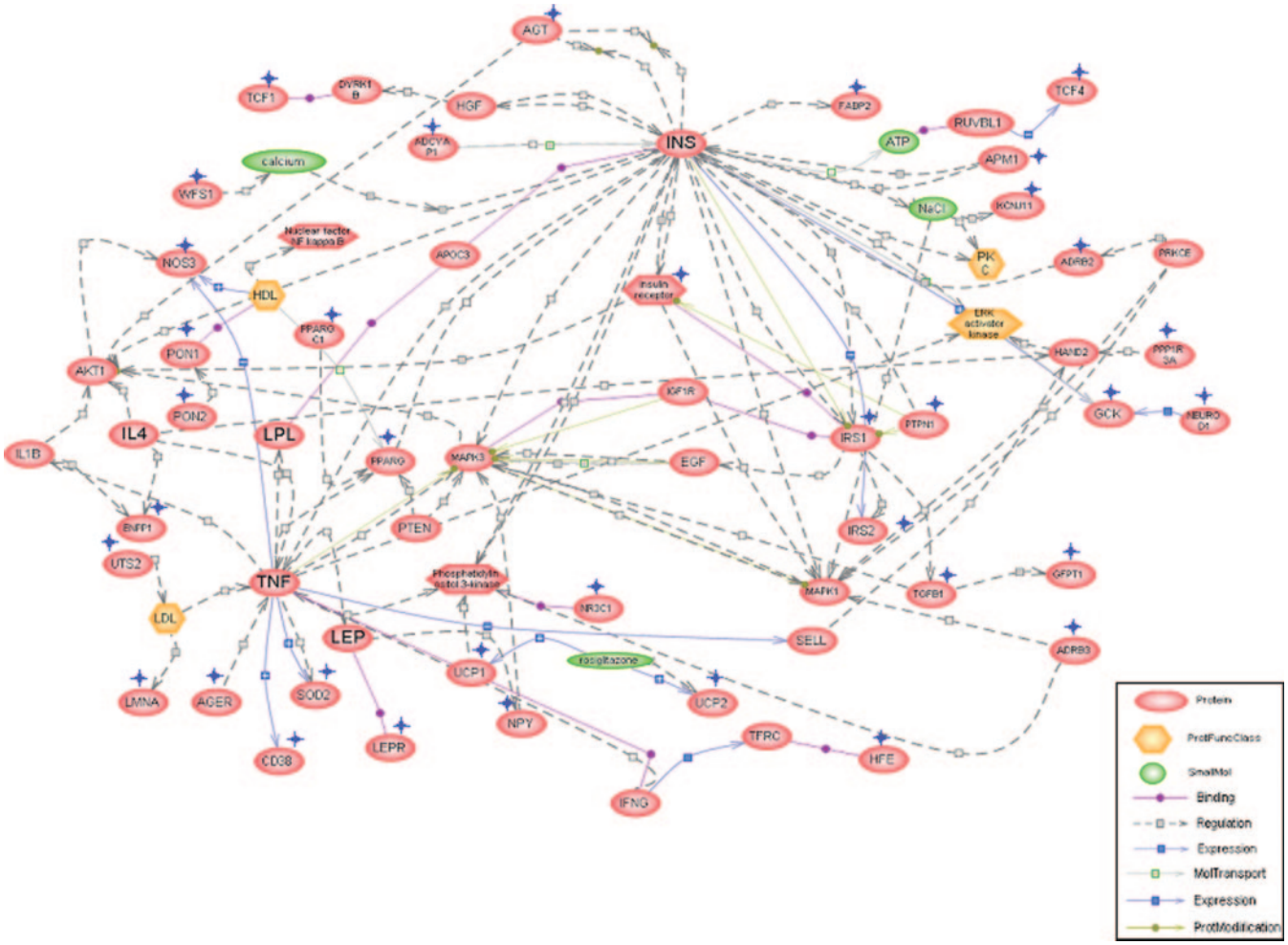

FIG. 1. **Association network among the genes harboring the DAVs associated with T2DM as obtained using Pathway Assist.** *Star* denotes genes harboring DAVs of T2DM. *NEUROD1*, neurogenic differentiation-1; *TCF4*, hepatic nuclear factor 4α; *TCF1*, hepatic nuclear factor 1α; *TNF*, tumor necrosis factor; *HGF*, hepatocyte growth factor; *EGF*, epidermal growth factor; *APOC3*, apolipoprotein C3; *INS*, insulin; *ERK*, extracellular signal-regulated kinase; *MAPK*, mitogen-activated protein kinase; *TGFB*, transforming growth factor β; *PPARG*, PPARγ; *LDL*, low density lipoprotein; *HDL*, high density lipoprotein; *PKC*, protein kinase C; *IFNG*, interferon γ; *IL1B*, interleukin-1β; *IL4*, interleukin-4; *FABP2*, fatty acid-binding protein 2; *GCK*, glucokinase; *NOS3*, nitric-oxide synthase isoform 3; *LPL*, lipoprotein lipase; *SOD2*, superoxide dismutase 2; *LEP*, leptin; *LEPR*, leptin receptor; *NPY*, neuropeptide Y; *Prot*, protein; *Func*, function; *SmallMol*, small molecule; *MolTransport*, molecular transport; *PRKCE*, protein kinase C epsilon; *ADRB3*, β 3 adrenergic receptor; *AGT*, angiotensin I; WFS1, wolframin; *AKT1*, RAC-serine/threonine protein kinase; *UTS2*, urotensin II; *PTEN*, phosphotidyl inositol-3,4,5-triphosphate 3-phosphatase; *LMNA*, lamin A/C; *AGER*, advanced glycosylation end product-specific receptor; *NR3C1*, glucocorticoid receptor; *TFRC*, transferrin receptor; *HFE*, hereditary hemachromatosis protein; *SELL*, selectin L; *ADRB2*, β 2 adrenergic receptor.

on structure and function of proteins PolyPhen (22) was used. PolyPhen is a World Wide Web server to automate functional annotation of nonsynonymous SNPs based on sequence-based characterization of the substitution site and structural parameters. This provides us with the position-specific independent count (PSIC) score calculated from the overall similarity of the sequences that share the amino acid type at this position with the help of statistical concepts and predicts whether a nonsynonymous variation is damaging, *i.e.* is supposed to affect the protein function, or benign, *i.e.* most likely lacking a profound phenotypic effect. Large differences in PSIC values (difference range above 1.5) for specific genetic variants might indicate that the substitution of interest is rarely or never observed in the protein family (23).

Variations in the protein core involving a change in the hydrophobic character of a buried residue may result in different degrees of protein destabilization (24). The hydrophobic effect is measured by solvent-accessible surface area of a protein that is part of a complex surface in direct contact with solvent. Solvent accessibility was predicted using RVPNET (25), which uses single residue information of neighbors and provides real predictions of accessible surface area. Hydrophobic interactions are considered to be the primary factor stabilizing β-sheets (26), therefore identification of secondary structure elements was done by using Chou-Fasman predictions (27).

Statistical Evaluation—To compare the DCVs and DAVs with CNVs during the assessment of their effect on the disease phenotype, χ^2 tests were performed, and p value was calculated.

RESULTS

Pathway Assist analysis established the products of the genes harboring DAVs to be potential interacting members of

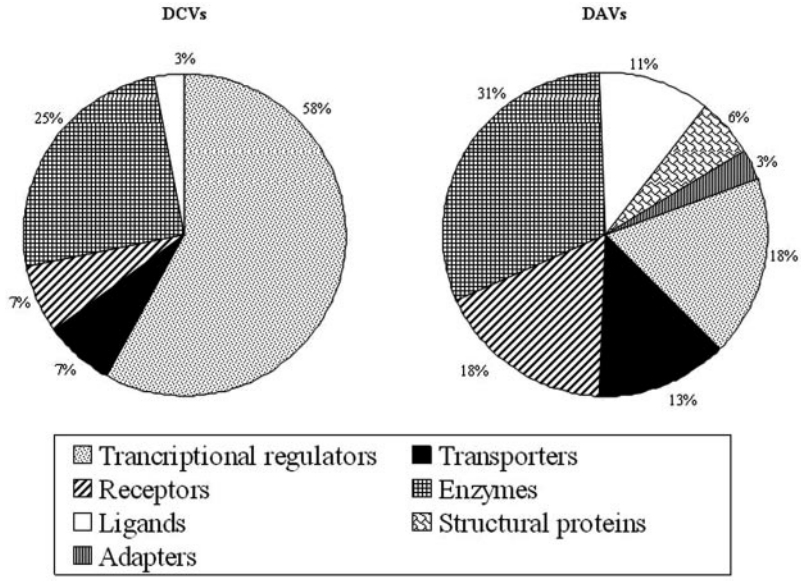

FIG. 2. **Functional classification of proteins harboring DCVs and DAVs of T2DM.** Classification of the proteins into different groups is done on the basis of function assigned in the Swiss-Prot Database.

the insulin signaling cascade (Fig. 1). Nevertheless it is to be noted that Pathway Assist connects any two input proteins, and some of the proteins identified by Pathway Assist during networking of the input proteins might not be involved in Type 2 diabetes as is understood at this point of time. Functional segregation of the proteins harboring the DAVs categorized enzymes as the major class (31%), whereas transcription regulators were the major class harboring DCVs (58%) (Fig. 2). Pfam analyses showed that most of the DCVs (67%), 49% of DAVs, and 63% of CNVs correspond to the functional domains of respective proteins (Supplemental Table 2). Therefore, of the total variations, an average of 56% lie in functional domains of proteins ($p = 0.02$). Furthermore in the total sequence space of the identified proteins, 60% is occupied by functional domains. eMATRIX analysis revealed that the majority of DCVs (50%) and DAVs (62%) corresponded to functional signatures in comparison to only 27% of CNVs. This clearly indicates that the DCVs and DAVs correspond significantly more to the functional signatures in comparison with the randomly picked CNVs ($p = 0.0002$). Scorecons analysis (Fig. 3) revealed that DCVs are more conservative changes (90% above the value of 0.5), whereas DAVs are radical (56% above 0.5) in comparison with CNVs (47% above 0.5), which are mostly changes in variable regions with a low Scorecons value ($p = 0.0003$).

Most of the patterns obtained from PROSITE for DCVs (51.7%) and DAVs (51.1%) represented consensus post-translational modification motifs for phosphorylation, glycosylation, and myristoylation (Supplemental Table 2) in contrast to only 37% of CNVs. Few peptides showed more than one post-translational motif. Phosphorylation changes predicted by NetPhos 2.0 for the patterns indicated a probable decrease in the phosphorylation of DCV T608R of IRS1 (<0.5) and DAVs P387L of protein tyrosine phosphatase 1B, D905Y of protein phosphatase, and P115Q of PPARγ, whereas in CNVs the difference was negligible except for A288T of lipoprotein lipase. In the DCV F635S of G3PD phosphorylation was strongly predicted for the Ser variant, whereas the phosphorylation site was absent when Phe was present (Fig. 4). However, certain predictions made by PROSITE and NetPhos 2.0 might not be logical because these are generated preserving certain statistical properties of biological sequences, emphasizing the need for experimental data to determine the accuracy of these predictions.

Experimental evaluations yielded interpretations similar to those of the predictions for P115Q of PPARγ (28), P387L of protein tyrosine phosphatase 1B (29), and T608R of IRS1 (30), thus providing confidence to the analysis. Ristow *et al.* (28) have reported that overexpression of the mutant PPARγ containing Gln-115 in murine fibroblasts led to the production of a protein in which the phosphorylation of Ser-114 was defective as well as to accelerated differentiation of the cells into adipocytes and greater cellular accumulation of triglyceride than with the Pro-115 (wild type PPARγ). These effects were similar to those of an *in vitro* mutation created directly at the Ser-114 phosphorylation site. Echwald *et al.* (29) have investigated the incorporation of γ-^{32}P-labeled radioactivity into the wild type peptide (Pro-387) *versus* the mutant peptide (Leu-387) of protein tyrosine phosphatase 1B in an *in vitro* kinase assay and reported that replacement of Pro by Leu reduced the phosphorylation of Ser-386 by 70%. Using NetPhos 2.0 we predicted that the likelihood of Ser-386 being a phosphorylation target was reduced from 0.89 to 0.13 when Pro-387 was replaced with Leu. These studies emphasize the importance of conserved residues adjoining the post-translational modification sites. Esposito *et al.* (30) have shown that when cells transfected with IRS1 Arg-608 were stimulated with insulin p85 coimmunoprecipitated with IRS1, and the

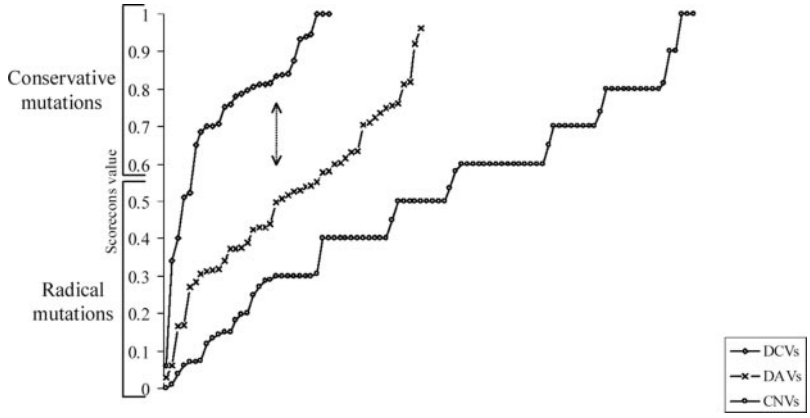

FIG. 3. **Scorecons analysis for DCVs, DAVs, and CNVs.** Entropy scores for each variant residue were calculated based on the multiple sequence alignment done on ClustalX (version 1.81). Higher Scorecons values correspond to lower relative entropy and hence higher conservation, whereas lower scores correspond to increasing sequence variability.

FIG. 4. **Sequence alignment of the protein sequence of G3PD.** Shown is the alignment of the peptide sequence (21-mer) encompassing the DCV F635S of G3PD. Sequence alignments were carried out using the ClustalW algorithm with the homologous sequences extracted from PolyPhen server. Strong phosphorylation is predicted for the Ser by NetPhos 2.0. *GPDM*, glyceraldehyde-3-phosphate dehydrogenase, mitochondrial.

associated phosphatidylinositol 3-kinase activity was ~50% less than that of cells transfected with IRS1 Thr-608 (wild type).

PSIC score calculation using PolyPhen (Fig. 5) illustrated that 88% of DCVs, 55% of DAVs, and 41% of CNVs lie in the range of 1.1–3.5 PSIC values, indicating that DCVs and DAVs are observed in significantly conserved positions in the protein families as compared with CNVs ($p = 0.0015$). Considering these values along with the biochemical and protein structure characteristics, PolyPhen predicted that 72% of DCVs are potentially damaging compared with 71% of DAVs and 80% of CNVs, which were benign, *i.e.* most likely lacking any profound effect. Combining the domain assignments from Pfam and predictions of PolyPhen, DCVs, DAVs, and CNVs can be categorized as domain-damaging (50, 20 and 10%, respectively) and domain-benign (14, 35, and 45%, respectively). Of those DCVs, DAVs, and CNVs that did not lie in a domain, 22, 9, and 10%, respectively, were damaging, whereas 14, 36, and 35%, respectively, were benign. Hence it is predicted that DCVs affect the function of the domains in which they lie to a large extent, whereas DAVs have a milder effect.

RVPNET predictions of solvent accessibility revealed that 62% of DCVs involved the residues affecting buried sites and thus the hydrophobicity of the functional molecule. In comparison, DAVs occurred mostly in exposed sites, and even when present in the buried site the change did not affect hydrophobicity. Compilation of the residue changes shows Arg, Thr, and Ile to be the most common DCVs (15.4% each), whereas in DAVs, Arg was the most common susceptible variation (15.6%), but Met was the most common variant (14%) in CNVs (Fig. 6). Interestingly Arg has been shown to be an energetic hotspot that is critically important for the affinities of a protein interface (31).

Chou-Fasman predictions revealed that most of the DCVs (47%) and DAVs (42%) lie in helices compared with only 27% of CNVs. Sheets harbor 24, 33, and 35%, and turns harbor 19, 17, and 6% of DCVs, DAVs, and CNVs, respectively. The structures for about 10% of DCVs, 8% of DAVs, and 32% of CNVs could not be predicted. The higher representation of CNVs in the "could not be predicted" class emphasizes the randomness of the variations selected as CNVs. Thus, most of the DCVs and DAVs of T2DM appear to lie in the helices contrary to the earlier reports of their concentration in the β-sheets (5, 6). A major role of the persistently conserved residues at the N and C termini of helices and sheets in helix initiation and termination has already been well established (32). About 17% of DCVs and 29% of DAVs that lie in helices appear to occur in these regions. Comparison of predictions with the available Protein Data Bank structures for proteins revealed that our secondary structure predictions are nearly 70% reliable.

Analysis of nucleotide substitutions in DCVs and DAVs indicated that in both cases the first base in the codon had undergone maximal change (58%). Of the total base substitutions 35% in DCVs and 41% in DAVs favor fixation of A. However, in the non-coding regions the gene conversion is biased toward G and C in accordance with earlier reports (33).

DISCUSSION

Type 2 diabetes mellitus is a multifactorial disorder involving a plethora of contributing factors with a strong genetic component as proved from twin studies (36). The widely practiced method to genetically dissect this disorder has been to perform association studies at a population level. Often as-

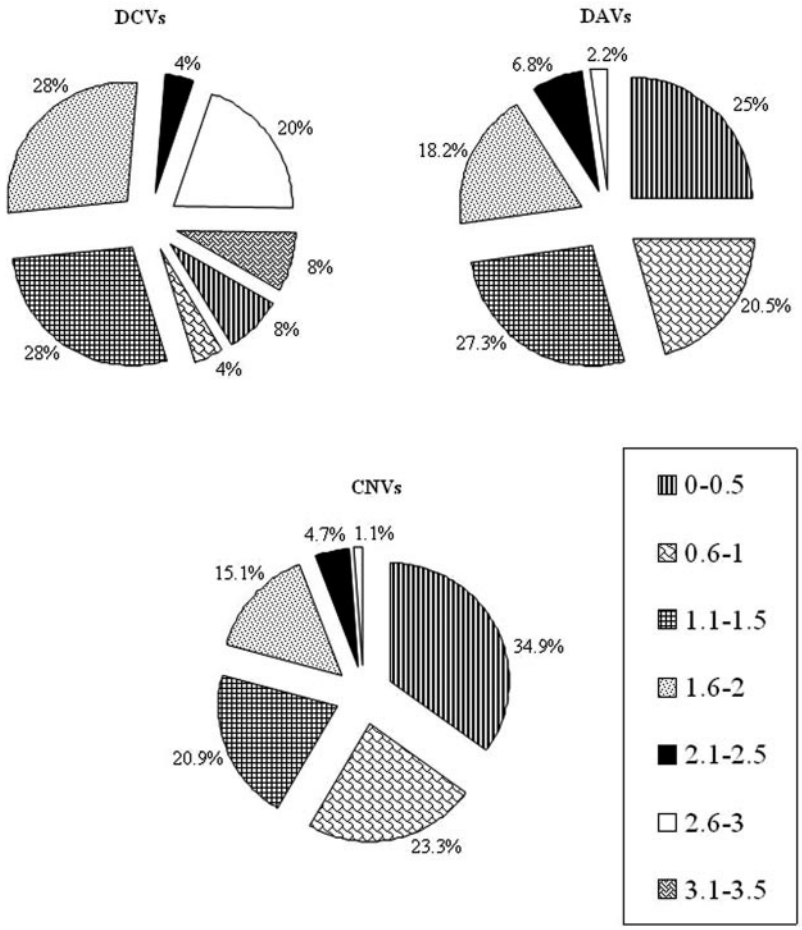

FIG. 5. **PSIC score for DCVs, DAVs, and CNVs from PolyPhen analyses.** Elements of the matrix (profile scores) are logarithmic ratios of the likelihood of a given amino acid occurring at a particular site to the likelihood of this amino acid occurring at any site.

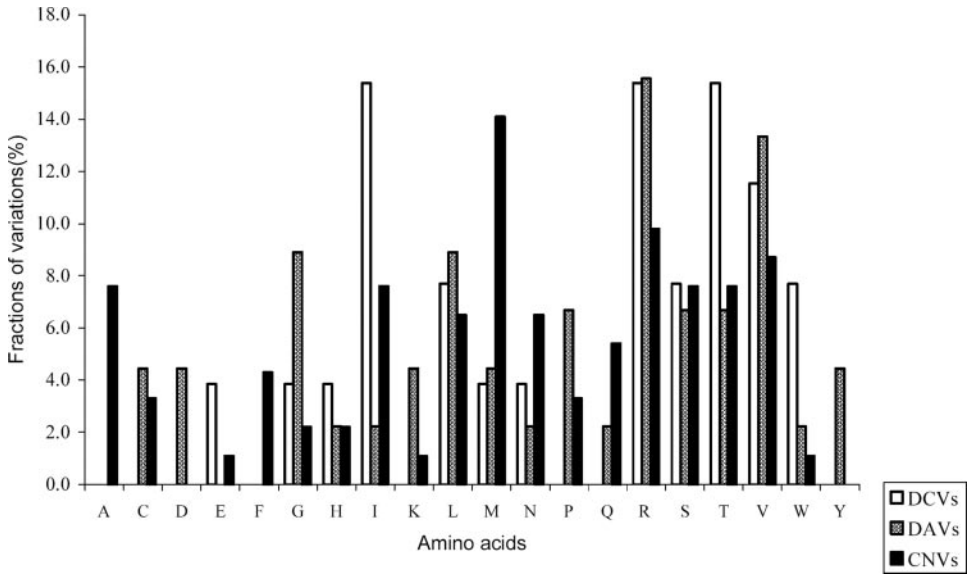

FIG. 6. **Contribution of variations at different amino acids to the overall spectrum.** The fraction of susceptible variations for categories DCVs, DAVs, and CNVs (normalized to 100% within each class) is shown.

sociation of alleles is dealt at the DNA level, and the results obtained are interpreted in a statistical sense. Here we report a systematic evaluation of determining the role of these variations in affecting protein structure and function because knowledge of these provide a more rational approach to the fight against the disease. A simple random model in which

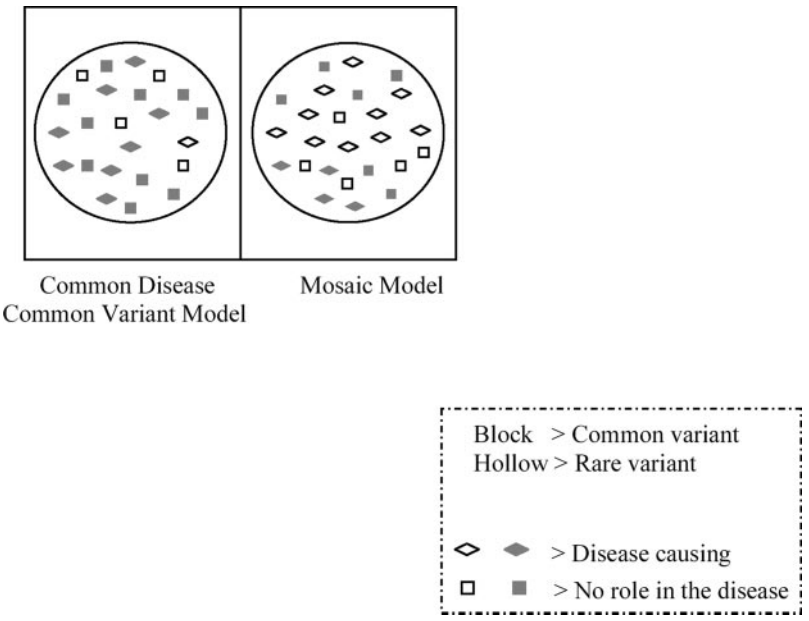

FIG. 7. **Allelic spectrum of Type 2 diabetes mellitus.** *Left*, common disease common variant model: the interaction of a large number of common variants at different loci to manifest the disease. *Right*, our hypothesis: mosaic model of interaction between a small number of common variants with a large number of rare variants.

variations at any position on the set of proteins found not to be associated to disease was used as a control data set to evaluate and distinguish our observations on the DCVs and DAVs from that occurring purely by chance.

The presence of most of the DCVs in the domains and their high Scorecons values and PSIC scores indicate that these lie at highly conserved sites, which have high selection pressure. The occurrence of most of these variations at the buried sites proves that the changes affect the protein stability and are deleterious. Thus it is comprehensible that these would be damaging (as predicted by PolyPhen) and individually have the ability to cause the disease.

The majority of the DAVs lie in the functional signatures, and in pattern analysis it is evident that most of them exist in consensus post-translational modification patterns. The relative entropy data from Scorecons clearly indicate that the DAVs lie in moderately variable positions, but it is interesting to note that the corresponding amino acids near the variants have low relative entropies, meaning high conservation. Some of the well characterized protein interaction domains involved in insulin signaling include the PH, PTB, SH2, SH3, LIM, PDZ, NOTCH, and WW. These interaction domains are either created by post-translational covalent modification of the protein, which includes phosphorylation of exposed Tyr, Ser, or Thr residues, or they exist in the natural tertiary structure of proteins (34) that is highly influenced by the core residues. Pattern search revealed that as many as 51% of DAVs appear to occur in potential post-translational modification sites. Certain phosphorylation changes predicted as an effect of these variations also have experimental evidence (28–30).

Our emphasis is more on conserved neighboring residues than the variation alone because the specificity of protein kinases is dominated by acidic, basic, or hydrophobic residues adjacent to the phosphorylated residue. We have also shown that the major class of proteins harboring DAVs is enzymes, the most common susceptible amino acid change is arginine, and most of the DAVs are exposed. Pathway Assist networking converges all the genes toward the insulin signaling cascade. It is well known that in any signaling cascade the cross-talk between proteins to bring about transduction is done mostly through post-translational modification mechanisms. All these facts provoke us to propose that T2DM must mostly be a result of disturbed protein-protein interactions.

Low PSIC scores and prediction of most of the DAVs as benign by PolyPhen indicate that the contribution of DAVs is more in terms of quantitative nature than qualitative in bringing about the disease, although no clear cut mechanism of measuring the extent of contribution is known as of today. We believe that the approach of considering a stretch of peptide sequence involving polymorphisms would be a better method of defining the role of a polymorphism in the manifestation of this disease. This is relevant when considering the possibility of using these properties to predict the pathological character of a mutation from only the knowledge of protein sequence. This in turn would help in the appropriate selection of polymorphisms for association studies adding value to the hunt for pieces of this complicated puzzle.

The control data set was important for this study because it contains mostly variations with no evidence of disease-causing or strong phenotypic changes. Although it is possible that some of the nonsynonymous SNPs in this set may turn out to be disease-related if more vigorous genetic, biochemical, and clinical studies are carried out, we expect that the variations contained in CNVs have little phenotypic changes. Our study suggests that there is a significant statistical difference in distribution of properties between disease-related variations and CNVs. It is provocative to say that the evaluation of these

parameters can help us predict whether newly discovered nonsynonymous variations would be associated with the disease or not.

Initial association studies led to the proposition that T2DM is an example of common disease common variant hypothesis (35) (well in accordance with the thrifty genotype hypothesis). Our observations reveal that more than 80% of the DAVs associated with T2DM are rare variants. Of the overall 113 polymorphisms considered in the data set, 70% of SNPs associated with the disease qualify as rare variants, whereas the rest are common variants. Based on this, we propose that T2DM fits a mosaic model (Fig. 7), which results from complex interaction between a low number of rare alleles from a large number of loci with a low number of common alleles at a small number of loci and the environment. This also implies that the disease phenotype has evolved recently. One of the main reasons for the failure to replicate these associations across different populations must be variable expressivity resulting from selection pressure that has occurred in accordance with the temporal differences in lifestyle adoption. These findings may complicate the understanding of already complicated complex disorder T2DM gene hunting because the occurrence of many rare variants would create statistical disturbances. However, our findings would help researchers to look from new perspectives in a highly competitive field.

Our study suggests that some DAVs in the protein coding regions are not of mere statistical significance alone but may have functional relevance with regard to their effects on protein-protein interactions through disrupting modular domain interactions of motifs involved in post-translational modifications. Existing experimental evidence and data that are rapidly emerging from research usually done as isolated case studies are providing a sound platform in establishing this paradigm. Coding sequence changes are not the only candidates for functional variation, and SNPs in proximal regulatory regions can have a large impact. A complete evaluation of the noncoding disease-associated SNPs would help us to better understand the complex interaction between different genes and their products in the manifestation of this multifactorial disorder.

Acknowledgments—We are extremely thankful to Dr. Vani Brahmachari (Dr. B. R. Ambedkar Centre for Biomedical Research), Prof. S. K. Brahmachari, Director Institute of Genomics and Integrative Biology, for support and critical comments on the manuscript and Tamal Das for computational assistance. We thank the anonymous reviewer for the critical evaluation and helping us improve the scientific impact of our work.

* This work was done under the project "Genetic and Proteomic Studies in Diabetes Mellitus" (Grant OLP0030) funded by CSIR. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ Both authors contributed equally to this work.

¶ Supported by CSIR, Government of India predoctoral fellowship.

** To whom correspondence should be addressed: Functional Genomics Unit, Inst. of Genomics and Integrative Biology (CSIR), Mall Rd., Delhi 110 007, India. Tel.: 91-11-2766-6156/6157; Fax: 91-112766-7471; E-mail: db@igib.res.in.

REFERENCES

1. Korc, M. (2003) Diabetes mellitus in the era of proteomics. *Mol. Cell. Proteomics* **2,** 399–404
2. Risch, N., and Merikangas, K. (1996) The future of genetics studies of complex human diseases. *Science* **273,** 1516–1517
3. Collins, F. S., Brooks, L. D., and Chakravarti, A. A (1998) DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8,** 1229–1231
4. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. O., and Lander, E. S. (1999) Characterization of single nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22,** 231–238
5. Sunyaev, S., Ramensky, V., and Bork, P. (2000) Towards a structural basis of human nonsynonymous single nucleotide polymorphisms. *Trends Genet.* **16,** 198–200
6. Wang, Z., and Moult, J. (2001) SNPs, protein structure and disease. *Hum. Mutat.* **7,** 263–270
7. Chasman, D., and Adams, R. M. (2001) Predicting the functional consequences of nonsynonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307,** 683–706
8. Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* **315,** 771–786
9. Vitkup, D., Sander, C., and Church, G. M. (2003) The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* **4,** R72
10. Stitziel, N. O., Tseng, Y. Y., Pervouchine, D., Goddeau, D., Kasif, S., and Liang, J. (2003) Structural location of disease-associated single-nucleotide polymorphisms. *J. Mol. Biol.* **327,** 1021–1030
11. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370
12. Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. (2003) Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* **19,** 1–3
13. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002) The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280
14. Bennett, S. P., Lu, L., and Brutlag, D. L. (2003) 3MATRIX and 3MOTIF: a protein structure visualization system for conserved sequence. *Nucleic Acids Res.* **31,** 3328–3332
15. Wu, T. D., Nevill-Manning, C. G., and Brutlag, D. L. (1999) Minimal-risk scoring matrices for sequence analysis. *J. Comput. Biol.* **6,** 219–235
16. Valdar, W. S. (2002) Scoring residue conservation. *Proteins* **48,** 227–241
17. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, a position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680
18. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8,** 275–282
19. Guo, H. H., Choe, J., and Loeb, L. A. (2004) Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9205–9210
20. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30,** 235–238
21. Blom, N., Gammeltoft, S., and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294,** 1351–1362
22. Ramensky, V., Bork, P., and Sunyaev, S. (2002) Human nonsynonymous

SNPs: server and survey. *Nucleic Acids Res.* **30,** 3894–3900
23. Rebbeck, T. R., Spitz, M., and Wu, X. (2004) Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet.* **5,** 589–597
24. Pakula, A. A., and Sauer, R. T. (1990) Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. *Nature* **344,** 363–364
25. Ahmad, S., Gromiha, M. M., and Sarai, A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics* **19,** 1849–1851
26. Minor, D. L., and Kim, P. S. (1994) Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367,** 660–663
27. Chou, P. Y., and Fasman, G. D. (1974) Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* **13,** 211–222
28. Ristow, M., Muller-Wieland, D., Pfeiffer, A., Krone, W., and Kahn, C. R. (1998) Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. *N. Engl. J. Med.* **339,** 953–959
29. Echwald, S. M., Bach, H., Vestergaard, H., Richelsen, B., Kristensen, K., Drivsholm, T., Borch-Johnsen, K., Hansen, T., and Pedersen, O. (2002) A P387L variant in protein tyrosine phosphatase-1B (PTP-1B) is associated with type 2 diabetes and impaired serine phosphorylation of PTP-1B in vitro. *Diabetes* **51,** 1–6
30. Esposito, D. L., Li, Y., Vanni, C., Mammarella, S., Veschi, S., Della Loggia, F., Mariani-Costantini, R., Battista, P., Quon, M. J., and Cama, A. (2003) A novel T608R missense mutation in insulin receptor substrate-1 identified in a subject with type 2 diabetes impairs metabolic insulin signaling. *J. Clin. Endocrinol. Metab.* **88,** 1468–1475
31. Bogan, A. A., and Thorn, K. S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280,** 1–9
32. Friedberg, I., and Margalit, H. (2002) Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. *Protein Sci.* **11,** 350–360
33. Marais, G. (2003) Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* **19,** 330–338
34. Virkamaki, A., Ueki, K., and Kahn, C. R. (1999) Protein-protein interaction in insulin signaling and the molecular mechanisms of insulin resistance. *J. Clin. Investig.* **103,** 931–943
35. Reich, D. E., and Lander, E. S. (2001) On the allelic spectrum of human disease. *Trends Genet.* **17,** 502–510
36. Newman, B., Selby, J. V., King, M. C., Slemenda, C., Fabsitz, R., and Friedman, G. D. (1987) Concordance for Type 2 (non-insulin dependent) diabetes mellitus in male twins. *Diabetologia* **30,** 763–768