Towards a structural basis of human non-synonymous single nucleotide polymorphisms

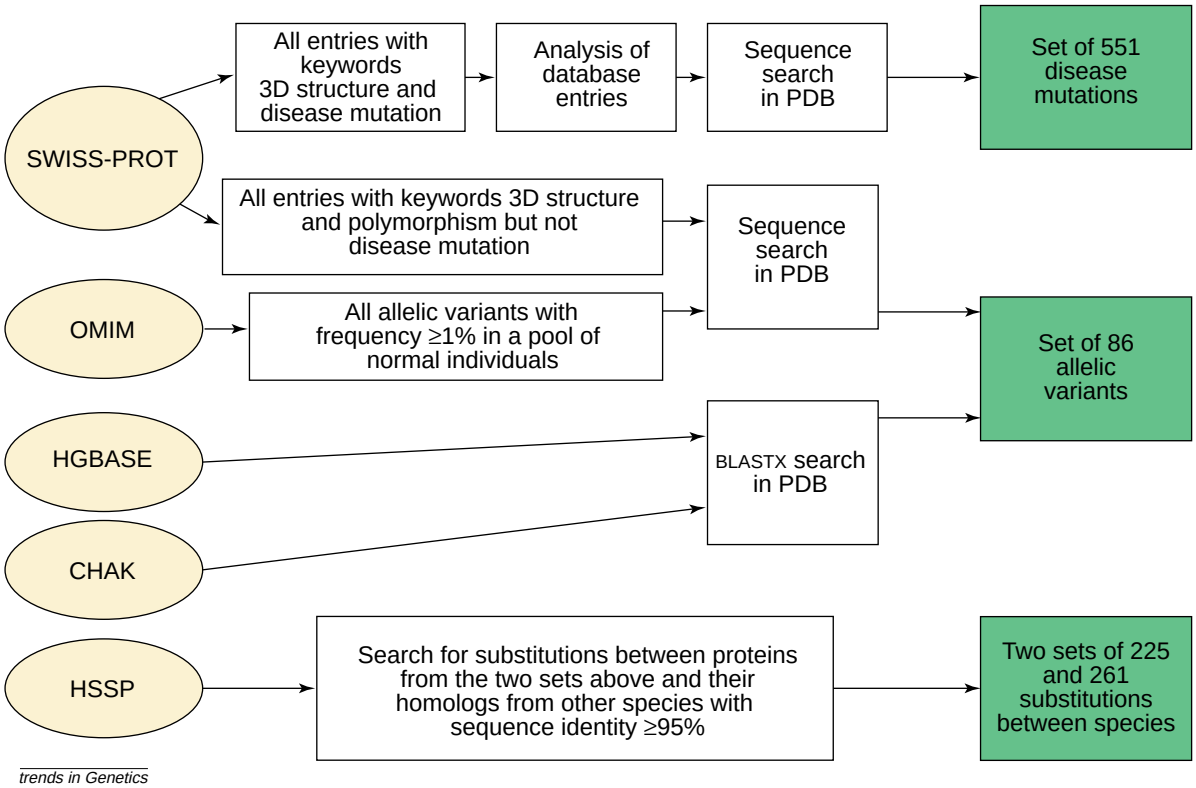
Shamil Sunyaev[+*¶] sunyaev@embl-heidelberg.de
Vasily Ramensky[+*¶] ramensky@imb.ac.ru
Peer Bork[+*] bork@embl-heidelberg.de

[+]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany.
[*]Max Delbrueck Center for Molecular Medicine (MDC), Robert-Roessle-Strasse 10, D-13122 Berlin, Germany.
[¶]Engelhardt Institute of Molecular Biology, Vavilova 32, 117984 Moscow, Russia.

About 90% of human genetic variation has been ascribed to single nucleotide polymorphism (SNP) allelic variants that occur at a frequency of >1% (Ref. 1). Owing to the application of high-throughput SNP detection techniques, the number of identified SNPs is growing rapidly, enabling detailed statistical studies[2–5]. These include studies of SNPs that affect the amino acid sequence of a gene product (non-synonymous SNPs); they complement the large body of literature on mutations that cause mendelian diseases, which represent the usually rare non-synonymous mutations with an allele frequency far below one percent[3].

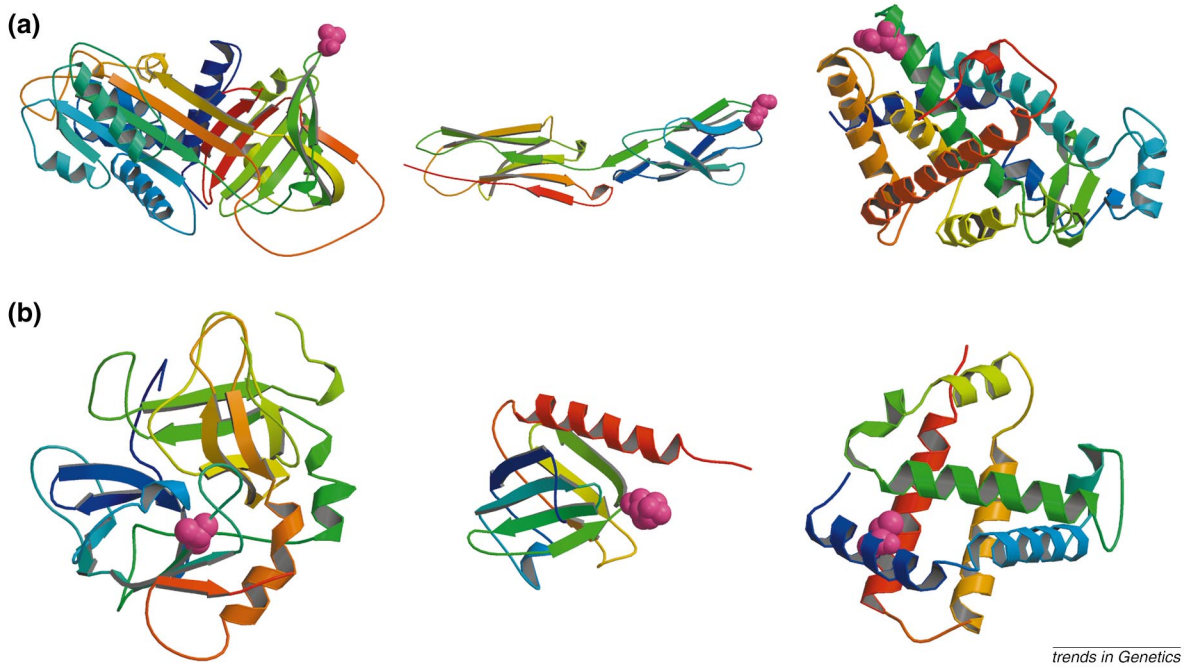
To understand the relationship between genetic and phenotypic variation, it is essential to assess the structural consequences of the respective non-synonymous mutations in proteins. To quantify how often a disease phenotype can be explained by a destructive effect on protein structures or functions, we have mapped known disease mutations onto known three-dimensional structures of proteins. The results were compared with a control set of substitutions observed between these proteins and their closely related homologs from other species that are unlikely to cause severe effects on the phenotype. With the knowledge about the structural properties of these two sets, we have also mapped a large number of non-synonymous SNPs (which are usually thought to be neutral[6], or to be the cause of only minor phenotypic effects) onto protein structures. This enables us to identify non-synonymous SNPs with

FIGURE 1. Protocol for mutation data mining



The set of disease-causing mutations in human proteins with known structures was extracted from the SWISS-PROT database by keyword search followed by an analysis of the references to the Online Mendelian Inheritance in Man[13] (OMIM) database. Four sources were used to form a set of human allelic variants [(non-synonymous single nucleotide polymorphisms (SNPs)]. They were extracted by keyword searches from SWISS-PROT (Release 38), OMIM (July 1999), Human Genic Bi Allelic Sequences (HGBASE, Release 4)[14] and from a set of SNPs reported by Halushka *et al.*[2] via the respective website of the Chakravarti group (shown on the diagram as CHAK). The variation sites identified were mapped onto the respective protein structures via detection of identical human protein sequence in Protein Data Bank (PDB)[15]. Two sets of substitutions between human proteins and homologs from closely related species were compiled on the basis of the Homology Derived Secondary Structure Of Proteins (HSSP) database[12]. The first set represents the between-species variation for proteins with mapped disease-associated mutations, whereas the second set reflects this variation for proteins with mapped non-synonymous SNPs.

FIGURE 2. Polymorphic sites for the most rare and the most frequent variants



Structural locations of variable residues are shown for selected examples of (a) the most frequent (minor allele frequency >20%) and (b) the most rare (minor allele frequency <3%) variants. The MOLSCRIPT[16] and RASTER3D[17] programs have been used to prepare the cartoons of protein structures, which are colored in 'rainbow' colors from N- (red) to C- (violet) termini. Polymorphic sites are shown in pink. The following Protein Data Bank (PDB) entries contain structures presented in the figure. Upper row from left to right: 2psi (α1-antitrypsin); 1iam (intercellular adhesion molecule-1); 4prgA (peroxisome proliferator activated receptor γ). Lower row from left to right: 1qfkH (coagulation factor viia, heavy chain); 1irsA (insulin receptor substrate 1); 1a00A (deoxyhemoglobin a). These sites reflect the tendency of frequent variants to occur on the protein surface and of rare variants to occur in the hydrophobic core; 1irsA as an outlier highlights that it is only a trend and not a rule. In total, in the subset of non-synonymous SNPs with known allele frequencies the fraction of rare (minor allele frequency between 1% and 5%) variants located in structurally and functionally important sites is 62%. The corresponding fraction of frequent (minor allele frequency >5%) variants is only 33%.

structural locations avoided in the set of substitutions between species because of selection over evolutionary time (although these substitutions are frequent in the set of disease-causing mutations). The number of polymorphic sites with such structural locations would give a lower limit estimate for the quantity of non-synonymous SNPs that might have phenotypic effects, providing an important baseline for current efforts to identify SNPs that are associated with multifactorial human disorders[7–9].

The extraction of the three data sets needed for the comparative analysis, (1) disease-causing mutations, (2) substitutions between close homologs in human and other species, and (3) human non-synonymous SNPs from public databases, is detailed in Fig. 1. Structural characteristics were extracted from STRIDE[10], SWISS-PROT[11] and Homology-Derived Secondary Structure of Proteins (HSSP)[12] databases.

As a result of the comparison of disease-causing mutations with between-species substitutions in the same set of proteins, we found that disease-causing mutations are much more likely to occur at sites with low solvent accessibility. In fact, 35% of 551 disease-causing mutations from our dataset affect buried sites, whereas only 9% of 225 substitutions between species do. This indicates that disease-causing mutations often affect intrinsic structural features of proteins. To increase the discrimination between the two sets, we also took into account possible interaction sites. Overall, ~70% of the disease-causing mutations are located in sites likely to be structurally and functionally important, namely sites with <5% solvent accessibility or in β-strands, active sites, sites involved in disulphide bonds or evolutionarily conservative sites (defined as sites with HSSP variability parameter VAR <10). By contrast, in the same set of proteins only 17% of substitutions observed between human sequences and closely related homologs from mammalian species are located at these sites.

Unexpectedly, the fraction of polymorphic sites located in structurally and functionally important regions (as described above) was 45%, which is significantly higher than the 24% in the case of the interspecies variation when considering proteins from the dataset of polymorphic sites (P value of the χ^2 test = 0.00013. In this set, we observe the abundance of proteins with high β-strand content; this explains the 17% vs 24% difference for two protein sets). This result suggests that a significant fraction of human protein allelic variants is represented by amino acid substitutions that might have a strong impact on protein structure, function, stability or folding. These variants are normally eliminated over long evolutionary periods, as can be seen from the comparison with the interspecies variation. One would expect that variants under pressure of purifying selection would have a lower allele frequency[6]. Indeed, for non-synonymous SNPs, we observe a correlation between allele frequency and fraction of occurrence in structurally and functionally important regions (examples shown in Fig. 2). The observation that many non-synonymous SNPs might have a phenotypic effect could be considered as indirect evidence that common amino acid variants contribute to genetic risk of common

human disorders (the so-called common disease–common variant hypothesis[7–9]).

In summary, there is a surprisingly high fraction of non-synonymous SNPs that affect the structure and, probably, function of proteins. This implies that a considerable fraction of the non-synonymous SNPs do indeed have some (probably negative) effect on phenotype. The allele frequency distribution makes it evident that variants in structurally important sites are not selectively neutral. Taking these observations into account, and given the progress in structural genomics and in large-scale SNP discovery, the comparative analysis of structural properties of protein allelic variants, such as the one described here, should have an important role to play in the pre-selection of candidates for disease-association studies and in the explaining of phenotypic effects.

References

1 Collins, F.S. *et al.* (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* 8, 1229–1231
2 Halushka, M.K. *et al.* (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22, 239–247
3 Cargill, M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238
4 Buetow, K.H. *et al.* (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* 21, 323–325
5 Sunyaev, S.R. *et al.* Individual variation in protein-coding sequences of the human genome. *Adv. Protein Chem.* (in press)
6 Li, W.H. (1997) *Molecular Evolution*, Sinauer
7 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517
8 Collins, F.S. *et al.* (1997) Variations on a theme: cataloging human DNA sequence variation. *Science* 278, 1580–1581
9 Lander, E.S. (1996) The new genomics: global views of biology. *Science* 274, 536–539
10 Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23, 566–579
11 Bairoch, A. and Apweiler, R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* 27, 49–54
12 Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56–68
13 Online Mendelian Inheritance in Man, OMIM™ (1999) Center for Medical Genetics, Johns Hopkins University, Baltimore, MD, and National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD. (http://www.ncbi.nlm.nih.gov/omim)
14 Brookes, A.J. *et al.* (2000) HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res.* 28, 356–360
15 Abola, E.E. *et al.* (1997) Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol.* 277, 556–571
16 Kraulis, P.J. (1991) MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 24, 946–950
17 Merritt, E.A. and Bacon, D.J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol.* 277, 505–524

Chromatin becomes hot in cancer pathogenesis and therapy

AACR Special Conference in Cancer Research on Transcription Factor Pathogenesis of Cancer at the Millennium, Dana Point, California, 26–30 January 2000

Although the role of chromatin structure in gene transcription seems obviously important, it became a hot topic in the transcription field only in recent years. Not surprisingly, perhaps, the study of chromatin-mediated transcription in cancer is now at the forefront of cancer research. Chromatin structure changes can be achieved by several different events, including post-translational modifications of histones, DNA methylation, remodeling protein complexes, and propagation of eu- or hetero-chromatin states by the trithorax and polycomb groups of genes. In this first meeting on transcription factor and cancer at the new millennium, exciting progress was reported in understanding the role of each of these chromatin events in cancer pathogenesis and treatment.

Many transcription factors recruit histone acetylation or deacetylation enzymes in their tumorigenic actions. The study of the retinoic acid receptor, RARα, in the pathogenesis of acute promyelocytic leukemia (APL) revealed that transformation might result from enhanced deacetylation by the translocation fusion protein PML–RAR (Ron Evans, Salk Institute, La Jolla, USA; Pier Paolo Pandolfi, Memorial Sloan-Kettering Cancer Center, New York, USA; Samuel Waxman, Mount Sinai School of Medicine, New York, USA). Accordingly, histone deacetylase inhibitors (HDACI) showed great promise in treating APL. However, side effects on normal gene expression was a concern raised by several meeting participants and remains to be addressed by ongoing studies.

DNA methylation plays an important role in inactivation of tumor suppressor genes and DNA repair genes. Methylation is also known to facilitate histone deacetylation, and direct interaction between 5-methylcytosine transferase (dnmt1) and HDAC1 has recently been demonstrated. Several speakers discussed the link between methylation and cancer. The study of *fos* oncogene transformation mechanisms revealed an important target gene of *fos* as *dnmt1*. Fos-transformed cells had 3-fold elevated levels of *dnmt1* and contained ~20% more 5-methylcytosine than normal fibroblasts. Importantly, transfection of the *dnmt1* gene induced morphological transformation, whereas inhibition of *dnmt1* expression or activity resulted in reversion of transformation (Tom Curran, St Jude Children's Research Hospital, Memphis, USA). In addition, histone deacetylation was shown to be important for Fos-mediated transformation. Novel cancer therapy approaches were discussed that were based on overcoming transcription repression in cancer cells (Samuel Waxman). A combination of HDACIs with azacytidine and retinoic acid appeared effective in reactivating RARβ expression in breast cancer cell lines. The fact that none of these compounds were effective alone provided evidence for the close cooperation between methylation and deacetylation in gene repression.

The role of chromatin-remodeling complexes in cancer was shown convincingly by the frequent mutation of the *SNF5/INI1* gene in a subset of sporadic human cancers and a new familial cancer syndrome tentatively termed rhabdoid predisposition syndrome (Olivier Delattre, Institute Curie, Paris). However, mutations of two other

Shi Huang
shuang@burnham.org
The Burnham Institute, La Jolla, CA92037, USA.