HUMAN PATHOGENIC MUTATIONS IN PROTEIN DOMAINS

Ilkka Lappalainen

ACADEMIC DISSERTATION

*To be presented with permission of the Faculty of Biosciences,
University of Helsinki for public criticism
in the auditorium 1041 at Viikki Biocenter, Viikinkaari 5, Helsinki,
on September, 24th, at 12 o'clock noon.*

For Kati

Supervisor: Professor Mauno Vihinen
Institute of Medical Technology
University of Tampere
Tampere
Finland

Reviewers: Docent Jari Ylänne
Biocenter Oulu and Department of Biochemistry
University of Oulu
Oulu
Finland

Docent Heikki Lehväslaiho
EMBL Outstation
European Bioinformatics Institute
Hinxton
United Kingdom

Opponent: Rudy W. Hendriks
Department of Immunology
Erasmus MC Rotterdam
Rotterdam
The Netherlands

ISSN 1239-9469
ISBN 952-10-1916-6
ISBN (e-thesis) 952-10-1917-4
Helsinki 2004
Yliopistopaino

Contents

ORIGINAL PUBLICATIONS

Thesis is based on the following original publications, referred to in the text by their Roman numerals I-VI, and on unpublished results presented in the text.

I. *Ollila, J., *Lappalainen, I., and Vihinen, M. (1996). Sequence specificity in CpG mutation hotspots, FEBS Lett *396*, 119-22.

II. Vihinen, M., Brandau, O., Branden, L. J., Kwan, S. P., Lappalainen, I., Lester, T., Noordzij, J. G., Ochs, H. D., Ollila, J., Pienaar, S. M., Riikonen, P., Saha, B. K., and Smith C. I. (1998). BTKbase, mutation database for X-linked agammaglobulinemia (XLA), Nucleic Acids Res *26*, 242-7.

III. Mattsson, P. T., Lappalainen, I., Backesjo, C. M., Brockmann, E., Lauren, S., Vihinen, M., and Smith, C. I. (2000). Six X-linked agammaglobulinemia-causing missense mutations in the Src homology 2 domain of Bruton's tyrosine kinase: phosphotyrosine-binding and circular dichroism analysis, J Immunol *164*, 4170-7.

IV. *Lappalainen, I., *Giliani, S., Franceschini, R., Bonnefoy, J. Y., Duckett, C., Notarangelo, L. D., and Vihinen, M. (2000). Structural basis for SH2D1A mutations in X-linked lymphoproliferative disease, Biochem Biophys Res Commun *269*, 124-30.

V. Lappalainen, I., and Vihinen, M. (2002). Structural basis of ICF-causing mutations in the methyltransferase domain of DNMT3B, Protein Eng *15*, 1005-14.

VI. Lappalainen, I., Shen, B., and Vihinen, M. Predicting the effects of pathogenic mutations on SH2 domain structures, manuscript.

ABBREVIATIONS

A	adenine
BCC	Basal-cell carcinoma
BTK	Bruton tyrosine kinase
C	cytosine
CD	circular dichroism
CFTR	cystic fibrosis transmembrane conductance regulator
CpG	CG dinucleotide
CSH2	carboxy terminal SH2 domain
G	guanine
HGMD	Human Genome Mutation Database
HGP	Human Genome Project
ICF	Immunodeficiency, Centromeric instability and Facial anomalies
LINEs	long interspersed repeating elements
MuStar	Mutation Storage and Retrieval software
NSH2	amino terminal SH2 domain
PID	primary immunodeficiency
PI3K	phosphatidyl inositol 3 kinase
PLCγ	phospholipase gamma
pY	phosphotyrosine
RASA1	Ras GTPase activating protein
SINEs	short interspersed repeating elements
SH2	Src homology 2 domain
SH3	Src homology 3 domain
SLAM	signal lymphocyte-activator molecule
SNP	single nucleotide polymorphism
SRS	Sequence Retrieval System software
T	thymine
UMD	Universal Mutation Database software
XLA	X-linked agammaglobulinemia
XLP	X-linked lymphoproliferative disease

SUMMARY

A large number of human DNA sequence variations have been identified and categorized as pathogenic or non-pathogenic based on their influence to the phenotype. Both types of variations have been collated into registries that are typically distributed through the Internet. The primary immunodeficiencies (PIDs) form a distinct group of mainly rare syndromes. More than 2700 patients have been diagnosed and the mutation and patient data collected into locus-specific databases. This study has concentrated on increasing the quality of the PID information on several levels.

Using a novel database format developed during the study, a number of locus-specific mutation databases were constructed and maintained. The data in the registries was used to analyse the underlying mutation mechanisms, especially deamination of methylated cytosines. As primary sequence of the affected proteins cannot be used to predict the putative changes in the biophysical properties of mutated structures, a bioinformatical method was developed for mutational analyses. The method applies structural homology when experimental three-dimensional structure of the defective protein is not available. By using structure-derived rules, the structure-function consequences of missense mutations in two distinct protein module families, Src homology 2 (SH2) and DNA methyltransferase domains, were analysed. In addition, pathogenic mutations were introduced into the SH2 domain of Bruton tyrosine kinase and analysed by using various biochemical methods. The experimental results verified the bioinformatical predictions for the pathogenic mutations in Bruton tyrosine kinase.

INTRODUCTION

The human genome sequence has been revealed and an enormous amount of variations mapped onto it. Majority of the DNA sequence variations results from short insertions, deletions or changes of single nucleotides. The variations can be categorized as pathogenic and non-pathogenic based on their influence to the phenotype. Today, more than 1500 different genes have been linked to a disease.

Primary immunodeficiencies (PIDs) are a group of mainly rare syndromes affecting various parts of the immune system. Although the symptoms of several PIDs are similar, more than hundred distinct phenotypes have been characterized. After diagnosis of the disease, a proper treatment is available for many of the PIDs and the patients may live fairly normal life. IMT Bioinformatics maintains and develops a knowledge base for PIDs including more than 80 different locus-specific mutation databases with roughly 2700 patients. The knowledgebase also provides curated disease information for the scientists, physicians and patients and software for mutation analyses and data distribution. The present study has concentrated on increasing the quality of the PID information on several levels.

A number of locus-specific mutation databases were created to store the mutation and patient information. In the first phase, a novel database format was developed for the BTKbase following the guidelines published by Human Genome Variation Initiative. The format was then applied to other constructed locus-specific mutation databases. In addition, a generic registry for Src homology 2 (SH2) domain mutations was created. The registry provides tools for accessing mutation and patient data from the individual locus-specific mutation databases and allows further studies, such as genotype-phenotype correlations. Secondly, the data in the registries was used to analyse the effect of the neighbouring nucleotides for the mutation process, especially in deamination of methylcytosine into thymine.

Currently, all the PID related locus-specific mutation registries describe the effects of a particular mutation to the mRNA and protein levels directly from the analyses of genomic DNA. Although the biophysical properties of proteins are determined by the amino acid sequence, it is not possible to predict the biophysical properties of the mutated protein structure directly from its primary sequence. Therefore, the third aim of this study was to develop a bioinformatical method that could be applied to a range of protein domains comprising thousands of PID causing mutations. The approach exploits structural homology among the family members when structural information is not available. Comparative modelling was used to build the defective protein domain structure based on a homological experimentally solved structure, and the structural consequences of the pathogenic mutations were analysed based on set of structure-derived rules and sequence entropy, *e.g.* the introduced side chain χ-angles were rotated to study if it can adopt a known rotamer on the corresponding structure.

The method was applied to two diverse protein domains, the Src homology 2 (SH2) and DNA methyltransferase domains, to study the structure-function consequences of eighty-nine different pathogenic amino acid substitutions. SH2 domains are a well-characterized protein module family that recognize phosphorylated tyrosines almost invariably in specific sequence contexts. These domains have been shown to mediate protein-protein interactions in many signal transduction pathways or intramolecular contacts that regulate enzyme activity. Pathogenic mutations affecting seven different SH2 domains have been identified from nine disease phenotypes. The methyltransferase domains catalyse the transfer of a methyl group from S-adenosyl-L-methione to the target cytosine in DNA. The effects of DNA methylation are widespread including *e.g*. transcriptional repression by methylation of promoter regions and X-chromosome inactivation. Mutations in the gene encoding for a DNMT3B, lead to an autosomal recessive Immunodeficiency, Centromeric instability and Facial anomalies (ICF).

To validate the method, six disease-causing mutations were cloned into the SH2 domain of Bruton tyrosine kinase (BTK). The mutated proteins were analysed for their consequences to the protein structure and function by using circular dichroism (CD) spectroscopy, and for their ability to bind to phosphotyrosine. Three of the mutants were also introduced into full-length BTK protein and transiently expressed in COS-7 cells to analyse the differences in stability between isolated SH2 domain mutants and BTK *in vivo*. The biochemical analyses verified the bioinformatical predictions of the mutation consequences on BTK SH2 domain structure model.

REVIEW OF THE LITERATURE

1 The Human Genome

The sequencing of the human genome was completed April 2003. As we learn more about human biology, additional layers of information will be mapped to the genome. One such layer consists of all types of DNA variations identified during the Human Genome Project (HGP) and by various groups studying polymorphisms and human diseases. The mapping of the genome has not only accelerated the cloning of disease-associated genes but also increased our understanding of how disease-causing variations differ from normal polymorphisms. The detailed discussion of genome composition appeared in the published draft sequences (McPherson et al., 2001; Venter et al., 2001).

1.1 DNA Structure

The genetic information is stored in the structure of the deoxyribonucleic acid (DNA). In 1953, Watson and Crick described how two complementary DNA chains could coil around each other to form the helical structure (Watson and Crick, 1953). In their structure, the nucleotides are inside of the helix, perpendicular to the common axis. The adenine (A) and guanine (G) are aromatic heterocyclic purines, whereas the cytosine (C) and thymine (T) consist of a single aromatic ring and are pyrimidines. As a result, the nucleotides can only fit inside the helix, if a purine bonds with a pyrimidine from the opposite DNA chains. Specific hydrogen bonds between G and C as well as between T and A generate complementary base pairing. The backbone is formed of phosphodiester bonds between the deoxyribose groups. The negative phosphate groups remain on the outside the helical structure and are available to interact with surrounding molecules. The two DNA chains run in opposite directions.

1.2 Genomic organization

The human genome consists of approximately three billion DNA base pairs organized into 23 chromosome pairs. 22 of these are autosomes and the remaining pair is formed by the sex chromosomes. The composition reflects both functional and structural elements of the genome (Figure 1A).

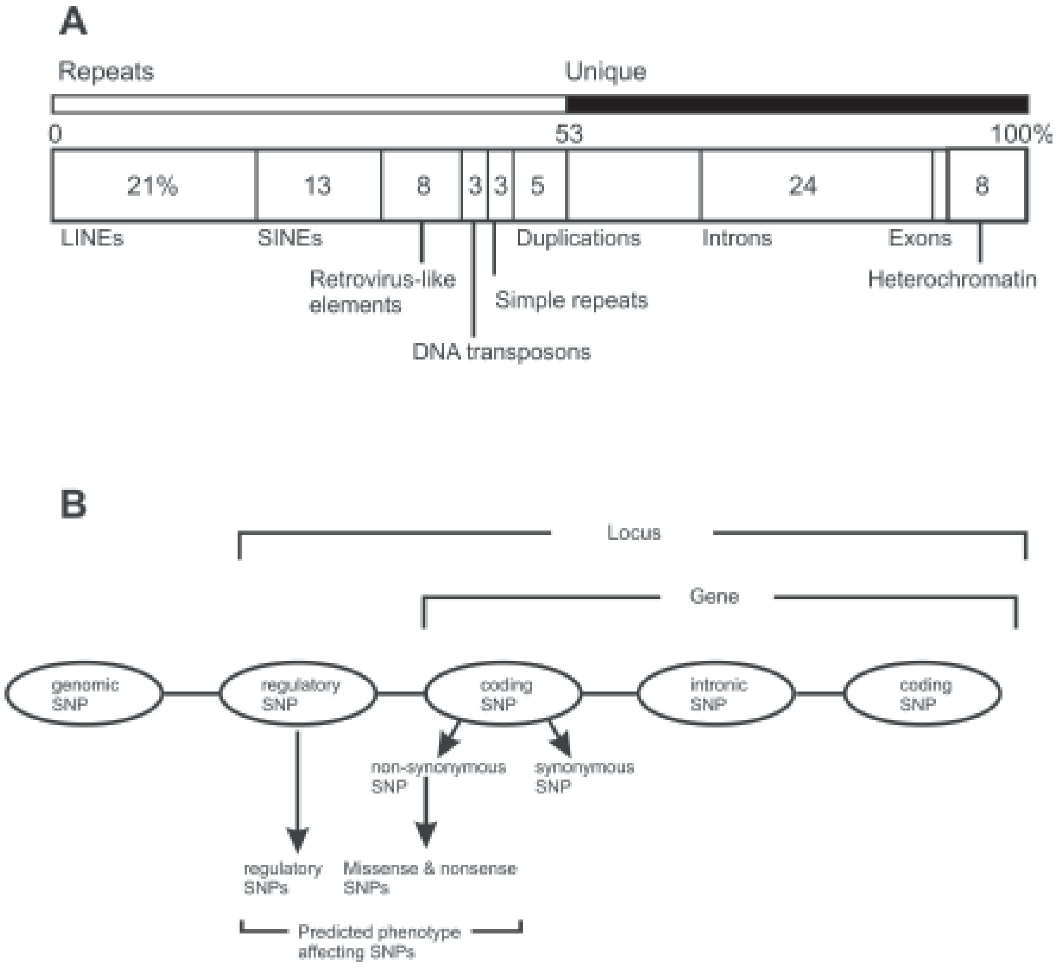

Figure 1 A - Although the 30000-35000 human genes comprise a quarter of our genome, only 1.5% of the DNA encodes for proteins. Majority of the genome consists of different type of repeating sequences. The figure was modified from the original one appearing in (Dennis and Gallagher, 2001)**.** The data was published in (McPherson et al., 2001) **B** - The SNPs affecting phenotype are either located in the regulatory or coding regions. Variants affecting splice sites are included either in exon or intron categories.

1.2.1 Repeating sequences

More than half of the nucleotides in our genome form repetitive sequences, with the vast majority of these accounted for by repeats derived from parasitic DNA sequences, known as transposons. Long interspersed sequences (LINEs) are the most ancient repeating unit in human genome. These transposons are roughly 6000 bp of length encoding the machinery for copying itself, whereas the short interspersed elements (SINEs, roughly 100-400 bp) implement the LINEs machinery for transposition. The most abundant repeating unit with a million copies in the genome is the Alu element belonging to the latter group of transposons. The observation of Alus near genes in GC and AT rich regions may be explained by a their role in protein translation regulation under conditions of stress. Dispersed Alu segments also exhibit significant differences in tissue-specific cytosine methylation levels (reviewed in Schmid, 1998). Of the transposons, only LINE1 and Alu are still active in our genome.

Roughly 3% of the human genome consists of repeats of just a few bases and 5% of duplications of larger segments. With the exception of Alus, repetitive DNA is enriched in AT rich regions. These areas are thought to be involved in the structure and reshaping of the chromosome by rearranging it to create new genes or modify the existing ones. The repeating sequences enclose a large number of DNA variations.

1.2.2 Unique sequences

A gene consists of a specific sequence of bases containing information to build protein(s). Genes are further split into exons and introns, the former encoding for proteins. Interestingly, only 1.5% of our genome encodes for proteins. The 30-35000 genes are unevenly distributed among the genome forming large gene-rich segments.

Normal males have X and Y-chromosomes, whereas females have two copies of the X chromosome. Hence, genes located outside sex chromosomes are available as two alleles situated in a locus that describes the chromosomal location of the gene. Typically genes are located in segments with higher C+G content than the genome average of 41%. This is partly due to a high selection pressure to preserve the nucleotide composition in the coding regions undamaged. The human promoter regions have also been shown to be associated with CpG islands, segments of DNA with a very high concentration of CpG dinucleotides (Bird, 1986; Larsen et al., 1992). These islands are involved in regulation of gene transcription in the germline and early embryonic cells. The majority of cytosines in CpG dinucleotides are methylated, whereas cytosines in CpG islands are unmethylated. The promoters without CpG islands are methylated in sperm and are always associated with tissue-specific genes (Antequera, 2003). The spontaneous deamination of methylated cytosines to thymine underlies in many human diseases.

1.2.3 From genes to proteins

The concept of a single gene encoding a particular native protein structure with one *in vivo* function is an over-simplification. Proteins may have more than one function. Most proteins consist of a variety of domains, independently folding modules with an evolutionary conserved function(s). Interactions between domains can affect the protein structure, stability and function (*e.g.* Altroff et al., 2001). As an example, a transient attachment of a small and highly negative phosphate group has been shown to act as a switch between the inactive and active enzyme conformations or by locating the molecule to its correct pathway (reviewed in Hubbard and Till, 2000).

The protein diversity is further increased by utilization of alternative promoters, multiple transcription start sites, modified polyadenylation or alternative splicing (reviewed in Landry et al., 2003 and Shabalina and Spiridonov, 2004). The encoded protein isoforms may differ in function, tissue-specific expression profile, cellular location or involvement in human diseases (Mironov et al., 1999; Caceres and Kornblihtt, 2002; Roberts and Smith, 2002). Recently, splice-variants have been shown to either insert or delete complete protein domains or target functional residues more frequently than expected (Kriventseva et al., 2003).

Furthermore, high conservation in alternative and constitutive splice sites between the human and mouse transcripts has been observed (Thanaraj et al., 2003). A large number of splice-variants, however, introduce a termination codon and the encoded protein product is likely to be highly unstable. These aberrant transcripts are detected and degraded rapidly by specific nonsense-mediated mRNA decay machinery (Maquat, 2002).

1.3 Genetic variation

The humans are almost identical to each other in their genomic DNA sequences. On average, our genomes differ only by 0.1% from each other and we are approximately 98.8% identical to chimpanzees at the nucleotide level (Clark et al., 2003). The pattern of variation in modern populations is dependent on our past. Historic population size, structure and genetic drift influence the pattern of variation across the whole genome. Natural selection, on the other hand, affects specific regions at particular loci through mutation and recombination. Variations between individuals form the genetic background responsible for biological and physical differences, such as colour of hair, susceptibility to a disease and response to treatment.

New alleles are introduced to gene loci by spontaneous endogenous processes or induced by various exogenous agents, such as UV radiation or tobacco smoke. Although these processes are rare, they constantly create new variations in the human population. The fate of the new mutation depends on its effect on the phenotype. Types of genetic variation vary in length, frequency and distribution. Chromosomal rearrangements involve duplications, inversions, translocations or deletions of large genomic segments. Most genetic variation, however, results from short insertions, deletions or changes of single nucleotides.

1.3.1 Single Nucleotide Polymorphims (SNPs)

By a strict definition, a SNP is a site where two nucleotides have been found in a specific population with the minor allele present in greater than 1%. An analogous disease-causing mutation typically has an allele frequency of below 1% with highly penetrating phenotype. However, the SNP definition is not applied strictly in public variation databases, and the allele frequency also depends on inheritance pattern.

The probability of a nucleotide position being heterozygous when comparing two chromosomes chosen randomly from the population is represented by a normalized heterozygosity (π). Depending of the number of chromosomes and the ratio of analyzed populations, π is approximately $5 \cdot 10^{-4}$ (Cargill et al., 1999; Halushka et al., 1999; Sachidanandam et al., 2001; Venter et al., 2001). The heterozygosity is relatively constant among the autosomes, but decreases in sex chromosomes. The lower nucleotide diversity in the X- and Y-chromosomes may be explained as a combination of smaller effective population size and strong selection due to hemizygosity in males (Sachidanandam et al., 2001).

The SNP allele frequency has been shown to correlate with the allele age, population specificity and functional class (Halushka et al., 1999). The majority of SNPs have a minor allele frequency of less than 10%. These are relatively new variations found only in specific populations. Individual genes also differ in their nucleotide diversity (Halushka et al., 1999; Sachidanandam et al., 2001). As an example, particular non–coding regions in the HLA locus show extremely high sequence variation owing to the balancing selection, whereas non-coding regions of seven X-linked loci have low nucleotide diversity (Horton et al., 1998). Generally, SNPs are less abundant in the exons than in the non-coding regions (Cargill et al., 1999; Halushka et al., 1999; Salisbury et al., 2003). However, these publications focused on coding regions and surveyed only limited portions of non-coding sequences.

SNPs with minor allele frequency of at least 1% have been shown to occur at a rate of 200-300 bp through the genome (Kruglyak and Nickerson, 2001; Stephens et al., 2001), suggesting as many as 15 million common SNPs in the human genome. These SNPs can be classified based on their genomic location (Figure 1B). The coding SNPs can be further divided into three categories based on their effect on the protein structure. Synonymous SNPs have no effect at the protein level as the new codon still encodes for the same residue. Non-synonymous SNPs may either lead to an amino acid substitution (missense) or a truncated protein (nonsense). Comprehensive variation studies including a large number of genes have found approximately four coding SNPs per gene with 40% of them being non-synonymous (Cargill et al., 1999; Halushka et al., 1999). Hence, the total number of non-synonymous SNPs is expected to be 48-56 000. These SNPs, together with still an unknown number of regulatory and other functional non-coding polymorphisms, are considered to form the pool of potential phenotype altering variations in the human genome.

1.3.2 Databases of normal variation

The SNP consortium, formed by several companies and academic institutions, was established in 1999 to produce a public resource of human SNPs (Thorisson and Stein, 2003). As the Human Genome was published, the consortium released more than 1.4 million SNPs collected from 24 ethnically diverse individuals (Sachidanandam et al., 2001). During the past two years, a number of other large-scale analyses of SNPs in specific populations or genes related to specific diseases have been published (Martin et al., 2000; Hirakawa et al., 2002; Lee et al., 2003). In addition to public variation databases, companies, such as Celera Genomics, provide access to their private databases lifting the number of non-redundant human genetic variations over six million. The publicly available information is deposited into two public databases, the Human Genome Variation database (HGVbase) and dbSNP (Wheeler et al., 2003; Fredman et al., 2004). The databases are listed together with some disease-causing mutation databases in the Table 1.

The non-pathogenic sequence variation databases include several problems caused by the massive increase of data in a short period of time as well as our inaccurate methods to model complex human biology. As an example, the precise exon structure on the genome is still likely to change as new genomes are sequenced and the algorithms detecting exons are refined to improve accuracy. Furthermore, public databases include a number of sequencing errors and SNPs located in pseudogenes (Ng and Henikoff, 2002). Some SNPs may also be associated with complex diseases. The current version of HGVbase (v.15) contains 2.8 million DNA variants, only 6.5% of the variations have been verified and 1.4% includes information about allele frequency.

Table 1. A partial list of human variation societies and databases.

Name	Description	Address
HGVS	The Human Genome Variation Society	http://www.hgvs.org/
HGVbase	Human Genome Variation Database	http://hgvbase.cgb.ki.se/
dbSNP	Main repository for normal variation	http://www.ncbi.nlm.nih.gov/SNP
OMIM	Online Mendelian Inheritance in Man	http://www.ncbi.nlm.nih.gov/omim/
HGMD	Human Gene Mutation Database	http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.htm
MITOMAP	Human mitochondrial genome database	http://www.mitomap.org
IARC TP53 database	Largest database of somatic mutations in TP53 gene	http://www.iarc.fr/p53
IDbases	Immunodeficiency related databases	http://bioinf.uta.fi/base_root/

Partly to correct the current situation, the International HapMap Project was initiated in 2002 (Consortium, 2003). The aim of the project is to provide publicly available set of common patterns of DNA sequence variation from three populations originating from parts of Africa, Asia and Europe by determining the allele frequencies and the degree of association between the variations. The resulting haplotype map can be used to *e.g.* identify association between a specific variant and a disease indirectly by comparing a group of affected individuals with a group of unaffected controls (Collins et al., 1997).

2 Genetics in Human Diseases

All types of sequence variations in germline DNA have been shown to cause diverse phenotypes. Chromosomal rearrangements affect the copy number of genes and disease results from a gene dosage effect. In contrast, the coding SNPs may be involved in the change of function or biophysical properties of the encoded protein (reviewed in Inoue and Lupski, 2002). However, the disease phenotype dominates the normal phenotype only if DNA variations affect the overall fitness of the organism. A clear relationship between proteins with an essential *in vivo* function and damaging phenotype has been observed (Jeong et al., 2001; Krylov et al., 2003). Moreover, the relatively dispensable proteins evolve more rapidly during the evolution as deleterious changes to the protein structure and function are subject to weaker selection (Hirsh and Fraser, 2001).

2.1 Patterns of Inheritance

The inherited diseases, in which a change in a single gene causes a distinct phenotype, are characterized as Mendelian syndromes. The pathogenic phenotypes can be further divided based on the chromosomal location and penetrating effect of the affected gene. In autosomal dominant disorders a single copy of the mutated gene is sufficient for the expression of the disease phenotype, such as in Huntington's disease. In autosomal recessive syndromes, *e.g.* cystic fibrosis, only individuals homozygous for the particular mutant allele or heterozygous for two different alleles develop the disease. Individuals with one healthy allele are phenotypically normal carriers of the syndrome. Males and females are equally likely to be affected, whereas sex-linked diseases show different pattern. The inheritance patterns of mtDNA are unique as the mitochondrial DNA is inherited maternally.

The majority of human genetic disorders, however, are of complex type. Variants in different parts of the genome together with environmental factors and, for example, aging may lead to a predisposition to complex diseases such as asthma, diabetes or depression.

2.1.1 Allelic spectra in rare diseases

The rare disorders, such as most Mendelian type diseases, are caused by panoply of diverse highly penetrable disease alleles with minor allele frequency of less than 1%. As an example, 461 different mutation types have been identified in the gene encoding for Bruton tyrosine kinase (http://protein.uta.fi/BTKbase). These mutations lead to an X-linked agammaglobulinemia (XLA) by disrupting the B-cell maturation process (Sideras and Smith, 1995). The varied mutational spectrum of XLA is typical of X-linked and autosomal dominant syndromes. The disease-associated alleles are eliminated rapidly by natural selection, whereas new mutations replenish the disease class leading to a rapid turnover and mutation-selection equilibrium (Reich and Lander, 2001).

Some recessive autosomal Mendelian diseases may have common alleles as a result of mild selection against disease alleles or because of selective heterozygous advantage. Cystic fibrosis is a fairly common disease resulting from a loss or dysfunction of a CF transmembrane conductance regulator (CFTR) Cl^- channel (Riordan et al., 1989). In contrast to XLA causing mutations, the cystic fibrosis is associated with few common alleles together with a large number of rare alleles (Estivill et al., 1997). It has been suggested that the alleles with high frequency are involved with resistance to *Salmonella typhi* among heterozygous individuals (Pier et al., 1998).

In addition to heterozygotic advantage, simpler allele spectra may also originate from historic or geographic reasons. A recent population bottleneck in Finland enriched certain disease alleles that are rare elsewhere, whereas the number of patients with *e.g.* cystic fibrosis is extremely low in the Finnish population (Kere, 2001).

2.2 Databases related to human diseases

In 1957, Ingram described the first defect in the gene encoding for human haemoglobin leading to severe anemia (Ingram, 1957). The first mutation database for haemoglobin mutations was published in 1976 (Lehman and Kynoch, 1976). Since then, the number and variety of databases cataloguing human disease variations has grown enormously. Majority of disease-causing mutations still exist in locus-specific databases maintained by the laboratories or consortia studying the gene affected. However, several generic mutation databases have emerged as a result of incompatible database formats for large-scale mutation analyses.

2.2.1 Locus-specific databases

The locus-specific databases can be categorized as mutation or patient based registries (Claustres et al., 2002). Both registries typically include a unique identifier for the disease allele and reference(s) either to the published article or submitting physician. The effect of the disease-causing mutations is described from the genomic DNA level through mRNA to protein level. The patient based databases also include information related to the phenotype, family history, patient data and response to treatment (Lappalainen et al., 1997). The most comprehensive listing of locus-specific databases is available from the Human Genome Variation Society web site.

The information and registry formats have gone through many changes during the last 15 years. The recommendations for the description of a particular mutation (den Dunnen and Antonarakis, 2001) or database format (Scriver et al., 1999) by the Human Genome Variation Initiative have united the field and allowed development of several generic tools for the maintenance and analyses of the databases. The Mutation Storage and Retrieval (MuStar) (Brown and McKie, 2000), Universal Mutation Database (UMD) (Beroud et al., 2000) and MUTbase (Riikonen and Vihinen, 1999) software have been successfully applied to create a number of locus-specific databases. Importantly, the database format for describing the mutation and various clinical data is highly structured in all programs permitting high data integrity. The programs verify that the submitted or manually included data refers to the correct position in the right reference sequence; a welcomed feature as the published patient data often includes errors at all levels. The UMD and MUTbase programs also generate several web pages showing statistical analyses of mutations in the corresponding gene or distribution of the mutation types at the exon/intron or protein domain levels. The MuStar and UMD distribute data either in spreadsheets or in relational database format, whereas MUTbase generates flat files. Another essential tool for searching data from various databases simultaneously and analyses is the Sequence Retrieve System (SRS) (Zdobnov et al., 2002). In SRS 3D, sequence features extracted from other databases can be simultaneously mapped onto structures. All the described programs are flexible and allow addition of tailored tools *e.g.* using Bioperl or The European Molecular Biology Open Software Suite (EMBOSS) (Rice et al., 2000; Stajich et al., 2002).

Today, analyses of disease-causing mutations include either a large number of mutations in a single affected gene or thousands of mutations from several different locus-specific databases. At the same time, the number of publications describing individual mutations has decreased leading to an increasing number of deposited mutations in the databases that are not publicly available. Roughly 4% of the mutations leading to various immunodeficiencies are hidden in the locus-specific databases. The number of confidential mutations is likely to vary according to the disease prevalence and curating database consortia as large estimates have been described for other diseases (Cotton, 2000).

2.2.2 General databases

In contrast to locus-specific databases, common repositories contain less detailed information of mutations from multiple loci. The Mendelian Inheritance in Man (MIM) was the first attempt to list all the inherited monogenic human diseases (McKusic, 1998). The current online version (OMIM) is available at the National Centre for Biotechnology Information website. OMIM only lists the most important or first mutation(s) identified in the corresponding disease. Hence, a second attempt to catalogue quantitatively all types of DNA variations associated with diseases was initiated by Cooper and Krawczak in 1990. The Human Genome Mutation Database (HGMD) is comprehensive collection of all types of germline mutations associated with human inherited diseases. The current version of HGMD contains 39415 different mutations affecting 1516 genes (Stenson et al., 2003). Each mutation has been logged only once to the database to avoid the problem of separating recurrent lesions from mutations identified in a descent. As these two main depositories contain only nuclear mutations, the human mitochondrial disease related mutations are collected into *e.g*. MITOMAP (Kogelnik et al., 1998).

Somatic mutations have also been collected into several databases. The largest of them describes almost 19 000 tumorigenic *TP53* mutations from a gene encoding for p53 protein. The tumour suppressor function of p53 protein is lost in more than half of human cancers. 75% of these mutations occur as missense mutations rather than deletions, insertions or frameshifts (Olivier et al., 2002).

2.2.3 Disease-centred platforms

The primary immunodeficiencies (PIDs) are a group of mainly rare syndromes affecting the function of immune system. As a result, patients with these intrinsic defects have increased susceptibility to recurrent and persistent infections. More than 100 different PIDs have been classified and a large number of disease-associated variants collected into a central registry by the European Society for Immunodeficiencies (ESID) or locus-specific databases (Fahrer et al., 2001; Vihinen et al., 2001).

As the symptoms of several PIDs are similar, the diagnosis is still largely based on analysis of the genetic defect(s). After correct diagnosis, however, many patients may live quite normal life, *e.g.* intravenous immunoglobulin can be used for treatment in XLA. Recently, the Immunodeficiency Diagnostics registry (IDdiagnostics) and Immunodeficiency Resource (IDR) were developed to help physicians to contact laboratories analysing these rare genetic defect(s), as well as to collect verified information related to immunodefiencies (Väliaho et al., 2002; Samarghitean et al., 2004). As rapidly accumulating information from the HGP has lead to cloning of a large number disease associated genes, knowledge bases providing curated information of a particular disease for scientists, physicians and patients are likely to become more important than locus-specific or generic databases.

2.3 Cellular mechanisms behind mutations

DNA is a reactive molecule modified continuously by a range of chemicals and enzymes inside the cell nucleus or mitochondria. Exogenous agents, such as UV radiation or chemical carcinogens in food, may induce variations at the DNA level. However, majority of the inherited disease-causing mutations are caused by errors in the endogenous procedures involved in genomic stability (Cooper and Krawczak, 1993). Cells have an extremely efficient capacity to suppress the generation of alterations to the DNA sequence. Errors escaping the proofreading machinery become substrates of mismatch, base extinction or nucleotide extinction repair systems (Jiricny, 1998). The efficiency and specificity of these processes is DNA sequence dependent (Cooper and Krawczak, 1993). As a result, variations occur non-randomly throughout the genome and each type of variation shows a pattern of hotspot and cold-spot sites in a given sequence.

The spectrum of single-base-pair substitutions in the HGMD was found to be highly hierarchical in their propensity to undergo substitution. The transitions (*e.g.* where a purine is substituted by a another purine) and transversions (*e.g.* where purine is substituted by a pyrimidine) occur at frequencies of 62,5% and 37,5%, respectively (Krawczak et al., 1998). The CpG transversions comprise 23% of all human hereditary disease-associated mutations (Waters and Swann, 2000). Furthermore, the mutation site is clearly affected by its surrounding nucleotide sequence, though it extends only by a few bases. A clear bias for the immediately flanking nucleotides for most of the 12 possible changes was shown (Krawczak et al., 1998).

The molecular mechanisms of spontaneous mutagenesis occurring during replication, recombination and repair processes were first investigated in bacteria and yeast (reviewed in Maki, 2002). The genes involved have then been shown to be highly conserved among various organisms (Reenan and Kolodner, 1992; Morrison and Sugino, 1994). Based on the genetic, biochemical and structural results, several models of how spontaneous mutations arise have been introduced.

2.3.1 Misincorporation

An insertion of a non-complementary nucleotide at the end of the primer by DNA polymerases results in a single nucleotide change (Figure 2A). There are at least three possible check points for the proper geometric alignment during base insertion: initial dNTP binding and forming of correct hydrogen bonds based on the Watson-Crick model (Galas and Branscomb, 1978; Clayton et al., 1979), selection for the correct geometry after binding of the dNTP by an induced-fit mechanism (Echols, 1982; Kuchta et al., 1987; Kuchta et al., 1988; Sloane et al., 1988; Wong et al., 1991), and the chemical step leading to formation of phosphodiester bond. Insertion of a non-complementary nucleotide has been shown to restrain primer extension, thereby allowing translocation of the primer terminus into the active site of the proofreading 3'->5' exonuclease (Kunkel and Bebenek, 2000).

The DNA polymerases differ in their interactions with the minor groove of the template-primer duplex and there are significant differences in the extent to which different polymerases use methods for recognizing the correct nucleotide. In some cases a non-complementary nucleotide may by-pass the proofreading. The efficiency of the proofreading varies as a function of the mismatch type and the sequence context in which it is embedded (10^{-5} to $>10^{-8}$)(Kunkel and Bebenek, 2000). For example, the common G/T mispair is stabilized by two hydrogen bonds causing only a small distortion in the helical structure of the DNA (Hunter et al., 1987). Local imbalances of dNTP pools have also been shown to increase the probability of misincorporation and lead to a disease phenotype (Bebenek et al., 1992; Martomo and Mathews, 2002; Song et al., 2003). In addition, the dNTP pools may be contaminated with unnatural nucleotides as oxygen radicals attack free nucleotides more readily than double helical DNA (Park et al., 1992). One such compound, 8-oxodGTP, can be inserted opposite to either cytosine or adenine of template DNA with almost equal efficiency resulting G/C to A/T tranversion during the next DNA replication process (Maki and Sekiguchi, 1992).

2.3.2 DNA Slippage

In 1966 Streisinger proposed a hypothesis for transient misalignment of the primer and template during the polymerization process (Sreisinger, 1966). This premutational intermediate is stabilized by correct base pairs between the nucleotides surrounding the misaligned nucleotide (Figure 2B). The following polymerization leads to a deletion if the unpaired nucleotide is in the template strand. An insertion occurs if the unpaired nucleotide is located in the primer strand. The error rates for insertion and deletion increase as the length of the repeating sequence increases. The opposite has been observed if the repeats are either interrupted or eliminated (Kunkel, 1985; Bebenek et al., 1993). A strand slippage may also lead to a single nucleotide substitution if the slippage is followed by a complementary nucleotide incorporation and immediate realignment before further polymerization (Figure 2C).

The initiation of template-primer slippage may occur via multiple pathways. The extension of the primer from a non-complementary nucleotide is highly inefficient (Benkovic and Cameron, 1995). Therefore, Kunkel suggested that primer relocation might occur after misinsertion to create correct terminal base pairing that allows further polymerization (Kunkel and Soni, 1988). This model is not limited to single-base pair errors and may occur at any template location. In a similar way, damaged templates might also cause frameshift by primer relocation. The model is supported by studies with several polymerases with different lesions (Schaaper et al., 1990; Lambert et al., 1992; Garcia et al., 1993). Slippage may also occur during enzyme dissociation or reassociation as has been observed for the polymerases with low processivity (Kunkel, 1985; Kunkel, 1986). Short deletions or insertions comprise the second most common type of mutation associated with human inherited diseases. In the HGMD, all gene deletions either overlap or flank with a two base pair repeat (Antonarakis et al., 2000).

2.3.3 Deamination of methylcytosine into thymine

In eukaryotic genomes, the methylated cytosines predominantly occur in the CpG dinucleotide (Bird, 1999). This dinucleotide undergoes germline transition to TpG (and CpA in the complementary strand) at frequencies six to seven times the base mutation rate (Cooper et al., 1995) as a result of spontaneous deamination of methylcytosine (Figure 2D). Although two human thymine DNA glycosylases have been identified, this repair pathway is clearly inadequate (Brown and Jiricny, 1987; Hendrich et al., 1999). Subsequently, CpG dinucleotides are only present at 20% of the expected frequency in human genome (Brown and Jiricny, 1987; Hendrich et al., 1999).

The CpG dinucleotides are significantly biased by the 5' flanking nucleotide on the non-coding DNA strand, whereas the nucleotide immediately downstream of CpG is significant irrespectively of the strand (Krawczak et al., 1998). The methylated cytosines are also known to occur within CpNpG triplets (where N is any nucleotide) at low frequency (Woodcock et al., 1988; Clark et al., 1995; Kay et al., 1997). The CpApG trinucleotide was shown to undergo transition to TpApG at a 50% higher rate than any other triplet on both strands (Krawczak et al., 1998). The data clearly indicate biased nucleotide neighbourhood surrounding the methylated CpG dinucleotide in human inherited diseases. However, the frequency of CpG mutations may differ between the male and female germ-lines owing to profound differences in DNA methylation. The oocyte DNA is markedly undermethylated, whereas sperm DNA is heavily methylated (Monk et al., 1987; Rideout et al., 1990).

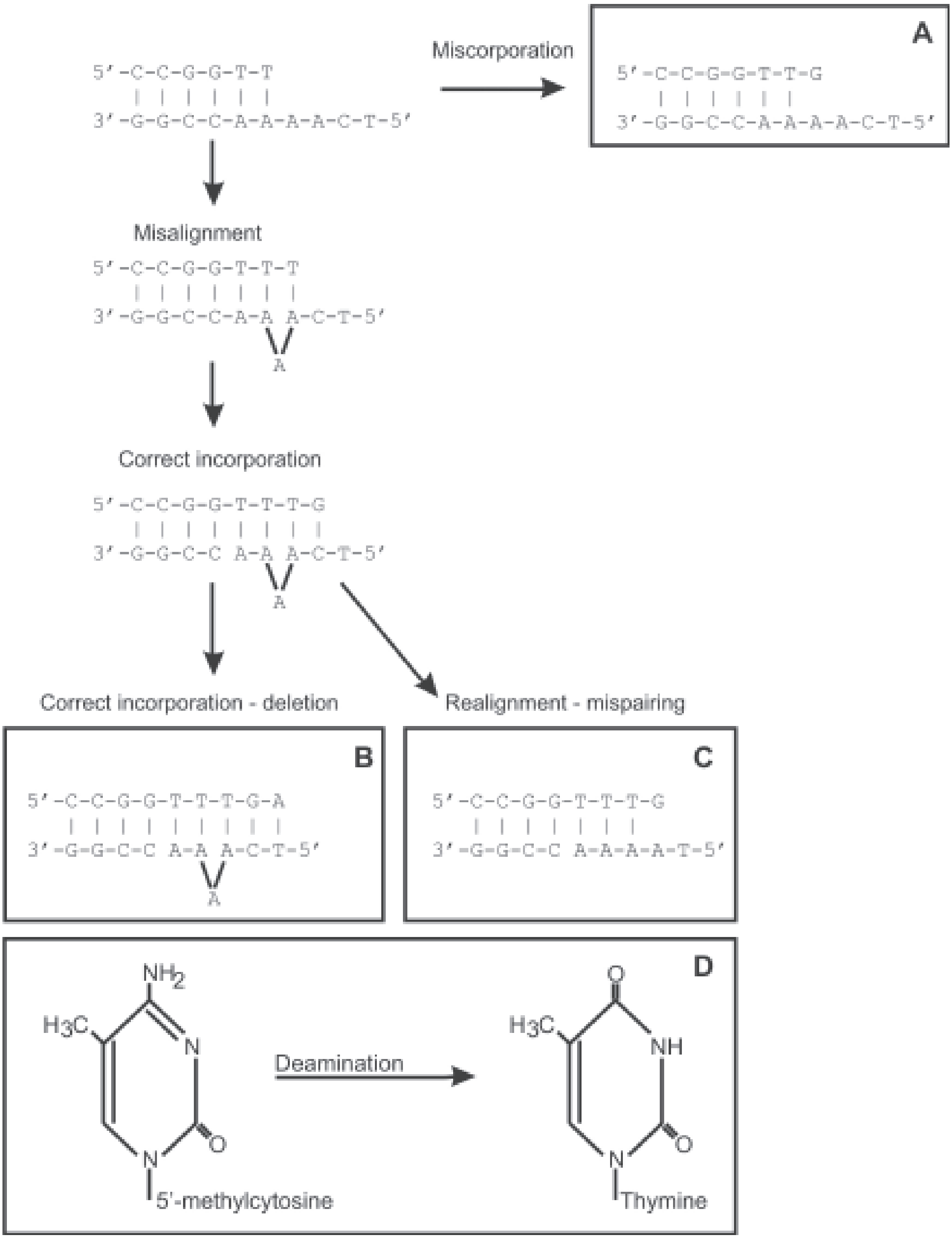

Figure 2. Proposed reaction mechanisms for mutations. **A** - incorporation of incorrect dNTP to the template. **B** - DNA slippage as a result of misalignment and correct incorporation. **C** - Mispairing initiated first by misalignment and followed by a correct incorporation and realignment of the polymerized DNA strand. **D** - Spontaneous deamination of 5'methylcytosine results in thymine. The figure was adapted from (Cooper and Krawczak, 1993).

2.4 Pathogenic variations affect biophysical properties of proteins

DNA variations located at the gene loci may cause pathological consequences by either affecting the cell specific expression profile or biophysical properties of the encoded protein. Currently, variations found in the regulatory positions comprise less than 1% of the inherited disease-causing mutations deposited in the HGMD. The number of these mutations is likely to increase together with our understanding of complex diseases and gene regulation. Changes leading to a loss or increase in number of active genes, such as an extra chromosome in Down syndrome, or complex rearrangements and large deletions spanning the whole disease loci, cover only 8% of disease-causing mutations registered in the HGMD. Vast majority of somatic and inherited pathogenic mutations are, therefore, small deletions and insertions or point mutations located at the protein-coding region (Olivier et al., 2002; Stenson et al., 2003). These genetic alterations specifically influence the features of the encoded polypeptide.

2.4.1 Characteristics of pathogenic SNPs

The nucleotide diversity at the coding sequence is dependent on the functional class of a SNP. The silent SNPs show approximately 2.5 times more diversity compared to that of nonsynonymous SNPs (Cargill et al., 1999; Halushka et al., 1999). In the majority of the non-synonymous SNPs, the minor allele frequency falls below 5% (Cargill et al., 1999; Stephens et al., 2001). The non-conservative SNPs leading to a dramatic change or termination codon have the lowest minor allele frequencies and the natural selection clearly acts against them (Figure 3B).

In most databases, the effect of a disease-causing SNP on the mRNA or on the protein level is predicted directly from the genomic DNA analyses. Translationally silent mutations have been shown to occur rarely (Figure 3A) and are assumed to affect mRNA splicing (*e.g.* Sumazaki et al., 2001). Although missense and nonsense mutations have also been shown to cause aberrant splicing, these SNPs are generally interpreted to change only the affected codon (reviewed in Cartegni et al., 2002). Point mutations introducing a premature termination codon are removed by nonsense-mediated mRNA decay (Maquat, 2002), whereas missense mutations accumulate to human genome depending on the consequences to the protein function, thermodynamic stability and folding *in vivo*.

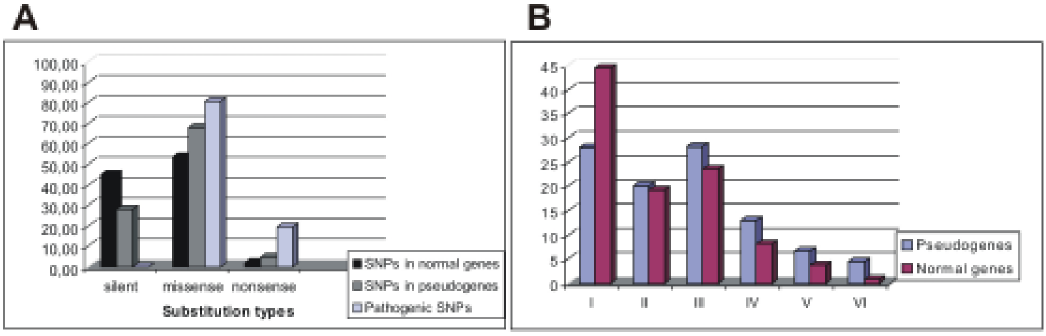

Figure 3. **A** - Natural selection acts against mutations with an increasing radical effect on the protein structure. **B** - The substitutions identified from the pseudogenes and SNPs at the exons were analysed based on Grantham's scale (Grantham, 1974) (I = silent, II conservative, III moderately conservative, IV moderately radical, V radical, and VI nonsense). The figure was created by using data from the HGMD database and results either described or referred in (Stephens et al., 2001).

2.4.2 Pathogenic mutations affect conserved positions

Several methods have been applied to analyse the differences between pathogenic and non-pathogenic missense mutations at the protein level. These methods have implemented sequence entropy together with various structural parameters derived from experimental structures (Sunyaev et al., 1999; Chasman and Adams, 2001; Ng and Henikoff, 2001; Ferrer-Costa et al., 2002; Saunders and Baker, 2002; Shen and Vihinen, 2004) or developed simple rules for predicting damaging amino acid substitutions (Sunyaev et al., 2001; Wang and Moult, 2001; Steward et al., 2003).

The disease-causing mutations are over-abundantly located at conserved positions, whereas normal variation is more randomly distributed (Miller and Kumar, 2001). At the secondary structural level, the normal variations are located in the exposed (solvent accessible surface >5%) α-helical and coil structures, whereas disease-associated substitutions are more likely to occur in the buried structures (Ferrer-Costa et al., 2002; Steward et al., 2003). Interestingly, 83% of disease-associated mutations were predicted to affect the protein stability whereas majority of the normal variations had no influence when similar rules were applied (Wang and Moult, 2001). Analysis of 63 disease associated protein structures assigned a functional role for only 29% of the analysed disease-causing mutations (Steward et al., 2003). Recently, pathogenic mutations were also shown to affect covariantly conserved positions (Shen and Vihinen, 2004).

The mutation types also differ between the disease variations and substitutions occurring between species or non-pathogenic SNPs (Miller and Kumar, 2001). There is a clear negative selection against SNPs leading to dramatic changes at the protein sequence based on the Grantham's physico-chemical score (Grantham, 1974; Cargill et al., 1999; Halushka et al., 1999; Stephens et al., 2001). The difference in physico-chemical properties of amino acid substitutions affecting the phenotype is larger for disease-associated substitutions (Figure 3B). The most severe substitutions were not observed, as they are more likely to result in lethal phenotypes (Miller and Kumar, 2001; Steward et al., 2003). The severity of the substitution has also been shown to correlate with the likelihood of observing patients clinically (Krawczak et al., 1998).

2.4.3 Two roads to disease

Protein evolution is primarily governed by protein function. As a result, proteins must be at least marginally stable and fold fast enough to prevent aggregation. Based on their structural consequences, disease-causing mutations can be categorized into two main classes: loss of protein function, which is often accompanied by improper localization and rapid degradation of defective product, and, mutations causing the pathological phenotype by affecting thermodynamic stability or kinetic pathway of the mutated protein. In this case the disease is generally associated with toxic properties of aggregation-prone folding intermediate (reviewed in Gregersen et al., 2000).

Disease-causing mutations influencing the balance between folding and misfolding pathways are likely to affect proteins with already small kinetic preferences for the folding pathway. One such protein is CFTR, where mutations have been shown to cause cystic fibrosis by impairing folding and biosynthetic processing of nascent molecules (reviewed in Kopito, 1999). However, maturation of wild-type CFTR protein has also been shown to be inefficient, less than 50% of synthesized CFTR folds correctly during its passage to the cell surface (Ward and Kopito, 1994).

The result of a missense mutation to protein structure and function, however, cannot be predicted simply by sequence entropy as has been illustrated for p53 mutations. Majority of somatic mutations affecting *TP53* gene are located in the DNA-binding domain, with six hot spots clustering to the DNA-binding surface, and three residues involved in binding of Zn^{2+} (Bullock et al., 2000). Based on the crystal structure, two of the residues at the DNA-binding surface contact DNA directly and four stabilize the surrounding structure (Cho et al., 1994). Mutations removing crucial interactions between the protein and its ligand had no effect on protein folding, but failed to bind an artificial p53 specific promoter DNA sequence. The reduced protein stability and capacity to bind DNA by the four other functional mutations varied. In contrast, mutations affecting hydrophobic core or Zn^{2+} binding residues destabilized the protein structure dramatically (Bullock et al., 2000). Interestingly, a number of core mutations could still bind DNA with 40-80% of the wild-type affinity. It may be possible to rescue these mutations by binding of a small molecule (reviewed in Bullock and Fersht, 2001), whereas functional mutations would all require their own ligand.

2.4.4 Theoretical and experimental analyses of missense mutations

Currently, there is no *de novo* method to calculate the correct three-dimensional structure of a protein from its primary sequence. Small perturbations caused by amino acid substitutions, however, can be predicted by using molecular modelling and molecular dynamic simulations from an experimentally solved structure (Leach, 2001). Comparative modelling exploits the structural similarities between proteins by constructing a three-dimensional structure based upon the known structure of one or more related proteins. In molecular dynamic simulations, successive configurations of the system are generated by integrating the Newton's laws of motion. The calculations are broken into a series of very short time steps (1-2 femtoseconds), and forces acting on each atom are recalculated at each step by using empirical force field. The resulting trajectory specifies the positions and velocities of the particles in the system as a function of time. However, there are limitations of how far consequences of missense mutations to the protein structure can be predicted. The current bioinformatical methods rely heavily on structural and biophysical data of a relatively small number of model proteins.

Protein folding occurs through an ensemble of structures that are transiently occupied and share an increasing number of wild-type contacts towards the native conformation (Fersht, 2002). The role of a particular position in protein folding can be studied by using ϕ-value analyses (Fersht et al., 1992), where a number of non-disruptive mutations removing specific interactions are created in several positions of the analysed protein. The value of ϕ is defined as a ratio of change in transition state energy compared to the change in stability on mutation. The difference in transition state energy on mutation can be analysed by measuring the folding rates of wild type and mutated proteins. Positions sharing wild type interactions have ϕ-values close to one as the mutation affects the transition state and wild type conformation identically. Protein denaturation by heat or chemical denaturants, such as guanidine hydrochloride and urea, is used for measuring the stability. The change in protein structure is typically monitored by using fluorescence or circular dichroism spectroscopy. The structure of the denatured and native states can be obtained with NMR spectroscopy or X-ray crystallography.

3 SH2 DOMAINS

At present, the results of mutations at the protein level are typically described as amino acid substitutions predicted directly from the genomic DNA analyses. To analyse the structural and functional consequences of pathogenic mutations at the protein level, we have concentrated on a distinct well-characterized protein domain family. The Src homology 2 (SH2) domains are about 100 residues in length. More than 100 different SH2 domains have been identified or predicted with an average of 28% pairwise residue identity (Pawson et al., 2002, Pfam code PF00017). SH2 domains mediate intramolecular recognition and intermolecular protein-protein association almost invariably by binding to phosphorylated tyrosine residues in specific sequence contexts. Structures of many individual SH2 domains have been solved and their binding to ligand studied (reviewed in Kuriyan and Cowburn, 1997). A number of disease-causing mutations have been described from the SH2 domains.

3.1 SH2 domain function

Tyrosine phosphorylated (pY) regions in proteins function as specific binding sites for the SH2 domains containing cellular signalling proteins. Binding of SH2 domains to their *in vivo* targets recruits the SH2 domain-containing protein to its proper signalling complex regulating downstream signalling cascades (reviewed in Schlessinger and Lemmon, 2003).

In addition to their role in assembling activated complexes, particular SH2 domains are involved in intramolecular interactions that control enzyme activity. A loop from the N-terminal SH2 domain binds to the catalytic cleft of the phosphatase domain in the same SHP-2 molecule leading to an autoinhibited configuration (Hof et al., 1998). The Src SH2 domain has been shown to bind a phosphorylated tyrosine at the C-terminus of the same molecule resulting inactivation of enzyme activity by rearrangement of catalytic center in the kinase domain (reviewed in Hubbard et al., 1998). In both examples, the high affinity ligands can compete with the intramolecular interactions and release the catalytic domains for their *in vivo* targets.

3.2 SH2 domain structure

Structures of a significant number of SH2 domains both in isolation and bound to various target molecules have been determined by X-ray crystallography and NMR spectroscopy. All the analysed SH2 domains have a typical SH2 domain fold consisting of a large anti-parallel β-sheet sandwiched between two α-helices The central β-sheet divides the domain into two functionally separate sides. The αA-helix borders the face binding to phosphotyrosine. Residues from αB-helix and the EF and BG-loops are involved in binding of side chains C-terminal to phosphotyrosine in the ligand. The βD', βE and βF strands form an additional β-sheet that closes off one part of this side (Figure 4 and the notation used for describing the secondary structures).

3.2.1 Residues involved in ligand-binging

The ligand binds in an extended conformation lying across the surface of the domain orthogonal to the central β-sheet in most experimentally solved SH2-ligand structures. SH2 domains make specific interactions with the phosphotyrosine and 3-6 residues immediately following it (reviewed in Kuriyan and Cowburn, 1997). There are only limited contacts formed between the domain and the side chains of the ligand residues upstream from the phosphorylated tyrosine, apart from SHP-2 and SH2D1A (Huyer et al., 1995; Poy et al., 1999).

The residues interacting with phosphotyrosine are generally conserved and form a positively charged binding pocket on the SH2 domain surface (reviewed in Kuriyan and Cowburn, 1997). The only invariant residue among the SH2 domains, an arginine at the fifth position of βB strand (and therefore coded as RβB5), extends from the bottom of the pocket to recognize the phosphate group from the phosphotyrosine. This interaction determines the binding specifically to phosphotyrosines as the arginine side chain is not long enough to interact with phosphorylated serine or threonine.

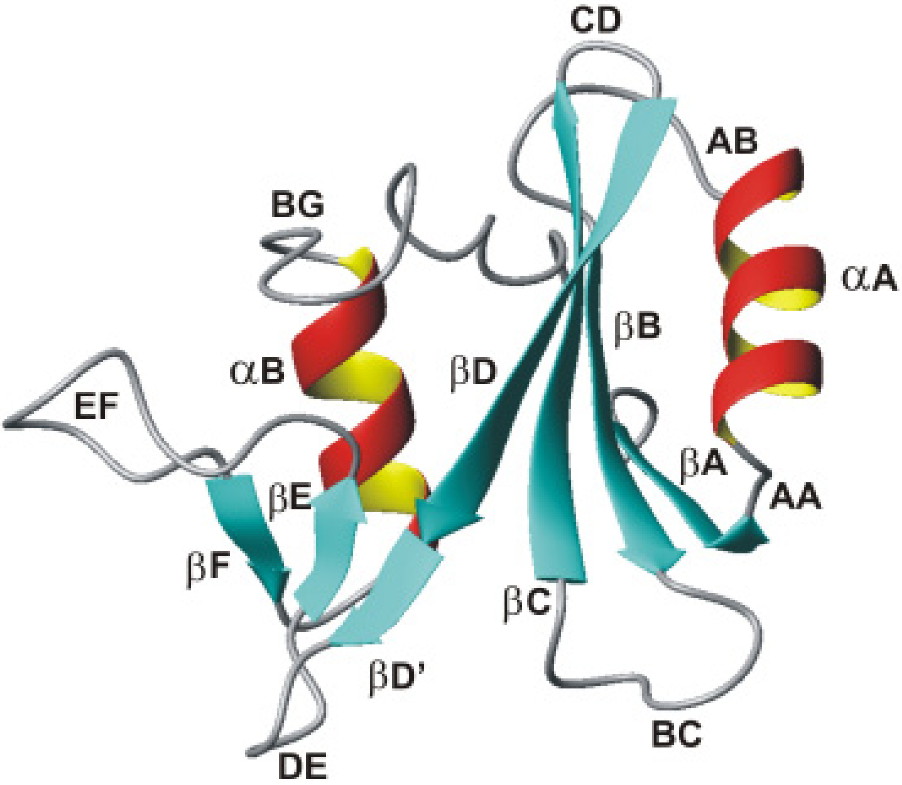

Figure 4 - A ribbon model of the SH2 domain of SH2D1A (PDB code 1D1Z). The large β-sheet (blue) is flanked by two α-helices (red). The secondary structures are indicated as was first introduced in (Eck et al., 1993).

The residues involved in binding of the third residue following pY (pY+3) are located in EF- and BG-loops. These residues are highly variable and respond to individual SH2 domain specificity. In the SH2 domain of the Src tyrosine kinase, the ligand-binding residues come close together forming another binding pocket on the SH2 domain surface. The majority of SH2 domains bind to their ligands as Src SH2 domain. In two phosphatase enzymes, SHP-2 and phospholipase Cγ-1 (PLCγ-1), the ligand-binding residues move away from each other opening up a binding groove on the SH2 domain surface (Lee et al., 1994; Pascal et al., 1994). The interactions between the ligand and SH2 domains extend beyond pY+3 position. In the structure of N-terminal SH2 domain of SHP-2 complexed with a ligand, the phenylalanine side chain at the pY+5 position in ligand is bound between BG and EF loops closing the isoleusine at the pY+3 position.

3.3 SH2 domain specificity

The SH2 domains bind to their ligands with only modest affinities. The dissociation constants range from 0.2 to 5 µM for SH2 domain-ligand interactions, whereas SH2 domains have been shown to associate with a random peptide only 10-fold lower affinity (Piccione et al., 1993; Ladbury et al., 1995). Unphosphorylated peptides do not bind to SH2 domain, with the only known exception of SH2D1A (Lemmon and Ladbury, 1994; Hwang et al., 2002). Furthermore, the phosphotyrosine alone shows very weak association (Kd > 1mM).

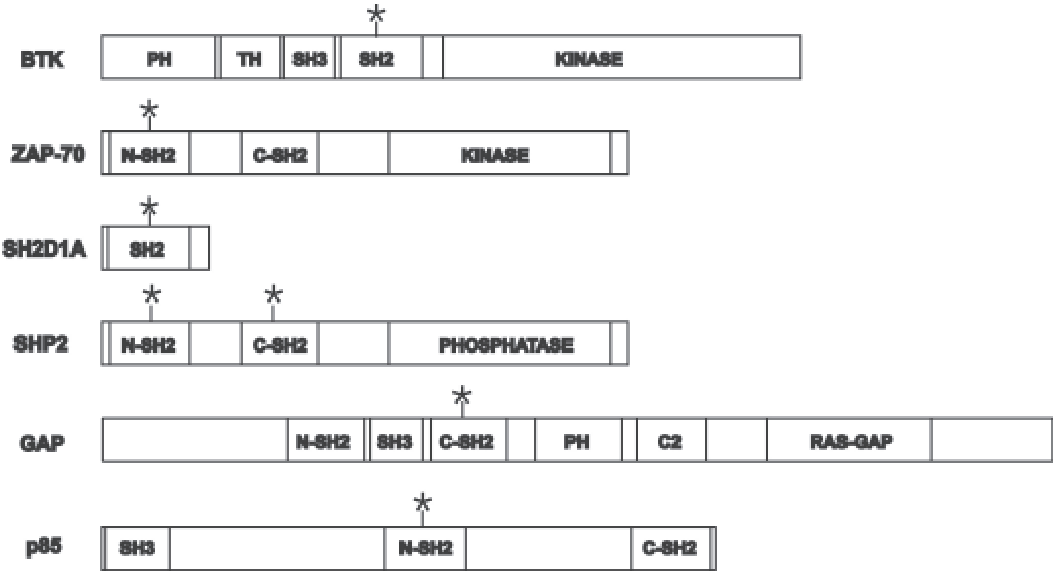

Figure 5. Domain organization of the defected proteins. Pleckstrin homology (PH); Tec homology (TH); Src homology 3 (SH3) and Ras GTPase activating (RASA1) domain. SH2 domains with disease-causing mutations are shown with an asterisk.

The specificity *in vivo* may be increased dramatically as a result of cooperative binding together with other signalling domains in the same molecule. PI3K, Zap-70, Syk, SHP-2 and PLCγ-1 contain two SH2 domains separated by a linker of varying length. All five proteins have been shown to associate with a conserved tyrosine-based activation motif (ITAM) in the cytoplasmic tail of different receptors. The binding of biologically relevant ITAM showed Kd of 0.3-3.0 nM, whereas alternative ITAMs were bound with 1000-10000-fold lower affinity (Ottinger et al., 1998). Recent work has also demonstrated the role of water molecules at the Src SH2 domain-ligand interface (Chung et al., 1998; Henriques and Ladbury, 2001).

Majority of the peptide motifs interacting with SH2 domains have been identified by using *in vitro* oriented phosphopeptide library assays (Songyang et al., 1993; Songyang and Cantley, 1995). Based on these results, together with structural analyses of different ligand-binding models, it is apparent that SH2 domains bind distinct but overlapping sequence motifs. The selectivity of an individual SH2 domain is not sharply defined, and a range of residues is typically tolerated at each site following the phosphotyrosine. Supporting these findings, different SH2 domains have been shown to compete for same binding target *in vivo* (*e.g.* Nishimura et al., 1993; Sayos et al., 2001).

3.4 Diseases related to SH2 domains

Mutations in the SH2 domains of Bruton tyrosine kinase (BTK), SH2D1A, Ras GTPase activating protein (RASA1), Zap-70, SHP-2 and the p85α subunit of the PIP3 kinase (PI3-K) cause nine distinct clinical phenotypes (Table 2). The domain organization of the proteins is given in Figure 5. Currently, 168 unique molecular events in 325 unrelated patients have been reported. The mutation types range from large gross deletions of the whole gene to single point mutations. Missense mutations comprise the most common mutational event (57%). Previously, proteins with an essential function have been shown to possess a more damaging phenotype (Jeong et al., 2001; Krylov et al., 2003). In agreement, proteins with defective SH2 domains either have a crucial role during cell development process or regulate multiple signaling cascades.

Table 2. Diseases related to SH2 domains[a].

Affected gene	Disease	Inheritance	Phenotypes
Btk	X-linked agammaglobulinemia (XLA)	X-linked	Hypogammaglobulinemia, antibody deficiency, and recurrent bacterial infections
SH2D1A	X-linked lymphoproliferative disease (XLP)	X-linked	Fatal infectious mononucleosis, malignant B-cell lymphomas and dysgammaglobulinemia
Zap-70	Severe combined immunodeficiency (SCID)	autosomal recessive	Severe pulmonary infection, chronic diarrhea, failure to thrive and persistent candidiasis
PTPN11	Noonan syndrome	autosomal dominant	Short stature, facial dysmorphia and wide spectrum of congenital heart defects
PTPN11	LEOPARD syndrome	autosomal dominant	Lentigines, ECG abnormalities, ocular hypertelorism, pulmonary stenosis, genitalia abnormalities, growth retardation, deafness.
PTPN11	Noonan-like/multiple giant-cell lesion syndrome	autosomal dominant	In addition to main Noonan syndrome phenotypes, giant-cell lesions of bone and soft tissues
PTPN11	Juvenile myelomonocytic leukaemia	n. a	~30% of myelodysplastic syndrome and 2% of leukaemia patients
p85alpha	Severe insulin deficiency	n. a	Acanthosis nigricans, hyperinsulinemia and diabetes mellitus at the later stage
RASA1	Basal cell carcinoma (BCC)	Sporadic	Clusters of basal cell carcinoma and development of tumours on the chest.

[a] not available.

The majority of the disease-causing mutations are found in BTK, SH2D1A and SHP-2. Analyses of missense mutations in these proteins have provided information of functionally and structurally important residues (Tzeng et al., 2000; Morra et al., 2001b; Hwang et al., 2002; Li et al., 2003a). However, no correlation has been described between the type and position of the mutations and clinical XLA (Vihinen et al., 1999) or XLP phenotype (Engel et al., 2003). In fact, identical mutations within the family have been shown to result in different phenotype (Kornfeld et al., 1997; Coffey et al., 1998; Nichols et al., 1998; Sumazaki et al., 2001). Furthermore, defects in *BTK* or *SH2D1A* have been mistakenly diagnosed with common variable immunodeficiency (CVID) (Spickett et al., 1997; Morra et al., 2001a; Nistala et al., 2001; Arico et al., 2002). CVID is the most common primary immunodeficiency with a highly heterogeneous clinical presentation and unknown genetic basis (Conley et al., 1999). Mutations in *BTK* and *SH2D1A* show a typical X-linked inheritance pattern without any genetic heterogeneity, and emphasize the importance of understanding how genetic defects cause clinical phenotype at the protein level.

On the contrary, mutations in *PTPN11* encoding the SHP-2 protein have been shown to cause at least four distinct diseases. The clinical and genetic heterogeneity of these disorders suggests a possible relation between different *PTPN11* mutations and distinct clinical features. Analyses of large cohort of individuals with Noonan syndrome (Tartaglia et al., 2001; Tartaglia et al., 2002) showed that *PTPN11* mutations are more likely to be found when pulmonary stenosis is present, whereas hypertrophic cardiomyopathy is less prevalent among these patients. In another study, this correlation was not found (Sarkozy et al., 2003). However, the location of mutations within the *PTPN11* gene correlated with different heart defects in Noonan and LEOPARD syndromes.

3.4.1 Mutations in BTK lead to X-linked agammaglobulinemia

BTK participates in immune cell signal transduction pathways regulating activation, proliferation, differentiation and apoptosis with the exception of T lymphocytes (Smith et al., 2001). Mutations in all five domains of BTK have been shown to cause X-linked agammaglobulinemia (XLA) by disrupting the pre-B cell receptor signal cascade (reviewed in Kurosaki, 2002). As a result, B-cell maturation is arrested between pro- and pre-B-cell stages and the complete lack of mature B-lymphocytes leads to extreme susceptibility to bacterial infections in patients (reviewed in Vihinen et al., 2001).

Although BTK has been shown to associate with a large number of proteins (Smith et al., 2001), the SH2 domain has been reported to interact only with the B-cell linker protein (BLNK) *in vivo* (Hashimoto et al., 1999; Su et al., 1999). B-cell receptor (BCR) engagement leads to phosphorylation of several BLNK tyrosines and, thereby, formation of an active complex as BTK, PLCγ2, Grb2 and Vav bind to BLNK. Recruitment of BTK and PLCγ2 proteins close together allows BTK to phosphorylate PLCγ2, which then leads to a sustained calcium release from the storage vesicles (Fluckiger et al., 1998). Calcium concentration has various general effects in B-lymphocytes *e.g.* regulation of transcription factors related to proliferation (Tan et al., 2001). Furthermore, BCR stimulated B-cells from XLA patients did not show elevated calcium mobilization (Genevier and Callard, 1997). Currently, 58 different XLA mutations in 102 patients have been reported from the SH2 domain (http://bioinf.uta.fi/BTKbase).

3.4.2 Genetic cause of X-linked Lymphoproliferative Disease

SH2D1A is a small lymphocyte-specific signalling molecule that is defective or absent in patients with X-linked Lymphoproliferative Disease (XLP) (Coffey et al., 1998; Nichols et al., 1998; Sayos et al., 1998). Unlike typical signalling proteins, SH2D1A is comprised of a single SH2 domain followed by a short tail. A total of 100 disease-causing mutations have been reported from 85 unrelated families (http://bioinf.uta.fi/SH2D1Abase). All missense mutations affect the SH2 domain.

SH2D1A has a dual role in regulation of the initial signal transduction events induced by at least six members of the SLAM (signal lymphocyte-activator molecule) family of cell-surface receptors. These receptors function in the immune synapse, between T lymphocytes or natural killer cells and antigen presenting cells (reviewed in Engel et al., 2003). SH2D1A binds to the cytoplasmic tail of SLAM family receptors through a conserved T-(I/V)-pY-X-X-(V/I) motif (where X is any amino acid). The structural basis for the specific recognition of SLAM by SH2D1A has been unravelled by both X-ray crystallography and NMR methods (Poy et al., 1999; Hwang et al., 2002). In addition to conventional SH2-ligand interactions, SH2D1A forms also specific interactions to the residues preceding the phosphotyrosine in the ligand. These interactions allow this protein to bind SLAM receptor independently of its phosphorylation status, and thereby, block the recruitment of SH2-containing signal-transduction molecules, such as SHP-2 (Sayos et al., 1998; Sayos et al., 2001). SH2D1A has also been shown to function as an adaptor molecule. The SH2 domain surface formed by positively charged residues in βF strand, N-terminal end of the αB helix and the intervening loop associates with an electrostatically complementary interface on the Fyn SH3 domain. Furthermore, the buried surface does not involve the phosphotyrosine binding site, whereas the bound surface on the SH3 domain overlaps the surface that is expected to participate in auto-inhibitory interactions in the Fyn kinase (Latour et al., 2001; Chan et al., 2003; Latour et al., 2003; Li et al., 2003b). The interaction between these domains results in recruitment of an active Fyn kinase close to active receptors in the immune synapse, and subsequently, phosphorylation of tyrosines in the cytoplasmic tails of these receptors. A number of missense mutations locate on the conventional ligand-binding surface, whereas none have been found from the Fyn binding surface. However, mutations leading to unstable SH2D1A may cause XLP by preventing the initial mechanism in which an adaptor molecule is required to link a receptor devoid of intrinsic catalytic activity to a cytoplasmic tyrosine kinase.

3.4.3 Mutations affecting ZAP-70

ZAP-70 protein consists of two SH2 domains followed by a C-terminal kinase domain. Association with both SH2 domains to the ζ chain of activated T cell antigen receptor (TCR) have been shown to regulate multiple downstream pathways after receptor activation (Chan et al., 1992). Genetic alterations in the *ZAP-70* gene lead to an extremely rare autosomal recessive form of severe combined immunodeficiency (SCID), also named as ZAP-70 deficiency. Only one of the reported fourteen patient mutations affects αB helix of the N-terminal SH2 domain (http://bioinf.uta.fi/ZAP70base). Although, the mutated protein associated with the ζ chain of TCR in a wild type manner *in vitro*, it is degraded rapidly *in vivo* (Matsuda et al., 1999). The loss of ZAP-70 function leads to selective inability to produce CD8+ T lymphocytes and abolishes TCR stimulation in mature CD4+ T lymphocytes (Arpaia et al., 1994; Elder et al., 1994). ZAP-70 deficiency is ultimately fatal unless patients undergo bone marrow transplantation.

Recently, a spontaneous missense mutation in the βB strand of C-terminal SH2 domain was shown to cause chronic autoimmune arthritis in mice that resembles human rheumatoid arthritis (Sakaguchi et al., 2003). Altered signal transduction from T-cell antigen receptor through the aberrant ZAP-70 is likely to change the threshold of T lymphocytes to thymic selection, leading to positive selection of otherwise negatively selected autoimmune T cells.

3.4.4 PI3K mutation is associated with severe insulin deficiency

Phosphatidylinositol 3-kinase (PI3K) plays a pivotal role in signal transduction pathways linking insulin with many of its specific cellular responses, including GLUT4 vesicle translocation to the plasma membrane and inhibition of glycogen synthase kinase-3 (Shepherd et al., 1998). Moreover, PI3K is necessary for the insulin-stimulated increase in glucose uptake, and glycogen synthesis in insulin-sensitive tissues (Holman and Kasuga, 1997). The structure of PI3K is heterodimeric, consisting of a catalytic subunit (p110) and a regulatory subunit (p85α) (Antonetti et al., 1996).

Recently, a missense mutation was found in the N- terminal SH2 domain of p85α leading to severe insulin resistance (Almind et al., 2002). The R409Q mutation is located in the C-terminus of αB helix, and is not involved in the normal ligand-binding surface. However, when binding of N-SH2 domain with mono and double phosphorylated ligands was studied with NMR spectroscopy, the doubly phosphorylated peptide showed nearly 10-fold higher binding to the isolated SH2 domain. From the NMR structure, it appears that the second phosphotyrosine is coordinated by the residues in BG-loop and C-terminal part of the αB-helix (Weber et al., 2000).

3.4.5 Sporadic mutations leading to Basal-cell carcinoma

Basal-cell carcinoma (BCC) is the most frequent skin cancer in the white population (Miller, 1991). BCCs mostly occur sporadically in relation to sun exposure, although their incidence is increased significantly in some rare genetic disorders (Gorlin, 1987; Bodak et al., 1999). Somatic mutations at the phosphotyrosine-binding pocket of the C-terminal SH2 domain of GTPase-activating protein RASA1 have also been found in a subset of BCCs (Friedman, 1995). RASA1 acts by enhancing the intrinsic GTPase activity of Ras, leading to hydrolysis of bound GTP to GDP and down regulation of Ras activity (Gold et al., 1993; Lazarus et al., 1993; Scheffzek et al., 1998). The structure of the defective SH2 domain has not been solved.

3.4.6 Mutations affecting PTPN11 gene

Mutations in the *PTPN11* have been found from patients suffering from Noonan syndrome (NS), LEOPARD syndrome or juvenile myelomonocytic leukaemia (JMML) (Tartaglia et al., 2001; Digilio et al., 2002; Tartaglia et al., 2002; Loh et al., 2003). The gene encodes SHP-2 protein, a ubiquitously expressed cytosolic non-receptor tyrosine phosphatase (PTP). SHP-2 is a key molecule in the cellular response to growth factors, hormones, cytokines and cell adhesion molecules (reviewed in Neel et al., 2003).

The SHP-2 is composed of two tandem N-terminal SH2 domains, a PTP domain, and a C-terminal tail. The structural data revealed the functional role of the N-terminal SH2 (N-SH2) domain in regulating the enzyme activity. The D'E loop and flanking βD' and βE strands of the N-SH2 domain extend deep into the catalytic cleft of the PTP domain blocking the enzyme active site. An intricate intra- and interdomain hydrogen-bonding network together with charged interactions stabilize the D'E loop conformation in the enzyme active site (Hof et al., 1998). Binding of N-SH2 domain to its phosphorylated ligand induces a conformational change that prevents PTP domain binding at a second site (Lee et al., 1994; Eck et al., 1996). The NS-causing *PTPN11* mutations cluster in the interacting portions of the N-SH2 and PTP domains (Tartaglia et al., 2001). Most of the residues mutated in NS are either directly involved in these interdomain interactions or in close spatial proximity leading to constitutively active enzyme.

4 Methyltransferase domains

The human family of DNA cytosine 5-methyltransferases (m^{5}C-MTases) consists of five family members (reviewed in Bestor, 2000). These enzymes catalyse the transfer of a methyl group from S-adenosyl-L-methione (AdoMet) to the target cytosine in DNA, with the exception of DNMT2 that is yet to be established as a catalytically active enzyme (Okano et al., 1998). DNMT1 acts as the classical maintenance methyltransferase being responsible for preservation of methylation pattern during DNA replication (Bestor et al., 1988), whereas DNMT3A, DNMT3B and DNMT3L participate in establishment of *de novo* methylation patterns during early embryonic development in a sex-specific manner (La Salle et al., 2004). The effects of DNA methylation are widespread including transcriptional repression by methylation of promoter regions (Jones, 1996), formation of compact chromatin structures (Kass et al., 1997), X-chromosome inactivation (Panning and Jaenisch, 1998) and imprinting control (Li et al., 1993).

4.1 Methyltransferase domain structure

In addition to the crystal structures of bacterial DNA methyltransferases from *Haemophilus haemolyticus* (M. *Hha*I) and *Haemophilus aegypitius* (M. *Hae*III) (Cheng et al., 1993; Klimasauskas et al., 1994; Reinisch et al., 1995), only the structure of human DNMT2 m^{5}C-Mtase domain has been solved (Dong et al., 2001). At the structural level, the bacterial and the human m^{5}C-Mtase structures have a common two-domain structure (Figure 6). The target DNA segment is bound between the two domains on a surface having positive electrostatic potential.

The experimental structures have revealed high sequence and structural similarity among the larger subdomains, whereas the similarity decreases in smaller subdomains. The core of the larger subdomain is composed of a six-stranded β-sheet sandwiched between two α-helices (C and D) on one side and two on the other (A and G). The αB-helix runs across the sheet in front of the sandwich. In contrast to larger subunits, all three known protein structures have different number and organization of secondary structures (Cheng et al., 1993; Reinisch et al., 1995; Dong et al., 2001). The conformational variety is required for recognition and binding of the specific target DNA sequences.

4.2 Methyltransferase domain function

The biochemical and structural analyses with M. *Hha*I have revealed the common mechanism of cytosine methylation. The interaction between the M. *Hha*I and DNA is dynamic. The target cytosine has two conformations; it is either flipped out from the double helix to the catalytic pocket near the AdoMet binding site or remains in the stacked state (Klimasauskas et al., 1998). The residues involved in this process are located in the smaller subdomain forming the target recognition domain (TRD) loop, the only conserved segment in the smaller subdomains among the M. *Hha*I, M. *Hae*III and DNMT2. The TRD loop runs parallel with the DNA strand and serves as a scaffold for conformational processing of the bound substrate (Cheng and Blumenthal, 1996). Binding of the cofactor induces a conformational change in active-site loop located in the larger subdomain. The loop folds on top of the cytosine locking it into the catalytic site. In the M.*Hha*I-DNA-AdoMet structure, arginine 165 and glutamate 119 bind to the cytosine allowing the cysteine 81 to attack the sixth carbon of the cytosine. This results in the addition of a methyl group to the fifth carbon of the cytosine followed by elimination of the proton from the fifth carbon and release of the covalent intermediate (reviewed in Kumar et al., 1994).

Interestingly, the m^{5}C-Mtases do not show binding specificity for the flippable base itself (Klimasauskas and Roberts, 1995; Yang et al., 1995). Instead, a number of specific interactions occur between the smaller subdomains and the major groove of the DNA. The residues involved are generally not conserved and the types of contacts differ between the known structures, with the exception of the threonine 250 of M. *Hha*I. This residue is conserved among the m5C-Mtases and the threonine is involved in conformational change of the target cytosine backbone as the base flipping occurs (Vilkaitis et al., 2000).

4.3 Diseases related to methyltransferase domain

Mutations in the gene encoding for a DNMT3B lead to a rare autosomal recessive immunodeficiency, centromeric instability and facial anomalies (ICF) syndrome (OMIM, 242860, Wijmenga et al., 1998. The chromosomes 1, 9 and or 16 are instable in all known ICF patients (Jeanpierre et al., 1993). These DNA regions contain classical satellites II and III, which are the major components of constitutive heterochromatin and are normally heavily methylated (Schuffenhauer et al., 1995; Miniou et al., 1997). In addition to the ICF syndrome, the expression levels of DNMT3B and its splice variants also play a role in different forms of cancer (Robertson et al., 1999; Kanai et al., 2001).

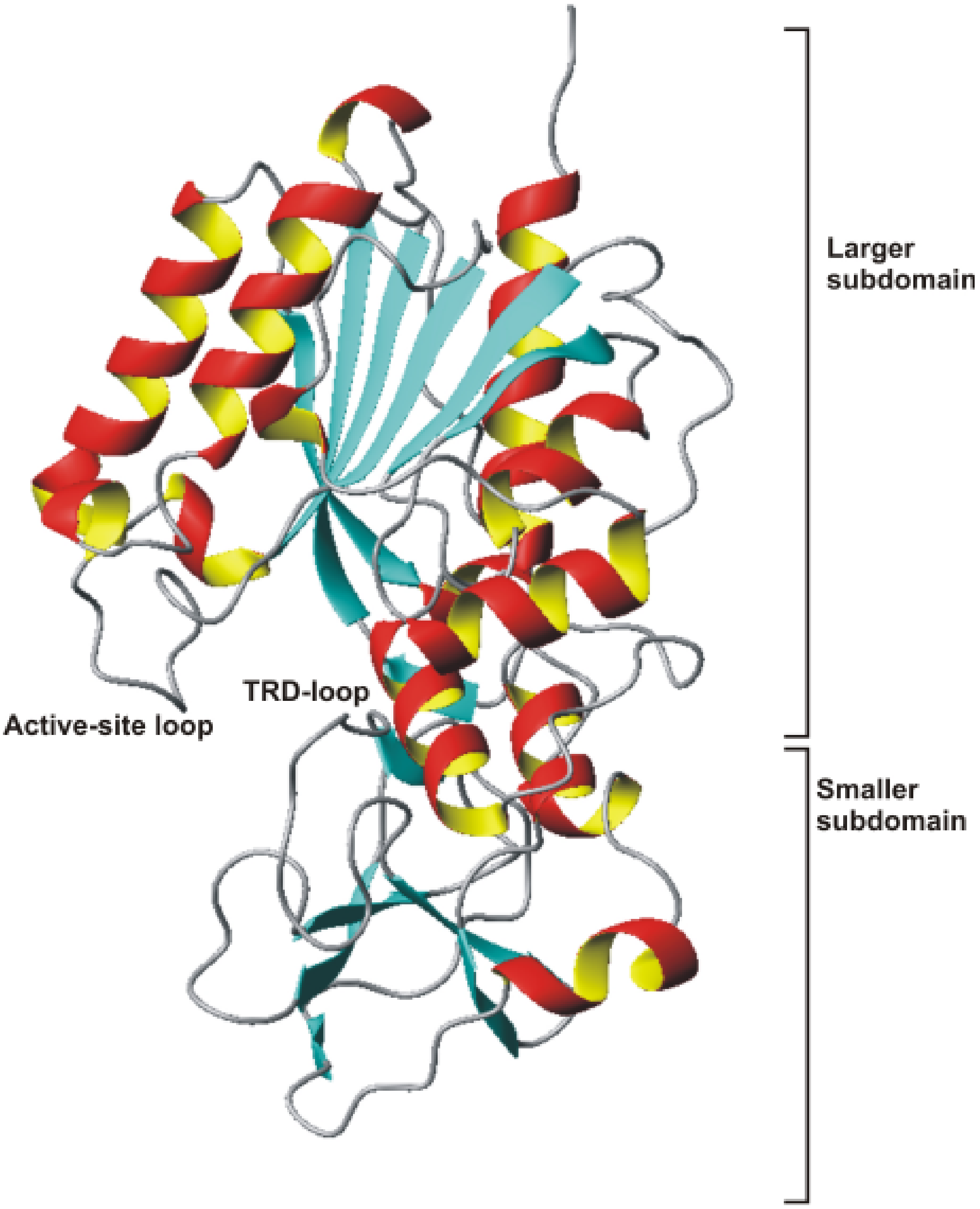

Figure 6. A ribbon model of the methyltransferase domain of *Haemophilus haemolyticus* (PDB code 6MHT) showing the secondary structures as in Figure 4. The structure has a typical two subdomain structure. The target DNA segment binds between the subdomains on a large positively charged surface. The larger subdomain contains conserved binding sites for the target cytosine and cofactor, S-adenosyl-L-methione, whereas the smaller subdomain consists of a target recognition (TRD) loop involved in flipping the cytosine from the double helical DNA conformation into the catalytic pocket of the protein. The cytosine is locked into catalytic site by active-site loop. The conformation of the larger subdomain is similar in all known methyltransferase structures, whereas the smaller subdomains differ in sequence and structure.

5 Aims of the study

Roughly hundred genes have been shown to underlie in various forms of immunodeficiency. Over the years, we have collated more than 2700 mutations together with patient data into locus-specific mutation databases. The purpose of the present study was therefore:

- to collate pathogenic mutations into locus-specific databases and to create methods for maintenance and analysis of the databases.

- to study if the nucleotides surrounding mutation site affect the frequency of mutation.

- to find a computationally suitable way that could be applied for predicting the effect of small local perturbations occurring on defective protein structures. The method was applied on two distinct protein domains, Src homology 2 (SH2) and methyltransferase domains.

- to validate bioinformatical results, X-linked agammaglobulinemia-linked mutations were tested by investigating the properties of cloned and produced recombinant proteins by using various biochemical methods.

6 Materials and Methods

Detailed description of materials and methods are found in the original publications.

Materials and methods	Original publication
Creation and maintenance of mutation databases	II, V, VI
Analyses of mutation mechanisms	I, II, VI
SCCP screening of XLA and XLP patients	III, IV
Construction of SH2 domain plasmids	III
Expression and purification of recombinant proteins	III
In vitro and *in vivo* solubility analyses	III
Binding of SH2 domains to pY-sepharose	III
Circular dichroism analyses	III
Sequence analyses and comparative modeling	IV, V, VI
Calculation of electrostatic surface potential	V
Analyses of mutation effects on corresponding structure	III, IV, V, VI
Analyses of mutation effects based on sequence entropy	VI

7 RESULTS

7.1 Locus-specific mutation databases (II, IV, V, VI)

In the present study, we have collected mutations from five different genes, *BTK* (II)*, SH2D1A* (IV) as well as *SHP-2, p85α* of *PIP3K* and *RASA1* (VI)*,* into locus-specific mutation registries. All encoded proteins include SH2 domain(s) allowing analyses of mutation consequences on a distinct protein fold. A generic mutation database, SH2base, was created to link results from the mutational analyses to the locus-specific mutation databases (VI). Furthermore, ICF causing mutations were collated into DNMT3Bbase (V). All mutation databases have been updated regularly, and provide the largest publicly available number of different mutations identified from 1071 patients. The databases are available at http://bioinf.uta. fi/ base_root/.

The database format used in the locus-specific mutation registries follows the guidelines first adopted in the BTKbase (Vihinen et al., 1999; Vihinen et al., 2001). Each patient forms an individual entry with a unique accession number and patient identification number (PIN). The accession number cannot change in any circumstance. PIN is used for indicating the type and location of a mutation in a simple way, *e.g.* the first described mutation of arginine to glycine at position 302 is R302G(1). Family members carrying an identical mutation are indicated with alphabets following the PIN, *e.g.* R302G(1a) and R302G(1b). The mutation consequences are depicted at the corresponding reference sequences by analysing the effects of the mutation at the genomic DNA through mRNA to protein level. In addition to mutations, registries contain information about symptoms, age at diagnosis and various parameters from the patients. Data security protects the patient's identity, which is not coded into the registry and not even known by the coordinators of databases. The database format is understandable without a computer program, but allows development of specific programs for maintenance and analyses. All databases are currently distributed as flatfiles.

The data in locus-specific registries was used to analyse the different mutation types (II, IV, V, VI). Majority of intron, deletion and insertion mutations in the databases lead to an introduction of a termination codon. Furthermore, intron mutations were found to affect mainly the classical splice sites causing aberrant splice-variants. Generally, insertions and deletions (indels) are only 1-5 bp of length. Direct repeats were found to appear in the immediate vicinity of all indels suggesting that these mutations have arisen through DNA slippage method. Nonsense mutations are located predominantly in the first and third codon positions with transitions and transversions occuring equally. The majority of missense mutations (85%) appear in the first two positions within the codon as would be expected. These two positions primarily determine the coded amino acid. Transitions (56%) dominate over transversions (42%), although the rate is typically higher for disease-causing mutations (Stenson et al., 2003).

The CpG dinucleotides form the single most mutated site among the analysed patients (II, IV, VI). The SH2 domains consist of 69 CpG spots comprising roughly 5% of the coding regions, with the exception of N-terminal SH2 domain of Zap-70 (21%). Although mutations were found in only eight sites, deamination of CG to TG or CA constitutes 23 % of all single base substitutions affecting the SH2 domains. The vast majority of the CpG mutations affect arginines required for phosphotyrosine binding. A high number of patients result from a common DNA mutation mechanism causing a severe enough phenotype to be clinically observed.

The early version of BTKbase was maintained manually (II). Mutations were collated from the literature or sent to the BTKbase study group via email in various formats. Partly to assist and standardize database maintenance, the MUTbase program package was created (Riikonen and Vihinen, 1999). At the present, the disease-causing mutations are usually submitted to the particular database coordinators by physicians or directly by the laboratories performing gene tests using a registry-specific form available in the Internet. The submission program verifies that the mutation is correctly located at the genomic reference sequence and determines the consequences at the RNA and protein levels automatically. The mutation is then emailed to the study group in a complete database format for confirmation. Therefore, mutations are never included into databases without human verification.

In addition, separate programs were created for SH2base to connect the mutation analyses and locus-specific mutation databases including patient and mutation data (VI). The programs separate only SH2 domain affecting mutations from the locus-specific mutation registries and automatically update tables describing the number of unique mutations and total number of analysed patients in each SH2 domain. It will also be possible to search mutations affecting certain positions among the SH2 domains, or to link all mutations causing particular phenotype to a certain gene or position in the gene. In the near future, the results of mutation analyses in SH2 domains (VI) will be connected to MultiDisp program (http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi), allowing *e.g.* further studies of residue types at particular position and grouping of proteins based on position specific amino acid frequency.

7.2 Analyses of pathogenic mutations in the DNMT3B (V)

The human DNMT3B belongs to the family of DNA cytosine-5-methyltransferases (m^{5}C-Mtases) (reviewed in Bestor, 2000). These enzymes catalyse the transfer of a methyl group from S-adenosyl-L-methinine to the C5 position of cytosine. The effects of DNA methylation are widespread including *e.g*. transcriptional repression by methylation of promoter regions and X-chromosome inactivation (Panning and Jaenisch, 1998). Mutations in the gene encoding for a DNMT3B, lead to an autosomal recessive Immunodeficiency, Centromeric instability and Facial anomalies (ICF). The project aim was three-fold: to collate all ICF-causing mutations into a database (DNMT3Bbase), to analyse the structure-function consequences of the mutations on the modelled domain structure and to develop a systematic method for detecting disruptive mutations at the structural level.

The sequence similarity between the DNMT3B and the known methyltransferase enzymes suggests a similar two-subdomain fold for the DNMT3B as seen in the experimentally resolved methyltransferase domains. However, this does not unambiguously prove that the DNMT3B adopts a typical methyltransferase structure. Based on the hypothesis, the DNMT3B methyltransferase domain was modelled by using *Haemophilus haemolyticus* protein structure (PDB code 6MHT) as a template. The structure has been solved with X-ray crystallography at 2.05 Å resolution (Kumar et al., 1997). As a result of low sequence similarity between the DNMT3B and template in the region corresponding to smaller subdomain (amino acids 740-800 and 174-274, respectively), only the larger subdomain was modeled. The target recognition domain (TRD) –loop was created to illustrate the single ICF-causing mutation identified in the smaller domain.

In all known methyltransferase domain structures, the DNA is bound to a large positively charged surface between the larger and smaller subdomains (Cheng et al., 1993; Reinisch et al., 1995; Dong et al., 2001). The cytosine is flipped from the target DNA sequence to the catalytic pocket on the surface of the larger subdomain. The methyl group is transferred to the cytosine from the cofactor, bound to another binding site next to the cytosine pocket. The catalytically important residues are located in the larger subunit, whereas all residues required for recognition of the target DNA sequence and for the flipping of the cytosine from the double helical DNA structure into the catalytic pocket are situated in the smaller subdomain. The DNMT3B model has a large positive surface corresponding to the DNA-binding surface of the larger subunit in the template. The majority of the catalytically important residues or amino acids required for binding of the cofactor are conserved (Figure 2, V). Although it was not possible to model the smaller subdomain, the sequence similarity in TRD loop region between the DNMT3B and the known methyltransferases suggests a common flip mechanism also for the DNMT3B methyltransferase domain.

Based on the analyses, the disease-causing mutations were found to disrupt the local structure or remove hydrophobic interface between the domain and the cofactor. Two mutations, G663S and L664T, affect the flexibility and packing of the active-site loop that folds on top of the cytosine burying it to the catalytic pocket. Three ICF-causing mutations, H814R, D817G and V818M appear on the border of a large positively charged surface next to the DNA and cofactor binding areas. Based on the modelled structure, these mutations seem not to cause structural alterations and most likely affect either solubility or an unknown intra/inter molecular binding site. Furthermore, charged residues without a direct interaction with ligand have been shown to be important for neutralization of residues in the ligand binding site or making long-range electrostatic interactions with the ligand (Bradshaw et al., 2000; Sheinerman et al., 2003). To reveal the role of these three residues in DNMT3B function or stability clearly requires biochemical analyses. The region preceding the β6 strand in the larger subunit of the DNMT3B model illustrates another example of the limitations in comparative modelling. The secondary structure predictions indicate only short β5 strand in DNMT3B, and the program detected no α-helical sequence following the β5 strand. Furthermore, this region faces the smaller subdomain in all known methyltransferase structures. Therefore, the model was not used to analyse the effects of the V726G mutation situated in the loop between β5 and β6 strands.

7.3 Nucleotide neighbourhood in CpG mutations (I)

Majority of methylated cytosines appear in CpG dinucleotides in the human genome. The CpG dinucleotides are also a known hotspot in many diverse diseases as a result of spontaneous deamination of 5-methylcytosine to thymine. To examine the effect of nucleotide neighbourhood surrounding the CpG dinucleotides, mutations affecting CpG sites in five different diseases were analysed. The genes encode for Bruton tyrosine kinase, factor VIII, tumour suppressor p53, phenylalanine hydroxylase and protein C. Only protein C contains a CpG island at the promoter region of the gene. The CpG dinucleotides are highly suppressed in all genes and they comprised at least a third of disease-causing missense and nonsense mutations.

Excluding mutations affecting protein C, the most frequently mutated tetra- and heptanucleotides sequences have a common pattern of YYCGRY/R, where Y denotes for pyrimidine and R purine (I, tables 2 and 3). The vast majority of the mutated CpG sites in the protein C are located inside the CpG island where cytosines are typically found unmethylated (Bird, 1986). These mutated CpG dinucleotides did not share the YCGR pattern. No sequence preferences were found for CpNpG triplets (N is any nucleotide but guanine), although CpNpG sites are stably methylated in mammalian cells at low frequency.

7.4 Putative effects of pathogenic mutations in the SH2 domains (III, IV, VI)

Mutations in the SH2 domains of Bruton tyrosine kinase (BTK), SH2D1A, Ras GTPase activating protein (RASA1), Zap-70, SHP-2 and the p85α subunit of the PIP3 kinase (PI3-K) have been shown to cause nine distinct clinical phenotypes (Table 1, VI). Currently, 168 unique molecular events in 325 unrelated patients have been reported (Table 2, VI). To examine the putative structural and functional consequences, we have analysed the mutations on corresponding protein structures. By using sequence entropy, disease-causing mutations were found to affect conserved positions, or to introduce an unnaturally occurring side chain in the disease position. Moreover, disease-causing mutations affected covariant positions among the SH2 domains (Figure 3, VI).

The structural information for the SH2D1A and RASA1 defective SH2 domains was not available. The RASA1 C-terminal SH2 was modelled based on the crystal structure of N-terminal SH2 domain of SHP-2 (1AYA, 2 Å resolution) (Lee et al., 1994) (VI). The SH2D1A SH2 domain was modelled based on the crystal structure of Abl SH2 domain (2ABL, 2.5Å) (Nam et al., 1996) (IV). Both modelled structures showed typical structural features when evaluated with PROCHECK program (Laskowski et al., 1993), also used to validate experimental structures.

The model of the SH2D1A differs from the crystal and NMR structures of the SH2D1A in two locations (Figure 7). As a result of more open BC-loop conformation in the template Abl SH2 domain, the model of the SH2D1A was interpreted to interact also with an unphosphorylated tyrosine in the SLAM receptor (IV). However, the SH2D1A crystal structure revealed an ordered water molecule replacing the interaction between phosphate group and RβB5. Furthermore, the interactions with the residues preceding the phosphotyrosine in the ligand provide additional binding energy (Poy et al., 1999; Hwang et al., 2002). The structural model also has an incorrect EF-loop structure as the loop in the model follows the Abl SH2 domain conformation. However, the altered loop conformation had no effect on mutation analyses. The domain model for SH2D1A was originally used to analyse the structural consequences of the known XLP mutations (IV), whereas the crystal structure was used for the analyses in (VI).

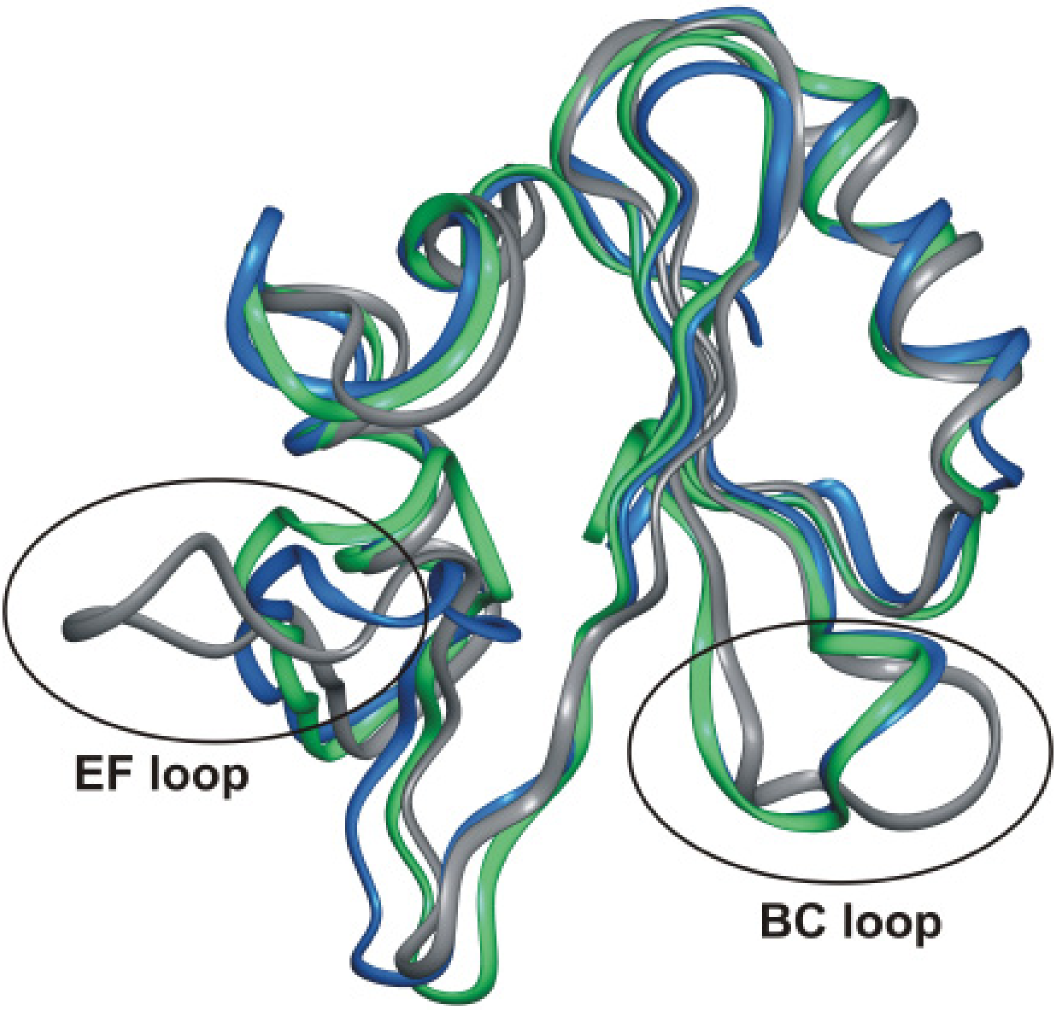

Figure 7. The model of the SH2D1A SH2 domain (grey) superimposed with crystal structure (1D1Z, green) and averaged NMR structure (1KA7, blue). The major differences occur in EF and BC loops. The backbone RMSD deviation between the NMR and crystal structures is 1.6 Å. SH2D1A model has a RMSD of 3.2 Å with the NMR structure and 3.5 Å with the crystal structure and a backbone RMSD for the secondary structures 1.4 Å and 1.6Å, respectively.

One third of the pathogenic mutations affecting SH2 domains either introduce a termination codon or lead to out of frame translation. Should these variants bypass the nonsense-mediated mRNA decay machinery, the variant proteins are likely to be extremely unstable. The XLA and XLP patients also have a large number of gross deletions (20%) causing deletion of the whole gene locus. Roughly half of the mutations are single point mutations at the coding region. Of the 79 different missense mutations, 59% are located on β-strands. The remaining mutations are situated equally in α-helices and loop structures. Based on our analyses of the SH2 domain structures, the large majority of missense mutations were found to affect protein stability (59%). Although missense mutations appear to destabilize the SH2 domain structure, majority of the mutations affecting identical positions in several SH2 domains are located on the binding surface. These missense mutations affect highly conserved residues involved in phosphotyrosine binding as well as unconserved positions related to SH2 domain specificity (Figure 2, VI).

7.5 Biochemical analyses of XLA-causing mutations in the SH2 domain of BTK (III)

The structure of BTK SH2 domain has not been solved and a structural model based on Src SH2 domain (Vihinen et al., 1994) was used instead to choose six XLA-causing mutations for the analyses. The R307G, Y334S, H358F, Y361C and H362Q mutations affect residues involved in ligand-binding, whereas G302E mutant was predicted to destabilize the protein (Figure 1A, III). Furthermore, a non-patient mutation of C337S was introduced to prevent observed *in vitro* dimerization of native SH2 domains. The cysteine is located close to the ligand-binding surface, is largely exposed to solvent and was not predicted to interact with the ligand. The mutations were analysed for their consequences to the protein structure and function by using circular dichroism (CD) spectroscopy, and for their ability bind to phosphotyrosine. The expression of mutated proteins *in vitro* was studied by creating several constructs for the analyses. Three of the mutants (R307G, Y334S and G302E) were also introduced into full-length BTK protein and transiently expressed in COS-7 cells to analyse the differences in stability between isolated SH2 domain mutants and BTK *in vivo*.

Circular dichroism spectrum of the wild type BTK SH2 domain showed a typical pattern of α+β protein. The native SH2 and R307G mutated proteins have identical spectra with broad minimum in the range of 212-220 nm as a result of overlapping α-helix and β-sheet bands. The C337S and Y361C CD spectra are highly similar to the spectrum of native BTK SH2 domain, whereas the other mutants studied differed clearly from the native SH2 (Figure 2, III). The spectra of G302E, Y334S and H362Q have only one minimum at 207 nm and the spectrum of L358F has two minima at 207 and 214 nm. Although the expression levels of mutated proteins were increased in lower temperatures, proteins with XLA mutation in isolated SH2 domains shared reduced solubility and binding of phosphotyrosine (Tables 1 and 2, III). However, the expression levels of full-length mutated BTK proteins were nearly identical to the wild type and the proteins were not prone to aggregation during analyses.

8 DISCUSSION

8.1 Creation and analyses of locus-specific mutation databases

In 1952, Dr. Bruton reported the first immunodeficiency in an eight-year-old boy suffering from agammaglobulinemia (Bruton, 1952). Almost forty years later, the gene affected in XLA was identified from the X-chromosome, and found to encode a novel cytoplasmic tyrosine kinase (BTK) (Tsukada et al., 1993; Vetrie et al., 1993). Today, more than 100 different genes have been linked to immunodeficiences, and roughly 1500 genes with other syndromes. In the present study, mutations from six different genes, *BTK, SH2D1A, SHP-2, p85α* of *PIP3K, RASA1,* and *DNMT3B* were collated into locus-specific mutation registries (II, IV, V, VI). In addition, a generic mutation database, SH2base, was created (VI). The databases were used to analyse the mutation types and mechanisms, especially deamination of cytosines in CG dinucleotides (I).

The analyses of mutation types in the locus-specific registries show similar distribution as has been shown for many inherited human diseases (Stenson et al., 2003) (II, IV, V, VI). The transitions occur more frequently than tranversions, although the rate is not as high as reported in the HGMD. Furthermore, missense mutations account for roughly half of all pathogenic mutations in the analysed databases. The insertions or deletions are typically of 1-5 bp of length with direct repeats at the vicinity favouring the DNA slippage model. Although the CG dinucleotides were suppressed in almost all analysed genes, point mutations at the CG dinucleotides comprised 23% of all point mutations (II, IV, VI). Further analyses of CpG sites in five different pathogenic mutation databases affecting not only SH2 domains clearly showed that the 5-methylcytosine is more prone to mutate when it occurs in a specific nucleotide environment (I). In four of the analysed diseases, the most frequently mutated CpG dinucleotide was preceded by a pyrimidine and followed by a purine. The environment also differs for CpG dinucleotides located inside a CpG island, where cytosines are predominantly unmethylated. Krawczak and Cooper have also found the same nucleotide environment surrounding the CpG site in their analyses for a larger dataset (Krawczak et al., 1998).

8.2 Mutations affecting SH2 domains

8.2.1 Biochemical analysis of six XLA-causing mutations

The structure of BTK SH2 domain has been modeled based on the Src SH2 domain. The pairwise sequence identity between BTK SH2 domain and Src SH2 domain is 30%, and the residues involved in ligand-binding are highly conserved in both domains. Furthermore, BTK SH2 domain has been shown to interact with the same high affinity peptide pYEEI *in vitro* that was complexed with the Src SH2 domain in template structure (Waksman et al., 1992; Tzeng et al., 2000).

The effect for function of six XLA-causing mutations in the SH2 domain BTK was studied (III). Five of these mutations (R307G, Y334S, L358F, Y361C and H362Q) were predicted to affect the ligand binding based on a structural model of the domain. One mutation, G302E, is located at the AB loop connecting αA-helix to βB-strand. This position is highly conserved among the SH2 domain and mutation of glycine to glutamate was predicted to disrupt the protein structure enabling us to clearly discriminate between functional and structural mutations. A non-patient mutation, C337S, was originally made to prevent dimerization of native SH2 domains during purification of this construct for structural studies. The cysteine is solvent exposed and not involved in ligand binding based on the SH2 domain model.

All XLA-causing mutations were found to reduce solubility and pY binding ability *in vitro*, whereas XLA-causing mutations introduced to full-length BTK were shown to share almost identical *in vivo* stability and expression levels in COS-7 cells (III). The *in vitro* stability of BTK determined by equilibrium denaturation is only 2.9 kcal/mol (Tzeng et al., 2000). The results indicate that the isolated BTK SH2 domain is only marginally stable, and is therefore intolerable to even slightly destabilizing mutations.

Changes in CD spectra were used to predict structural and functional consequences of each mutation (III). The R307G, C337S and Y362C mutated proteins had similar CD spectra with the native BTK SH2 domain. The R307G mutation corresponds to the invariant RβB5 recognizing phosphotyrosine from the ligand. The R307G mutation does not alter the domain structure and causes XLA by disrupting the interaction between BTK and BLNK proteins. In Src SH2 domain structure, the tyrosine corresponding to Y361 of BTK is solvent accessible and the hydroxyl group extends to the pY+3 binding pocket making a hydrogen bond to the isoleucine at the pY+3 position of ligand. Based on the BTK SH2 model and biochemical data of the mutated protein, Y361C mutation causes a mild case of XLA by destabilizing protein and reducing ligand binding. The spectra of other mutants differed from the wild type indicating conformational changes of secondary structures in relation to each other. Furthermore, G302E and L358F mutations have been shown to decrease BTK protein level *in vivo* (Saffran et al., 1994; Futatani et al., 1998). Both mutantions introduce a larger side chain likely to clash with the surrounding residues leading to unpredictable structural alterations.

8.2.2 Comparison of disease-causing mutations on SH2 domain structures

The analysed missense mutations can be categorized into two groups (III, VI). Mutations abolishing only SH2 domain function may either affect the enzyme activity as in the case of *PTPN11*, or lead to improperly located active enzyme in the cells. The mutations predicted to be functional deleted crucial hydrogen bonds or lead to loss of hydrophobic interactions between the SH2 domain and ligand or between the N-terminal SH2 domain of SHP-2 and the phosphatase domain (VI). As an example, two identical positions were affected in BTK and SH2D1A SH2 domains leading to XLA and XLP, respectively. The R32T (corresponding to RβB5) mutation in SH2D1A was predicted to cause XLP by removing the crucial interaction between the domain and phosphotyrosine (IV, VI). An identical mutation (R307T) has also been indentified from the BTK SH2 domain. The R307G mutation shared similar CD spectrum with the wild type BTK SH2 domain but the mutated SH2 domain was unable to bind phosphotyrosine (III). Furthermore, the R307G mutation was not found to influence the stability i*n vivo* (Vorechovský et al., 1997).

Mutations affecting the thermodynamic stability or kinetic pathway of folding often lead to aggregation, as the mutated form either possesses aggregation prone motif on the surface of the protein or the mutated protein is not able to fold fast enough. A large number of pathogenic mutations in SH2 domains are not involved in ligand binding. These mutations were found to influence conserved positions either in the central β-sheet or in the loops mainly connecting αA helix and βB strand or between αB helix and βG strand (VI). As an example, the G302E and G302R mutations in BTK and the corresponding mutation G27F in SH2D1A affect strictly conserved position in the loop αA helix to βB strand (III, IV). The glycine is important for the tight turn and correct positioning of the following βB-strand required for the phosphotyrosine binding. Mutations at this position were predicted to lead to over-packing and most likely to disrupt the protein structure (VI). In agreement with the predictions, the CD spectrum of the BTK G302E mutant differed from the wild type and had a decreased stability *in vitro* (III). Furthermore, the G302E mutation has also been shown to decrease the amount of BTK protein *in vivo* (Futatani et al., 1998).

8.3 Disease-causing mutations affecting methyltransferase domain of DNMT3B

The sequence analyses and model of the methyltransferase domain of DNMT3B were used to predict the putative consequences of ICF-causing mutations, possible role of the different splice variants and specific target DNA sequence for methylation (V). As a result of low structural homology and sequence identity, only the larger subunit of the methyltransferase domain could be modeled. Luckily, the majority of the pathogenic point mutations are located in the larger subunit affecting mainly DNA, cytosine or cofactor binding. The amino acids in the cofactor and target cytosine binding sites as well as the electrostatic properties of the binding pockets are conserved.

Based on the sequence similarity, the arginine 764 (R764) of DNMT3B corresponds to R240 and R227 of the two bacterial methyltransferases, *Haemophilus haemolyticus* and *Haemophilus aegyptius*, respectively. Structural and functional analyses have revealed that the arginine is involved in recognition of guanine 5’ to the flipped cytosine (Reinisch et al., 1995; Kumar et al., 1997). The DNMT3B is, therefore, likely to methylate cytosines in the GCGX sequence context that differs from the environment found for the CpG mutation hotspots in patients.

Five different DNMT3B splice variants have been described with a tissue-specific expression behavior. Several of these forms are upregulated in tumours. Of these isoforms, DNMT3B3 has been shown to have important role *in vivo* (Robertson et al., 1999). The variant protein contains a 64 residues deletion in the methyltransferase domain excluding specifically the smaller subdomain. The deletion could represent a natural method to abolish methyltransferase activity allowing other functions of the protein to predominate.

The DNMT3B project also acted as a pilot project for analyzing a small number of pathogenic mutations on the corresponding modeled structure before applying the method to a larger data set affecting SH2 domains (VI). The method and possible future improvements are discussed in the next chapter.

CONCLUDING REMARKS

A number of locus-specific mutation databases together with a generic database for the Src homology 2 (SH2) domains were created, maintained and analysed during the present study. The results support previously published experimental analyses of the endogenous mutation mechanisms underlying in pathogenic sequence variations. Interestingly, our results indicate a role for the flanking nucleotides in the pathogenic CpG hotspots, an observation that has subsequently been verified by others (Krawczak et al., 1998). The observed CpG dinucleotides in disease-causing mutations also differ from the predicted target sequence for the human *de novo* DNA methyltransferase, DNMT3B. Clearly more analyses are required to fully understand the role of neighboring nucleotides in mutation mechanisms, and, especially how nucleotides surrounding CpG sites affect the methylation process, repair pathway and mutability of the specific CpG spot. In the future, transfer of our locus-specific mutation databases into a real database format would allow linking of mutation data with several reference sequence coordinations permitting *e.g.* comparison of disease data with normal sequence variations in dbSNP or HGVbase. A large number of sequence variations together with the published Human Genome sequence would also allow analyses of nucleotide neighborhood surrounding pathogenic and non-pathogenic sequence variations.

Missense mutations form the most common mutation type known to cause human genetic diseases. In the present study, a number of structure related rules for assigning the consequences of a particular mutation on structural level were introduced. The putative consequences on corresponding protein structures were predicted for a large number of different missense mutations in seven different SH2 domains and in the methyltransferase domain of DNMT3B. As bioinformatical tools are heavily dependent of the biochemical and biophysical results, the effects of six disease-causing mutations on BTK SH2 domain were analysed. The developed method predicted correctly XLA mutations for the full length BTK protein, but failed with some of the functional XLP mutations. Based on the biochemical analyses, transiently expressed full-length BTK protein with a XLA causing mutation behaves differently than the seperate BTK SH2 domain carrying identical mutation *in vitro*. The BTK SH2 domain alone has been shown to be only marginally stable, and is therefore more prone to mutations. Similarly, the SH2D1A is comprised of only a SH2 domain followed by a short C-terminal tail and appears to be intrinsically unstable for mutations. Hence, the effects of disease-causing mutations on several SH2 domains have to be analysed in detail to fully understand how the mutation affects the stability of the defective protein domain *in vitro* and *in vivo*. A set of stable model proteins could be used for mutational analyses in case the disease related SH2 domains are intolerable for mutations. This method, however, excludes analyses of functional mutations.

It should be possible to automate the current calculations for hundreds of primary immunodeficiency causing mutations provided that structural information is available or structural homology exists. Together with sequence-based analyses developed in our laboratory, these two methods standardize the structure-function analyses of pathogenic mutations leading to an increase of data quantity and quality in locus-specific mutation databases. In the future, the next steps for improving the methods involve simulating the dynamic effects of the protein structures. It would be interesting to use molecular dynamic simulation techniques with implicit water models to obtain an ensemble of mutated protein structures for further analyses (reviewed in Lazaridis and Karplus, 2000). It may also be possible to analyse the effects of disease-causing mutations to the ligand binding by using methods such as molecular mechanics Poisson-Bolzmann surface area simulations (*e.g.* Huo et al., 2002). Again, the defective SH2 domains provide an excellent model for studying mutations involved in ligand-binding as number of disease-causing mutations in several different SH2 domains affect identical positions.

ACKNOWLEDGEMENTS

This study has been carried out at the Department of Bio and Environmental Sciences, Division of Biochemistry, University of Helsinki and at the Institute of Medical Technology, University of Tampere under the supervision of professor Mauno Vihinen. The last year of my studies was spent as an academic visitor in the Dr. Jane Clarke's laboratory in the Department of Chemistry, Cambridge University. This work has been financially supported by The National Graduate School of Information and Structural Biology, Jenny and Antti Wihuri Foundation, Eemil Aaltonen Foundation, and Magnus Ehrnrooth Foundation.

I wish to start by thanking my supervisor and friend, Mauno Vihinen, for his advice and support, and for encouraging me to always believe in my own ideas, though frequently unrealistic. He has provided an exceptional environment for me to grow and develop as an undergraduate and later as a young scientist.

I also want to thank professor Carl G. Gahmberg for providing me access to the facilities at the Division of Biochemistry and for all the support he has given me during my studies.

I am grateful for Jari Ylänne and Heikki Lehväslaiho for reviewing this thesis and for their constructive comments on it.

I wish to thank Dr. Jane Clarke for accepting me as part of her group while I was finishing this project, and for the introduction to the magic of protein folding.

My special thanks to Juha Ollila. I simply could not have wished for a better friend to share many years of studying and working.

I have been blessed with two brilliant research groups while working on my thesis. I would like to thank all the past or present members of the Vihinen Academia, especially Pena, Tomi, Kaj and Jukka. Also thanks to all members of the Clarke group as well as to my friends in the UK: Spoon, Matt, Peter, Max, Karyn, Gernot and Chieh as well as Michael and Namhee for the many tea breaks, dinners of all sorts and for the days when all we do is play games or tennis.

Thanks to all my friends in the Division of Biochemistry for the help and friendship, particularly Tiina, Susanna, Lessu and Pauli for some unforgettable evenings. I am also grateful for the secretarial help in the Department and at the graduate school as well as the technical staff of the department, especially Leena Kuoppasalmi.

I am indebted to my parents for all they have done to me, and to my sister and brother for the many years of laughter. I would also like to thank Ata and Mummi for their love and for warm afternoons spent together, Hannu and Varpu for numerous nights of good company and great food, Anja for the many past and future adventures around the world, and Kimmo for his friendship and exceptional computer skills.

And last but not least, I would like to thank Kati for her everlasting encouragement and love. It all started in 1980 as little first graders with tiny cartoons of orange juice. I have absolutely no idea how I could have survived this long journey without you.

Cambridge, May 2004

REFERENCES

Almind, K., Delahaye, L., Hansen, T., Van Obberghen, E., Pedersen, O., and Kahn, C. R. (2002). Characterization of the Met326Ile variant of phosphatidylinositol 3-kinase p85α, Proc Natl Acad Sci U S A *99*, 2124-8.

Altroff, H., van der Walle, C. F., Asselin, J., Fairless, R., Campbell, I. D., and Mardon, H. J. (2001). The eighth FIII domain of human fibronectin promotes integrin α5β1 binding via stabilization of the ninth FIII domain, J Biol Chem *276*, 38885-92.

Antequera, F. (2003). Structure, function and evolution of CpG island promoters, Cell Mol Life Sci *60*, 1647-58.

Antonarakis, S. E., Krawczak, M., and Cooper, D. N. (2000). Disease-causing mutations in the human genome, Eur J Pediatr *159*, S173-8.

Antonetti, D. A., Algenstaedt, P., and Kahn, C. R. (1996). Insulin receptor substrate 1 binds two novel splice variants of the regulatory subunit of phosphatidylinositol 3-kinase in muscle and brain, Mol Cell Biol *16*, 2195-203.

Arico, M., Allen, M., Brusa, S., Clementi, R., Pende, D., Maccario, R., Moretta, L., and Danesino, C. (2002). Haemophagocytic lymphohistiocytosis: proposal of a diagnostic algorithm based on perforin expression, Br J Haematol *119*, 180-8.

Arpaia, E., Shahar, M., Dadi, H., Cohen, A., and Roifman, C. M. (1994). Defective T cell receptor signaling and CD8+ thymic selection in humans lacking zap-70 kinase, Cell *76*, 947-58.

Bebenek, K., Roberts, J. D., and Kunkel, T. A. (1992). The effects of dNTP pool imbalances on frameshift fidelity during DNA replication, J Biol Chem *267*, 3589-96.

Bebenek, K., Abbotts, J., Wilson, S. H., and Kunkel, T. A. (1993). Error-prone polymerization by HIV-1 reverse transcriptase. Contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots, J Biol Chem *268*, 10324-34.

Benkovic, S. J., and Cameron, C. E. (1995). Kinetic analysis of nucleotide incorporation and misincorporation by Klenow fragment of Escherichia coli DNA polymerase I, Methods Enzymol *262*, 257-69.

Beroud, C., Collod-Beroud, G., Boileau, C., Soussi, T., and Junien, C. (2000). UMD (Universal mutation database): a generic software to build and analyze locus-specific databases, Hum Mutat *15*, 86-94.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases, J Mol Biol *203*, 971-83.

Bestor, T. H. (2000). The DNA methyltransferases of mammals, Hum Mol Genet *9*, 2395-402.

Bird, A. (1999). DNA methylation *de novo*, Science *286*, 2287-8.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation, Nature *321*, 209-13.

Bodak, N., Queille, S., Avril, M. F., Bouadjar, B., Drougard, C., Sarasin, A., and Daya-Grosjean, L. (1999). High levels of patched gene mutations in basal-cell carcinomas from patients with xeroderma pigmentosum, Proc Natl Acad Sci U S A *96*, 5117-22.

Bradshaw, J. M., Mitaxov, V., and Waksman, G. (2000). Mutational investigation of the specificity determining region of the Src SH2 domain, J Mol Biol *299*, 521-35.

Brown, A. F., and McKie, M. A. (2000). MuStaR and other software for locus-specific mutation databases, Hum Mutat *15*, 76-85.

Brown, T. C., and Jiricny, J. (1987). A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine, Cell *50*, 945-50.

Bruton, O. C. (1952). Agammaglobulinemia, Pediatrics *9*, 722-8.

Bullock, A. N., Henckel, J., and Fersht, A. R. (2000). Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy, Oncogene *19*, 1245-56.

Bullock, A. N., and Fersht, A. R. (2001). Rescuing the function of mutant p53, Nat Rev Cancer *1*, 68-76.

Caceres, J. F., and Kornblihtt, A. R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease, Trends Genet *18*, 186-93.

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. Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes, Nat Genet *22*, 231-8.

Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing, Nat Rev Genet *3*, 285-98.

Chan, A. C., Iwashima, M., Turck, C. W., and Weiss, A. (1992). ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR zeta chain, Cell *71*, 649-62.

Chan, B., Lanyi, A., Song, H. K., Griesbach, J., Simarro-Grande, M., Poy, F., Howie, D., Sumegi, J., Terhorst, C., and Eck, M. J. (2003). SAP couples Fyn to SLAM immune receptors, Nat Cell Biol *5*, 155-60.

Chasman, D., and Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation, J Mol Biol *307*, 683-706.

Cheng, X., Kumar, S., Posfai, J., Pflugrath, J. W., and Roberts, R. J. (1993). Crystal structure of the HhaI DNA methyltransferase complexed with S-adenosyl-L-methionine, Cell *74*, 299-307.

Cheng, X., and Blumenthal, R. M. (1996). Finding a basis for flipping bases, Structure *4*, 639-45.

Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations, Science *265*, 346-55.

Chung, E., Henriques, D., Renzoni, D., Zvelebil, M., Bradshaw, J. M., Waksman, G., Robinson, C. V., and Ladbury, J. E. (1998). Mass spectrometric and thermodynamic studies reveal the role of water molecules in complexes formed between SH2 domains and tyrosyl phosphopeptides, Structure *6*, 1141-51.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. (2003). Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios, Science *302*, 1960-1963.

Clark, S. J., Harrison, J., and Frommer, M. (1995). CpNpG methylation in mammalian cells, Nat Genet *10*, 20-7.

Claustres, M., Horaitis, O., Vanevski, M., and Cotton, R. G. (2002). Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases, Genome Res *12*, 680-8.

Clayton, L. K., Goodman, M. F., Branscomb, E. W., and Galas, D. J. (1979). Error induction and correction by mutant and wild type T4 DNA polymerases. Kinetic error discrimination mechanisms, J Biol Chem *254*, 1902-12.

Coffey, A. J., Brooksbank, R. A., Brandau, O., Oohashi, T., Howell, G. R., Bye, J. M., Cahn, A. P., Durham, J., Heath, P., Wray, P., Pavitt, R., Wilkinson, J., Leversha, M., Huckle, E., Shaw-Smith, C. J., Dunham, A., Rhodes, S., Schuster, V., Porta, G., Yin, L., Serafini, P., Sylla, B., Zollo, M., Franco, B., Bentley, D. R., and et al. (1998). Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene, Nat Genet *20*, 129-35.

Collins, F. S., Guyer, M. S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation, Science *278*, 1580-1.

Conley, M. E., Notarangelo, L. D., and Etzioni, A. (1999). Diagnostic criteria for primary immunodeficiencies. Representing PAGID (Pan-American Group for Immunodeficiency) and ESID (European Society for Immunodeficiencies), Clin Immunol *93*, 190-7.

Consortium, T. I. H. (2003). The International HapMap Project, Nature *426*, 789-96.

Cooper, D. N., and Krawczak, M. (1993). Human Gene Mutation (Oxford, BIOS Scientific Publishers Limited).

Cooper, D. N., Antonarakis, S. E., and Krawczak, M. (1995). The nature and mechanisms of human gene mutation, 7 ed (New York, McGraw-Hill).

Cotton, R. G. H. (2000). Progress of the HUGO Mutation Database Initiative: A brief Introduction to the Human Mutation MDI special Issue, Hum Mutat *15*, 4-6.

den Dunnen, J. T., and Antonarakis, S. E. (2001). Nomenclature for the description of human sequence variations, Hum Genet *109*, 121-4.

Dennis, C., and Gallagher, R. (2001). The Human Genome, first ed (New York, Palgrave).

Digilio, M. C., Conti, E., Sarkozy, A., Mingarelli, R., Dottorini, T., Marino, B., Pizzuti, A., and Dallapiccola, B. (2002). Grouping of multiple-lentigines/LEOPARD and Noonan syndromes on the PTPN11 gene, Am J Hum Genet *71*, 389-94.

Dong, A., Yoder, J. A., Zhang, X., Zhou, L., Bestor, T. H., and Cheng, X. (2001). Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA, Nucleic Acids Res *29*, 439-48.

Echols, H. (1982). Mutation rate: some biological and biochemical considerations, Biochimie *64*, 571-5.

Eck, M. J., Shoelson, S. E., and Harrison, S. C. (1993). Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck, Nature *362*, 87-91.

Eck, M. J., Pluskey, S., Trub, T., Harrison, S. C., and Shoelson, S. E. (1996). Spatial constraints on the recognition of phosphoproteins by the tandem SH2 domains of the phosphatase SH-PTP2, Nature *379*, 277-80.

Elder, M. E., Lin, D., Clever, J., Chan, A. C., Hope, T. J., Weiss, A., and Parslow, T. G. (1994). Human severe combined immunodeficiency due to a defect in ZAP-70, a T cell tyrosine kinase, Science *264*, 1596-9.

Engel, P., Eck, M. J., and Terhorst, C. (2003). The SAP and SLAM families in immune responses and X-linked lymphoproliferative disease, Nat Rev Immunol *3*, 813-21.

Estivill, X., Bancells, C., and Ramos, C. (1997). Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium, Hum Mutat *10*, 135-54.

Fahrer, A. M., Bazan, J. F., Papathanasiou, P., Nelms, K. A., and Goodnow, C. C. (2001). A genomic view of immunology, Nature *409*, 836-8.

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*, 771-86.

Fersht, A. (2002). Structure and Mechanism in Protein Science: a Guide to Enzyme Catalyses and Protein Folding, fourth ed (New York, W.H. Freeman and Company).

Fersht, A. R., Matouschek, A., and Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding, J Mol Biol *224*, 771-82.

Fluckiger, A. C., Li, Z., Kato, R. M., Wahl, M. I., Ochs, H. D., Longnecker, R., Kinet, J. P., Witte, O. N., Scharenberg, A. M., and Rawlings, D. J. (1998). Btk/Tec kinases regulate sustained increases in intracellular Ca^{2+} following B-cell receptor activation, EMBO J *17*, 1973-85.

Fredman, D., Munns, G., Rios, D., Sjoholm, F., Siegfried, M., Lenhard, B., Lehvaslaiho, H., and Brookes, A. J. (2004). HGVbase: a curated resource describing human DNA variation and phenotype relationships, Nucleic Acids Res, D516-9.

Friedman, E. (1995). The role of ras GTPase activating protein in human tumorigenesis, Pathobiology *63*, 348-50.

Futatani, T., Miyawaki, T., Tsukada, S., Hashimoto, S., Kunikata, T., Arai, S., Kurimoto, M., Niida, Y., Matsuoka, H., Sakiyama, Y., Iwata, T., Tsuchiya, S., Tatsuzawa, O., Yoshizaki, K., and Kishimoto, T. (1998). Deficient expression of Bruton's tyrosine kinase in monocytes from X-linked agammaglobulinemia as evaluated by a flow cytometric analysis and its clinical application to carrier detection, Blood *91*, 595-602.

Galas, D. J., and Branscomb, E. W. (1978). Enzymatic determinants of DNA polymerase accuracy. Theory of coliphage T4 polymerase mechanisms, J Mol Biol *124*, 653-87.

Garcia, A., Lambert, I. B., and Fuchs, R. P. (1993). DNA adduct-induced stabilization of slipped frameshift intermediates within repetitive sequences: implications for mutagenesis, Proc Natl Acad Sci U S A *90*, 5989-93.

Genevier, H. C., and Callard, R. E. (1997). Impaired Ca^{2+} mobilization by X-linked agammaglobulinaemia (XLA) B cells in response to ligation of the B cell receptor (BCR), Clin Exp Immunol *110*, 386-91.

Gold, M. R., Crowley, M. T., Martin, G. A., McCormick, F., and DeFranco, A. L. (1993). Targets of B lymphocyte antigen receptor signal transduction include the p21ras GTPase-activating protein (GAP) and two GAP-associated proteins, J Immunol *150*, 377-86.

Gorlin, R. J. (1987). Nevoid basal-cell carcinoma syndrome (Baltimore), pp. 98-113.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution, Science *185*, 862-4.

Gregersen, N., Bross, P., Jorgensen, M. M., Corydon, T. J., and Andresen, B. S. (2000). Defective folding and rapid degradation of mutant proteins is a common disease mechanism in genetic disorders, J Inherit Metab Dis *23*, 441-7.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis, Nat Genet *22*, 239-47.

Hashimoto, S., Iwamatsu, A., Ishiai, M., Okawa, K., Yamadori, T., Matsushita, M., Baba, Y., Kishimoto, T., Kurosaki, T., and Tsukada, S. (1999). Identification of the SH2 domain binding protein of Bruton's tyrosine kinase as BLNK—functional significance of Btk-SH2 domain in B-cell antigen receptor-coupled calcium signaling, Blood *94*, 2357-64.

Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites, Nature *401*, 301-4.

Henriques, D. A., and Ladbury, J. E. (2001). Inhibitors to the Src SH2 domain: a lesson in structure—thermodynamic correlation in drug design, Arch Biochem Biophys *390*, 158-68.

Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., and Nakamura, Y. (2002). JSNP: a database of common gene variations in the Japanese population, Nucleic Acids Res *30*, 158-62.

Hirsh, A. E., and Fraser, H. B. (2001). Protein dispensability and rate of evolution, Nature *411*, 1046-9.

Hof, P., Pluskey, S., Dhe-Paganon, S., Eck, M. J., and Shoelson, S. E. (1998). Crystal structure of the tyrosine phosphatase SHP-2, Cell *92*, 441-50.

Holman, G. D., and Kasuga, M. (1997). From receptor to transporter: insulin signalling to glucose transport, Diabetologia *40*, 991-1003.

Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J., and Beck, S. (1998). Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC, J Mol Biol *282*, 71-97.

Hubbard, S. R., Mohammadi, M., and Schlessinger, J. (1998). Autoregulatory mechanisms in protein-tyrosine kinases, J Biol Chem *273*, 11987-90.

Hubbard, S. R., and Till, J. H. (2000). Protein tyrosine kinase structure and function, Annu Rev Biochem *69*, 373-98.

Hunter, W. N., Brown, T., Kneale, G., Anand, N. N., Rabinovich, D., and Kennard, O. (1987). The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution, J Biol Chem *262*, 9962-70.

Huo, S., Massova, I., and Kollman, P. A. (2002). Computational alanine scanning of the 1:1 human growth hormone-receptor complex, J Comput Chem *23*, 15-27.

Huyer, G., Li, Z. M., Adam, M., Huckle, W. R., and Ramachandran, C. (1995). Direct determination of the sequence recognition requirements of the SH2 domains of SH-PTP2, Biochemistry *34*, 1040-9.

Hwang, P. M., Li, C., Morra, M., Lillywhite, J., Muhandiram, D. R., Gertler, F., Terhorst, C., Kay, L. E., Pawson, T., Forman-Kay, J. D., and Li, S. C. (2002). A “three-pronged” binding mechanism for the SAP/SH2D1A SH2 domain: structural basis and relevance to the XLP syndrome, EMBO J *21*, 314-23.

Ingram, V. M. (1957). Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin, Nature *180*, 326-8.

Inoue, K., and Lupski, J. R. (2002). Molecular mechanisms for genomic disorders, Annu Rev Genomics Hum Genet *3*, 199-242.

Jeanpierre, M., Turleau, C., Aurias, A., Prieur, M., Ledeist, F., Fischer, A., and Viegas-Pequignot, E. (1993). An embryonic-like methylation pattern of classical satellite DNA is observed in ICF syndrome, Hum Mol Genet *2*, 731-5.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks, Nature *411*, 41-2.

Jiricny, J. (1998). Replication errors: cha(lle)nging the genome, EMBO J *17*, 6427-36.

Jones, P. A. (1996). DNA methylation errors and cancer, Cancer Res *56*, 2463-7.

Kanai, Y., Ushijima, S., Kondo, Y., Nakanishi, Y., and Hirohashi, S. (2001). DNA methyltransferase expression and DNA methylation of CPG islands and peri-centromeric satellite regions in human colorectal and stomach cancers, Int J Cancer *91*, 205-12.

Kass, S. U., Landsberger, N., and Wolffe, A. P. (1997). DNA methylation directs a time-dependent repression of transcription initiation, Curr Biol *7*, 157-65.

Kay, P. H., Harmon, D., Fletcher, S., Ziman, M., Jacobsen, P. F., and Papadimitriou, J. M. (1997). Variation in the methylation profile and structure of Pax3 and Pax7 among different mouse strains and during expression, Gene *184*, 45-53.

Kere, J. (2001). Human population genetics: lessons from Finland, Annu Rev Genomics Hum Genet *2*, 103-28.

Klimasauskas, S., Kumar, S., Roberts, R. J., and Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix, Cell *76*, 357-69.

Klimasauskas, S., and Roberts, R. J. (1995). Disruption of the target G-C base-pair by the HhaI methyltransferase, Gene *157*, 163-4.

Klimasauskas, S., Szyperski, T., Serva, S., and Wuthrich, K. (1998). Dynamic modes of the flipped-out cytosine during HhaI methyltransferase-DNA interactions in solution, Embo J *17*, 317-24.

Kogelnik, A. M., Lott, M. T., Brown, M. D., Navathe, S. B., and Wallace, D. C. (1998). MITOMAP: a human mitochondrial genome database - 1998 update, Nucleic Acids Res *26*, 112-5.

Kopito, R. R. (1999). Biosynthesis and degradation of CFTR, Physiol Rev *79*, S167-73.

Kornfeld, S. J., Haire, R. N., Strong, S. J., Brigino, E. N., Tang, H., Sung, S. S., Fu, S. M., and Litman, G. W. (1997). Extreme variation in X-linked agammaglobulinemia phenotype in a three-generation family, J Allergy Clin Immunol *100*, 702-6.

Krawczak, M., Ball, E. V., and Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes, Am J Hum Genet *63*, 474-88.

Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., and Sunyaev, S. (2003). Increase of functional diversity by alternative splicing, Trends Genet *19*, 124-8.

Kruglyak, L., and Nickerson, D. A. (2001). Variation is the spice of life, Nat Genet *27*, 234-6.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution, Genome Res *13*, 2229-35.

Kuchta, R. D., Mizrahi, V., Benkovic, P. A., Johnson, K. A., and Benkovic, S. J. (1987). Kinetic mechanism of DNA polymerase I (Klenow), Biochemistry *26*, 8410-7.

Kuchta, R. D., Benkovic, P., and Benkovic, S. J. (1988). Kinetic mechanism whereby DNA polymerase I (Klenow) replicates DNA with high fidelity, Biochemistry *27*, 6716-25.

Kumar, S., Cheng, X., Klimasauskas, S., Mi, S., Posfai, J., Roberts, R. J., and Wilson, G. G. (1994). The DNA (cytosine-5) methyltransferases, Nucleic Acids Res *22*, 1-10.

Kumar, S., Horton, J. R., Jones, G. D., Walker, R. T., Roberts, R. J., and Cheng, X. (1997). DNA containing 4'-thio-2'-deoxycytidine inhibits methylation by HhaI methyltransferase, Nucleic Acids Res *25*, 2773-83.

Kunkel, T. A. (1985). The mutational specificity of DNA polymerases-alpha and -gamma during in vitro DNA synthesis, J Biol Chem *260*, 12866-74.

Kunkel, T. A. (1986). Frameshift mutagenesis by eucaryotic DNA polymerases in vitro, J Biol Chem *261*, 13581-7.

Kunkel, T. A., and Soni, A. (1988). Mutagenesis by transient misalignment, J Biol Chem *263*, 14784-9.

Kunkel, T. A., and Bebenek, K. (2000). DNA replication fidelity, Annu Rev Biochem *69*, 497-529.

Kuriyan, J., and Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling, Annu Rev Biophys Biomol Struct *26*, 259-88.

Kurosaki, T. (2002). Regulation of B cell fates by BCR signaling components, Curr Opin Immunol *14*, 341-7.

La Salle, S., Mertineit, C., Taketo, T., Moens, P. B., Bestor, T. H., and Trasler, J. M. (2004). Windows for sex-specific methylation marked by DNA methyltransferase expression profiles in mouse germ cells, Dev Biol *268*, 403-15.

Ladbury, J. E., Lemmon, M. A., Zhou, M., Green, J., Botfield, M. C., and Schlessinger, J. (1995). Measurement of the binding of tyrosyl phosphopeptides to SH2 domains: a reappraisal, Proc Natl Acad Sci U S A *92*, 3199-203.

Lambert, I. B., Napolitano, R. L., and Fuchs, R. P. (1992). Carcinogen-induced frameshift mutagenesis in repetitive sequences, Proc Natl Acad Sci U S A *89*, 1310-4.

Landry, J. R., Mager, D. L., and Wilhelm, B. T. (2003). Complex controls: the role of alternative promoters in mammalian genomes, Trends Genet *19*, 640-8.

Lappalainen, I., Ollila, J., and Vihinen, M. (1997). Registries of immunodeficiency patients and mutations, Hum Mutat *10*, 261-7.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome, Genomics *13*, 1095-107.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures, J Appl Cryst *26*, 283-291.

Latour, S., Gish, G., Helgason, C. D., Humphries, R. K., Pawson, T., and Veillette, A. (2001). Regulation of SLAM-mediated signal transduction by SAP, the X-linked lymphoproliferative gene product, Nat Immunol *2*, 681-90.

Latour, S., Roncagalli, R., Chen, R., Bakinowski, M., Shi, X., Schwartzberg, P. L., Davidson, D., and Veillette, A. (2003). Binding of SAP SH2 domain to FynT SH3 domain reveals a novel mechanism of receptor signalling in immune regulation, Nat Cell Biol *5*, 149-54.

Lazaridis, T., and Karplus, M. (2000). Effective energy functions for protein structure prediction, Curr Opin Struct Biol *10*, 139-45.

Lazarus, A. H., Kawauchi, K., Rapoport, M. J., and Delovitch, T. L. (1993). Antigen-induced B lymphocyte activation involves the p21ras and ras.GAP signaling pathway, J Exp Med *178*, 1765-9.

Leach, A. R. (2001). Molecular Modelling: Principles and Applications (Harlow, Pearson Education Limited).

Lee, C. H., Kominos, D., Jacques, S., Margolis, B., Schlessinger, J., Shoelson, S. E., and Kuriyan, J. (1994). Crystal structures of peptide complexes of the amino-terminal SH2 domain of the Syp tyrosine phosphatase, Structure *2*, 423-38.

Lee, J. K., Kim, H. T., Cho, S. M., Kim, K. H., Jin, H. J., Ryu, G. M., Oh, B., Park, C., Kimm, K., Jo, S. A., Jung, S. C., Kim, S., In, S. M., Lee, J. E., and Jo, I. (2003). Characterization of 458 single nucleotide polymorphisms of disease candidate genes in the Korean population, J Hum Genet *48*, 213-6.

Lehman, H., and Kynoch, P. A. M. (1976). Human haemoglobin variants and their characteristics (Amsterdam, Noth-Holland Publishing).

Lemmon, M. A., and Ladbury, J. E. (1994). Thermodynamic studies of tyrosyl-phosphopeptide binding to the SH2 domain of p56lck, Biochemistry *33*, 5070-6.

Li, C., Iosef, C., Jia, C. Y., Gkourasas, T., Han, V. K., and Shun-Cheng Li, S. (2003a). Disease-causing SAP mutants are defective in ligand binding and protein folding, Biochemistry *42*, 14885-92.

Li, C., Iosef, C., Jia, C. Y., Han, V. K., and Li, S. S. (2003b). Dual functional roles for the X-linked lymphoproliferative syndrome gene product SAP/SH2D1A in signaling through the signaling lymphocyte activation molecule (SLAM) family of immune receptors, J Biol Chem *278*, 3852-9.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting, Nature *366*, 362-5.

Loh, M. L., Vattikuti, S., Schubbert, S., Reynolds, M. G., Carlson, E., Lieuw, K. H., Cheng, J. W., Lee, C. M., Stokoe, D., Bonifas, J. M., Curtiss, N. P., Gotlib, J., Meshinchi, S., Le Beau, M. M., Emanuel, P. D., and Shannon, K. M. (2003). Somatic mutations in PTPN11 implicate the protein tyrosine phosphatase SHP-2 in leukemogenesis, Blood *26*, 26.

Maki, H., and Sekiguchi, M. (1992). MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis, Nature *355*, 273-5.

Maki, H. (2002). Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses, Annu Rev Genet *36*, 279-303.

Maquat, L. E. (2002). Nonsense-mediated mRNA decay, Curr Biol *12*, R196-7.

Martin, E. R., Lai, E. H., Gilbert, J. R., Rogala, A. R., Afshari, A. J., Riley, J., Finch, K. L., Stevens, J. F., Livak, K. J., Slotterbeck, B. D., Slifer, S. H., Warren, L. L., Conneally, P. M., Schmechel, D. E., Purvis, I., Pericak-Vance, M. A., Roses, A. D., and Vance, J. M. (2000). SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease, Am J Hum Genet *67*, 383-94.

Martomo, S. A., and Mathews, C. K. (2002). Effects of biological DNA precursor pool asymmetry upon accuracy of DNA replication in vitro, Mutat Res *499*, 197-211.

Matsuda, S., Suzuki-Fujimoto, T., Minowa, A., Ueno, H., Katamura, K., and Koyasu, S. (1999). Temperature-sensitive ZAP70 mutants degrading through a proteasome-independent pathway. Restoration of a kinase domain mutant by Cdc37, J Biol Chem *274*, 34515-8.

McKusic, V. A. (1998). Mendelian Inheritance in Man: catalogs of human genes and genetic disorders (Baltimore, John Hopkins University Press).

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner-McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel, G., Whittaker, A., Holden, J. L., McCann, O. T., Dunham, A., Soderlund, C., Scott, C. E., Bentley, D. R., Schuler, G., Chen, H. C., Jang, W., Green, E. D., Idol, J. R., Maduro, V. V., Montgomery, K. T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J. H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P. J., Catanese, J. J., Nowak, N., Osoegawa, K., Qin, S*., et al.* (2001). A physical map of the human genome, Nature *409*, 934-41.

Miller, J. H. (1996). Spontaneous mutators in bacteria: insights into pathways of mutagenesis and repair, Annu Rev Microbiol *50*, 625-43.

Miller, M. P., and Kumar, S. (2001). Understanding human disease mutations through the use of interspecific genetic variation, Hum Mol Genet *10*, 2319-28.

Miller, S. J. (1991). Biology of basal cell carcinoma (Part I), J Am Acad Dermatol *24*, 1-13.

Miniou, P., Bourc'his, D., Molina Gomes, D., Jeanpierre, M., and Viegas-Pequignot, E. (1997). Undermethylation of Alu sequences in ICF syndrome: molecular and in situ analysis, Cytogenet Cell Genet *77*, 308-13.

Mironov, A. A., Fickett, J. W., and Gelfand, M. S. (1999). Frequent alternative splicing of human genes, Genome Res *9*, 1288-93.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development, Development *99*, 371-82.

Morra, M., Silander, O., Calpe, S., Choi, M., Oettgen, H., Myers, L., Etzioni, A., Buckley, R., and Terhorst, C. (2001a). Alterations of the X-linked lymphoproliferative disease gene SH2D1A in common variable immunodeficiency syndrome, Blod *98*, 1321-5.

Morra, M., Simarro-Grande, M., Martin, M., Chen, A. S., Lanyi, A., Silander, O., Calpe, S., Davis, J., Pawson, T., Eck, M. J., Sumegi, J., Engel, P., Li, S. C., and Terhorst, C. (2001b). Characterization of SH2D1A missense mutations identified in X-linked lymphoproliferative disease patients, J Biol Chem *276*, 36809-16.

Morrison, A., and Sugino, A. (1994). The 3'—>5' exonucleases of both DNA polymerases delta and epsilon participate in correcting errors of DNA replication in Saccharomyces cerevisiae, Mol Gen Genet *242*, 289-96.

Nam, H. J., Haser, W. G., Roberts, T. M., and Frederick, C. A. (1996). Intramolecular interactions of the regulatory domains of the Bcr-Abl kinase reveal a novel control mechanism, Structure *4*, 1105-14.

Neel, B. G., Gu, H., and Pao, L. (2003). The 'Shp'ing news: SH2 domain-containing tyrosine phosphatases in cell signaling, Trends Biochem Sci *28*, 284-93.

Ng, P. C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions, Genome Res *11*, 863-74.

Ng, P. C., and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function, Genome Res *12*, 436-46.

Nichols, K. E., Harkin, D. P., Levitz, S., Krainer, M., Kolquist, K. A., Genovese, C., Bernard, A., Ferguson, M., Zuo, L., Snyder, E., Buckler, A. J., Wise, C., Ashley, J., Lovett, M., Valentine, M. B., Look, A. T., Gerald, W., Housman, D. E., and Haber, D. A. (1998). Inactivating mutations in an SH2 domain-encoding gene in X-linked lymphoproliferative syndrome, Proc Natl Acad Sci U S A *95*, 13765-70.

Nishimura, R., Li, W., Kashishian, A., Mondino, A., Zhou, M., Cooper, J., and Schlessinger, J. (1993). Two signaling molecules share a phosphotyrosine-containing binding site in the platelet-derived growth factor receptor, Mol Cell Biol *13*, 6889-96.

Nistala, K., Gilmour, K. C., Cranston, T., Davies, E. G., Goldblatt, D., Gaspar, H. B., and Jones, A. M. (2001). X-linked lymphoproliferative disease: three atypical cases, Clin Exp Immunol *126*, 126-130.

Okano, M., Xie, S., and Li, E. (1998). Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells, Nucleic Acids Res *26*, 2536-40.

Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C., and Hainaut, P. (2002). The IARC TP53 database: new online mutation analysis and recommendations to users, Hum Mutat *19*, 607-14.

Ottinger, E. A., Botfield, M. C., and Shoelson, S. E. (1998). Tandem SH2 domains confer high specificity in tyrosine kinase signaling, J Biol Chem *273*, 729-35.

Panning, B., and Jaenisch, R. (1998). RNA and the epigenetic regulation of X chromosome inactivation, Cell *93*, 305-8.

Park, E. M., Shigenaga, M. K., Degan, P., Korn, T. S., Kitzler, J. W., Wehr, C. M., Kolachana, P., and Ames, B. N. (1992). Assay of excised oxidative DNA lesions: isolation of 8-oxoguanine and its nucleoside derivatives from biological fluids with a monoclonal antibody column, Proc Natl Acad Sci U S A *89*, 3375-9.

Pascal, S. M., Singer, A. U., Gish, G., Yamazaki, T., Shoelson, S. E., Pawson, T., Kay, L. E., and Forman-Kay, J. D. (1994). Nuclear magnetic resonance structure of an SH2 domain of phospholipase C-gamma 1 complexed with a high affinity binding peptide, Cell *77*, 461-72.

Pawson, T., Raina, M., and Nash, P. (2002). Interaction domains: from simple binding events to complex cellular behavior, FEBS Lett *513*, 2-10.

Piccione, E., Case, R. D., Domchek, S. M., Hu, P., Chaudhuri, M., Backer, J. M., Schlessinger, J., and Shoelson, S. E. (1993). Phosphatidylinositol 3-kinase p85 SH2 domain specificity defined by direct phosphopeptide/SH2 domain binding, Biochemistry *32*, 3197-202.

Pier, G. B., Grout, M., Zaidi, T., Meluleni, G., Mueschenborn, S. S., Banting, G., Ratcliff, R., Evans, M. J., and Colledge, W. H. (1998). Salmonella typhi uses CFTR to enter intestinal epithelial cells, Nature *393*, 79-82.

Poy, F., Yaffe, M. B., Sayos, J., Saxena, K., Morra, M., Sumegi, J., Cantley, L. C., Terhorst, C., and Eck, M. J. (1999). Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition, Mol Cell *4*, 555-61.

Reenan, R. A., and Kolodner, R. D. (1992). Isolation and characterization of two Saccharomyces cerevisiae genes encoding homologs of the bacterial HexA and MutS mismatch repair proteins, Genetics *132*, 963-73.

Reich, D. E., and Lander, E. S. (2001). On the allelic spectrum of human disease, Trends Genet *17*, 502-10.

Reinisch, K. M., Chen, L., Verdine, G. L., and Lipscomb, W. N. (1995). The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing, Cell *82*, 143-53.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet *16*, 276-7.

Rideout, W. M., 3rd, Coetzee, G. A., Olumi, A. F., and Jones, P. A. (1990). 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes, Science *249*, 1288-90.

Riikonen, P., and Vihinen, M. (1999). MUTbase: maintenance and analysis of distributed mutation databases, Bioinformatics *15*, 852-9.

Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J. L., and et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA, Science *245*, 1066-73.

Roberts, G. C., and Smith, C. W. (2002). Alternative splicing: combinatorial output from the genome, Curr Opin Chem Biol *6*, 375-83.

Robertson, K. D., Uzvolgyi, E., Liang, G., Talmadge, C., Sumegi, J., Gonzales, F. A., and Jones, P. A. (1999). The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors, Nucleic Acids Res *27*, 2291-8.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., and Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature *409*, 928-33.

Saffran, D. C., Parolini, O., Fitch-Hilgenberg, M. E., Rawlings, D. J., Afar, D. E., Witte, O. N., and Conley, M. E. (1994). Brief report: a point mutation in the SH2 domain of Bruton's tyrosine kinase in atypical X-linked agammaglobulinemia, N Engl J Med *330*, 1488-91.

Sakaguchi, N., Takahashi, T., Hata, H., Nomura, T., Tagami, T., Yamazaki, S., Sakihama, T., Matsutani, T., Negishi, I., Nakatsuru, S., and Sakaguchi, S. (2003). Altered thymic T-cell selection due to a mutation of the ZAP-70 gene causes autoimmune arthritis in mice, Nature *426*, 454-60.

Salisbury, B. A., Pungliya, M., Choi, J. Y., Jiang, R., Sun, X. J., and Stephens, J. C. (2003). SNP and haplotype variation in the human genome, Mutat Res *526*, 53-61.

Samarghitean, C., Väliaho, J., and Vihinen, M. (2004). Online Registry of Genetic and Clinical Immunodeficiency Diagnostic Laboratories, IDdiagnostics, J Clin Immunol *24*, 53-61.

Sarkozy, A., Conti, E., Seripa, D., Digilio, M. C., Grifone, N., Tandoi, C., Fazio, V. M., Di Ciommo, V., Marino, B., Pizzuti, A., and Dallapiccola, B. (2003). Correlation between PTPN11 gene mutations and congenital heart defects in Noonan and LEOPARD syndromes, J Med Genet *40*, 704-8.

Saunders, C. T., and Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction, J Mol Biol *322*, 891-901.

Sayos, J., Wu, C., Morra, M., Wang, N., Zhang, X., Allen, D., van Schaik, S., Notarangelo, L., Geha, R., Roncarolo, M. G., Oettgen, H., De Vries, J. E., Aversa, G., and Terhorst, C. (1998). The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM, Nature *395*, 462-9.

Sayos, J., Martin, M., Chen, A., Simarro, M., Howie, D., Morra, M., Engel, P., and Terhorst, C. (2001). Cell surface receptors Ly-9 and CD84 recruit the X-linked lymphoproliferative disease gene product SAP, Blood *97*, 3867-74.

Schaaper, R. M., Koffel-Schwartz, N., and Fuchs, R. P. (1990). N-acetoxy-N-acetyl-2-aminofluorene-induced mutagenesis in the lacI gene of *Escherichia coli*, Carcinogenesis *11*, 1087-95.

Scheffzek, K., Ahmadian, M. R., and Wittinghofer, A. (1998). GTPase-activating proteins: helping hands to complement an active site, Trends Biochem Sci *23*, 257-62.

Schlessinger, J., and Lemmon, M. A. (2003). SH2 and PTB domains in tyrosine kinase signaling, Sci STKE *15*.

Schmid, C. W. (1998). Does SINE evolution preclude Alu function?, Nucleic Acids Res *26*, 4541-50.

Schuffenhauer, S., Bartsch, O., Stumm, M., Buchholz, T., Petropoulou, T., Kraft, S., Belohradsky, B., Hinkel, G. K., Meitinger, T., and Wegner, R. D. (1995). DNA, FISH and complementation studies in ICF syndrome: DNA hypomethylation of repetitive and single copy loci and evidence for a trans acting factor, Hum Genet *96*, 562-71.

Scriver, C. R., Nowacki, P. M., and Lehväslaiho, H. (1999). Guidelines and recommendations for content, structure, and deployment of mutation databases, Hum Mutat *13*, 344-50.

Shabalina, S. A., and Spiridonov, N. A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences, Genome Biol *5*, 25.

Sheinerman, F. B., Al-Lazikani, B., and Honig, B. (2003). Sequence, structure and energetic determinants of phosphopeptide selectivity of SH2 domains, J Mol Biol *334*, 823-41.

Shen, B., and Vihinen, M. (2004). Conservation and covariance in PH domain sequences: Physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain, Protein Eng Des Sel *13*, 13.

Shepherd, P. R., Withers, D. J., and Siddle, K. (1998). Phosphoinositide 3-kinase: the key switch mechanism in insulin signalling, Biochem J *333*, 471-90.

Sideras, P., and Smith, C. I. (1995). Molecular and cellular aspects of X-linked agammaglobulinemia, Adv Immunol *59*, 135-223.

Sloane, D. L., Goodman, M. F., and Echols, H. (1988). The fidelity of base selection by the polymerase subunit of DNA polymerase III holoenzyme, Nucleic Acids Res *16*, 6465-75.

Smith, C. I., Islam, T. C., Mattsson, P. T., Mohamed, A. J., Nore, B. F., and Vihinen, M. (2001). The Tec family of cytoplasmic tyrosine kinases: mammalian Btk, Bmx, Itk, Tec, Txk and homologs in other species, BioEssays *23*, 436-46.

Song, S., Wheeler, L. J., and Mathews, C. K. (2003). Deoxyribonucleotide pool imbalance stimulates deletions in HeLa cell mitochondrial DNA, J Biol Chem *278*, 43893-6.

Songyang, Z., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G., King, F., Roberts, T., Ratnofsky, S., Lechleider, R. J., and et al. (1993). SH2 domains recognize specific phosphopeptide sequences, Cell *72*, 767-78.

Songyang, Z., and Cantley, L. C. (1995). SH2 domain specificity determination using oriented phosphopeptide library, Methods Enzymol *254*, 523-35.

Spickett, G. P., Farrant, J., North, M. E., Zhang, J. G., Morgan, L., and Webster, A. D. (1997). Common variable immunodeficiency: how many diseases?, Immunol Today *18*, 325-8.

Sreisinger, G. (1966). Frameshift mutations and the genetic code, Cold Spring Harb symp quant Biol *31*, 77-84.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences, Genome Res *12*, 1611-8.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., Abeysinghe, S., Krawczak, M., and Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 update, Hum Mutat *21*, 577-81.

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, C. J., Chew, A., Han, J. H., Duan, J., Carr, J. L., Lee, M. S., Koshy, B., Kumar, A. M., Zhang, G., Newell, W. R., Windemuth, A., Xu, C., Kalbfleisch, T. S., Shaner, S. L., Arnold, K., Schulz, V., Drysdale, C. M., Nandabalan, K., Judson, R. S., Ruano, G., and Vovis, G. F. (2001). Haplotype variation and linkage disequilibrium in 313 human genes, Science *293*, 489-93.

Steward, R. E., MacArthur, M. W., Laskowski, R. A., and Thornton, J. M. (2003). Molecular basis of inherited diseases: a structural perspective, Trends Genet *19*, 505-13.

Su, Y. W., Zhang, Y., Schweikert, J., Koretzky, G. A., Reth, M., and Wienands, J. (1999). Interaction of SLP adaptors with the SH2 domain of Tec family kinases, Eur J Immunol *29*, 3702-11.

Sumazaki, R., Kanegane, H., Osaki, M., Fukushima, T., Tsuchida, M., Matsukura, H., Shinozaki, K., Kimura, H., Matsui, A., and Miyawaki, T. (2001). SH2D1A mutations in Japanese males with severe Epstein-Barr virus—associated illnesses, Blood *98*, 1268-70.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S., and Bork, P. (2001). Prediction of deleterious human alleles, Hum Mol Genet *10*, 591-7.

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., and Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations, Protein Eng *12*, 387-94.

Tan, J. E., Wong, S. C., Gan, S. K., Xu, S., and Lam, K. P. (2001). The adaptor protein BLNK is required for b cell antigen receptor-induced activation of nuclear factor-kappa B and cell cycle entry and survival of B lymphocytes, J Biol Chem *276*, 20055-63.

Tartaglia, M., Mehler, E. L., Goldberg, R., Zampino, G., Brunner, H. G., Kremer, H., van der Burgt, I., Crosby, A. H., Ion, A., Jeffery, S., Kalidas, K., Patton, M. A., Kucherlapati, R. S., and Gelb, B. D. (2001). Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome, Nat Genet *29*, 465-8.

Tartaglia, M., Kalidas, K., Shaw, A., Song, X., Musat, D. L., van der Burgt, I., Brunner, H. G., Bertola, D. R., Crosby, A., Ion, A., Kucherlapati, R. S., Jeffery, S., Patton, M. A., and Gelb, B. D. (2002). PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity, Am J Hum Genet *70*, 1555-63.

Thanaraj, T. A., Clark, F., and Muilu, J. (2003). Conservation of human alternative splice events in mouse, Nucleic Acids Res *31*, 2544-52.

Thorisson, G. A., and Stein, L. D. (2003). The SNP Consortium website: past, present and future, Nucleic Acids Res *31*, 124-7.

Tsukada, S., Saffran, D. C., Rawlings, D. J., Parolini, O., Allen, R. C., Klisak, I., Sparkes, R. S., Kubagawa, H., Mohandas, T., Quan, S., and et al. (1993). Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia, Cell *72*, 279-90.

Tzeng, S. R., Pai, M. T., Lung, F. D., Wu, C. W., Roller, P. P., Lei, B., Wei, C. J., Tu, S. C., Chen, S. H., Soong, W. J., and Cheng, J. W. (2000). Stability and peptide binding specificity of Btk SH2 domain: molecular basis for X-linked agammaglobulinemia, Protein Sci *9*, 2377-85.

Waksman, G., Kominos, D., Robertson, S. C., Pant, N., Baltimore, D., Birge, R. B., Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., and et al. (1992). Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides, Nature *358*, 646-53.

Wang, Z., and Moult, J. (2001). SNPs, protein structure, and disease, Hum Mutat *17*, 263-70.

Ward, C. L., and Kopito, R. R. (1994). Intracellular turnover of cystic fibrosis transmembrane conductance regulator. Inefficient processing and rapid degradation of wild-type and mutant proteins, J Biol Chem *269*, 25710-8.

Waters, T. R., and Swann, P. F. (2000). Thymine-DNA glycosylase and G to A transition mutations at CpG sites, Mutat Res *462*, 137-47.

Watson, J. D., and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, Nature *171*, 737-8.

Weber, T., Schaffhausen, B., Liu, Y., and Gunther, U. L. (2000). NMR structure of the N-SH2 of the p85 subunit of phosphoinositide 3-kinase complexed to a doubly phosphorylated peptide reveals a second phosphotyrosine binding site, Biochemistry *39*, 15860-9.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., *et al.* (2001). The sequence of the human genome, Science *291*, 1304-51.

Vetrie, D., Vorechovský, I., Sideras, P., Holland, J., Davies, A., Flinter, F., Hammarström, L., Kinnon, C., Levinsky, R., Bobrow, M., and et al. (1993). The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases, Nature *361*, 226-33.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., and Wagner, L. (2003). Database resources of the National Center for Biotechnology, Nucleic Acids Res *31*, 28-33.

Vihinen, M., Nilsson, L., and Smith, C. I. (1994). Structural basis of SH2 domain mutations in X-linked agammaglobulinemia, Biochem Biophys Res Commun *205*, 1270-7.

Vihinen, M., Kwan, S. P., Lester, T., Ochs, H. D., Resnick, I., Väliaho, J., Conley, M. E., and Smith, C. I. (1999). Mutations of the human BTK gene coding for Bruton tyrosine kinase in X-linked agammaglobulinemia, Hum Mutat *13*, 280-5.

Vihinen, M., Arredondo-Vega, F. X., Casanova, J. L., Etzioni, A., Giliani, S., Hammarström, L., Hershfield, M. S., Heyworth, P. G., Hsu, A. P., Lähdesmaki, A., Lappalainen, I., Notarangelo, L. D., Puck, J. M., Reith, W., Roos, D., Schumacher, R. F., Schwarz, K., Vezzoni, P., Villa, A., Väliaho, J., and Smith, C. I. (2001). Primary immunodeficiency mutation databases, Adv Genet *43*, 103-88.

Wijmenga, C., van den Heuvel, L. P., Strengman, E., Luyten, J. A., van der Burgt, I. J., de Groot, R., Smeets, D. F., Draaisma, J. M., van Dongen, J. J., De Abreu, R. A., Pearson, P. L., Sandkuijl, L. A., and Weemaes, C. M. (1998). Localization of the ICF syndrome to chromosome 20 by homozygosity mapping, Am J Hum Genet *63*, 803-9.

Vilkaitis, G., Dong, A., Weinhold, E., Cheng, X., and Klimasauskas, S. (2000). Functional roles of the conserved threonine 250 in the target recognition domain of HhaI DNA methyltransferase, J Biol Chem *275*, 38722-30.

Wong, I., Patel, S. S., and Johnson, K. A. (1991). An induced-fit kinetic mechanism for DNA replication fidelity: direct measurement by single-turnover kinetics, Biochemistry *30*, 526-37.

Woodcock, D. M., Crowther, P. J., Jefferson, S., and Diver, W. P. (1988). Methylation at dinucleotides other than CpG: implications for human maintenance methylation, Gene *74*, 151-2.

Vorechovský, I., Luo, L., Hertz, J. M., Froland, S. S., Klemola, T., Fiorini, M., Quinti, I., Paganelli, R., Ozsahin, H., Hammarström, L., Webster, A. D., and Smith, C. I. (1997). Mutation pattern in the Bruton’s tyrosine kinase gene in 26 unrelated patients with X-linked agammaglobulinemia, Hum Mutat *9*, 418-25.

Väliaho, J., Pusa, M., Ylinen, T., and Vihinen, M. (2002). IDR: the ImmunoDeficiency Resource, Nucleic Acids Res *30*, 232-4.

Yang, A. S., Shen, J. C., Zingg, J. M., Mi, S., and Jones, P. A. (1995). HhaI and HpaII DNA methyltransferases bind DNA mismatches, methylate uracil and block DNA repair, Nucleic Acids Res *23*, 1380-7.

Zdobnov, E. M., Lopez, R., Apweiler, R., and Etzold, T. (2002). The EBI SRS server-new features, Bioinformatics *18*, 1149-50.