CRINA SAMARGHITEAN

Primary Immunodeficiency Information Knowledge Services

ACADEMIC DISSERTATION
To be presented, with the permission of
the board of Institute of Biomedical Technology of the University of Tampere,
for public discussion in the Jarmo Visakorpi Auditorium,
of the Arvo Building, Lääkärinkatu 1, Tampere,
on August 10th, 2012, at 12 o'clock.

UNIVERSITY OF TAMPERE

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere, Institute of Biomedical Technology
Tampere Graduate Program in Biomedicine and Biotechnology (TGPBB)
Finland

Supervised by
Professor Mauno Vihinen
Lund University
Sweden
Affiliated research group,
Institute of Biomedical Technology
University of Tampere
Finland

Reviewed by
Professor Lennart Hammarström
Karolinska Institute
Sweden
Professor Olli Lassila
University of Turku
Finland

Distribution
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 40 190 9800
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Mikko Reinikka

Acta Universitatis Tamperensis 1738
ISBN 978-951-44-8824-5 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1209
ISBN 978-951-44-8825-2 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2012

In memory of my father, Constantin Samarghitean (1945-1994) and for all PIDs patients and their caregivers.

CONTENTS

ABSTRACT

New technologies allow researchers to produce large amounts of data (e.g. from textual and multimedia sources), which represents a challenge for the scientific community. Bioinformatics fills the gaps by creating algorithms, tools and methods to process the increasing quantity of information.

The main contribution of this research project was to introduce new and improved biomedical informatics methods in the field of primary immunodeficiency diseases (PIDs). Patients with these diseases have an increased rate of infections but also autoimmune and malignant manifestations. Many of these diseases are very rare with a fatal end and often they are misdiagnosed or have a delayed diagnosis. Developing software systems within the domain of primary immunodeficiencies is a highly challenging task.

In this study two databases were created and a new classification for PIDs was developed. The studies described here use an interdisciplinary approach, based on database and data mining technologies, artificial intelligence, machine learning and combined data from different disciplines such as molecular biology, genetics, immunology, bioinformatics.

The wide ranging domain of PIDs was investigated at the protein, genetic, and clinical level and, based on the analyses of different PIDs. Two databases, ImmunoDeficiency Resources (IDR) and IDdiagnostics were developed. IDR is a comprehensive knowledge base for PIDs, which includes tools for clinical, biochemical, genetic, structural and computational analyses as well as links to related information maintained by others. IDdiagnostics is a directory of laboratories performing genetic and clinical tests for PIDs. A concept map for the bioinformatics study of PIDs was designed for different types of users. The model can be used for different types of hereditary diseases.

Several computational methods for the classification and clustering of PIDs have developed a novel classification of 11 groups, which revealed previously unknown features and relationships of PIDs. These methods aim at automating the classification of PIDs and therefore would be very useful for the PID research and clinical community. Comparison of the classification to independent features such as severity and therapy of the diseases, functional classification of proteins, and network vulnerability, indicated a strong statistical support. The method can be applied to any group of diseases.

Keywords: primary immunodeficiency, data mining, registries, databases, medical expert systems, classification, network analysis, systems biology

ABBREVIATIONS

AAAAI	American Academy of Allergy, Asthma and Immunology
ACAAI	The American College of Allergy, Asthma and Immunology
AI	Artificial intelligence
AT	Ataxia telangiectasia
ATM	Ataxia telangiectasia mutated gene
CBR	Case-based reasoning
CVID	Common variable immunodeficiency
CGD	Chronic granulomatous disease
DGS	DiGeorge syndrome
DSS	Decision support system
DT	Decision tree
DTD	Document Type Definition
EDDNAL	European Directory of DNA Diagnostic Laboratories
EMBL	Genetic sequence database by European Molecular Biology Laboratory
GA	Genetic algorithm
GDB	Genome Database
GenBank	Genetic sequence database by National Center for Biotechnology Information
EPR	Electronic patient record
ES	Expert system
ESID	European Society for Immunodeficiencies
HIES	The hyper-IgE syndromes
HTML	Hypertext Markup Language

IDML	Inherited Disease Markup Language
IDR	ImmunoDeficiency Resource
IGAD	IgA deficiency
IE	Information extraction
IFNγ	Interferon γ
LAD	Leukocyte adhesion defect
MAI	Medical artificial intelligence
MDSS	Medical decision support system
MES	Medical expert system
MHCII	Major histocompatibility class II
ML	Machine learning
NLP	Natural language processing
NN	Neural network
OMIM	Online Mendelian Inheritance in Man
PID	Primary immunodeficiency
PIDexpert	Primary immunodeficiency expert system
PDB	Protein DataBank
ProDom	Protein Domain Database
PROSITE	Database of Protein Families and Domains
SCID	Severe combined immunodeficiency
SGML	Standard Generalized Markup Language
SMART	Simple Modular Architecture Research Tool
SNP	Single nucleotide polymorphism
SOAP	Simple Object Access Protocol
Swiss-Prot	Protein knowledgebase
URL	Unified Resource Locator
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol
WAS	Wiskott Aldrich Syndrome
WHIM	Warts-Hypogammaglobulinemia-Infections-Myelokathexis
WHO	World Health Organization

XHIM	X-linked hyper IgM
XLA	X-linked agammaglobulinemia
XLP	X-linked lymphoproliferative syndrome
XML	Extensible Markup Language
XSL	Extensible Style Language
XSLT	XSL Transformations

LIST OF ORIGINAL COMMUNICATIONS

The thesis is based on the following original articles, which are referred to in the text by their Roman numerals:

I. **Crina Samarghitean**, Jouni Väliaho, Mauno Vihinen: Online registry of genetic and clinical immunodeficiency diagnostic laboratories. IDdiagnostics. J Clin Immunol (2004) 24**:**53-61.

II. **Crina Samarghitean**, Jouni Väliaho, Mauno Vihinen: ImmunoDeficiency Resource, knowledge base for primary immunodeficiencies. Imm Res (2007) 3(1):6.

III. **Crina Samarghitean**, Csaba Ortutay, Mauno Vihinen: Systematic classification of primary immunodeficiencies based on clinical, pathological and laboratory parameters. J Immunol (2009) 183(11), 7569- 7575.

1. INTRODUCTION

Primary immunodeficiencies (PIDs) are a large and heterogenous group of mainly rare hereditary disorders of the immune system that often have serious consequences. These diseases represent a challenge in their diagnosis and treatment due to overlapping symptoms and similarities between diseases. The main manifestations include infections but also autoimmune and cancer diseases, granulomatosis, hemophagocytic syndrome, angioedema, autoinflammation, thrombotic microangiopathy, or a predisposition to allergies.

Bioinformatics provides a vast amount of different tools that are designed to collect the data, analyze it and make predictions when for some reason there is a shortage of experimentally defined data or there is a lack of consensus in classification or diagnosis. Various databases for collecting information about PIDs exist on the Internet, starting from the proteomic, mutational level and ending with the national PID patient registry. The results from these databases were evaluated periodically to see whether there are differences among different populations and the epidemiological data collected, and how the different statistics differ from each other. Classification and network analysis methods have been used to characterize, e.g., the spread of epidemics, to determine ways to control them, and to identify novel target genes for prostate cancer.

The aim of this study was to develop new comprehensive knowledge bases for PIDs which could integrate all the available PIDs resources on the Internet tailored for different types of users. Using the knowledge accumulated in these databases and datamining techniques, a mathematical, systematic classification of PIDs was proposed which was useful in developing of a decision support system for PIDs. These approaches bridge the knowledge bases with diagnostic and therapeutic protocols to be applied directly in patient care. These specific applications have been helpful in gene discovery, development of other databases, diagnosis protocols and of ICD 11.

2. REVIEW OF LITERATURE

2.1 Primary immunodeficiency diseases

2.1.1 Background on PIDs

Integrity of the immune system is essential for defence against infectious organisms and their toxic products and for the survival of all individuals. Defects in one or more components of the immune system can lead to serious and often fatal disorders, which are collectively called immunodeficiency diseases. These diseases have been classified since the beginning into two main groups. The congenital or primary immunodeficiencies are genetic defects transmitted hereditary that result in an increased susceptibility to infectious, autoimmune and cancer manifestations, frequently manifested early in infancy and childhood but sometimes clinically detected later in life. Acquired or secondary immunodeficiencies develop as a consequence of e.g. malnutrition, disseminated cancer, immunosuppressive drugs, infections, especially HIV (Abbas and Lichtman 2005).

Primary immunodeficiencies (PIDs) are challenging for both scientists and clinicians and may manifest with a wide range of clinical symptoms including susceptibility to infections, allergy, autoimmune and inflammatory diseases, lymphoproliferation and cancer (Marodi and Notarangelo 2007). Currently, more than 200 PIDs and some 170 PID-related genes are known, many of them reported during the last few years. Diagnosis of PIDs may be demanding due to symptoms and signs often being discreet, overlapping and variable in many PIDs. Also, the large number of PIDs makes the field difficult for non-experts. PIDs (Notarangelo et al. 2004; Stiehm et al. 2004) are relatively rare, with an incidence varying between 1:500-1:500000 and in some cases there are just a few patients globally. Most PIDs present in childhood, but also in the second and third decades of life, for example common variable immunodeficiency (CVID).

The infections in immunodeficiency patients can be mainly categorized in two classes. Patients with defects in immunoglobulins, complement proteins or phagocytes are very susceptible to recurrent infections with encapsulated bacteria, such as *Haemophilus influenzae*, *Steptococcus pneumonia* and *Staphylococcus aureus*. Patients with defects in cell-mediated immunity are susceptible to infections with opportunistic microorganisms, such as yeast and chickenpox (Roitt et al. 2001). Early detection, before serious infections have compromised patient's general condition, is important for prognosis and genetic counseling of the family (Stiehm et al. 2004).

Primary immunodeficiency disorders were first identified after the introduction of antibiotics. Several syndromes of immunodeficiency with characteristic clinical features were described before 1940, including mucocutaneous candidiasis by Thorpe and Handley in 1929, ataxia-telangiectasia by Syllaba and Henner in 1926, and Wiskott-Aldrich syndrome (WAS) by Wiskott in 1937 (Stiehm, Ochs et al. 2004). The first patient with cellular deficiency was initially described by Glanzmann and Riniker in 1950. The seminal article by Bruton in 1952 reported the first patient with congenital agammaglobulinemia who had an excellent response to immunoglobulin replacement therapy (Bruton 1952). This discovery, considered to be the birth of the primary immunodeficiency field, opened the door to numerous other achievements. Several other immune defects were recognized on the basis of their consistent clinical findings or inheritance patterns (Buckley 2002). The first case of phagocytic defect was reported in 1956 (Kostmann 1956). The combination of antibody deficiency with defective cellular immunity was first identified in a Swiss infant, Swiss-type agammaglobulinemia (Hitzig et al. 1958). The first case of complement deficiency was reported in 1966 (Klemperer et al. 1966). One of the most puzzling conditions was described by Janeway and his colleagues in patients with pyogenic infections and enlarged lymph nodes, hepatosplenomegaly and hypergammaglobulinemia, which proved to be later X-linked chronic granulomatous disease (Buckley 2002).

From 1961 to 1972 the role of the thymus gland in immune system function began to be elucidated with animal models (Miller 1961; Good et al. 1962; Cooper et al. 1965).

Different studies *in vivo* showed cellular immunity to be impaired in patients who had congenital hypothyroidism and no thymic tissue at autopsy, pathologic presentation of thymic hypoplasia (DiGeorge syndrome) (Miller 1961, Cooper et al. 1967). The first insights into the defects of the oxidative metabolism of the phagocytic cells were another important discovery during this period (Holmes et al. 1966). A series of technological advances in the 1960's and 1970s helped the investigation of the immune system enormously. The 1980s led to the cloning of the genes responsible for many of the known primary immunodeficiencies (Kwan et al. 1986; Gatti et al. 1988; Kwan et al. 1988; Puck et al. 1989).

Treatment milestones include the first use of immunoglobulin for immune deficiency (Bruton 1952), the first successful hematopoietic cell transplants in humans in 1968, in three patients with primary immunodeficiencies who received grafts from HLA-matched siblings (two with SCID and one with Wiskott-Aldrich syndrome) (Gatti et al. 1968). The first use of a cytokine for the treatment for chronic granulomatous disease was reported in 1991 and the first success of gene therapy for X-linked severe combined immunodeficiency (Cavazzana-Calvo et al. 2000).

2.1.2 Genetics of PIDs

Within the past 20 years major advances in the understanding of the genetic basis and molecular mechanism of immunodeficiencies has been made. Increasing understanding of these molecular defects has influenced both basic and translational research. More than 170 distinct genes have been associated with PIDs. Different mechanisms may cause PID pathologies, including transcription factors, cytokines and their receptors, cell surface and cytoplasmic signaling mediators, cell cycle regulators, DNA modifying enzymes, intracellular chaperones and transport proteins. The genetic defects that cause PIDs can affect the expression and function of proteins involved in different biological processes, such as immune development, effector-cell functions, signaling cascades and maintenance of immune homeostasis (Marodi and Notarangelo 2007). In recent years, the genetic basis of more than 200 primary immunodeficiencies has been identified and their mode of inheritance as X-linked, autosomal recessive, or autosomal dominant has been elucidated (Table 1).

The first diagnostic measure when considering any heritable disorder is to find out the full pedigree and family health history. Families can provide details about potentially affected members in past generations. Ethnic and geographical ancestry should be also recorded as part of every family history because they may suggest consanguinity or increased associated population risks (Ochs et al. 2006).

The X-linked (XL) immunodeficiency diseases occur in high frequency in males, because mutations in genes of the sole copy of chromosome X are uncompensated and a single mutation causes overt disease (Table 1). Recessive diseases require defects in both alleles of a gene in order for a disease phenotype to manifest. One way is through consanguineous marriages such as the interferon γ receptor mutations causing susceptibility to mycobacterial infections (Al-Muhsen and Casanova 2008). Another risk factor for homozygous recessive diseases is to belong to a genetically isolated population in which the same mutation may be transmitted through both maternal and paternal lines to the patient because of inbreeding (Table 1). There are few dominant immunodeficiency phenotypes. They are characterized by incomplete penetrance, skipping generations in family pedigree and/or variable expressivity, and a large spectrum of severity among relatives with the same genetic mutation (Table 1).

Table 1. Chromosome locations and inheritance of genes related to PIDs

Disorder	**Gene symbol**	**Gene location**
X linked Disorders		
XL chronic granulomatous disease	*CYBB*	Xp26
XL SCID	*IL2RG*	Xp13
XL lymphoproliferative syndrome	*SH2D1A* *BIRC4*	Xq25
Wiskott-Aldrich syndrome	*WASP*	Xp11.22

Disorder	**Gene symbol**	**Gene location**
XL agammaglobulinemia	*BTK*	Xq21.3
XL hyper-IgM syndrome	*CD40L*	Xq27
Glucose 6-phosphate dehydrogenase deficiency	*G6PD*	Xq28
Ectodermal dysplasia with NEMO mutation	*IKBKG*	Xq28
Properdin deficiency	*PFC*	Xp11
Hoyeraal-Hreidarsson syndrome	*DKC1*	Xq28
Barth syndrome	*TAZ*	Xq28
IPEX syndrome	*FOXP3*	Xp11.23
AR Disorders		
AR chronic granulomatous disease	*p22-phox*	16q24
	p47-phox	7q11.23
	p67-phox	1q25
Janus kinase 3 deficiency	*JAK3*	19p13.1
CD40 deficiency	*CD40*	20q12-q13.2
RAG1 deficiency	*RAG1*	11p13
RAG2 deficiency	*RAG2*	11p13
Artemis deficiency	*DCLRE1C/*	10p
ZAP70 deficiency	*ZAP70*	8q12
Adenosine deaminase deficiency	*ADA*	20q13
Purine nucleoside phosphorylase deficiency	*PNP*	14q13
MHC class II deficiency	*CIITA*	16p13
	RFX-5	1q21
	RFXAP	13q
MHC class I deficiency	*TAP1*	6p21.3
	TAP2	6p21.3
	TAPBP	6p21.3
μ heavy-chain deficiency	*IGHM*	14q32.3

Disorder	**Gene symbol**	**Gene location**
λ5 surrogate light-chain deficiency	*IGLL1*	22q11.22
BLNK deficiency	*BLNK*	10q23.2
Ig α deficiency	*CD79A*	19q13.2
Ig β deficiency	*CD79B*	17q23
ICOS deficiency	*ICOS*	2q33
TNFRSF13B deficiency	*TNFRSF13B*	17p11.2
CD19 deficiency	*CD19*	16p11.2
AID deficiency	*AICDA*	12p13
UNG deficiency	*UNG*	17q11.2
DOCK8 deficiency	*DOCK8*	9p24.3
Leukocyte adhesion deficiency I	*ITGB2*	21q22.3
Leukocyte adhesion deficiency II	*SLC35C1*	11p11.2
Leukocyte adhesion deficiency III	*RASGRP2*	11q13
Griscelli syndrome, type 1	*MYO5A*	15q21
Griscelli syndrome, type 2	*RAB27A*	15q21
Griscelli syndrome, type 3	*MLPH*	2q37
Myeloperoxidase deficiency	*MPO*	17q23.1
Glycogen storage disease Ib	*G6PC*	11q21
Shwachman syndrome	*SBDS*	7q11
MAPBPIPdeficiency	*MAPBPIP*	1q22
Hermansky-Pudlak syndrome 2	*AP3B1*	Chr.5
Familial haemophagocytic lymphohistiocytosis type 2	*PRF1*	10q22
Familial haemophagocytic lymphohistiocytosis type 3	UNC13D	17q25.3
Familial haemophagocytic lymphohistiocytosis type 4	STX11	6q24
Interleukin-12 receptor β1 deficiency	*IL12RB1*	19p13.1
Interleukin-12 (IL-12) p40 deficiency	*IL12B*	5q31.1

Disorder	**Gene symbol**	**Gene location**
STAT5b deficiency	*STAT5B*	17q11.2
IRAK4 deficiency	*IRAK4*	12q12
UNC93B deficiency	*UNC93B1*	11q13
Autoimmune polyendocrinopathy with candidiasis and ectodermal dystrophy	*AIRE*	21q22.3
Cartilage-hair hypoplasia	*RMRP*	9p21-p12
Epidermodysplasia verruciformis type 1	*TMC6*	17q25
Epidermodysplasia verruciformis type 2		
	TMC8	17q25
Natural killer deficiency	*FCGR3A*	1q23
Transcobalamin II deficiency	*TCN2*	22q11.2
Hepatic veno-oclussive disease with immunodeficiency syndrome	*SP110*	2q37.1
Ataxia-telengiectasia	*ATM*	11q22-q23
Nijmegen-breakage syndrome	*NBS*	8q21
Ataxia-telangiectasia-like disorder	*MRE11A*	11q21
DNA ligase I deficiency	*LIG1*	19q13.2-
DNA ligase deficiency IV	*LIG4*	13q22-q34
Bloom syndrome	*BLM*	15q26.1
Immunodeficiency, centromere instability and facial abnormalities syndrome (ICF)	*DNMT3B*	20q11.2
C1q α-polypeptide deficiency	*C1QA*	*1p36.3-*
C1q β-polypeptide deficiency	*C1QB*	*1p36.3-*
C1q γ-polypeptide deficiency	*C1QC*	*1p36.3-*
C1r deficiency	*C1R*	*12p13*
C1s deficiency	*C1S*	*12p13*
C2 deficiency	*C2*	*6p21.3*
C3 deficiency	*C3*	*19p13.3-*
C4A deficiency	*C4A*	*p13.2*

Disorder	**Gene symbol**	**Gene location**
C4B deficiency	*C4B*	*6p21.3*
C5 deficiency	*C5*	*6p21.3*
C6 deficiency	*C6*	*9q32-q34*
C7 deficiency	*C7*	*5p13*
C8 α-polypeptide deficiency	*C8A*	*5p13*
C8 β-polypeptide deficiency	*C8B*	*1p32*
C8 γ-polypeptide deficiency	*C8G*	*1p32*
C9 deficiency	*C9*	*9q34.3*
Factor B deficiency	*BF*	*5p13*
Factor D deficiency	*CFD*	*6p21.3*
Factor H1 deficiency	*CFH*	*19p13.3*
C4 binding protein α deficiency	*C4BPA*	*1q32*
C4 binding protein β deficiency	*C4BPB*	*1q32*
Decay-accelerating factor (CD55) deficiency	*CD55*	*1q32*
Factor I deficiency	*CFI*	*1q32*
MAC inhibitor (CD59) deficiency	*CD59*	*4q25*
Familial Mediterranean fever	*MEFV*	*16p13*
Hyperimmunoglobulinemia D with periodic fever syndrome	*MVK*	*12q24*
AD disorders		
WHIM syndrome	*CXCR4*	2q21
LRRC8 deficiency	*LRRC8A*	9q34.13
Severe congenital neutropenias, including Kostmann syndrome	*CSF3R* *ELA2*	1p35-p34.3
Cyclic neutropenia		19p13.3
GFI1 deficiency	*GFI1*	1p22
LAD with RAC2 deficiency	*RAC2*	22q12.3-q13.2

Disorder	**Gene symbol**	**Gene location**
Autoimmune lymphoproliferative syndrome type II	*CASP10*	2q33-q34
Autoimmune lymphoproliferative syndrome type IIB	*CASP8*	2q33-q34
ALPS type III	*NRAS*	1p13.2
TLR3 deficiency	*TLR3*	4q35
Autosomal dominant anhidrotic ectodermal dysplasia and T-cell immunodeficiency	*NFKBIA/IKBA*	14q13
DiGeorge-anomaly	*DGCR*	22q11
Tumor necrosis factor receptor-associated periodic syndrome	*TNFRSF1A*	12p13.2
Familial cold urticaria and Muckle-Wells syndrome	*CIAS1*	1q44
Chronic infantile neurological cutaneous and articular syndrome	*CIAS1*	1q44
Granulomatous sinovitis with uveitis and cranial neuropathies	*CARD15*	16q12

A molecular genetic focus adds precision to the diagnosis and allows genotype and phenotype correlations. Genetic studies can confirm a suspected diagnosis of PIDs and pinpoint a specific genetic etiology in the absence of other data. They can help in genetic counseling.

Gene testing in PIDs in the past was limited to expert academic laboratories; nowadays it is easily available to physicians with a broad range of clinical expertise. Such testing can establish or confirm a suspected diagnosis and also may predict future disease risk in advance of clinical signs and symptoms, inform reproductive decision making, and guide clinicians in selecting the most appropriate therapeutic options (Morra et al. 2008).

Many PIDs are the result of a single gene mutation, making a molecular diagnosis for a patient straightforward. More difficult is to find information on the Internet related to a specific lab doing a genetic test for a primary immunodeficiency. Over 170 PID genes have been identified and the prevalence of most of the genetic defects is still low. Different genetic defects can have a similar clinical presentation, and patients with different defects in the same gene can have diverse clinical pictures (Burg et al. 2009). For example, Omenn syndrome can result from defects in *RAG1*, *RAG2*, *DCLERC1* (Artemis), *RMRP*, *IL/R*, *IL2RG*, or *ADA* (Villa et al. 2008) (Table 2).

Table 2. Genetic heterogeneity of PIDs (modified with permission from Notarangelo and Sorensen 2008)

PID phenotype	**Associated gene defects**
T-B+NK-SCID	*ILDRG, JAK3*
T-B+NK+SCID	*RAG1, RAG2, DCLERC1*
Omenn syndrome	*RAG1, RAG2, DCLERC1, RMRP, IL7R, IL2RG, ADA*
Agammaglobulinemia	*BTK, IGHM, IGLL1; CD79A, CD79B, BLNK*
CVID	*TNFRSF13B (TACI), ICOS, TNFRSF13C (BAFF-R), CD40L, CD19, SH2D1A*
Hyper-IgM syndrome	*CD40L, CD40, AICDA, UNG*
XL lymphoproliferative syndrome	*SH2D1A, XIAP*
Familial haemophagocytic lymphohistiocytosis	*PRF1, MUNC13B, STX11*
Chronic granulomatous disease	*CYBB, CYBA, NCF1, NCF2*
Severe congenital neutropenias	*ELA2, HAX1, GFI1, MAPBP, WASP*
Mendelian susceptibility to mycobacterial disease	*IL12B, IL12RB1, IFNGR1, IFNGR2, STAT1*
Herpes simplex encephalitis	*UNC93B1, TLR3*
The hyper-IgE syndromes (HIES)	*TYK2, DOCK8, STAT3*

Certain mutations in PID genes can lead to atypical clinical and immunologic presentations. For example, null mutations in *RAG1* or *RAG2* genes cause T-B-NK+SCID, hypomorphic mutations in the same gene can associate with Omenn syndrome, leaky SCID, or combined immunodeficiency with granulomas (Schuetz et al. 2008). There is a certain level of phenotypic and genotypic heterogeneity in PIDs (Table 3), making mutation analysis a useful tool to solve differential diagnosis of complex presentations (Notarangelo and Sorensen 2008).

Table 3. Phenotypic heterogeneity of PIDs (modified with permission from Notarangelo and Sorensen 2008)

Mutated genes	**Associated phenotypes**
RAG1, *RAG2*, *DCLERC1*	SCID, OS, leaky SCID
LIG4	SCID, LIG4 syndrome
WASP	WAS, XLT, XLN
SH2D1A	XLP, CVID
RMRP	CHH, SCID, OS
ELA2	SCN, cyclic neutropenia

Molecular diagnostics is important for patients and families and to better understand the pathophysiology of the disease. Mutation analysis can provide a definitive and precise diagnosis, may help in genetic counseling and treatment compliance, and also help to establish a diagnosis in an atypical presentation. Molecular diagnostics enable the development of long-term preventive strategies to limit complications and irreversible organ damage and permit presymptomatic identification of individuals affected with potentially lethal forms of PIDs and prompt timely life-saving interventions, such as hematopoietic cell transplantation (HCT).

2.1.3 Classes of PIDs

Originally, PIDs were referred to as rare Mendelian traits associated with multiple, recurrent, opportunistic infections in early childhood with fatal course. There are still numerous controversies and discussions regarding an accepted definition of PIDs. There is no single system of classification for PIDs useful for both educational and clinical purposes. Most texts use a functional classification based on the immunological mechanism disrupted. These types of descriptive functional categories may overlap to varying degrees.

In 1973, a committee of the World Health Organization (WHO) published the first in a series of classifications for PIDs (Cooper et al. 1973) followed by revisions (WHO 1983; WHO 1986; WHO 1992; WHO 1995; WHO 1997; WHO 1999). The definition and classification of PIDs, based on immunological phenotypes, is now reviewed every 2 years by WHO in conjunction with the International Union of Immunological Societies (IUIS) Expert Committee (Notarangelo, Casanova et al. 2004; Notarangelo et al. 2006; Geha et al. 2007; Notarangelo et al. 2009). Others have used and proposed different classification systems (Bonilla et al. 2005; Bonilla and Geha 2005; Casanova et al. 2005).

2.1.4 Decision making in diagnosis of PIDs

Identification of PIDs is difficult. General practitioners and pediatricians need to consider PIDs as a potential diagnosis in any child or adult with repeated infections resistant to treatments. Recently, different protocols have been published for the diagnosing of PIDs (Folds and JL. 2003; Bonilla, Bernstein et al. 2005; de Vries et al 2006). The 10 warning signs of primary immunodeficiency and the 4 stages of testing for PIDs have been developed by the Jeffrey Modell Foundation Medical Advisory Board (http://info4pi.org). Recently, 12 warning signs for infants were proposed (Carneiro-Sampaio et al. 2010).

All these guidelines and protocols are based on the traditional classification of antibody, T lymphocytes, phagocyte and complement deficiencies and require some knowledge of the immune system and its defects (Folds and JL. 2003; Bonilla et al. 2005). These guidelines are aimed at both pediatricians and adult physicians (de Vries et al 2006). It is important to take a stepwise approach for the diagnosis of PIDs. First, the patient need to be evaluated clinically and immunologically using different diagnostic protocols (Conley et al. 1999; Bonilla et al. 2005; de Vries et al 2006) and checked for all the warning signs for PIDs (step 1). To define the immunological defects further flow cytometric immunophenotyping and functional studies (when applicable) of blood, bone marrow, or other specimens are done (step 2). These steps are of great value for guiding the selection of candidate genes for molecular diagnostics (step 3). Finally, genetic counseling and prenatal diagnostics are facilitated with the identification of a genetic defect (step 4).

Guidelines for the diagnosis of PIDs have been developed jointly by the European Society for Immune Deficiencies and the Pan-American Group for Immune Deficiencies (Conley et al. 1999). Mutation analysis performed through DNA sequencing is supported by analysis of mRNA/cDNA and protein expression and by functional assays to prove the disease-causing role of mutations. If these are not available or inconclusive, there are bioinformatics tools available that might help estimate the significance of the observed DNA changes (Fleisher and Notarangelo 2008).

2.2 Bioinformatics tools for PIDs

Bioinformatics tools helpful in PID diagnosis can be divided into several categories (Figure 1) (Samarghitean and Vihinen 2009):

1. General PID resources
2. Classifications of PIDs helpful for diagnosis
3. Laboratories for genetic and clinical tests for PIDs
4. National and international PID patient registries
5. Mutation databases

6. Tools for prediction or prioritization of novel PID candidate genes, and
7. Decision support systems in PID diagnosis (Table 4).

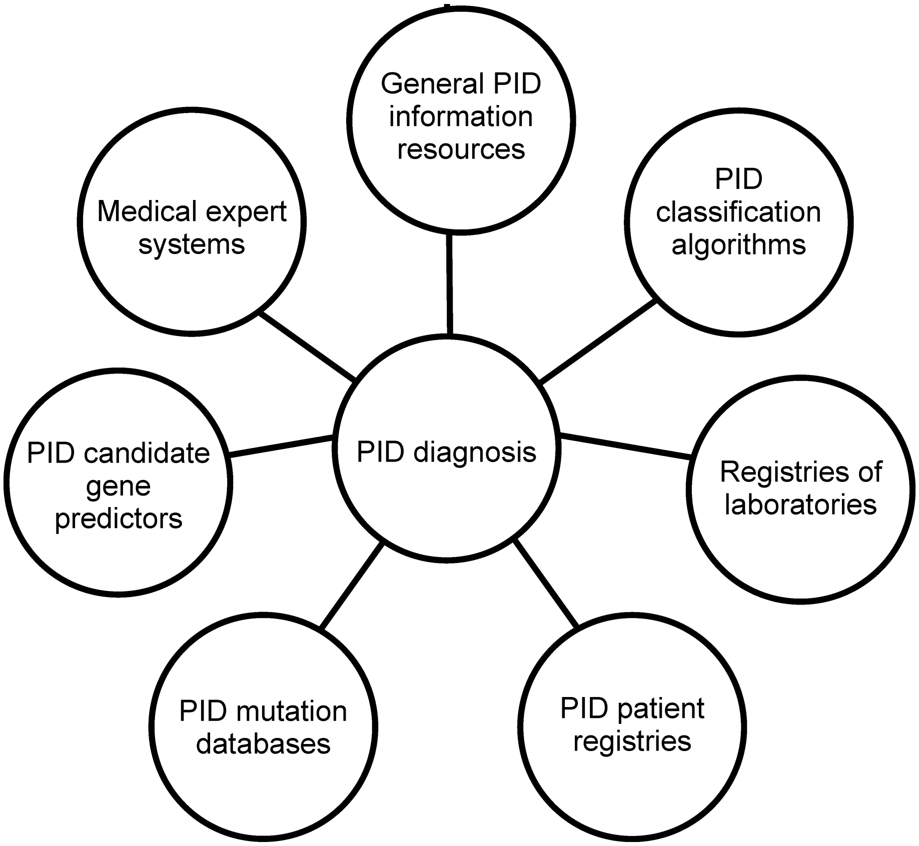

General PID information resources
Medical expert systems
PID classification algorithms
PID candidate gene predictors
PID diagnosis
Registries of laboratories
PID mutation databases
PID patient registries

Figure 1. Categories of bioinformatics tools for PIDs

2.2.1 General PID resources

Dedicated PID information resources on the web include INFO4PI, RAPID (Keerthikumar et al. 2009) and general resources on rare diseases such as ORPHANET. INFO4PI is a web service that includes information for over 50 diagnostic and research centers worldwide and a registry of experts worldwide. It is the official webpage of the Jeffrey Modell Foundation (JMF) and designed for diverse user groups. The medical advisory group of the foundation has developed '10 warning signs of PIDs', recently updated, and '4 stages for immunologic testing', which have been adopted in many countries. Another new development launched by the foundation is the software for primary immunodeficiency recognition, intervention and tracking (SPIRIT) which matches patient ICD-9 codes with the 10 warning signs of PIDs.

Resource of Asian PIDs (RAPID) is a newly developed database which contains information about PID genes, protein-protein interactions, mouse knockout studies and microarray gene expression profiles in various cells and organs of the immune system (Keerthikumar et al. 2009). The website also hosts information on sequence variations and expression at the mRNA and protein levels of all genes reported to be involved in PID patients.

ORPHANET provides information on rare diseases for healthcare professionals, patients, and their relatives. ORPHANET includes expert-authored and peer-reviewed information for rare, mostly genetic, diseases. There is a directory of clinics, clinical laboratories, research activities and patient organizations. ORPHANET provides up-to-date information about rare diseases in many languages.

Table 4. Resources on the web for PIDs-related information

Bioinformatics tools	**URL**
General PID resources	
INFO4PI	http://www.info4pi.org
JMF	http://www.jmfworld.com
ORPHANET	http://www.orpha.net
PID classifications	
AAAAI	http://www.aaaai.org/professionals/resources/pdf/immunodeficiency2005.pdf
ESID/IUIS	http://www.esid.org/downloads/ESID_Diseases_2009_0.pdf
WHO (ICD10)	http://apps.who.int/classifications/apps/icd/icd10online/
Registries of diagnostic laboratories	
EDDNAL	http://www.eddnal.com
GeneTests	http://www.ncbi.nlm.nih.gov/sites/GeneTests

Patient registries	
ASCIA	http://www.immunodeficiency.org.au
CEREDIH	http://www.ceredih.fr/
ESID	http://www.esid.org/
IPINET	http://www.aieop.org
USIDnet	http://www.usidnet.org
Variation databases	
IDbases	http://bioinf.uta.fi/IDbases
Other PID mutation databases	http://bioinf.uta.fi/base_root/mutation_databases_list2.php
RAPID	http://rapid.rcai.riken.jp/RAPID
Novel PIDs candidate genes	
PID candidates	(Ortutay and Vihinen 2009)
RAPID	http://rapid.rcai.riken.jp/RAPID/SVM
PID diagnosis tools	
UKPIN	http://www.ukpin.org.uk/ESID/index.htm

2.2.2 Classifications of PIDs

Traditionally, PIDs have been classified into wide categories. The PIDs Classification Committee of the International Union of Immunological Societies (IUIS) has provided classifications for many years. The essential textbook in the field, Primary Immunodeficiency Diseases (Ochs et al. 2006), contains another grouping of PIDs as well as the International Classification of Diseases (ICD10). The classification in the European Immunodeficiency Society (ESID) online patient database further expands the IUIS classification. All these classifications have been developed by experts in the field and updated regularly. The methodology used for the classification has not been described and many newly discovered PIDs do not fit well within these classifications.

Different classification schemes have been developed to discriminate patients with common variable immunodeficiency disease, a heterogeneous group of immunological disorders characterised by low serum immunoglobulin IgG, IgA and/or IgM, defective specific antibody production and an increased susceptibility to bacterial infections. The Freiburg classification discriminates patients with a disturbed germinal centre of memory B cells and patients with defects of early peripheral B cell differentiation by analysing CD21 expression (Warnatz and Schlesier 2008). Another CVID classification differentiates patients based on the assessment of class-switched memory B cells and total CD27+ B cells including the marginal zone like B cells (Piqueras et al. 2003). The EUROClass classification (Wehr et al. 2008) distinguishes patients with less than 1% of B cells of lymphocytes (B-) from patients with a higher percentage (B+). A new analytical computational approach for phenotype analysis of CVID patients was recently described (Kalina et al. 2009). This unsupervised hierarchical clustering analysis enabled the definition of clusters of individuals with similar B-cell profiles.

2.2.3 Registries of diagnostic laboratories

The diagnosis of immunodeficiencies can be difficult because several disorders may have similar symptoms. In some cases early and reliable diagnosis is crucial for efficient treatment since delayed diagnosis and management can lead to severe and irreversible complications. For many PIDs the definitive diagnosis can be obtained only based on both genetic and clinical tests. The physical signs may be non-specific, very discreet, or absent. Due to the rareness of PIDs there may not be many laboratories analyzing a particular disease.

Some PIDs laboratory information can be found from general services for diagnostic laboratories. The GeneTests service provides an international directory of genetic testing laboratories and genetic and prenatal diagnosis clinics. There are also expert-authored, peer-reviewed disease descriptions called GeneReviews. The European Directory of DNA Diagnostic Laboratories (EDDNAL) has information about DNA-based diagnostic services.

2.2.4 Databases and patient registries

A database is a collection of related data with some inherent meaning. A database is designed, built, and populated with data from different sources for a specific purpose. In medicine databases have been established for some diseases (Cotton et al. 1996). Because biological and medical data have many special characteristics, the management of this kind of information is a highly challenging problem (Elmasri and Navathe 2004). A specific design tailored for these types of data is necessary for the management of database systems.

In the case of primary immunodeficiencies, existing population-based databases are limited. Available data are derived from case-based disease registries that collect patient-specific information from multiple sources.

Case-based registries usually are designed to improve patient care but can be helpful for studying rare diseases. In 1992, the Immune Deficiency Foundation (IDF) initiated a registry of U.S. patients with CGD and 5 years later expanded the project to include seven other disorders: HIGM, XLA, CVID, WAS, SCID, LAD, and DGS (Winkelstein et al. 2000). The registry is also used to collect data related to natural history and clinical course, including the response to treatment. IDF assembled a group of investigators under the name The United States Immunodeficiency Network (USIDNET) to implement an on-line system of the IDF registry.

Several PID patient registries have been released in different countries during the last two decades (Rezaei et al. 2008). The information in them can help in making a diagnosis for a patient and provide valuable epidemiological information. For example, countries, such as Australia (Baumgart et al. 1997), Spain (Matamoros Flori et al. 1997), Switzerland (Ryser et al. 1988), Italy (Luzi et al. 1983), Sweden (Fasth 1982), and Norway (Stray-Pedersen et al. 2000) have developed their own registry-based estimates of the frequency of PIDs. The ASCIA PID Register of Australia and New Zealand has already described the prevalence of PIDs in Australia for 1209 patients in 88 centers (Kirkpatrick and Riminton 2007). The Italian Primary Immunodeficiency Network (IPINET) collects patient information, including the pedigree, date of diagnosis, immunological data and clinical manifestations, laboratory data and information about replacement therapy (Plebani et al. 2004). The Spanish Registry for Primary Immunodeficiencies (REDIP) has quite similar contents (Matamoros Flori et al. 1997). In 2004, the Spanish registry counted 2607 cases from 82 centers. The Iranian PID Registry has followed 930 patients over a period of 30 years (Rezaei et al. 2006).

The ESID patient registry is the largest patient registry (Eades-Perner et al. 2007). In December 2011, it contained 15052 patient entries from 89 centers for 142 PIDs. For each PID there is an individual database; they all share a common core dataset with information about diagnosis, therapy, quality of life and some laboratory data. There are also disease-specific data models for some of the most prevalent PIDs. Submission of mutation data to the patient registry is combined with IDbases (Piirilä et al. 2006). Access to the data is allowed only to the registered users who have obtained permission.

The same software is used by the United States Immunodeficiency Network (USIDnet) for registry of US patients. Currently, the USIDnet data is collected for eight PIDs, but soon there will be information for over 30 disorders. Another regional registry using the ESID online database system is the Latin American Group for Primary Immunodeficiency Diseases (LAGID), which so far includes information for 3321 patients in 14 countries (Leiva et al. 2007).

Considering the recent reports from four major registries, antibody deficiencies are the most common PID comprising more than half of all patients. Other well-defined immunodeficiencies, phagocytes defects, and combined T and B cell immunodeficiencies are also relatively common. CVID is the most common PID followed by selective IgA and/or IgG subclass deficiency, agammaglobulinemia with absent B cells, ataxia telangiectasia (AT), chronic granulomatous disease (CGD), and SCID (Rezaei et al. 2008).

Other important PID patient- and treatment-related databases include those by the Center for International Blood and Marrow Transplant Research (CIBMTR), a combined research program of the National Marrow Donor Program, the Medical College of Wisconsin, and the European Blood and Marrow Transplant Group (EBMT). These facilitate research into allogeneic hematopoietic stem cell transplantation outcomes (Horowitz 2008). SCETIDE is a specific stem cell transplantation database for PIDs, which has data going back to 1985. These databases allow addressing questions difficult to answer through clinical trials and may also aid the development of optimal designs for prospective clinical trials (Griffith et al. 2008; Gennery et al. 2010).

2.2.5 Variation databases

Other sources of case-based information are the Internet-based, locus-specific immunodeficiency mutation databases established by ESID and expanded by other investigators (Smith and Vihinen 1996; Vihinen et al. 1999; Väliaho et al. 2000; Piirilä et al. 2006). These databases contain information regarding specific mutations and certain clinical features of affected persons. The first Internet-based immunodeficiency mutation database, BTKbase, was initiated in 1994 to collect information related to mutations in the *BTK* gene (Bruton tyrosine kinase), which causes XLA. Some 130 similar locus-specific mutation databases have been developed since then (Piirilä et al. 2006). Mutation databases can be used to analyze the types of mutations and their distribution in exons and introns, including their location in protein domains. Mutation databases that contain clinical information can be helpful in assessing genotype-phenotype relations and determining the presence of gene variants in asymptomatic family members (Porter et al. 2000).

Data from disease and mutation registries can be used to estimate the minimal incidence of a disorder, characterize epidemiologic features, and define a range of clinical characteristics in a cohort of patients. Current registries provide incomplete population-based data regarding the burden of PIDs diseases.

Although many identified genetic variations in PIDs are novel it is important to analyze whether the variation in a patient has been previously described. Information about PID-related genetic variations is available from different mutation databases. The majority of these registries are in the ImmunoDeficiency mutation databases (IDbases) maintained at the IBT Bioinformatics group. Currently, there are 122 freely available IDbases, which contain mutation information for 5388 patients from 4513 families having altogether 2327 different mutational events. In many IDbases there is plenty of information in addition to the description of the actual mutation. The mutation entries are linked to sequence databanks, literature and OMIM (Amberger et al. 2009). The format of IDbases is standardized and uniform throughout. Currently, some important changes are being made to IDbases. Reference sequences at three levels (DNA, RNA, protein) have been developed in collaboration with RefSeqGene (http://www.ncbi.nlm.nih.gov/refseq/rsg/ and LRG projects (http://www.lrg-sequence.org). IDbases will have a stable genomic framework for reporting mutations that allows easy integration with other services. The IDbases can provide new insights into both genotype-phenotype correlations in patients as well as protein structure-function relationships for the encoded proteins. The IDbases are linked to the University of California Santa Cruz (UCSC) genome browser (Karolchik et al. 2008) from where the mutation data can be easily viewed with PhenCode (Giardine et al. 2007), along with other genetic and variation information.

PID mutation information can be found also in Asian Primary Immunodeficiency Diseases. For PID genes and proteins there is information about mRNA and protein expression as well as on protein-protein interactions and mouse studies. A tool can visualize mutations on protein three dimensional structures (Keerthikumar, Raju et al. 2009).

While identification of gene defects and variations has become easy and fast to perform, the interpretation of the effects and elucidation of the detailed molecular mechanisms of genetic diseases is much more difficult. Disease-related alterations may have diverse effects on the structure and function at DNA, RNA and protein levels (Thusberg and Vihinen 2009). Numerous methods can be used for predicting the effects of amino acid substitutions and are collected in the recently developed Pathogenic-Or-Not pipeline (PON-P), which is freely available at http://bioinf.uta.fi/PON-P.

2.2.6 Prediction of PID candidate genes

Some bioinformatics approaches have been applied to predict novel PID candidate genes, putative novel PID genes were prioritized (Ortutay and Vihinen 2009). The method combines information about protein interaction network properties and Gene Ontology terms. The analysis was based on a dataset for the immunome, the entirety of genes and proteins essential for mounting immune responses (Ortutay et al. 2007). The approach utilizes the protein interaction network information available in the Immunome Knowledge Base (IKB) (Ortutay and Vihinen 2009). The identified disease candidate genes are mainly involved in cellular signaling including receptors, protein kinases and adaptors and binding proteins, as well as enzymes (Ortutay and Vihinen 2009).

Another PID candidate list of altogether 1442 genes (Keerthikumar et al. 2009) is available from RAPID (Keerthikumar et al. 2009). Using a support vector machine learning approach the authors used 69 binary features of 148 known PID genes and 3162 non-PID genes as a training set. Six of the predicted genes have been experimentally confirmed to be PIDs genes.

2.2.7 PID diagnosis tools

Some services for PIDs diagnosis have been implemented on the web. The multi-stage diagnostic protocol designed for non-immunologists (de Vries et al 2006) has been converted to linked web pages on the United Kingdom Primary Immunodeficiency Network (UKPIN) site. Recommendations for diagnosis and treatment for some PIDs, both in English and in Italian, are provided by IPINET. In the CEREDIH, the French National Immunodeficiency Centre service, there is information about French PID diagnosis centers and laboratories. They also have a decision tree-like schema for diagnostic protocol. All the information in this service is written in French.

More advanced tools for diagnosis are called Medical Expert Systems (MESs). These are computer software systems that are based on a set of rules applied to knowledge originally extracted from human experts or generated by computational analyses. In addition to helping in diagnosis and report generation, MESs can improve consistency in decisions, as well as timeliness in decision-making and productivity (Samarghitean and Vihinen 2008). MESs can be integrated with other health care applications, such as electronic patient records, and systems for prescribing and dispensing medicines.

2.3 Medical expert systems (MES)

2.3.1 Definitions and short history of MES

There are various definitions for electronic decision support systems depending on the primary use and on the functions the systems have to perform. Expert Systems (ESs) (Gonzalez and Dankel 1993; Nikolopoulos 1997) are computer programs developed to contain and use knowledge from human experts. They represent and structure medical knowledge, and can supplement or replace man in decision-making. ESs uses a set of rules to solve problems that need human expertise and can simulate a dialogue with experts. They improve the abilities of experts and increase the consistency and quality of problem-solving activities. ESs use techniques from artificial intelligence, machine learning, and data mining (Samarghitean and Vihinen 2008).

Two approaches can be used to define an expert system. The first one is the human/artificial intelligence (AI) approach and the second one is the technology–oriented approach (Godall 1985; Patel et al. 2009). From the human/AI approach, an expert system is a computer system that performs functions similar to those normally performed by a human expert (Shortliffe 1993). From the technological approach, an ES is a computer system that operates by applying an inference mechanism to a body of specialist expertise represented in the form of 'knowledge' (Samarghitean and Vihinen 2008).

The first computerized decision support system in health care in 1959 was based on Bayesian statistics and decision theory. Both logic and probabilistic reasoning were shown to be essential components of medical reasoning (Ledley and Lusted 1959). During the 1960's and 1970's many data driven programs, which used pattern recognition techniques, were developed for diagnostic problems. Since the 1970's, expert systems have mainly utilized artificial intelligence approaches (Blum 1986; Shortliffe 1993). For the first generation medical expert systems, such as CASNET

(Weiss et al. 1978), INTERNIST-I (Miller et al. 1982) and MYCIN (Buchanan and Shortliffe 1984), the ability to reason about the diagnostic process was most important.

In recent years, artificial intelligence approaches have been integrated with multimedia and Internet technologies. Connectionist data-mining approaches (Hripcsak et al. 2003) including machine learning (Takeuchi and Collier 2005), artificial neural networks (Reategui et al. 1997), and genetic algorithms (Gaal et al. 2005), have been applied in health care and medicine for decision support.

Developed in recent years, Health-e-Child project, produced an integrated healthcare platform for European pediatrics (Freund et al. 2006; Branson et al. 2008; Jimeno-Yepes et al. 2009). Another goal of the project was to provide uninhibited access to universal biomedical knowledge repositories for personalized and preventive healthcare, large-scale information-based biomedical research and training, and informed policymaking. CaseReasoner, a Health-e-Child product tool, helps clinicians to search thousands of disease diagnoses, treatments and outcomes to find a child similar to their own patients (Freund et al. 2006). Another tool of the project is AITION, which uses semantic tools to search medical literature and interviews with clinicians as well as patient data (Freund et al. 2006).

2.3.2 Architecture of medical expert systems

The structure of expert systems can be described and understood in terms of the components of the system and in terms of the interchange of the information between the components. In the early years, expert systems were written in high-level programming languages, usually in LISP and PROLOG. The expert knowledge of the domain and the algorithms for applying the knowledge were interwoven and the systems could not be easily modified for other applications. From stand-alone consultation systems in 1980s, we assist to integration with other healthcare systems such as electronic medical records, provider order-entry systems, results reporting systems, e-prescribing systems, or tools for genomic/proteomic data management and analysis.

Current expert systems have three main components: a knowledge base, an inference engine, and a user interface (UI) (Figure 3) (Samarghitean and Vihinen 2008).

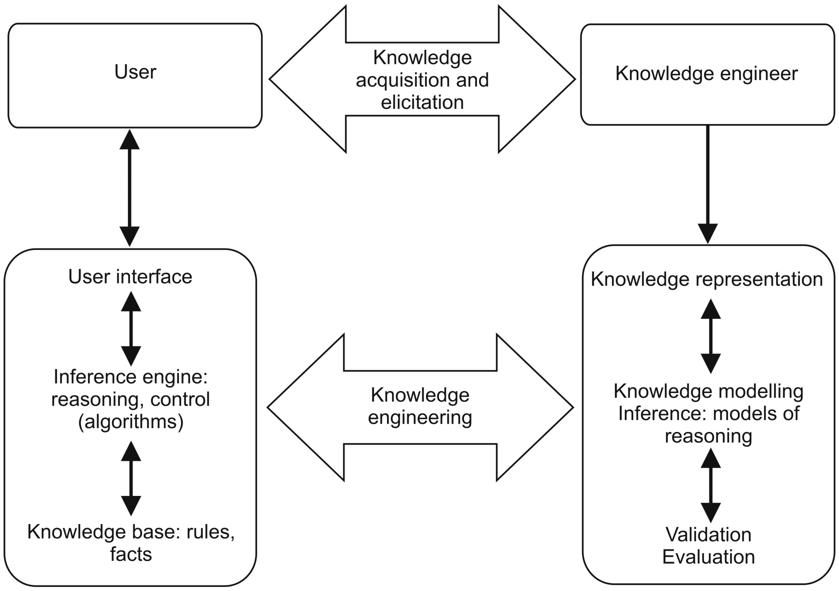

User
Knowledge acquisition and elicitation
Knowledge engineer
User interface
Inference engine: reasoning, control (algorithms)
Knowledge base: rules, facts
Knowledge engineering
Knowledge representation
Knowledge modelling Inference: models of reasoning
Validation Evaluation

Figure 2. Structure of an expert system

The knowledge base contains the domain-specific knowledge acquired from experts, including object descriptions, problem-solving behaviors, constraints, heuristics and uncertainties. The knowledge base is the heart of an expert system and usually is in the form of if-then rules. The inference engine finds a sequence in which inferences are made. The inference engine is used to reason with the knowledge base. Knowledge stored, based on published guidelines, databases, and custom knowledge bases, it is translated into active rules. These rules or facts are used for deduction and are part of the inference engine capable of fuzzy reasoning. Expert systems also facilitate the users by providing an explanation subsystem which explains its reasoning to the users. In some systems there is a knowledge base editor for writing and updating the knowledge base. Expert systems have the ability to separate problem specific knowledge from general-purpose reasoning. The user interface applies the knowledge, rules, and local patient and clinical data and acts as the front-end of the system for the user.

Modern expert systems are built in a special software environment, known as, expert system shells. These provide more general facilities and an easy way to enter necessary knowledge about the problem domain. Some of the popular software packages used include ESTA, EXSYS, XpertRule, ACQUIRE, and FLEX. They produce patient-specific and situation specific recommendations. Medical expert systems can be integrated with other applications, such as electronic patient records, systems for prescribing and dispensing medicines, and other information systems used in health settings (Figure 3) (Samarghitean and Vihinen 2008).

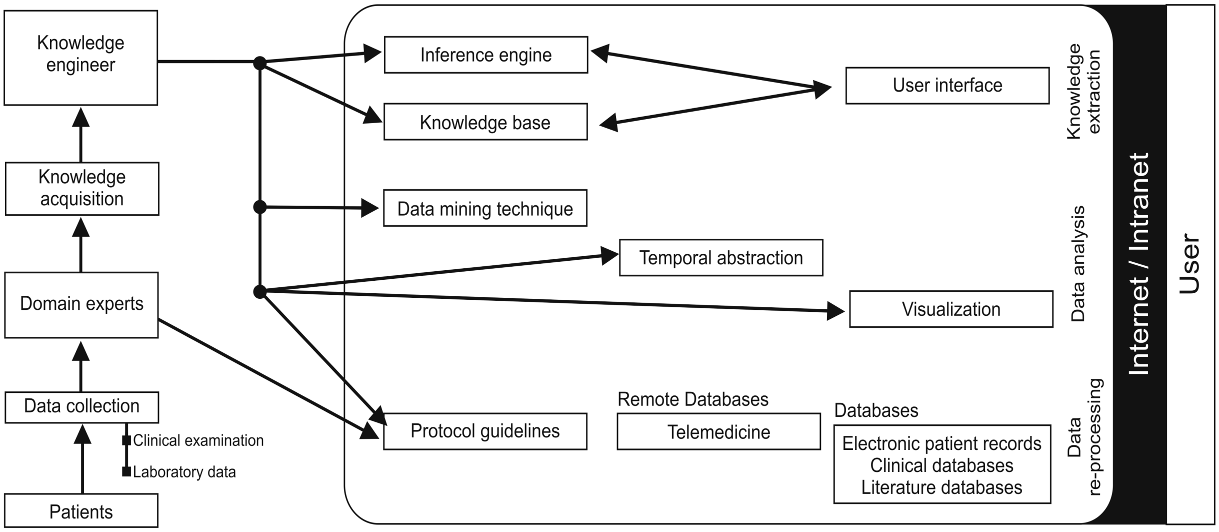


Figure 3. A modern architecture of a medical expert system

2.4 Data mining

2.4.1 Definitions

Data mining is the process of discovering hidden patterns and trends (information) in data (Westphal and Blaxton 1998; Han and Kamber 2000; Soukup and Davidson 2002). Data mining is seen as the extraction of implicit, previously unknown, and potentially useful information from large data sets or databases. Data mining is an umbrella term and is used with varied meaning in a wide range of contexts. Most data mining efforts are focused on developing a finely grained, highly detailed model of some large data set. Data mining gives information that wouldn't be available otherwise. When the data collected involves individual people, there are many questions concerning privacy, legality, and ethics.

Data mining is at the intersection of machine learning, statistics, visualization and databases. Statistics is more theory based and focuses on testing hypotheses, while machine learning is more heuristic and focuses on improving the performance of a learning agent. Data mining and knowledge discovery integrate theory and heuristics.

2.4.2 Steps involved in data mining

Data mining is used to extract valid, novel and potentially useful and understandable patterns from data. Preliminary steps include data collection, selection, preprocessing, and transformation. Appropriate data is first retrieved from the database, and then possibly cleaned and reduced before knowledge discovery. After the discovery step, the knowledge will be evaluated manually or on the basis of objective quality criteria. The overall flow of the data mining process from data sources to different visualizations and data mining tools (Soukup and Davidson 2002) is shown in Figure 4 (Samarghitean and Vihinen 2008). Data mining is frequently used for prediction and description.

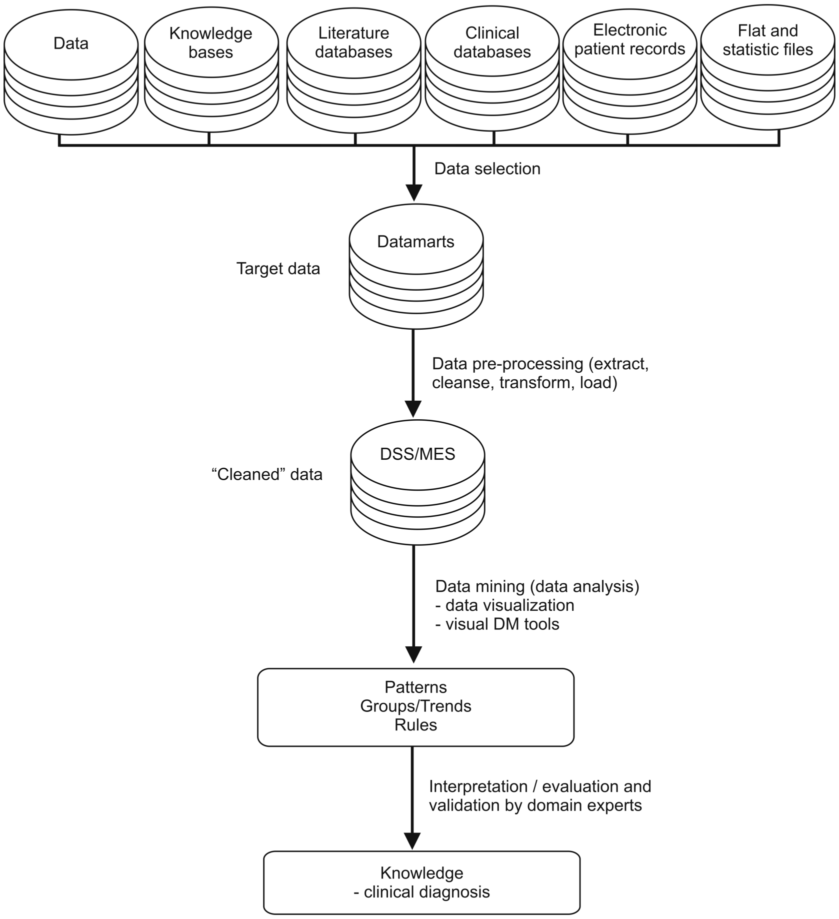


Figure 4. Steps involved in the data mining process

2.4.3 Applications of data mining

.

Knowledge acquisition and data mining are important application areas of ML. Knowledge acquisition, extracting the knowledge from experts, is the major bottleneck in the development of expert systems. Semi-automatic methods can support either the experts or knowledge engineers in knowledge acquisition. Data mining, by applying different machine learning techniques, allows nearly automatic knowledge acquisition. Decision trees (Quinlan 1993), neural networks (Swingler 1996), genetic algorithms (Pena-Reyes and Sipper 2000) and nearest neighbor classification are some examples of such ML methods.

Data mining has been extensively used in industry, banking, government, and health care delivery. In medicine it has been used for integrated sources of clinical data, inferential probabilistic prediction of risk of ventilator-associated pneumonia (Chapman et al. 2001), discrimination between ageing-related and non-ageing-related DNA repair genes (Freitas et al. 2011), and between cancerous and non-cancerous cells (Dexter et al. 2010), population surveillance for signs of bioterrorism, pharmacovigilance (Lindquist et al. 2000; Hauben et al. 2005) and predictive data mining (Bellazzi and Zupan 2008).

In medical data mining applications one of the most important features is a safety critical context in which decision making activities should always be supported by explanations. The value of each datum may be higher than in other contexts, the data sets can be small and report nonreproducible situations and may be further affected by several sources of uncertainty. Physicians and researchers deal with such difficulties by exploiting their knowledge of the domain. Data mining involves applying variable and model selection, evaluating the resulting models and encoding this knowledge and using it in data analysis. Data mining problems are often solved by using a mosaic of different approaches from computer science, such as multi-dimensional databases, machine learning, soft computing and data visualization, and from statistics, such as hypothesis testing, clustering, classification and regression techniques.

Data mining algorithms are developed, tested and/or used by health authorities, pharmaceutical companies and academic researchers also in the pharmacovigilance domain. A principle concern of pharmacovigilance is detection of adverse drug reactions that are novel by virtue of their clinical nature, severity and/or frequency. There is an increased interest in developing database screening tools to assist human reviewers in identifying associations worthy of further investigation (Hauben, Madigan et al. 2005).

Predictive data mining is becoming an essential instrument for researchers and clinical practitioners in medicine. Thanks to the integration of molecular and clinical data within data warehouses, epidemiological studies and emerging studies in genomics and proteomics, the area has gained a fresh impulse. Crucial to such data are those data mining approaches which allow the use of the background knowledge, discover interesting interpretable and non-trivial relationships, and construct rule-based and other symbolic-type models (Bellazzi and Zupan 2008).

3. AIMS OF THE STUDY

The aims of the study were to develop knowledge bases for PIDs which may include genetic, proteomic, molecular, epidemiological, clinical information about PIDs and could be easily updated and integrated with other knowledge services for PIDs. The specific aims were:

1. Build, maintain and update the information in a newly developed genetic and clinical test database for PIDs, IDdiagnostics (I).
2. Refine, maintain and update the information in the ImmunoDeficiency Resource (IDR), a knowledge base for PIDs (II).
3. Develop a new classification for PIDs based on clinical, pathological and laboratory parameters (III).

4. MATERIALS AND METHODS

4.1 Data sources

Different data sources were used to collect and update the information in IDR (II) and IDdiagnostics (I). These include:

1. For clinical data the best evidence on primary immunodeficiency were used: online scientific literature (PubMed, immunology journals), immunology books (Stiehm et al. 2004; Ochs et al. 2006; Rezaei et al. 2008).
2. For genetic information electronic databases were used: OMIM (Amberger, Bocchini et al. 2009), ORPHANET (Ayme and Schmidtke 2007), GeneTests (Pagon 2006), GeneCards (Safran et al. 2002), SOURCE (Diehn et al. 2003), PROSITE (Sigrist et al. 2002), Swiss-Prot (Boeckmann et al. 2003), Ensemble (Birney et al. 2004), eMedicine (2002; Lundberg 2006).
3. Direct submissions from clinicians were used for the genetic test and clinical test database for IDdiagnostics (I).
4. Direct communications with clinicians and PID scientists in conferences (ESID 2002-2010, IUIS 2009), meetings (ESID Prague meeting 2007, 2009, 2010; ESID Summer School Bled, Slovenia 2009; ESID Juniors workshop Florence, Italy 2010; PID Advanced School Sao Paolo, Brazil 2010; ESID Junior Symposium Tampere, Finland 2011; ESID Registry Workshop, Freiburg, Germany 2011).

4.2 Tools used for developing IDR and IDdiagnostics

For building, maintaining and updating the information in the IDR (II) and IDdiagnostics (I) different tools were used.

Web development tools:

1. eXtensible Markup Language (XML)
2. Inherited Disease Markup Language(IDML)

4.2.1 eXtensible Markup Language

The Extensible Markup Language (XML) is a standard created by the World Wide Web Consortium (W3C) for characterizing the content and structure of documents, designed to improve the functionality of the Web by enabling more flexible and adaptable information identification and presentation (Väliaho et al. 2005). Information encoded in XML is easy to read and understand, and easy to process by computers. In XML files, structured data is bounded by tags and attributes. XML tags, attributes and element structure provide context information that facilitates the interpretation of the meaning of content, thereby making it feasible to develop efficient search engines and agents and perform intelligent data mining, etc. The XML allows the separation of content, logic and presentation. There are many advantages for XML language (time saving, meta-data gives extra meaning to information, useful for indexing, future uses, documents exchange etc.). XML is increasingly becoming the preferred method of encoding structured data for exchange over the Internet.

4.2.2 Inherited Disease Markup Language (IDML)

The Inherited Disease Markup Language (IDML) provides a standard method for exchanging genetic and clinical data along with general disease descriptions, diagnostic information and links to other related resources (Väliaho et al. 2005). Separation of data from the presentation enables the seamless integration of data from diverse sources. The IDML format is a published, documented open format, offered especially for the purpose of data interchange between platforms and databases in the Internet. The data from the IDML format for different systems, e.g. to hypertext markup language (HTML), is transformed by using extensible stylesheet language transformations (XSLT) stylesheets. The IDML specification and document type definition (DTD) follow the XML standards of the World Wide Web Consortium (W3C).

The fact file data model and the Inherited Disease Markup Language (IDML) were developed to facilitate disease information integration, storage and exchange in the first place for immunodeficiencies, but in principle for any hereditary disease. The IDML is an XML specification and container for bioinformatics data on hereditary diseases. The fact file data model schema was defined according to W3C XML specification. The fact file data model can be depicted as a tree structure graph where a <*FactFile*> element is a root. Fact files make use of the following specifications, standards and databases: HUGO gene nomenclature (Wain et al. 2004), RefSeq (Pruitt and Maglott 2001), Swiss-Prot (Boeckmann et al. 2003) and SOURCE (Diehn et al. 2003).

4.3 Clustering and network analysis methods

In paper (III) we used a systems biological approach to develop a new classification for PIDs. Different clustering and network analysis methods were used.

Clustering is a process that groups a set of objects into clusters so that the similarity among the objects in a cluster is high. There are different types of data clustering algorithms: hierarchical and partitional. In hierarchical algorithms, successive clusters are found using previously established clusters, while partitioning algorithms determine all clusters at once. Hierarchical clusters are either agglomerative (bottom up) or divisive (top-down) methods. Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. The divisive algorithms begin with the whole set and divide it into successively smaller clusters (Wang et al. 2005). A K-means algorithm assigns each point to the clusters whose centroid is closest. The centroids are the points that have the coordinates on a central tendency. The main advantage of this algorithm is its simple implementation and computational speed. K-means maximizes inter-cluster variance and ensures a local-optimal solution on a local minimum of variance.

Network analysis is a methodological tool and a theoretical paradigm to describe, explore and understand structural and relational aspects. A network consists of entities (nodes) that may represent individuals, organizations, or programs connected by lines or arrows, which show some relationship between them. A network can be used for four different purposes: as a conceptual model, a description of an existing real-word structure or system, a mathematical model, or a simulation (Luke and Harris 2007). The network paradigm has four important features:

1. It is a structural approach focusing partly on patterns of linkages between nodes
2. It is grounded in empirical data
3. It makes frequent use of mathematical and computational models and

4 It is highly graphical.

Network approaches focus on relationships between subjects rather than relationships between subject attributes (i.e. variables). In network analyses, data is typically organized in an N-by N square matrix. The data entries represent a relationship between a pair of nodes. In many cases, network identification is straightforward, especially when boundaries are clear (Salganik and Heckathorn 2004).

5. RESULTS

We developed a procedure for the bioinformatics analysis of PIDs based on the experience in our group and previous studies on the development of different knowledge bases and systematic classifications for PIDs. The procedure was implemented and refined in the studies (I, II, III).

5.1 ImmunoDeficiency Resource IDR (II)

The ImmunoDeficiency Resource (IDR) is a knowledge base that integrates a wide spectrum of information including clinical, biochemical, genetic, genomic, proteomic, structural, and computational data. Over time, IDR has proved to be a valuable resource for the community, with an increased numbers of users every year, being integrated with other new resources such as RAPID and ORPHANET and cited in prestigious journals and books. At the core of the system are *Fact files* in which are stored in a structured form information regarding PIDs: genes, heredity, nosology, taxonomy, mutations, protein sequences, online resources, organizations and associations. The fact files are XML-based, validated data sources on PID-related information. Each fact file provides basic information on diseases and affected genes. The fact files also contain HTML hyperlinks to other Internet resources that are accepted to be reliable by the experts and are integrated within IDdiagnostics, IDbases and IDR. The fact files act as a quick portal to further information on diseases. At present there are 178 factfiles. Stand-alone IDML fact files have been generated for each PID (Figure 5) (Väliaho et al. 2005).

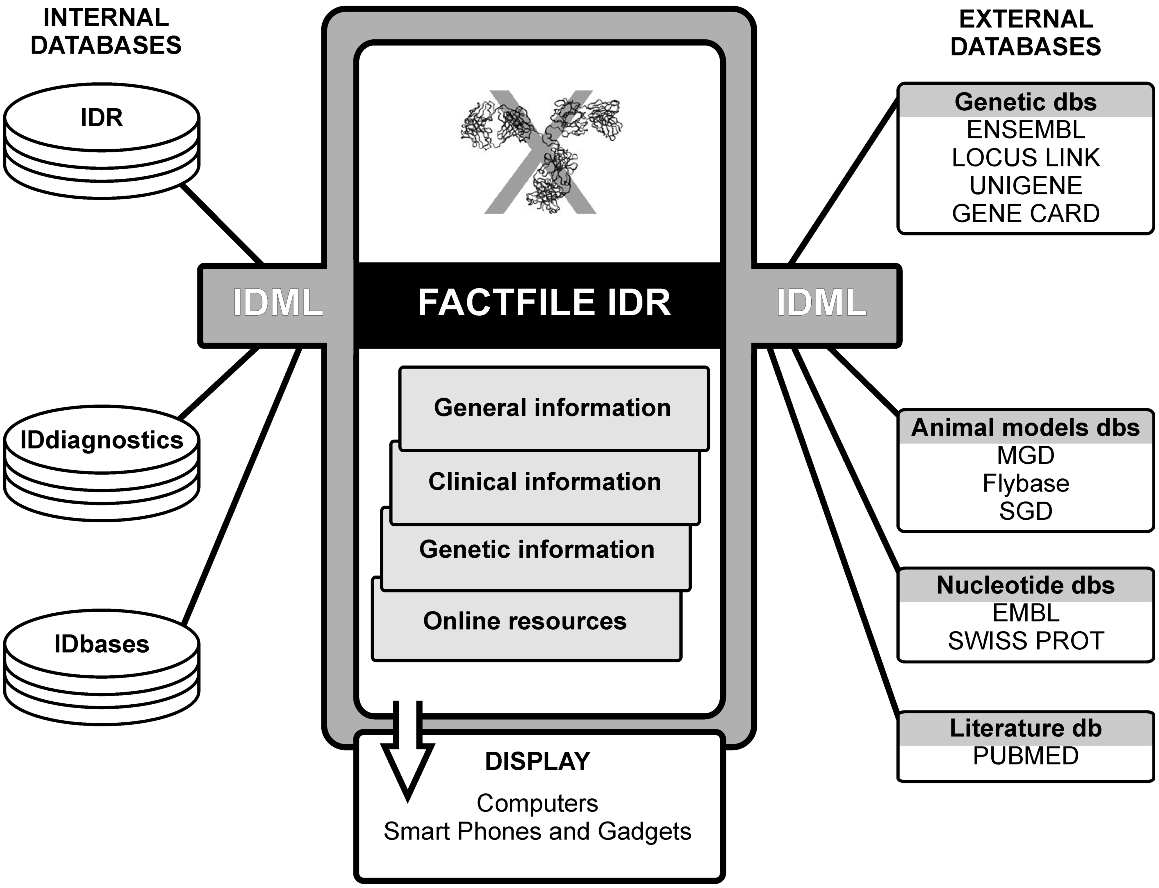


Figure 5. Factfiles at the core of IDR

The IDR includes in an organized manner also electronic articles, links to electronic journals, e-books, digital atlases, other online health services, instructional resources, analyses and visualization tools as well as advanced search routines.

The IDR contains extensive cross-referencing and links to other services. The IDR integrates numerous web-based services e.g. sequence databases (EMBL, GenBank, SWISS-PROT), genome information (Ensembl, GDB, UniGene, GeneCard, GenAtlas, euGene), protein structural database (PDB), diseases (OMIM), references (PubMed), clinical guidelines (ESID/PAGID recommendations), mutation data (IDbases), animal models (MGD, FlyBase, SacchDB) and information produced by the IDR team (Väliaho et al. 2002). The immunodeficiencies category includes, for example, an *Introduction to'*, and '*Classification of'*, immunodeficiencies web pages. Information about the affected genes and loci are linked to corresponding servers. ESID and PAGID recommendations for diagnostics criteria for immunodeficiences and other guidelines developed by different parties are also distributed. The immunology section includes immunology related data sources such as lectures on immunology and immunodeficiencies. The IDbases section contains links to some 122 mutation registries (Piirilä et al. 2006). Most of them are maintained by IBT Bioinformatics. The pages for animal models list links to knock-outs of immunodeficiency related genes in mice (MGD), *Drosophila melanogaster* (FlyBase) and *Saccharomyces cerevisiae* (SacchDB). Interest groups for immunologists, nurses and patients are listed. There are several societies related to immunodeficiency research, care and patients. The immunology laboratories section contains a list of home pages of laboratories which are active in the many fields of immunodeficiency research including diagnosis, treatment and basic research in such areas as genetic analyses, protein structure determination and signal transduction. It is also possible to read about immunodeficiency related research programs. Links to immunodeficiency related DNA and amino acid sequences are available as well as to three-dimensional structures of proteins. Furthermore, a picture gallery and a list of meetings and workshops are available.

IDR has an increasing number of users that demonstrate that the web is an effective way to deliver complex and updated information to the medical community. As it is shown in the figure below in the IDR there are more than 10000 hits each month (Figure 6).

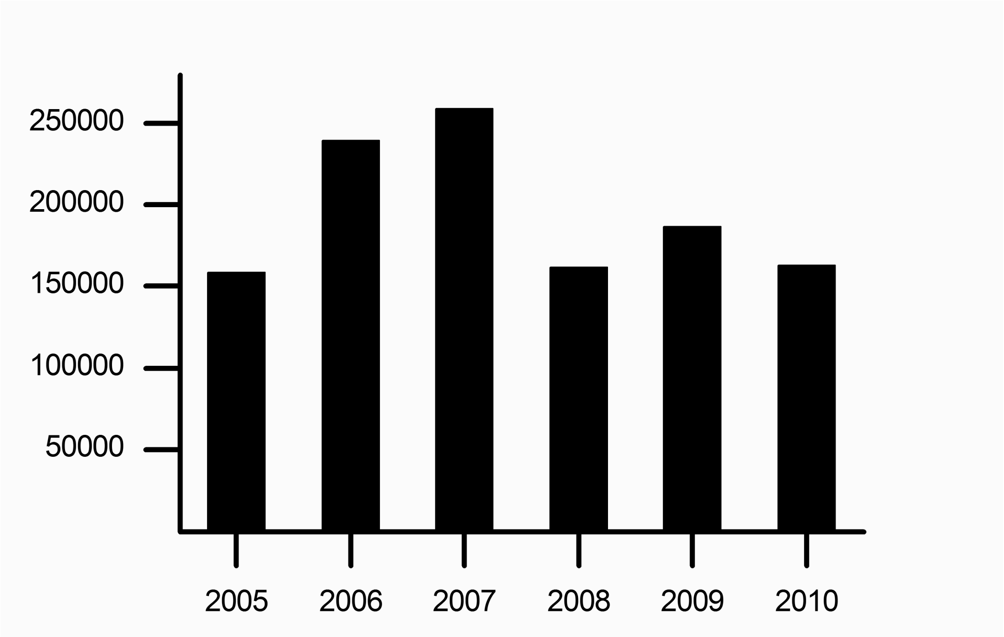


Figure 6. Number of pages requests per month/year in IDR over the time

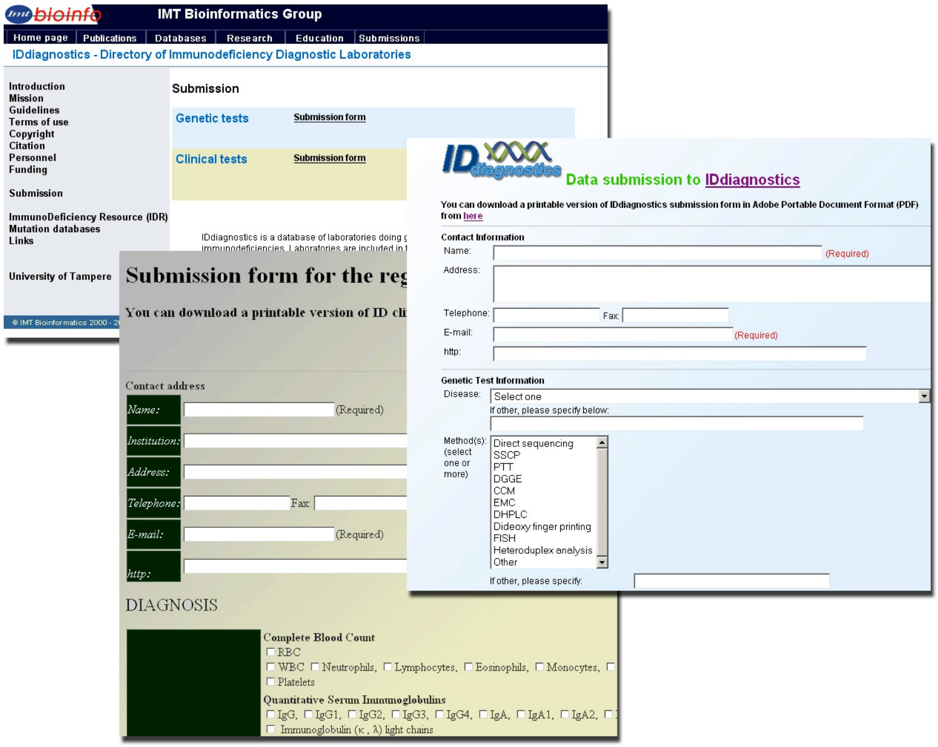
The Internet contains a large amount of pages, but only a few contain information about data validation. Many times search engines give thousands of links making it almost impossible to trawl through all this information. The most difficult task is usually not to find data but to differentiate useful and reliable data from other less important search results. In the IDR, the experts check all the data; especially links to external information sources, and approve only those sites with solid scientific and medical information.

The IDR is easy to navigate. The IDR pages are color coded for different interest groups: researchers, physicians, nurses, patients and families. By selecting the group of interest, the user can get specific pages, such as an introduction written for this particular group. This makes it easier for the user to find interesting and useful information for their own area. The IDR also provides an advanced text search facility, that can utilize boolean logic searches with multiple keywords. IDR is integrated with internal and external databases, which makes this site a valuable resource for PIDs and a quick tool to be used. The structured data from the IDR facfiles has been integrated as weblinks in ORPHANET and in the Asian database for immunodeficiencies (Keerthikumar et al. 2009).

5.2 IDdiagnostics (I)

IDdiagnostics is an online registry that includes worldwide data for genetic and clinical tests for PIDs. IDdiagnostics is available online at http://bioinf.uta.fi/IDdiagnostics (Figures 7, 8). The service is based on extensible Markup XML and is interlinked with internal databases: IDR (II) and IDbases (Piirilä et al. 2006) and with external databases OMIM (Amberger et al. 2009), GeneCards (Safran et al. 2002) .

The submission forms for clinical and genetic tests are available online and also in paper format (Figure 7, 8). The genetic test questionnaire includes contact information, diseases investigated and details about the test. The clinical test questionnaire contains contact information about doctors and laboratories that perform analyses, and details about immunological tests. The submission is sent by email to the curators.



IMT Bioinformatics Group

Home page | Publications | Databases | Research | Education | Submissions

IDdiagnostics - Directory of Immunodeficiency Diagnostic Laboratories

Introduction
Mission
Guidelines
Terms of use
Copyright
Citation
Personnel
Funding

Submission

ImmunoDeficiency Resource (IDR)
Mutation databases
Links

University of Tampere

© IMT Bioinformatics 2000 - 2

Submission

Genetic tests	Submission form
Clinical tests	Submission form

IDdiagnostics is a database of laboratories doing g
immunodeficiencies. Laboratories are included in t

Submission form for the reg

You can download a printable version of ID cli

Contact address

Name: (Required)
Institution:
Address:
Telephone: Fax:
E-mail: (Required)
http:

DIAGNOSIS

Complete Blood Count
□ RBC
□ WBC □ Neutrophils, □ Lymphocytes, □ Eosinophils, □ Monocytes, □
□ Platelets
Quantitative Serum Immunoglobulins
□ IgG, □ IgG1, □ IgG2, □ IgG3, □ IgG4, □ IgA, □ IgA1, □ IgA2, □
□ Immunoglobulin (κ , λ) light chains

Data submission to IDdiagnostics

You can download a printable version of IDdiagnostics submission form in Adobe Portable Document Format (PDF) from here

Contact Information
Name: (Required)
Address:
Telephone: Fax:
E-mail: (Required)
http:

Genetic Test Information
Disease: Select one
If other, please specify below:
Method(s): (select one or more)
Direct sequencing
SSCP
PTT
DGGE
CCM
EMC
DHPLC
Dideoxy finger printing
FISH
Heteroduplex analysis
Other
If other, please specify:

Figure 7. IDdiagnostics submission form for gene/clinical test information

After the curation process the data appears online in the Internet database. Searching facilities allow users to run text-based search queries. The search engine makes it easy to find laboratories for certain diseases, methods, and geographical location.

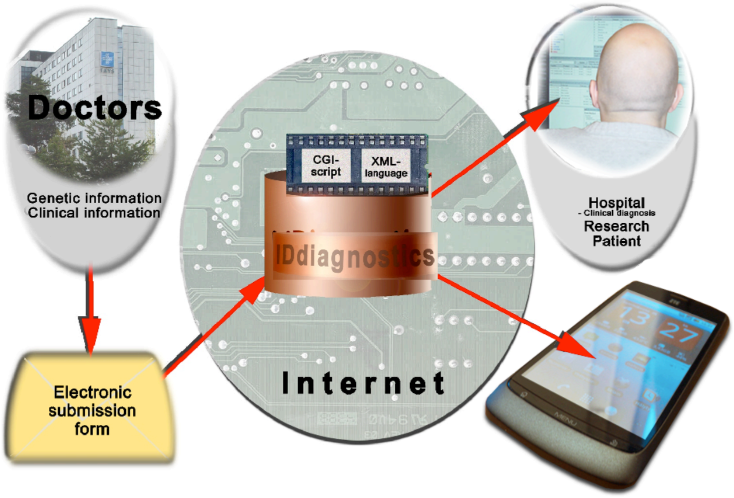


Figure 8. IDdiagnostics and different ways to use it

IDdiagnostics increases the availability of genetic and clinical methods for PIDs and provides possibilities for accurate diagnosis and fast communication between laboratories. Until now, within the genetic tests there are 222 entries for 43 diseases from 52 centers in 10 countries. In the clinical test database there are 30 entries from 29 centers in 12 countries.

Searches from IDdiagnostics are easy by using free text or by limiting the search to certain data fields. Gene test laboratories can be searched by disease name (also alternative names), gene symbol, OMIM code, laboratory, laboratory location, and free text. Similar searches are available for clinical laboratories. The search engine makes it easy to find laboratories for certain diseases, methods, and geographical locations.

5.3 Systematic classification of PIDs (III)

We developed a method and produced a novel systematic classification of PIDs based on clinical signs and features as well as laboratory parameters. Network and clustering methods were used to obtain a consensus that was shown to be valid according to a number of independent features.

PIDs related information was collected from several sources (Figure 9). Characteristic clinical, pathological and laboratory parameters were identified, combined and optimized to 87 informative parameters, which were used with an equal weight in analysis. A consensus of at least five independent methods (3 clustering and 3 network community methods) yielded a novel classification of 11 groups, which revealed previously unknown features and relationships of PIDs.

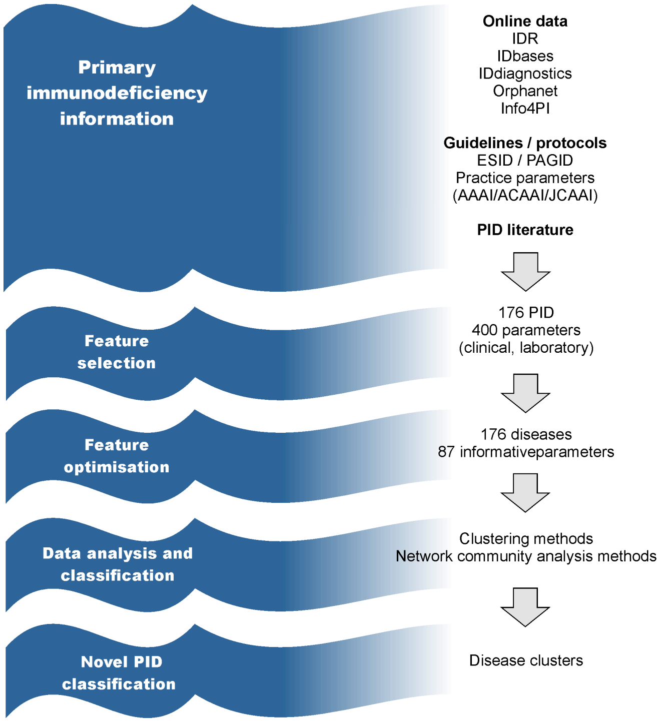


Figure 9. Approach for obtaining novel PID classification

11 well-defined disease clusters (DCs) of at least 4 diseases were identified. There is one giant cluster, which contains the majority of the PIDs and some separate clusters and singleton PIDs. 1285 pairs of diseases are grouped together by at least five methods out of a possible 18721, and out of the 12721 that are grouped by at least one method. The results are available in an interactive web page at http://bioinf.uta.fi/PID_classification.

Most of the clusters are very homogeneous and contain related diseases. These groups, which have been exclusively generated based on disease characteristics, indicate the power of the method. Other information for the PID genes, proteins and their functions further support the classification. In addition, these results are statistically significant.

Clusters III and VII contain (almost) exclusively SCIDs, whereas in cluster IX the diseases are related to the complement system and clusters I and XI to phagocyte functions; in cluster VIII are fever syndromes, and in cluster X Fanconi anemias. All of the known MHC I diseases are in cluster V and all of the MHC II genes are in cluster III. The classical complement pathway diseases are in clusters V and IX. In cluster II are diseases related to DNA instability and DNA damage repair, except for G6PC3. Clusters IV, V, and VI are the most heterogeneous. Cluster IV contains numerous receptor and signaling molecule-related diseases. Some of the proteins behind these disorders form transmembrane channels. Cluster V contains mainly antibody and complement deficiencies together with some SCIDs. The majority of the cluster VI diseases are related to phagocytosis and apoptosis.

The distribution of characteristics describing independent clinical, functional, genetic, and network properties of the diseases and the corresponding genes and proteins were investigated in order to test whether the etiologically rather homogeneous DCs shared any similarities These features were not used for the original clustering of the PIDs.

The PIDs were divided into four groups according to their severity. Severity was not among the symptoms used in our classification because it is not routinely used in the clinical description of the diseases. The severity shows a very homogeneous pattern in some of the clusters. Diseases in DCs II, VII, and IX belong to two categories, whereas in almost all of the remaining clusters there are almost exclusively moderate and severe or severe and life-threatening diseases.

PIDs are treated in a number of ways that can be grouped as: Ig treatment, the use of antibiotics, antifungals or antivirals, immunomodulators and hematopoietic stem cell transplantation. The data for the treatment of the diseases indicates that the majority of the clusters are very homogeneous in regard to treatments and there are clear differences between DCs. There are DCs in which all of the diseases are treated with the same battery of therapeutic modalities. DCs reliably reflect the properties of diseases, and the therapy applied to diseases within DCs is usually similar.

The distribution of the functional properties in the PID classification shown that the majority of the group I proteins are involved in inflammation and cellular immunity. Group II and VII proteins have functions as transcription factors involved in humoral and cellular immunity. Humoral immunity is most prevalent in DC IV, complement proteins in DC IX, and both humoral and complement functions in DC V. Inflammation is the function involved in DC VIII.

6. DISCUSSION

Health professionals such as primary immunodeficiency specialists are often forced to make decisions about diagnosis and treatment in imperfect conditions, under stressful circumstances with incomplete information, in the face of unclear pathology. Because new advances are made in the field, clinicians might accommodate new information and implement new practices. The diagnosis of PIDs is difficult due to overlapping symptoms, which often leads to misdiagnosis and delays in treatment or inappropriate treatments. Health professionals and researchers should embrace new technologies, such as databases, registries, biobanks and expert systems, in their daily practices as well as specific studies, and not be afraid of having their decisions being helped by technologies. Available data suggest that these technologies are merely tools that aid our memory limitations and facilitate more accurate decisions.

6.1.1 Databases, biobanks and registries

The IDR fact file is a user interface, which serves as a good starting point to explore information on PIDs. For some time now, there have been many advanced search facilities in the Internet such as Google, which are able to find very fast web pages that contain given keywords. However, the web searches typically turn up innumerable completely irrelevant findings, requiring much manual filtering by the user. Database searching and accessibility have some difficulties including inaccurate and redundant search results, nomenclature issues, lack of internal access, lack of customization and differing data formats. New methods are needed for improving search results.

Almost all the pages on the Internet have been written in HyperText Markup Language (HTML), which it is used for style description. It provides some possibilities for simple description about a document. It is able to use special metatags that contain simple keywords or more advanced descriptions, but they are very little utilized and only the most sophisticated search engines can exploit them.

There are some efforts to integrate heterogeneous biomedical databases. Some level of standardization is required for more automatic integration. In the future, Web services will use standard Internet protocols including SOAP, WSDL, and UDDI for interoperability with other resources. We developed IDR and IDdiagnostics using XML in accordance with the specific types of data. The PID dataset from IDR have been used to develop tools for finding new PID candidate genes (Ortutay and Vihinen 2009). IDR factfiles were helpful also in developing the PIDs network in paper (III) and RAPID (Keerthikumar et al. 2009).

Developments of intelligent databases (IDSs) and high-performance storage systems (HPSS) over the WWW have many advantages. Information sharing, collaboration between medical practitioners, on-line discussion, on-line treatment and diagnosis are among the features which enable doctors to share their knowledge and expertise. Standardized communication protocols facilitate the use of increasing datasets. Laboratory techniques, for example genome sequencing, and gene and protein expression analysis, will improve the accuracy of diagnoses and help in the discovery of novel molecular targets. New machine learning techniques will be developed in the future to automate knowledge acquisition and knowledge extraction. Future technological evolution may lead to powerful diagnostic tools based on classification algorithms developed today as stand-alone systems. They will be integrated into existing devices and hospital information systems, which will be part of large national/international databases. The current trend in disease diagnosis is to merge efforts towards a complete decision support system that also integrates modern tools from gene expression, protein expression, analysis of protein interactions, loss of function phenotyping, using RNAi, animal models, tissue microarrays, metabolomics, cellomics.

In Finland, over several decades large population survey approaches have produced datasets on especially non-communicable diseases and their risk factors. In these surveys, also biological material, mainly serum samples, have been collected. In addition to immediate laboratory analyses, samples have also been stored for future purposes. Reliable and comprehensive national administrative registers in Finland provide unique possibility for definition of disease endpoints for these cohorts.

Even though these data sources have been widely utilized and internationally recognized in epidemiological research, there are also several challenges in effective utilization of data. Instead of utilizing existing data sets, current funding opportunities easily drive researchers to new data collection. Resources available at the time of data collection restrict the focus of surveys and also sample sizes easily remain too small to examine complex associations and rare diseases and disorders. Especially, most of these data sets alone are too small to assess genetics of diseases and risk factors. Creation of larger, comprehensive datasets is complicated and resource consuming.

Well-developed epidemiological biobanks could markedly improve the quantity, quality and utilization of data and thus increase effectiveness and reduce costs of research on the etiology of complex diseases. They could also enlarge the scope and perspectives in research and increase national and international collaboration. In this development all different stakeholders such as policy makers, researchers, funders and the general public would need to have a common understanding of the principles of biobanking and legal systems and ethical review mechanisms applied to biobanks should be enabling and fit-for-purpose. Many practical challenges would need to be solved. On the other hand, lots of experience and skills already exist to enhance the development.

Data on PID patients in epidemiological biobanks in Finland are spread in different biobanks for other diseases such as the hematological registry and cancer registry. It would be beneficial to merge the data already collected in IDR, IDbases or IDdiagnostics with data from biobanks and a national registry specifically targeted at PIDs. At the moment there is no national PID registry and the development of such a registry would constitute a public health priority for Finland or other countries, such as Romania which don't have these kinds of healthcare systems yet for PIDs.

Data from IDR, IDdiagnostics and IDbases have been already integrated into the ESID registry, a registry which collects PID information from all over the Europe (Eades-Perner et al. 2006; Gathmann et al. 2009). Current efforts are under way to standardize the information from different national PID registries and centralize at the European level in the ESID registry.

Finland will benefit also by merging the bioinformatics tools developed locally already for PIDs with a national registry and a biobank specifically tailored for PIDs and also by connecting the data with the ESID registry (Gathmann, Grimbacher et al. 2009), other Scandinavian PID registries, the SCETIDE registry and the EBMT registry.

6.1.2 Systematic classification of PIDs

The knowledge accumulated in the IDR factfiles, IDbases and IDdiagnostics was also used beside most recent discoveries to develop a systematic classification for PIDs. This novel method is based on several clinical, pathological and physiological parameters. The new classification shares certain features with previous groupings. For example, the new classification indicates that cell type expression cannot be very reliably used for classification of PIDs. The obtained 11 disease clusters are very robust due to the consensus of at least five methods. The p values show the significance of the observations. More detailed subgrouping is available by using the consensus of all the six methods. The PID parameters could offer guidelines for medical descriptions of PIDs. The classification allows a novel and fresh look at the relationships of PIDs, the genes behind them and the encoded proteins. Based on the classification and the parameters it might be possible to develop novel diagnostic schemes for PIDs. The correlation with other independent information implies that the classification reflects numerous properties of the diseases and the genes and proteins mutated in them.

The previous PID classifications have been based on the cell types in which the disease-related genes are normally expressed. Antibody deficiencies, combined PIDs and diseases related to phagocytosis are widely scattered in the graph. These heterogenous diseases affect numerous parts of the immune system. The highest concentration of SCIDs is in DCs II, IV and VII, whereas antibody deficiencies mainly appear in clusters IV and V. Phagocyte diseases are exclusive to DCs I and XI but also appear in DCs IV, V and VI.

Clustering methods are widely used in many fields, such as in microarray data analysis. In medicine, cluster analysis has been used in the nosological splitting of different phenotypes, to classify patients with chronic pain and to develop a new taxonomy for airway diseases. We combined both network and clustering methods and used their consensus to obtain a robust grouping for PIDs. Considering the consensus of four, five, or six independent methods makes the results robust and independent from the individual methods.

The new PID classification can have several applications. Because it was generated independently from the existing classifications using solely mathematical analysis of clinical and laboratory parameters, it can be used in evaluation and development of other classifications. Disease groups defined by experts and found also by our independent approach have a strong indication of their existence. Some other possible applications of the new classification are:

1. Organizing PID and information about them in the IDR
2. Detailed demographics
3. Mortality records and
4. Development of a decision support for PID.

7. SUMMARY AND CONCLUSIONS

In this study we have developed two databases and a procedure for in depth analysis of PIDs. The analysis has provided interesting insights into different PIDs relationships at clinical and molecular level.

Online knowledge bases, such as IDR collect and distribute many kinds of information, and keep users up-to-date with the deluge of data. Resources for diagnosis laboratories, such as IDdiagnostics are frequently used by those wanting to verify a suspected PID case. Classification of PIDs and knowledge of candidate genes have many applications. New methods, algorithms and ontologies are needed to fully exploit all the available data for primary immunodeficiencies.

IDR, IDdiagnostics and PIDs systematic classification provide a comprehensive navigation point for anyone interested in these disorders, whether a physician, a nurse, a research scientist, a patient, a parent of a patient or general public. These bioinformatics tools contain in a systemized manner, continuously renewed information, which can be fast approach and updated. They have proved to be a real help for clinicians, medical students and researchers and may help in improving medical care. Newly discovered relationships found in biomedical databases like these will potentially lead to better understanding between observations and outcomes in PIDs and other fields of medicine. The method used for PID classification has built-in statistical output and iterative analysis using different inputs leads to alternative classifications useful in gene discovery, therapeutic trials, and the development of screening or diagnosis protocols. The flexibility and statistical power of these methods make them very suitable for the development of evidence-based protocols for diagnosis and therapy, not just of PID, but of any other group(s) of diseases. This classification approach bridges the existing knowledge bases with diagnostic and therapeutic protocols, which may be applied directly in patient care. The new approach proposed (using data warehousing and mining technology in the health care) can influence medical outcomes and merit further studies and funding.

With the rapidly changing landscape of healthcare, these kind of bioinformatics tools need to evolve quickly. They should be allowed to choose the functionalities that represent the best available ideas in healthcare and should be made compatible and interoperable with other IT healthcare systems.

8. ACKNOWLEDGEMENTS

The work described in this research project has been mainly carried out in the Bioinformatics group at the Institute of Biomedical Technology, University of Tampere. Finland. The sincerest appreciation and gratitude goes to Prof. Mauno Vihinen, Head of the Bioinformatics group for the opportunity to work in his group and for excellent teaching and guidance during all these years.

The Bioinformatics group at the Institute of Biomedical Technology has been a pleasant environment. All my gratitude goes to the entire personnel of the group. Special thanks are reserved for Jouni Väliaho, with whom I worked in many projects and have been co-author in some of my articles (I, II), as well as Csaba Ortutay, co-author in paper (III). Martti Tolvanen, Markku Siermala, Sofia Khan, Johanna Salonen, Heidi Ali, Janita Thursberg and Kathryn Rannikko have been wonderful colleagues with whom I spent many hours debating about life and science. Hannu Korhonen and Jukka Lehtiniemi provided valuable technical and creative support in many aspects of my research. I am deeply grateful for their help and positive, creative attitude.

Special thanks go to Kati Iltanen and Kirsi Varpa from Department of Computer Science, and the late Ayodeji Olatubosun’ from the Bioinformatics group, Institute of Biomedical Technology, University of Tampere that has been involved in the development of PIDexpert.

I would like to acknowledge the former director of the Institute of Medical Technology, Professor Olli Silvennoinen, M.D., Ph.D., and the Director of BioMediTech, Dr. Hannu Hanhijärvi, Ph.D., for providing the research facility for this study. Special thanks go to Prof. Howard Jacobs, Head of the Postgraduate Program, not only for his inspiring lectures about mitochondria and immunology but also for his precious support in the administrative matters of the doctoral thesis, as well as to

Henna Matilla Ph.D., the coordinator of the program, Prof. Jorma Isola, the Chair of the Doctoral studies committee and all the wonderful secretaries in IBT/UTA.

Minna Laine and Kaj Stenberg, members of the thesis committee provided valuable feedback all these years. I am deeply grateful to them.

Professors Olli Lassila and Lennart Hammarström are also acknowledged for engaging in the revision of this thesis and for their valuable comments and criticism. I am especially indebted to Professor Ole Lund who agreed to act as my opponent in the public examination of this thesis.

My clinical expertise in primary immunodeficiencies has been improved in the Children's Infectious Disease department, Tampere University Hospital, Finland; Children's Immunology Department at Newcastle General Hospital in the United Kingdom and in the Children's Hospital, University of Brescia, Italy. I am very grateful to all the wonderful clinicians I met there, especially Prof. Andrew Cant, Prof. Gavin Spickett, Dr. Andrew Genery, Doc. Merja Helminen, Dr. Silvia Gigliani, Prof. Alessandro Plebani and Dr. Annarosa Soresina. They have been such inspiring hosts for me.

My deepest gratitude goes to the European Society of Immunodeficiency (ESID) and the entire PID community, for continuous and useful interaction all these years. Special thanks for my 'golden mentors': Prof. Luigi Notarangelo, Prof. Anders Fasth, Prof. Hans Ochs, Prof. Alain Fisher, Prof. Steve Holland, Prof. Hellen Chapel and Dr. Esther de Vries for useful scientific/clinical discussion and for mentorship. ESID Juniors have a special place in my heart. The Advanced School in PIDs, Sao Paolo, Brazil 2010, excellently organized by Prof. Magda Carneiro enriched tremendously my knowledge in PIDs as well as ESID Registry workshops and ESID Juniors symposium/workshops. I am very grateful to all of them.

Special thanks go to Dr. Kaarina Järvenpää and her wonderful team from Helsinki who enriches tremendously my experience in pediatrics and provide a pleasant environment to grow, as well as to Dr. Tuomo Lunnikivi from Hämeenlinna who improved immensely my clinical skills in geriatrics.

I would like to thank also to Countess Bettina Bernadotte and the Lindau Nobel Laureates Meeting committee who allowed me in 2011 to have the unique chance to meet 20 Nobel laureates and the best 500 young researchers in the world. This meeting gave us not only inspiring lectures but also good models every young researcher should aspire to. I am very happy that I could represent successfully both Finland and Romania in this prestigious meeting, which contributed immensely to my knowledge and my future research.

The Academy of Finland, Sigrid Juselius Foundation, Tampere University Hospital Medical Research Fund, Tampere City Hall and the European Union supported the research, and are gratefully acknowledged. Different professional associations (CIS, ESID, EFIS, SSI), different pharmaceutical companies: Talecris, CSL Behring (Italian and Finnish branch), Sanquin (Finnish branch), Baxter, Octapharma, Kedrion, Grifols and the Tampere Graduate Program in Biomedicine and Biotechnology (TGPBB) helped the research through different travel awards and sponsorships and made possible the dissemination of the results. I am very grateful to all of them.

I am thankful to all these wonderful people (doctors, researchers, nurses, patients) for useful discussions, commitment to our services and their valuable articles, without them none of these would be possible.

In the end, life is not only work, medicine and science. Sometimes, it can be music, sometimes it can be dance and sometimes it can be just simply an empty canvas ready to paint on it. I am deeply grateful to all of my friends from Tampere and all over the world that make my life so colorful and enjoyable. You are all in my heart no matter where you are situated.

Finally, my whole love goes to my family, Nicoleta, Costantino and Luca who send me systematically infusions of sun and love. My mother, Ioana gave me strength, patience and a positive attitude that can overcome all the hardships of life. Without her unconditional love none of these would be possible.

Tampere, 11 May 2012

Crina Samarghitean

9. REFERENCES

Abbas A and Lichtman A (2005). Cellular and Mollecular Immunology, Saunders.

Al-Muhsen S and Casanova JL (2008). "The genetic heterogeneity of mendelian susceptibility to mycobacterial diseases." J Allergy Clin Immunol 122(6): 1043-51; quiz 1052-3.

Amberger J, Bocchini CA, Scott AF and Hamosh A (2009). "McKusick's Online Mendelian Inheritance in Man (OMIM)." Nucleic Acids Res 37(Database issue): D793-6.

Ayme S and Schmidtke J (2007). "Networking for rare diseases: a necessity for Europe." Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 50(12): 1477-83.

Baumgart KW, Britton WJ, Kemp A, French M and Roberton D (1997). "The spectrum of primary immunodeficiency disorders in Australia." J Allergy Clin Immunol 100(3): 415-23.

Bellazzi R and Zupan B (2008). "Predictive data mining in clinical medicine: current issues and guidelines." Int J Med Inform 77(2): 81-97.

Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez X, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehväslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M and Hubbard T (2004). "Ensembl 2004." Nucleic Acids Res(Database issue): D468-70.

Blum B (1986). Clinical information systems. New York, Springer Verlag.

Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin M, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." 31: 365-370.

Bonilla F, Bernstein I and Khan Dea (2005). "American Academy of Allergy, Asthma and Immunology; Joint Council of Allergy, Asthma and Immunology. Practice parameter for the diagnosis and management of primary immunodeficiency." Ann Allergy Asthma Immunol 94(5 Suppl. 1): S1-63.

Bonilla FA and Geha RS (2005). "Are you immunodeficient?" J Allergy Clin Immunol 116(2): 423-5.

Branson A, Hauer T, McClatchey R, Rogulin D and Shamdasani J (2008). "A data model for integrating heterogeneous medical data in the Health-e-Child project." Stud Health Technol Inform 138: 13-23.

Bruton OC (1952). "Agammaglobulinemia." Pediatrics 9(6): 722-8.

Buchanan B and Shortliffe E (1984). Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming. California, Addison-Wesley.

Buckley RH (2002). "Primary immunodeficiency diseases: dissectors of the immune system." Immunol Rev 185: 206-19.

Burg Mvd, Zelm MCv and Dongen JJMv (2009). "Molecular diagnostics of primary immunodeficiencies: benefits and future challenges." Adv Exp Med Biol. 634: 231-234.

Carneiro-Sampaio M, Jacob MC and Leone CR (2010). "IN PRESS." Pediatr Allergy Immunol.

Casanova JL, Fieschi C, Bustamante J, Reichenbach J, Remus N, von Bernuth H and Picard C (2005). "From idiopathic infectious diseases to novel primary immunodeficiencies." J Allergy Clin Immunol 116(2): 426-30.

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL, Bousso P, Deist FL and Fischer A (2000). "Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease." Science 288(5466): 669-72.

Chapman W, Fizman M, Chapman B and Haug P (2001). "A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia." J Biomed Inform 34: 4-14.

Conley M, Notarangelo L and Etzioni A (1999). "Diagnostic criteria for primary immunodeficiencies." Clin Immunol 93: 190-197.

Cooper MD, Faulk WP, Fudenberg HH, Good RA, Hitzig W, Kunkel H, Rosen FS, Seligmann M, Soothill J and Wedgwood RJ (1973). "Classification of primary immunodeficiencies." N Engl J Med 288(18): 966-7.

Cooper MD, Gabrielsen AE and Good RA (1967). "Role of the thymus and other central lymphoid tissues in immunological disease." Annu Rev Med 18: 113-38.

Cooper MD, Peterson RD and Good RA (1965). "Delineation of the Thymic and Bursal Lymphoid Systems in the Chicken." Nature 205: 143-6.

Cotton R, Scriver C and McKusick V (1996). "Locus-specific mutation databases: a resource." Genome Digest 3(1): 6-10.

de Vries E and Clinical Working Party of the European Society for Immunodeficiencies (ESID) (2006). "Patient-centred screening for primary immunodeficiency: a multi-stage diagnostic protocol designed for non-immunologists." Clin Exp Immunol 145(2): 204-14.

Dexter TJ, Sims D, Mitsopoulos C, Mackay A, Grigoriadis A, Ahmad AS and Zvelebil M (2010). "Genomic distance entrained clustering and regression modelling highlights interacting genomic regions contributing to proliferation in breast cancer." BMC Syst Biol 4: 127.

Diehn M, Sherlock G, Binkley G, Jin H, Matese J, Hernandez-Boussard T, Rees C, Cherry J, Botstein D, Brown P and Alizadeh A (2003). "SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data." Nucleic Acids Res 31: 219-223.

Eades-Perner A, Knerr V, Gathmann B, Guzman D, Veit D, Kindle G and Grimbacher B (2006). "The European internet-based patient and research database for primary immunodeficiencies: Results 2004-2006." Clin Exp Immunol: 1-5.

Eades-Perner AM, Gathmann B, Knerr V, Guzman D, Veit D, Kindle G and Grimbacher B (2007). "The European internet-based patient and research database for primary immunodeficiencies: results 2004-06." Clin Exp Immunol 147(2): 306-12.

Elmasri R and Navathe S (2004). Fundamentals of Database System, Pearson-Addison Wesley.

Fasth A (1982). "Primary immunodeficiency disorders in Sweden: cases among children, 1974-1979." J.Clin. Immunol. 2: 86-92.

Fleisher TA and Notarangelo LD (2008). "What does it take to call it a pathogenic mutation?" Clin Immunol 128(3): 285-6.

Folds J and JL. S (2003). "Clinical and laboratory assessement of immunity." J Allergy Clin Immunol 111 (2 Suppl): S702-11.

Freitas AA, Vasieva O and de Magalhaes JP (2011). "A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related." BMC Genomics 12(1): 27.

Freund J, Comaniciu D, Ioannis Y, Liu P, McClatchey R, Morley-Fletcher E, Pennec X, Pongiglione G and Zhou XS (2006). "Health-e-child: an integrated biomedical platform for grid-based paediatric applications." Stud Health Technol Inform 120: 259-70.

Friedman C, Elstein A, Wolf F, Murphy G, Franz T, Heckerling P, Fine P, Miller T and Abraham V (1999). "Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems." JAMA 282: 1851-1856.

Gaal B, Vassanyi I and Kozmann G (2005). "A novel artificial intelligence method for weekly dietary menu planning." Methods Inf Med 44: 655-664.

Gathmann B, Grimbacher B, Beaute J, Dudoit Y, Mahlaoui N, Fischer A, Knerr V and Kindle G (2009). "The European internet-based patient and research database for primary immunodeficiencies: results 2006-2008." Clin Exp Immunol 157 Suppl 1: 3-11.

Gatti RA, Berkel I, Boder E, Braedt G, Charmley P, Concannon P, Ersoy F, Foroud T, Jaspers NG, Lange K and et al. (1988). "Localization of an ataxia-telangiectasia gene to chromosome 11q22-23." Nature 336(6199): 577-80.

Gatti RA, Meuwissen HJ, Allen HD, Hong R and Good RA (1968). "Immunological reconstitution of sex-linked lymphopenic immunological deficiency." Lancet **2**(7583): 1366-9.

Geha RS, Notarangelo LD, Casanova JL, Chapel H, Conley ME, Fischer A, Hammarstrom L, Nonoyama S, Ochs HD, Puck JM, Roifman C, Seger R and Wedgwood J (2007). "Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee." J Allergy Clin Immunol 120(4): 776-94.

Gennery AR, Slatter MA, Grandin L, Taupin P, Cant AJ, Veys P, Amrolia PJ, Gaspar HB, Davies EG, Friedrich W, Hoenig M, Notarangelo LD, Mazzolari E, Porta F, Bredius RG, Lankester AC, Wulffraat NM, Seger R, Gungor T, Fasth A, Sedlacek P, Neven B, Blanche S, Fischer A, Cavazzana-Calvo M and Landais P (2010). "Transplantation of hematopoietic stem cells and long-term survival for primary immunodeficiencies in Europe: entering a new century, do we do better?" J Allergy Clin Immunol 126(3): 602-10 e1-11.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommarinh M, Patnaik SK, Blumenfeld O, Gottlieb B, Vihinen M, Valiaho J, Kent J, Miller W and Hardison RC (2007). "PhenCode: connecting ENCODE data with mutations and phenotype." Hum Mutat 28(6): 554-62.

Godall A (1985). The guide to expert systems. Oxford, England, Learned Information Ltd.

Gonzalez A and Dankel D (1993). The engineering of knowledge-based systems: Theory and practice. New Jersey, Prentice-Hall, Englewood Cliffs.

Good RA, Dalmasso AP, Martinez C, Archer OK, Pierce JC and Papermaster BW (1962). "The role of the thymus in development of immunologic capacity in rabbits and mice." J Exp Med 116: 773-96.

Griffith LM, Cowan MJ, Kohn DB, Notarangelo LD, Puck JM, Schultz KR, Buckley RH, Eapen M, Kamani NR, O'Reilly RJ, Parkman R, Roifman CM, Sullivan KE, Filipovich AH, Fleisher TA and Shearer WT (2008). "Allogeneic hematopoietic cell transplantation for primary immune deficiency diseases: current status and critical needs." J Allergy Clin Immunol 122(6): 1087-96.

Han J and Kamber M (2000). Data mining: concepts and technique. California, Morgan Kaufmann Publishers.

Hauben M, Madigan D, Gerrits CM, Walsh L and Van Puijenbroek EP (2005). "The role of data mining in pharmacovigilance." Expert Opin Drug Saf 4(5): 929-48.

Hitzig WH, Biro Z, Bosch H and Huser HJ (1958). "[Agammaglobulinemia & alymphocytosis with atrophy of lymphatic tissue.]." Helv Paediatr Acta 13(6): 551-85.

Holmes B, Quie PG, Windhorst DB, Pollara B and Good RA (1966). "Protection of phagocytized bacteria from the killing action of antibiotics." Nature 210(5041): 1131-2.

Horowitz M (2008). "The role of registries in facilitating clinical research in BMT: examples from the Center for International Blood and Marrow Transplant Research." Bone Marrow Transplant 42 Suppl 1: S1-S2.

Hripcsak G, Bakken S, Stetson P and Patel V (2003). "Mining complex clinical data for patient safety research: a framework for event discovery." J Biomed Inform 36: 120-130.

Jimeno-Yepes A, Jimenez-Ruiz E, Berlanga-Llavori R and Rebholz-Schuhmann D (2009). "Reuse of terminological resources for efficient ontological engineering in Life Sciences." BMC Bioinformatics 10 Suppl 10: S4.

Kalina T, Stuchly J, Janda A, Hrusak O, Ruzickova S, Sediva A, Litzman J and Vlkova M (2009). "Profiling of polychromatic flow cytometry data on B-cells reveals patients' clusters in common variable immunodeficiency." Cytometry A 75(11): 902-9.

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D and Kent WJ (2008). "The UCSC

Genome Browser Database: 2008 update." Nucleic Acids Res 36(Database issue): D773-9.

Keerthikumar S, Bhadra S, Kandasamy K, Raju R, Ramachandra YL, Bhattacharyya C, Imai K, Ohara O, Mohan S and Pandey A (2009). "Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach." DNA Res 16(6): 345-51.

Keerthikumar S, Raju R, Kandasamy K, Hijikata A, Ramabadran S, Balakrishnan L, Ahmed M, Rani S, Selvan LD, Somanathan DS, Ray S, Bhattacharjee M, Gollapudi S, Ramachandra YL, Bhadra S, Bhattacharyya C, Imai K, Nonoyama S, Kanegane H, Miyawaki T, Pandey A, Ohara O and Mohan S (2009). "RAPID: Resource of Asian Primary Immunodeficiency Diseases." Nucleic Acids Res 37(Database issue): D863-7.

Kirkpatrick P and Riminton S (2007). "Primary immunodeficiency diseases in Australia and New Zealand." J Clin Immunol 27(5): 517-24.

Klemperer MR, Woodworth HC, Rosen FS and Austen KF (1966). "Hereditary deficiency of the second component of complement (C'2) in man." J Clin Invest 45(6): 880-90.

Kostmann R (1956). "Infantile genetic agranulocytosis; agranulocytosis infantilis hereditaria." Acta Paediatr Suppl 45(Suppl 105): 1-78.

Kulikowski C (1985). History and development of artificial intelligence methods for medical decision making. The Biomedical Engineering Handbook. J. Bronzino. USA, CRC Press and IEEE Press: 2681-2698.

Kwan SP, Kunkel L, Bruns G, Wedgwood RJ, Latt S and Rosen FS (1986). "Mapping of the X-linked agammaglobulinemia locus by use of restriction fragment-length polymorphism." J Clin Invest 77(2): 649-52.

Kwan SP, Sandkuyl LA, Blaese M, Kunkel LM, Bruns G, Parmley R, Skarshaug S, Page DC, Ott J and Rosen FS (1988). "Genetic mapping of the Wiskott-Aldrich syndrome with two highly-linked polymorphic DNA markers." Genomics 3(1): 39-43.

Ledley R and Lusted L (1959). "Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason." Science 130: 9-21.

Leiva LE, Zelazco M, Oleastro M, Carneiro-Sampaio M, Condino-Neto A, Costa-Carvalho BT, Grumach AS, Quezada A, Patino P, Franco JL, Porras O, Rodriguez FJ, Espinosa-Rosales FJ, Espinosa-Padilla SE, Almillategui D, Martinez C, Tafur JR, Valentin M, Benarroch L, Barroso R and Sorensen RU (2007). "Primary immunodeficiency diseases in Latin America: the second report of the LAGID registry." J Clin Immunol 27(1): 101-8.

Lindquist M, Stahl M, Bate A, Edwards IR and Meyboom RH (2000). "A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database." Drug Saf 23(6): 533-42.

Luke DA and Harris JK (2007). "Network analysis in public health: history, methods, and applications." Annu Rev Public Health 28: 69-93.

Lundberg GD (2006). "WebMD, Medscape, eMedicine, and the relevance of a medical encyclopedia in 2006." MedGenMed 8(1): 32.

Luzi G, Businco L and Aiuti F (1983). "Primary immunodeficiency syndromes in Italy: a report of the national register in children and adults." J Clin Immunol 3(4): 316-20.

Marodi L and Notarangelo LD (2007). "Immunological and genetic bases of new primary immunodeficiencies." Nat Rev Immunol 7(11): 851-61.

Matamoros Flori N, Mila Llambi J, Espanol Boren T, Raga Borja S and Fontan Casariego G (1997). "Primary immunodeficiency syndrome in Spain: first report of the National Registry in Children and Adults." J Clin Immunol 17(4): 333-9.

Miller JF (1961). "Immunological function of the thymus." Lancet 2(7205): 748-9.

Miller R, Pople HJ and Myers J (1982). "Internist-1, an experimental computer-based diagnostic consultant for general internal medicine." N Engl J Med 307: 468-476.

Morra M, Geigenmuller U, Curran J, Rainville IR, Brennan T, Curtis J, Reichert V, Hovhannisyan H, Majzoub J and Miller DT (2008). "Genetic diagnosis of primary immune deficiencies." Immunol Allergy Clin North Am 28(2): 387-412.

Nikolopoulos C (1997). Expert Systems: Introduction to first and second generation and hybrid knowledge based systems. New York, Marcel Dekker.

Notarangelo L, Casanova J and Fischer Aea (2004). "Primary immunodeficiency diseases: un apdate." J Clin Immunol(114): 677-87.

Notarangelo L, Casanova JL, Conley ME, Chapel H, Fischer A, Puck J, Roifman C, Seger R and Geha RS (2006). "Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee Meeting in Budapest, 2005." J Allergy Clin Immunol 117(4): 883-96.

Notarangelo LD, Fischer A, Geha RS, Casanova JL, Chapel H, Conley ME, Cunningham-Rundles C, Etzioni A, Hammartrom L, Nonoyama S, Ochs HD, Puck J, Roifman C, Seger R and Wedgwood J (2009). "Primary immunodeficiencies: 2009 update." J Allergy Clin Immunol 124(6): 1161-78.

Notarangelo LD and Sorensen R (2008). "Is it necessary to identify molecular defects in primary immunodeficiency disease?" J Allergy Clin Immunol 122(6): 1069-73.

Ochs H, Smith C and Puck J (2006). Primary immunodeficiency disease: a molecular and genetic approach. New York, Oxford University Press.

Ortutay C, Siermala M and Vihinen M (2007). "Molecular characterization of the immune system: emergence of proteins, processes, and domains." Immunogenetics 59(5): 333-48.

Ortutay C and Vihinen M (2009). "Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies." Nucleic Acids Res 37(2): 622-8.

Ortutay C and Vihinen M (2009). "Immunome knowledge base (IKB): an integrated service for immunome research." BMC Immunol 10: 3.

Pagon RA (2006). "GeneTests: an online genetic information resource for health care providers." J Med Libr Assoc 94(3): 343-8.

Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R and Abu-Hanna A (2009). "The coming of age of artificial intelligence in medicine." Artif Intell Med 46(1): 5-17.

Pena-Reyes C and Sipper M (2000). "Evolutionary computation in medicine: an overview." Artif Intell Med 19: 1-23.

Piirilä H, Väliaho J and Vihinen M (2006). "Immunodeficiency mutation databases (IDbases)." Hum Mutat 27(12): 1200-8.

Piqueras B, Lavenu-Bombled C, Galicier L, Bergeron-van der Cruyssen F, Mouthon L, Chevret S, Debre P, Schmitt C and Oksenhendler E (2003). "Common variable immunodeficiency patient classification based on impaired B cell memory differentiation correlates with clinical aspects." J Clin Immunol 23(5): 385-400.

Plebani A, Soresina A, Notarangelo LD, Quinti I, Mattia DD, Moschese V, Rondelli R, Pession A and Ugazio AG (2004). "The italian network of primary immunodeficiencies." Iran J Allergy Asthma Immunol 3(4): 165-8.

Porter CJ, Talbot CC and Cuticchia AJ (2000). "Central mutation databases--a review." Hum Mutat 15(1): 36-44.

Pruitt K and Maglott D (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucleic Acids Res 29: 137-140.

Puck JM, Nussbaum RL, Smead DL and Conley ME (1989). "X-linked severe combined immunodeficiency: localization within the region Xq13.1-q21.1 by linkage and deletion analysis." Am J Hum Genet 44(5): 724-30.

Quinlan J (1993). C4.5 Programs for machine Learning. San Mateo, California, Morgan Kaufmann Publishers.

Reategui E, Campbell J and Leao B (1997). "Combining a neural network with case-based reasoning in a diagnostic system." Artif Intell Med 9: 5-27.

Rezaei N, Aghamohammadi A, Moin M, Pourpak Z, Movahedi M, Gharagozlou M, Atarod L, Ghazi BM, Isaeian A, Mahmoudi M, Abolmaali K, Mansouri D, Arshi S, Tarash NJ, Sherkat R, Akbari H, Amin R, Alborzi A, Kashef S, Farid R, Mohammadzadeh I, Shabestari MS, Nabavi M and Farhoudi A (2006). "Frequency and clinical manifestations of patients with primary immunodeficiency disorders in Iran: update from the Iranian Primary Immunodeficiency Registry." J Clin Immunol 26(6): 519-32.

Rezaei N, Aghamohammadi A and Notarangelo L (2008). Primary immunodeficiencies diseases. Definition, diagnosis and management, Springer.

Roitt I, Brostoff J and Male D (2001). Immunology, Mosby-Paperback.

Ryser O, Morell A and Hitzig WH (1988). "Primary immunodeficiencies in Switzerland: first report of the national registry in adults and children." J Clin Immunol 8(6): 479-85.

Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V and Lancet D (2002). "GeneCards 2002: towards a complete, object-oriented, human gene compendium." Bioinformatics 18(11): 1542-3.

Salganik MJ and Heckathorn DD (2004). "Sampling and estimation in hidden population using respondent-driven sampling." Sociol. Methodol. 34: 193-239.

Samarghitean C and Vihinen M (2008). "Medical expert systems." Curr Bioinf 3: 56-65.

Samarghitean C and Vihinen M (2009). "Bioinformatics services related to diagnosis of primary immunodeficiencies." Curr Opin Allergy Clin Immunol 9(6): 531-6.

Schuetz C, Huck K, Gudowius S, Megahed M, Feyen O, Hubner B, Schneider DT, Manfras B, Pannicke U, Willemze R, Knuchel R, Gobel U, Schulz A, Borkhardt A, Friedrich W, Schwarz K and Niehues T (2008). "An immunodeficiency disease with RAG mutations and granulomas." N Engl J Med 358(19): 2030-8.

Shortliffe E (1993). "The adolescence of AI in medicine: will the field come of age in the '90s?" Artif Intell Med 5: 93-106.

Sigrist C, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A and Bucher P (2002). "PROSITE: a documented database using patterns and profiles as motif descriptors." 3: 265-274.

Smith CI and Vihinen M (1996). "Immunodeficiency mutation databases-a new research tool." Immunol Today 17(11): 495-6.

Soukup T and Davidson I (2002). Visual data mining: techniques and tools for data visualization and mining. New York, Wiley.

Stiehm E, Ochs H and Winkelstein J (2004). Immunodeficiency disorders: general considerations. Immunologic disorders in infants and children. E. Stiehm, H. Ochs and W. JA. Philadelphia**:** 289-355.

Stray-Pedersen A, Abrahamsen TG and Froland SS (2000). "Primary immunodeficiency diseases in Norway." J Clin Immunol 20(6): 477-85.

Swingler K (1996). Applying Neural Networks: A Practical Guide. London, Academic Press.

Takeuchi K and Collier N (2005). "Bio-medical entity extraction using support vector machines." Artif Intell Med 33: 125-137.

Thusberg J and Vihinen M (2009). "Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods." Hum Mutat 30(5): 703-14.

Vihinen M, Lehvaslaiho H and Cotton R (1999). Immunodeficiency mutation databases. Primary immunodeficiency diseases: a molecular and genetic approach. H. Ochs, C. Smith and J. Puck. New York, Oxford University Press.

Villa A, Marrella V, Rucci F and Notarangelo LD (2008). "Genetically determined lymphopenia and autoimmune manifestations." Curr Opin Immunol 20(3): 318-24.

Väliaho J, Riikonen P and Vihinen M (2005). "Distribution of immunodeficiency fact files with XML-from Web to WAP." BMC Med Inform Decis Mak 5: 21.

Väliaho J, Pusa M, Ylinen T and Vihinen M (2002). "IDR: the ImmunoDeficiency Resource." Nucleic Acids Res 30(1): 232-4.

Väliaho J, Riikonen P and Vihinen M (2000). "Novel immunodeficiency data servers." Immunol Rev 178: 177-85.

Wain H, Lush M, Ducluzeau F, Khodiyar V and Povey S (2004). "Genew: the Human Gene Nomenclature Database, 2004 updates." Nucleic Acids Res(Database issue): D255-7.

Wang J, Zaki M, Toivonen H and Shasha D (2005). Data mining in bioinformatics, Springer.

Warnatz K and Schlesier M (2008). "Flowcytometric phenotyping of common variable immunodeficiency." Cytometry B Clin Cytom 74(5): 261-71.

Wehr C, Kivioja T, Schmitt C, Ferry B, Witte T, Eren E, Vlkova M, Hernandez M, Detkova D, Bos PR, Poerksen G, von Bernuth H, Baumann U, Goldacker S, Gutenberger S, Schlesier M, Bergeron-van der Cruyssen F, Le Garff M, Debre P, Jacobs R, Jones J, Bateman E, Litzman J, van Hagen PM, Plebani A, Schmidt RE, Thon V, Quinti I, Espanol T, Webster AD, Chapel H, Vihinen M, Oksenhendler

E, Peter HH and Warnatz K (2008). "The EUROclass trial: defining subgroups in common variable immunodeficiency." Blood 111(1): 77-85.

Weiss S, Kulikowski C, Amarel S and Safir A (1978). "A model-based method for computer-aided medical decision-making." Artif Intell 11: 145-172.

Westphal C and Blaxton T (1998). Data mining solutions - methods and tools for solving real-world problems. New York, Wiley.

WHO (1983). "Primary immunodeficiency diseases. Report prepared for the WHO by a scientific group on immunodeficiency." Clin Immunol Immunopathol 28(3): 450-75.

WHO (1986). "Primary immunodeficiency diseases. Report of a World Health Organization scientific group." Clin Immunol Immunopathol 40(1): 166-96.

WHO (1992). "Primary immunodeficiency diseases. Report of a WHO scientific group." Immunodefic Rev 3(3): 195-236.

WHO (1995). "Primary immunodeficiency diseases. Report of a WHO Scientific Group." Clin Exp Immunol 99 Suppl 1: 1-24.

WHO (1997). "Primary immunodeficiency diseases. Report of a WHO scientific group." Clin Exp Immunol 109 Suppl 1: 1-28.

WHO (1999). "Primary immunodeficiency diseases. Report of an IUIS Scientific Committee. International Union of Immunological Societies." Clin Exp Immunol 118 Suppl 1: 1-28.

Winkelstein J, Marino M, Johnston RB J and al. e (2000). "Chronic granulomatous disease. Report on a national registry of 368 patients." Medicine 79: 155-169.

10. ORIGINAL COMMUNICATIONS

Journal of Clinical Immunology, Vol. 24, No. 1, January 2004 (© 2004)

Online Registry of Genetic and Clinical Immunodeficiency Diagnostic Laboratories, IDdiagnostics

CRINA SAMARGHITEAN,[1] JOUNI VÄLIAHO,[1] and MAUNO VIHINEN[1–3]

Accepted: July 24, 2003

Primary immunodeficiencies (IDs) are caused by inherited genetic defects leading to intrinsic defects in cells of the immune systems. Most IDs are rare diseases and can be difficult to diagnose because similar symptoms characterize several disorders. Mutation detection is the most reliable method in such cases. These tests are not available at most centers and physicians can have difficulties in finding laboratories that could analyze the genetic defects because certain genes are possibly analyzed by just one laboratory. The IDdiagnostics registry has been established to provide information for physicians and other health care professionals. The database at http://bioinf.uta.fi/IDdiagnostics contains currently information for the analysis of defects in 30 ID-related genes. Another part of IDdiagnostics is a database of clinical tests. Laboratories performing these analyses, either gene or clinical tests, are asked to submit their information to the database by using a printed form or electronic submission at http://bioinf.uta.fi/cgi-bin/submit/IDClini.cgi. The clinical test database contains information about tests for clinical data, immune status, and studies of function, antibody response, cell function, enzyme assays, clinical function, and apoptosis assays. Both the services are freely available and regularly updated. The services aim at increasing the awareness of IDs and helping to obtain exact and early diagnosis.

KEY WORDS: Immunodeficiency; diagnostics; IDdiagnostics; diagnostic laboratories; genetic tests; clinical tests; DNA laboratories.

INTRODUCTION

Primary immunodeficiency disorders (IDs) impair the function of the immune system. Patients suffering these intrinsic defects have increased susceptibility to recurrent and persistent infections, but other symptoms are also common. More than 100 IDs affecting the immune system are known and several genes have been identified (1, 2). The immune system consists of a very large number of molecules and processes, and IDs can therefore be caused by genetic alterations at many loci (2, 3). A particular ID can be caused by defects in any one of several molecules that are required for certain responses, because a defect in any of the sequential steps may impair the whole system.

Early and exact diagnosis is essential for proper treatment of patients to improve the prognosis in the syndrome evolution, to minimize the expenses related to the treatment, to optimize health care efforts, and for genetic counseling. There is often also a need for prenatal diagnosis. In the case of certain IDs early diagnosis is crucial for lifesaving treatment.

IDs can be difficult to diagnose because similar symptoms can be caused by several disorders. A further complication is that many IDs are very rare. The incidences for IDs vary from 1 in 100 to less than 1 in 2,000,000, and for many of them the prevalence is below 1 in 100,000. Recently, diagnostic guidelines for some common IDs have been published (4). In many cases the definitive diagnosis is based on the analysis of a genetic a defect(s), because mutation detection is the most reliable method of making a diagnosis. Due to the rareness of several IDs there are generally not many centers analyzing a certain disease, and therefore physicians may have difficulties in finding laboratories that could analyze the genetic defects. New genes related to IDs are described frequently. It is difficult even for researchers within the field to keep track of all new inventions.

The internet contains plenty of medical information about IDs. We provide a validated knowledge base of practically all aspects of IDs in ImmunoDeficiency Resource (IDR) at http://bioinf.uta.fi/idr/. IDR aims at providing comprehensive integrated knowledge on ID in an easily accessible format offering data for clinical, biochemical,

[1] Institute of Medical Technology, FIN-33014 University of Tampere, Finland.

[2] Research Unit, Tampere University Hospital, FIN-33520 Tampere, Finland.

[3] To whom correspondence should be addressed at Institute of Medical Technology, FIN-33014 University of Tampere, Finland. Fax: +358-3-2157710; e-mail: Mauno.Vihinen@uta.fi.

0271-9142/04/0100-0053/0 © 2004 Plenum Publishing Corporation

genetic, structural, and computational analysis (5,6). In addition, we maintain some 50 ID mutation databases available on the Internet (3, 5, 7–9).

Here we describe IDdiagnostics, a new Internet-based service of registries for laboratories performing clinical and genetic tests for patients with heritable IDs. The database is designed for physicians, researchers, and other health care professionals. In addition to providing contact information, the registry also aims at promoting the appropriate use of genetic services inpatient care as well as increasing the general awareness of IDs.

IDdiagnostics

The aim of the IDdiagnostics registry (http://bioinf.uta.fi/IDdiagnostics/) is to collect, identify, describe, and disseminate information on diagnostic services for IDs. IDdiagnostics is formed of two independent registries for laboratories performing genetic and clinical tests, respectively, for IDs. The service is intended for physicians, researchers, and other medical genetics health professionals. Laboratories are included in IDdiagnostics on a voluntary basis. We have contacted numerous persons from laboratories performing these tests worldwide. There are in many countries contact persons who are in touch with local centers. The registry is not complete, which means that it does not include all possible laboratories or services. Only those willing to have their information posted on the Internet are included. Inclusion is at no charge. To be included, a completed registration form available in both electronic and paper form should be submitted.

According to the guidelines of the services, laboratories are regularly contacted to verify the accuracy of their information. The curators keep the right to remove information for a laboratory if there are problems, e.g., with time schedule or quality of information.

Submission

Data can be submitted to the registry in different ways. The preferred method is Web submission at http://bioinf.uta.fi/cgi-bin/submit/IDClini.cgi. The forms are easy to use and usually require just clicking the correct options from forms. Printable forms are also available (Figs. 1 and 2), which after completion can be mailed or faxed to the database curators. The paper and electronic forms are identical. If the electronic form is used, submission script formats the information from the filled form such that it can be added to a database in a similar fashion as in the MUTbase system for immunodeficiency mutation databases (10). New entries are added to databases only after curatorial review. The program allows the user to check the information in the database format before final submission by pressing the test button. The submissions are sent by email to the curators. An example of an entry in the genetic test database is shown in Fig. 3.

The submission questionnaires were constructed after consulting with prominent clinicians in the field. The genetic test form includes contact information, details about the disease investigated and applied methods, and also information about the frequency of the analysis and price details. The clinical test questionnaire contains contact information and tests performed for immune status, divided into complete blood count, quantitative serum immunoglobulins, enumeration of blood cell populations, and evaluation of functional molecules, and tests performed for studies of function: antibody response, cell function, enzyme assays, complement function, apoptosis assays, and others.

Database Management

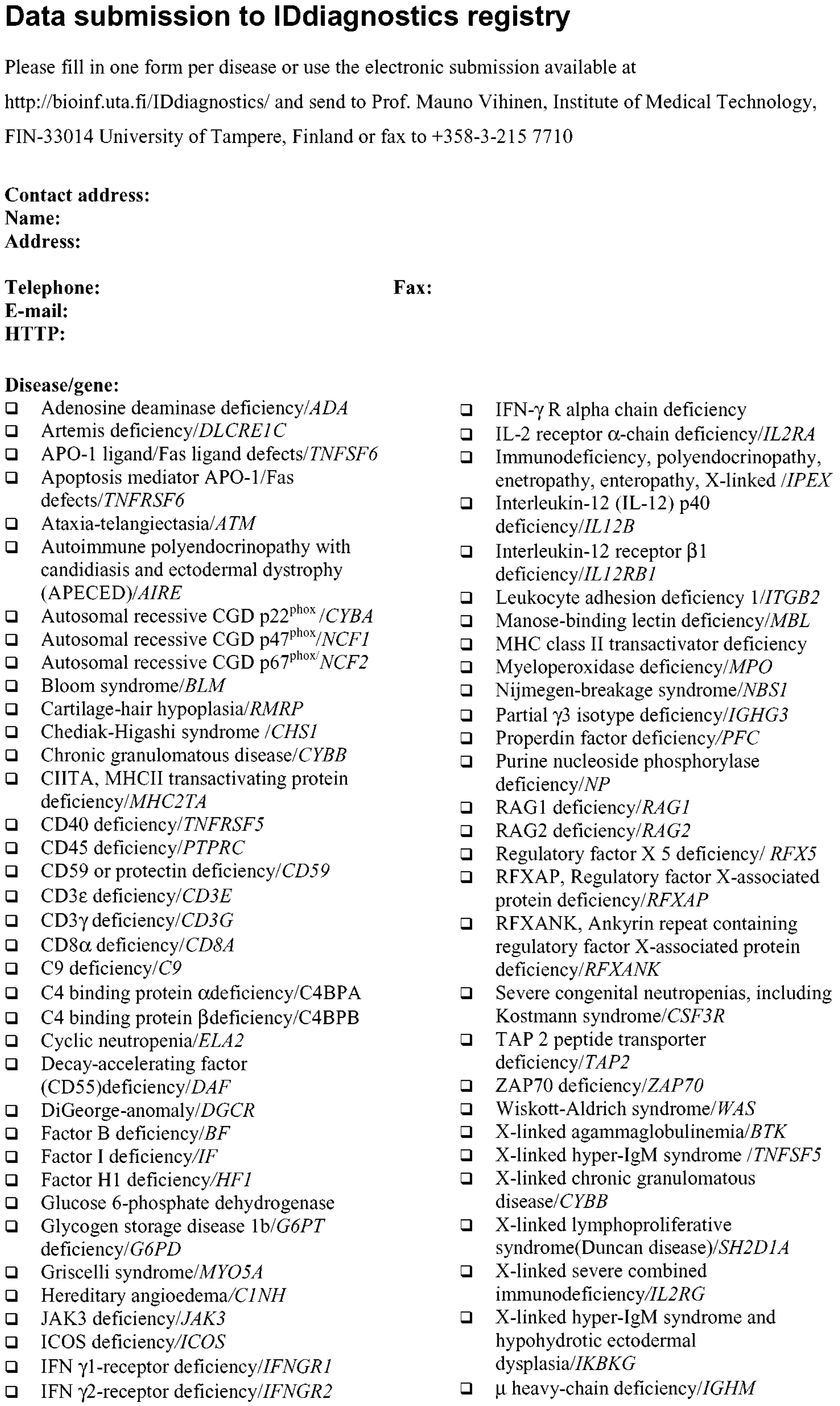
The IDdiagnostics registry is used and maintained on the Internet to provide up-to-date information at all times. Submission forms are programmed using CGI (common gateway interface) scripts. One of the key features of the Internet is that it allows different Internet-accessible databases to be connected to one another through hypertext links (11). These hyperlinks allow users to interactively retrieve related information from different Internet databases. However, there are some problems associated with hyperlinks (12). A large-scale integrative analysis needs more powerful tools for database integration and interoperation (13).

To address these problems, IDdiagnostics is based on eXtensible Markup Language (XML). Because of its simplicity and self-describing nature, XML has been proposed as a standard for exchanging data over the Internet. We created a data model, which follows W3C DOM (The World Wide Web Consortium Document Object Model, http://www.w3.org/DOM/) specification.

In addition to the Internet, IDdiagnostics can also be utilized with WAP (Wireless Application Protocol)-compliant devices such as mobile phones by using the BioWAP Service (14, 15). The IDdiagnostics system consists of electronic submission forms, Web-based maintenance routines, automated generation of the distribution version, a versatile search engine, and updated links to related services.

Contents of IDdiagnostics

IDdiagnostics is a resource that includes data for genetic and clinical tests on IDs worldwide. Currently in



Data submission to IDdiagnostics registry

Please fill in one form per disease or use the electronic submission available at http://bioinf.uta.fi/IDdiagnostics/ and send to Prof. Mauno Vihinen, Institute of Medical Technology, FIN-33014 University of Tampere, Finland or fax to +358-3-215 7710

Contact address:
Name:
Address:

Telephone: **Fax:**
E-mail:
HTTP:

Disease/gene:

- □ Adenosine deaminase deficiency/*ADA*
- □ Artemis deficiency/*DLCRE1C*
- □ APO-1 ligand/Fas ligand defects/*TNFSF6*
- □ Apoptosis mediator APO-1/Fas defects/*TNFRSF6*
- □ Ataxia-telangiectasia/*ATM*
- □ Autoimmune polyendocrinopathy with candidiasis and ectodermal dystrophy (APECED)/*AIRE*
- □ Autosomal recessive CGD p22^{phox} /*CYBA*
- □ Autosomal recessive CGD p47^{phox}/*NCF1*
- □ Autosomal recessive CGD p67$^{phox/}$*NCF2*
- □ Bloom syndrome/*BLM*
- □ Cartilage-hair hypoplasia/*RMRP*
- □ Chediak-Higashi syndrome /*CHS1*
- □ Chronic granulomatous disease/*CYBB*
- □ CIITA, MHCII transactivating protein deficiency/*MHC2TA*
- □ CD40 deficiency/*TNFRSF5*
- □ CD45 deficiency/*PTPRC*
- □ CD59 or protectin deficiency/*CD59*
- □ CD3ε deficiency/*CD3E*
- □ CD3γ deficiency/*CD3G*
- □ CD8α deficiency/*CD8A*
- □ C9 deficiency/*C9*
- □ C4 binding protein αdeficiency/C4BPA
- □ C4 binding protein βdeficiency/C4BPB
- □ Cyclic neutropenia/*ELA2*
- □ Decay-accelerating factor (CD55)deficiency/*DAF*
- □ DiGeorge-anomaly/*DGCR*
- □ Factor B deficiency/*BF*
- □ Factor I deficiency/*IF*
- □ Factor H1 deficiency/*HF1*
- □ Glucose 6-phosphate dehydrogenase
- □ Glycogen storage disease 1b/*G6PT* deficiency/*G6PD*
- □ Griscelli syndrome/*MYO5A*
- □ Hereditary angioedema/*C1NH*
- □ JAK3 deficiency/*JAK3*
- □ ICOS deficiency/*ICOS*
- □ IFN γ1-receptor deficiency/*IFNGR1*
- □ IFN γ2-receptor deficiency/*IFNGR2*
- □ IFN-γ R alpha chain deficiency
- □ IL-2 receptor α-chain deficiency/*IL2RA*
- □ Immunodeficiency, polyendocrinopathy, enetropathy, enteropathy, X-linked /*IPEX*
- □ Interleukin-12 (IL-12) p40 deficiency/*IL12B*
- □ Interleukin-12 receptor β1 deficiency/*IL12RB1*
- □ Leukocyte adhesion deficiency 1/*ITGB2*
- □ Manose-binding lectin deficiency/*MBL*
- □ MHC class II transactivator deficiency
- □ Myeloperoxidase deficiency/*MPO*
- □ Nijmegen-breakage syndrome/*NBS1*
- □ Partial γ3 isotype deficiency/*IGHG3*
- □ Properdin factor deficiency/*PFC*
- □ Purine nucleoside phosphorylase deficiency/*NP*
- □ RAG1 deficiency/*RAG1*
- □ RAG2 deficiency/*RAG2*
- □ Regulatory factor X 5 deficiency/ *RFX5*
- □ RFXAP, Regulatory factor X-associated protein deficiency/*RFXAP*
- □ RFXANK, Ankyrin repeat containing regulatory factor X-associated protein deficiency/*RFXANK*
- □ Severe congenital neutropenias, including Kostmann syndrome/*CSF3R*
- □ TAP 2 peptide transporter deficiency/*TAP2*
- □ ZAP70 deficiency/*ZAP70*
- □ Wiskott-Aldrich syndrome/*WAS*
- □ X-linked agammaglobulinemia/*BTK*
- □ X-linked hyper-IgM syndrome /*TNFSF5*
- □ X-linked chronic granulomatous disease/*CYBB*
- □ X-linked lymphoproliferative syndrome(Duncan disease)/*SH2D1A*
- □ X-linked severe combined immunodeficiency/*IL2RG*
- □ X-linked hyper-IgM syndrome and hypohydrotic ectodermal dysplasia/*IKBKG*
- □ μ heavy-chain deficiency/*IGHM*

Fig. 1. Submission page for IDdiagnostics. Laboratories performing genetic tests are requested to fill out one form per disease and submit it to the registry.

□ κ light-chain deficiency/*IGKC*
□ λ5 surrogate light-chain deficiency/*IGLL1*
□ γ1 isotype deficiency/*IGHG1*
□ γ2 isotype deficiency/*IGHG2*
□ γ4 isotype deficiency/*IGHG4*
□ α1 isotype deficiency/*IGHA1*
□ α2 isotype deficiency/*IGHA2*
□ ε isotype deficiency/*IGHE*
Other, specify:

Method:
□ Direct sequencing
□ SSCP
□ PTT
□ DGGE
□ CCM
□ EMC
□ DHPLC
□ Dideoxy finger printing
□ FISH
□ Heteroduplex analysis
Other, specify:

How often analysis is performed?

Turnaround time?

How many samples of this disease analysed/year

Fig. 1. (*Continued.*)

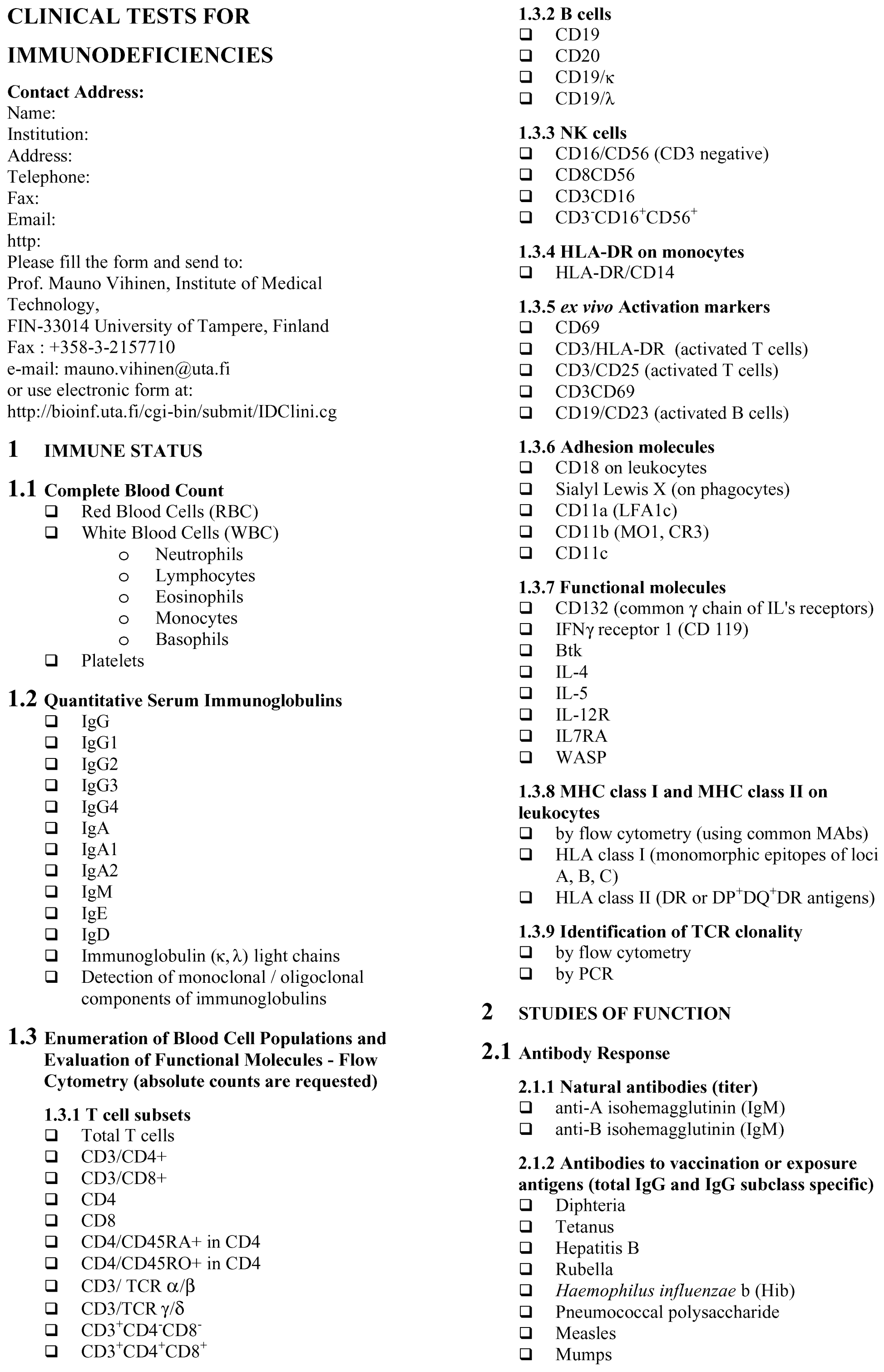
gene tests there are 79 entries for 30 diseases from 22 centers in nine countries. In the clinical test database there are 21 entries from 20 centers in 10 countries (Tables I and II). Because many of the IDs are so rare it is wise to perform analyses for certain diseases in only a few laboratories due to, e.g., laboratory and reagent costs. IDdiagnostics also provides details related to the analyses. The number of laboratories analyzing each defect varies from one to eight. The highest number of centers is available for the diagnosis of X-linked agammaglobulinemia (8), Wiskott–Aldrich syndrome (5), RAG1 (6), and RAG2 (6) deficiencies.

Mutation detection can be done by several methods (16). Many laboratories use direct sequencing, whereas others rely on fast screening methods followed by sequencing. Only with sequencing is it possible to analyze all the mutations, however, the majority of the ID-related genes are so large that primary scanning for the gene defect within an exon or intron can be more cost- and time-effective.

Since IDs are rare disorders, only big centers may have enough samples to run certain analyses on a constant basis. To get an idea about the time required for diagnosis, data are provided for details on the turnaround time, how often the samples are run, and how many samples are studied annually. The cost of the analyses varies depending, e.g., on the method used, the type of laboratory, and research interest in a particular disease. More common diseases are analyzed routinely and their diagnosis may bear substantial cost. Laboratories can have special requirements for samples and for their handling. Also, these data are available in the IDdiagnostics service. Physicians ought to consult the gene test laboratories before sending any samples to learn about conditions, shipment details, and, e.g., expected time for the test.

IDdiagnostics databases provide search facilities which allow users to run text-based search queries. IDdiagnostics searching is easy by using free text search or by limiting the search to certain data fields. Gene test laboratories can be searched by disease name (also alternative names), gene symbol (17), OMIM (On-Line Mendelian Inheritance in Man) (18) code, laboratory, laboratory location, and free text. Similar searches are available for clinical laboratories. The search engine makes it easy to find laboratories for certain diseases, methods, and geographical locations.

The definitive diagnostics of IDs depends on genetic and laboratory tests since the physical signs may be nonspecific, very discreet, or absent. As a first step in diagnosis,



CLINICAL TESTS FOR IMMUNODEFICIENCIES

Contact Address:
Name:
Institution:
Address:
Telephone:
Fax:
Email:
http:
Please fill the form and send to:
Prof. Mauno Vihinen, Institute of Medical Technology,
FIN-33014 University of Tampere, Finland
Fax : +358-3-2157710
e-mail: mauno.vihinen@uta.fi
or use electronic form at:
http://bioinf.uta.fi/cgi-bin/submit/IDClini.cg

1 IMMUNE STATUS

1.1 Complete Blood Count

- □ Red Blood Cells (RBC)
- □ White Blood Cells (WBC)
  - ○ Neutrophils
  - ○ Lymphocytes
  - ○ Eosinophils
  - ○ Monocytes
  - ○ Basophils
- □ Platelets

1.2 Quantitative Serum Immunoglobulins

- □ IgG
- □ IgG1
- □ IgG2
- □ IgG3
- □ IgG4
- □ IgA
- □ IgA1
- □ IgA2
- □ IgM
- □ IgE
- □ IgD
- □ Immunoglobulin (κ, λ) light chains
- □ Detection of monoclonal / oligoclonal components of immunoglobulins

1.3 Enumeration of Blood Cell Populations and Evaluation of Functional Molecules - Flow Cytometry (absolute counts are requested)

1.3.1 T cell subsets

- □ Total T cells
- □ CD3/CD4+
- □ CD3/CD8+
- □ CD4
- □ CD8
- □ CD4/CD45RA+ in CD4
- □ CD4/CD45RO+ in CD4
- □ CD3/ TCR α/β
- □ CD3/TCR γ/δ
- □ $CD3^+CD4^-CD8^-$
- □ $CD3^+CD4^+CD8^+$

1.3.2 B cells

- □ CD19
- □ CD20
- □ CD19/κ
- □ CD19/λ

1.3.3 NK cells

- □ CD16/CD56 (CD3 negative)
- □ CD8CD56
- □ CD3CD16
- □ $CD3^-CD16^+CD56^+$

1.3.4 HLA-DR on monocytes

- □ HLA-DR/CD14

1.3.5 *ex vivo* Activation markers

- □ CD69
- □ CD3/HLA-DR (activated T cells)
- □ CD3/CD25 (activated T cells)
- □ CD3CD69
- □ CD19/CD23 (activated B cells)

1.3.6 Adhesion molecules

- □ CD18 on leukocytes
- □ Sialyl Lewis X (on phagocytes)
- □ CD11a (LFA1c)
- □ CD11b (MO1, CR3)
- □ CD11c

1.3.7 Functional molecules

- □ CD132 (common γ chain of IL's receptors)
- □ IFNγ receptor 1 (CD 119)
- □ Btk
- □ IL-4
- □ IL-5
- □ IL-12R
- □ IL7RA
- □ WASP

1.3.8 MHC class I and MHC class II on leukocytes

- □ by flow cytometry (using common MAbs)
- □ HLA class I (monomorphic epitopes of loci A, B, C)
- □ HLA class II (DR or DP^+DQ^+DR antigens)

1.3.9 Identification of TCR clonality

- □ by flow cytometry
- □ by PCR

2 STUDIES OF FUNCTION

2.1 Antibody Response

2.1.1 Natural antibodies (titer)

- □ anti-A isohemagglutinin (IgM)
- □ anti-B isohemagglutinin (IgM)

2.1.2 Antibodies to vaccination or exposure antigens (total IgG and IgG subclass specific)

- □ Diphteria
- □ Tetanus
- □ Hepatitis B
- □ Rubella
- □ *Haemophilus influenzae* b (Hib)
- □ Pneumococcal polysaccharide
- □ Measles
- □ Mumps

Fig. 2. Submission page for IDdiagnostics.

- □ Rubella
- □ Poliomyelitis
- □ *Varicella-zoster*
- □ Other antiviral antibodies, please specify

2.1.3 Antibodies to neoantigens

- □ Keyhole limpet hemocyanin (KLH)
- □ Bacteriophage Φ X174

2.2 Cell Function

2.2.1 Study of in *vitro* lymphocyte activation and proliferation by flow cytometry upon stimulation with mitogens or antigens

- □ Surface activation markers/molecules on T cells
- □ CD69
- □ CD25
- □ CD71
- □ HLA-DR
- □ CD45RO
- □ CD45RA
- □ CD154 (CD40 ligand)
- □ IL-12 receptor
- □ Cell (intra-cellular detection) cytokine production
- □ Blastogenesis (forward light scatter FS/side light scatter SS)
- □ Cell proliferation (DNA content, BrDU incorporation)
- □ Cell proliferation (tritiated thymidine incorporation)

2.2.2 *In vitro* studies of phagocytic cell metabolism

- □ Oxidative burst (oxidation of dihydrorodamina-123 DHR stimulation with PMA or with *E.coli*) by flow cytometry
- □ Phagocytosis of (fluoresceinated) *E. coli* by flow cytometry
- □ Nitroblue tetrazolium (NBT) test
- □ Glucose metabolism
- □ Oxygen consumption
- □ Glucose monophosphate shunt activity
- □ Chemiluminescence

2.2.3 *In vitro* studies of phagocytic cell chemotaxis

- □ Chemotaxis under agarose
- □ Chemotaxis across a membrane filter (e.g., Boyden chamber)

2.3 Enzyme assays

- □ Adenosine deaminase (ADA)
- □ Purine nucleoside phosphorylase (PNP)
- □ Glucose-6-phosphate dehydrogenase (G6PD)
- □ Glutathione peroxidase (GSH-Px)
- □ Myeloperoxidase (MPO)

2.4 Complement function

- □ CH50 (Hemolytic activity, classical pathway)
- □ AH50 (Hemolytic activity, alternate pathway)
- □ Function of C1 esterase inhibitor
- □ Assay of components
  - ○ C1
  - ○ C1 esterase inhibitor
  - ○ C2
  - ○ C3
  - ○ C4
  - ○ C5
  - ○ C6
  - ○ C7
  - ○ C8
  - ○ C9
  - ○ B
  - ○ D
  - ○ Properdin
  - ○ Other, please specify

2.5 Apoptosis assays

- □ Fas/Fas-ligand (flow cytometry, other)
- □ BCL-1
- □ Caspase expression
- □ Cells with reduced DNA content (fluorescent DNA stain with cell permeabilization, flow cytometry)
- □ Strand breaks in cell DNA (TdT incorporation of fluorescence labelled nucleotides, e.g., TUNEL, flow cytometry)
- □ Translocated phosphatidylserine on cell membrane (adsorption of Annexin V, flow cytometry)
- □ functional assays of apoptosis

2.6 Other diagnostic tests

- □ Sweat Cl^-
- □ α1-antitrypsin
- □ Uric acid (serum)
- □ α -fetoprotein

3 COMMENTS AND ADDITIONAL TESTS

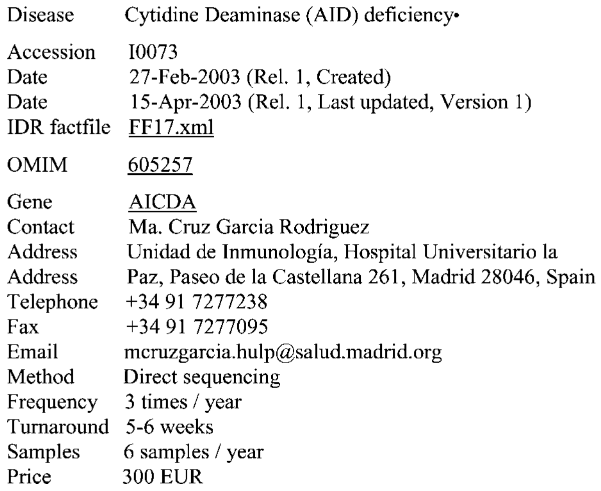
Fig. 2. (*Continued.*)

secondary ID and other rare nonimmunological disorders have to be excluded (2).

The laboratory tests included in IDdiagnostics are listed in Fig. 2. These tests are used to identify different abnormalities within the T and B cell systems, NK cells, phagocytes, and components of the complement systems.

When immunodeficiency is suspected the first tests are for complete blood count (RBC, WBC) with differential and platelet counts. The CBC will establish the presence of anemia, thrombocytopenia, neutropenia, or lymphopenia. Significant and persistent lymphopenia is one of the most relevant signs for primary ID.

The initial screening for B lymphocyte function is the measurement of immunoglobulins. Ig levels have to be interpreted with care because of marked alterations with age. However, concentrations of immunoglobulins cannot be used as the sole criterion for the diagnosis of primary ID, because reduced immunoglobulin levels may be due to the loss of Ig when antibody production is normal. Serum immunoglobulins (IgG, IgA, IgM, and IgE) measured by different methods including radial immunodiffusion, radioimmunoassay, and ELISA should be a part of the initial screening. Limited heterogeneity of immunoglobulins and abnormal κ/λ light-chain ratios have been observed



Disease	Cytidine Deaminase (AID) deficiency•
Accession	I0073
Date	27-Feb-2003 (Rel. 1, Created)
Date	15-Apr-2003 (Rel. 1, Last updated, Version 1)
IDR factfile	FF17.xml
OMIM	605257
Gene	AICDA
Contact	Ma. Cruz Garcia Rodriguez
Address	Unidad de Inmunología, Hospital Universitario la
Address	Paz, Paseo de la Castellana 261, Madrid 28046, Spain
Telephone	+34 91 7277238
Fax	+34 91 7277095
Email	mcruzgarcia.hulp@salud.madrid.org
Method	Direct sequencing
Frequency	3 times / year
Turnaround	5-6 weeks
Samples	6 samples / year
Price	300 EUR

Fig. 3. Example of an entry for the genetic test database.

Table I. Primary Immunodeficiencies in the IDdiagnostics Genetic Test Database

Disease	OMIM No.	Gene symbol	Number of entries
Deficiencies predominantly affecting antibody production			
X-linked agammaglobulinemia (XLA)	300300	*BTK*	8
μ-Heavy-chain gene deletions	147020	*IGLL1*	2
κ-Light-chain deficiency	147200	*IGKC*	1
Combined B and T cell immunodeficiencies			
X-linked lymphoproliferative syndrome (XLP)	308240	*SH2D1A*	6
Purine nucleoside phosphorylase deficiency	164050	*NP*	2
ZAP-70 deficiency	176947	*ZAP70*	2
CD3γ deficiency	186740	*CD3G*	1
RAG1 deficiency	179615	*RAG1*	6
RAG2 deficiency	179616	*RAG2*	6
Artemis deficiency	605988	*DCLRE1C*	1
X-linked hyper IgM syndrome (XHIM)	308230	*TNFSF5*	4
Cytidine deaminase (AID) deficiency	605257	*AICDA*	1
X-linked severe combined immunodeficiency (XSCID)	300400	*IL2RG*	6
Jak3 deficiency	600173	*JAK3*	2
MHCII deficiency (defect in CIITA)	60005	*MHC2TA*	1
MHCII deficiency (defect in RFX5)	601863	*RFX5*	1
CD4 deficiency	153390	*LCK*	1
CVID-ICOS deficiency	240500	*ICOS*	1
IL7R deficiency	146661	*IL7R*	1
Familial hemophagocytic lymphohistiocytosis	603552	*PRF1*	2
Other well-defined immunodeficiency syndromes			
Wiskott–Aldrich syndrome (WAS)	300392	*WASP*	7
Autoimmune lymphoproliferative syndrome (ALPS), defect in TNFRSF6	601859	*TNFRSF6*	2
Defects of phagocyte function			
Autosomal recessive CGD p22^{phox} deficiency	233690	*CYBA*	2
Autosomal recessive CGD p47^{phox} deficiency	233700	*NCF1*	2
Autosomal recessive CGD p67^{phox} deficiency	233710	*NCF2*	2
X-linked chronic granulomatous disease (XCGD)	306400	*CYBB*	3
Leukocyte adhesion deficiency 1 (LAD1)	600065	*ITGB2*	3
Chediak–Higashi syndrome	214500	*CHS1*	1
Griscelli syndrome	214450	*MYO5A*	1
Interferon γ-associated immunodeficiency			
IFN-γ receptor deficiency	107470	*IFNGR1*	1
Total			79

Table II. Numbers of Clinical and Genetic Tests and Laboratories, by Country in IDdiagnostics

Country	Clinical tests	Genetic tests	Number of centers
Australia	1		1
Finland	1		1
France	1	7	3
Germany		10	3
Greece	2		2
Hong Kong		5	2
Italy	3	14	5
Japan		3	2
The Netherlands		19	4
Portugal	1		
Spain	1	8	3
Sweden	2	2	4
United Kingdom	8		8
USA	1	11	4
Total	21	79	42

in ID syndromes (2). In addition to the measurement of serum immunoglobulin concentrations, the assessment of antibody function is a necessary part of humoral immunity testing. Titers of natural antibodies (anti-A isohemagglutinin and anti-B isohemagglutinin) and titers of antibodies after immunization with protein antigens (tetanus or diphtheria toxoids) and polysaccharide (e.g., pneumococcal capsular polysaccharides) are most convenient. Immunizations with live viral vaccines should be avoided whenever an ID is suspected. If immunoglobulin levels and/or antibody titers are decreased, the evaluation should proceed with more advanced tests of B lymphocyte numbers and functions. B cell enumeration is performed by assessing the percentage of lymphocytes (e.g., CD19, CD20) reacting with fluoresceinated antibodies to B cell-specific antigens as assessed by flow cytometry.

The most valuable tests in cellular ID are T cell, T subset enumeration, and NK cell enumeration (CD3, CD4, CD8, CD16, CD56) performed by flow cytometry. More specialized tests of T cell function include the assessment of lymphocyte activation and proliferation in response to stimulation with mitogens or antigens (CD69, CD25, CD71, CD45RO, CD45RA, CD40 ligand). In specialized laboratories, it is possible to measure the production of several cytokines and do other *in vitro* functional tests such as proliferation and cellular cytotoxicity assays. Cytotoxic defects are variably present in cellular ID. In some forms of severe combined ID, enzymes of the purine nucleotide synthesis pathway (adenosine deaminase, nucleoside phosphorylase) are deficient. The evaluation of phagocyte cells imposes the assessment of their number and function. The number of phagocytic cells can be detected by using a white blood cell count and differential. The assessment of phagocytic cell function requires a number of different assays. *In vitro* studies of phagocytic cell metabolism (oxidative burst, phagocytosis, glucose metabolism, oxygen consumption), directed cell movement (chemotaxis), and bactericidal activity test are usually available only in specialized laboratories. The NBT test measures the ability of phagocytic cells to respond with an oxidative burst by measuring the reduction of nitroblue tetrazolium (NBT). More recently, a method to measure the metabolic burst of activated neutrophils by flow cytometry with dihydrorhodamine-123 DHR has been determined.

The complement system can be evaluated by measuring the rate of hemolysis of sheep red cells (CH50) for the classical pathway and the rate of hemolysis of either rabbit or guinea pig cells (AH50) for the alternative pathway. If defects are seen in CH50 or AH50, individual complement components have to be analyzed.

DISCUSSION

IDdiagnostics was developed to provide health care professionals information about ID diagnostic laboratories, to increase the availability of genetic and clinical methods, and to provide accurate diagnosis for patients. IDdiagnostics is, to our knowledge, the first service for the diagnosis of IDs. There are two related services, EDDNAL (European Directory of DNA Laboratories; http://www.eddnal.com/) and GeneTests (http://www.genetests.org/), which provide related information about European and respective American genetic test centers in general. IDdiagnostics differs from these two by being focused on IDs instead of containing all diseases. The IDdiagnostics content and resultant entries were intended to systematically address the need, not met by the EDDNAL or GeneTests database, for directly applicable clinical information about genetic and clinical testing on IDs.

The number of entries in the genetic and clinical test databases is increasing constantly. The gene test database currently contains 79 entries, with the most records from The Netherlands, Italy, Spain, and Germany (see Table II). The highest numbers of tests are available for XLA, RAG2 deficiency, and Wiskott–Aldrich syndrome (see Table I).

We plan to improve the registries in the future by adding information about laboratory accreditation and participation in quality control schemes. The systems vary widely between countries. In some European countries, a laboratory can be certified by a governmental body to perform mutation analyses. In the United States laboratory testing is regulated through the Clinical Laboratory Improvements Amendments (CLIA), administrated by the Health Care Financing Administration (HCFA), the Centers for

Disease Control and Prevention (CDC), and the Food and Drug Administration (FDA).

The clinical information in IDdiagnostics is interfaced with other databases (OMIM, GeneCards, and IDR). The IDdiagnostics registry provides data that can help to find centers doing genetic and clinical tests on IDs and form the basis for future epidemiological and clinical studies related to the IDs.

The IDdiagnostics registry is also a useful tool for ID patients and their families, who can find the most recent genetic test for a specific disease. ID gene tests offer great opportunities for biomedical research and health care. They play an important role in the understanding of disease mechanisms. Studies of relationships between gene sequence variations and protein structure as well as protein expression and function are currently major aspects of biomedical research.

In health care genetic tests are performed for a number of purposes including prenatal diagnosis, newborn screening, carrier testing, diagnostic and prognostic testing, presymptomatic testing, predictive testing, and genetic counseling. Genetic screening has also become a promising tool for early detection of IDs. Once the defective genes are identified, methods can be devised to tailor treatment to the specific variation of the individual patient (pharmacogenomics) or even to attempt to repair or replace the affected gene (gene therapy) (19). In all cases of primary ID, there is a need for selected clinical tests to confirm or establish the diagnosis. Thus a database which contains information on clinical tests performed by different laboratories is a real help for clinicians and can improve medical care.

ACKNOWLEDGMENTS

We are very thankful to Drs. Gavin Spickett, Luigi D. Notarangelo, J. Castro e Melo, Teresa Español, C. I. Edvard Smith, and Alain Fischer for their help in refining the questionnaire for clinical tests. Financial support from the Finnish Academy, Medical Research Fund of the University Hospital Tampere, and European Union (QLRT-2000-01395) is gratefully acknowledged.

REFERENCES

1. Anonymous: Primary immunodeficiency diseases. Report of a WHO scientific group. Clin Exp Immunol 118 (Suppl 1):1–28, 1999
2. Ochs HD, Smith CIE, Puck JM: Primary immunodeficiency diseases: A molecular and genetic approach. New York, Oxford University Press, 1999
3. Vihinen M, Arredondo-Vega FX, Casanova J-L, Etzioni A, Giliani S, Hammarström L, Heyworth PG, Hershfield MS, Hsu AP, Lappalainen I, Lähdesmäki A, Notarangelo LD, Puck JF, Reith W, Roos D, Schumacher RF, Schwarz K, Vezzoni P, Villa A, Väliaho J, Smith CIE: Primary immunodeficiency mutation databases. Adv Genet 43:103–188, 2001
4. Conley ME, Notarangelo LD, Etzioni A: Diagnostic criteria for primary immunodeficiencies. Clin Immunol 93:190–197, 1999
5. Väliaho J, Riikonen P, Vihinen M: Novel data servers for immunodeficiencies. Immunol Rev 178:177–185, 2000
6. Väliaho J, Pusa M, Ylinen T, Vihinen M: IDR: ImmunoDeficiency resource. Nucleic Acids Res 30:232–234, 2002
7. Vihinen M, Lehväslaiho H, Cotton RGH: Immunodeficiency mutation databases. *In* Primary Immunodeficiency Diseases. A Molecular and Genetic Approach, HD Ochs, CIE Smith, JM Puck (eds). New York, Oxford University Press, 1999, pp 443–447
8. Vihinen M, Cooper MD, de Saint Basile G, Fischer A, Good RA, Hendriks RW, Kinnon C, Kwan S-P, Litman GW, Notarangelo LD, Ochs HD, Rosen FS, Vetrie D, Webster ADB, Zegers BJM, Smith CIE: BTKbase: A database of XLA causing mutations. Immunol Today 16:460–465, 1995
9. Vihinen M, Kwan S-P, Lester T, Ochs HD, Resnick I, Väliaho J, Smith CIE: Mutation of the human BTK gene coding for Bruton's tyrosine kinase in X-linked agammaglobulinemia. Hum Mutat 13:280–285, 1999
10. Riikonen P, Vihinen M: MUTbase: Maintenance and analysis of distributed mutation databases. Bioinformatics 15:852–859, 1999
11. Karp PD: Database links are a foundation for interoperability. Trends Biotechnol 14:273–279, 1996
12. Cheung KH, Deshpande AM, Tosches N, Nath S, Agrawal A, Miller P, Kumar A, Snyder M: A metadata framework for interoperating heterogeneous genome data using XML. In Proc AMIA Symp, 2001, pp. 110–114
13. Karp PD: A strategy for database interoperation. J Comput Biol 2:573–586, 1995
14. Riikonen P, Boberg J, Salakoski T, Vihinen M: BioWAP mobile Internet service for bioinformatics. Bioinformatics 17:855–856, 2001
15. Riikonen P, Boberg J, Salakoski T, Vihinen M: Mobile access to biological databases on the Internet. IEEE Trans Biomed Eng 49:1477–1479, 2002
16. Cotton RGH: Mutation Detection. New York, Oxford University Press, 1997
17. Povey S: Guidelines for human gene nomenclature. Community nomenclature: Standardized gene symbols. Genomics 79(4):463–463, 2002
18. Online Mendelian Inheritance in Man, OMIM (TM). McKusick–Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD), and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). http://www.ncbi.nlm.nih.gov/omim/
19. Telenti A, Aubert V, Spertini F: Individualising HIV treatment–Pharmacogenetics and immunogenetics. Lancet 359:722–723, 2002

Immunome Research

BioMed Central

Database

Open Access

IDR knowledge base for primary immunodeficiencies

Crina Samarghitean[1], Jouni Väliaho[1] and Mauno Vihinen*[1,2]

Address: [1]Institute of Medical Technology, FI-33014 University of Tampere, Finland and [2]Research Unit, Tampere University Hospital, FI-33520 Tampere, Finland

Email: Crina Samarghitean - crina.samarghitean@uta.fi; Jouni Väliaho - jouni.valiaho@uta.fi; Mauno Vihinen* - mauno.vihinen@uta.fi

* Corresponding author

Published: 29 March 2007

Immunome Research 2007, **3**:6 doi:10.1186/1745-7580-3-6

Received: 9 February 2007
Accepted: 29 March 2007

This article is available from: http://www.immunome-research.com/content/3/1/6

Abstract

Background: The ImmunoDeficiency Resource (IDR) is a knowledge base for the integration of the clinical, biochemical, genetic, genomic, proteomic, structural, and computational data of primary immunodeficiencies. The need for the IDR arises from the lack of structured and systematic information about primary immunodeficiencies on the Internet, and from the lack of a common platform which enables doctors, researchers, students, nurses and patients to find out validated information about these diseases.

Description: The IDR knowledge base, first released in 1999, has grown substantially. It contains information for 158 diseases, both from a clinical as well as molecular point of view. The database and the user interface have been reformatted. This new IDR release has a richer and more complete breadth, depth and scope. The service provides the most complete and up-to-date dataset. The IDR has been integrated with several internal and external databases and services. The contents of the IDR are validated and selected for different types of users (doctors, nurses, researchers and students, as well as patients and their families). The search engine has been improved and allows either a detailed or a broad search from a simple user interface.

Conclusion: The IDR is the first knowledge base specifically designed to capture in a systematic and validated way both clinical and molecular information for primary immunodeficiencies. The service is freely available at http://bioinf.uta.fi/idr and is regularly updated. The IDR facilitates primary immunodeficiencies informatics and helps to parameterise *in silico* modelling of these diseases. The IDR is useful also as an advanced education tool for medical students, and physicians.

Background

Primary immunodeficiency disorders (PIDs) impair the function of the immune system. Patients with these intrinsic defects have increased susceptibility to recurrent and persistent infections, and they may also have autoimmune and cancer related symptoms. Most PIDs are rare and the diagnosed patients for a condition are often randomly spread out around the world. More than 150 PIDs affecting the immune system have been described and more than 100 genes involved in PIDs have been identified [1]. The number of mutations, identified in unrelated families with different PIDs, totals over 4,500 [2].

There is plenty of information related to immunology and immunodeficiencies on the Internet. General immunome information can be found e.g. from IMGT [3], AntiJen [4] and Immunome [5] databases and more specific data e.g. in ImmTree [6], IDbases [2], SYFPEITHI [7], and Immune

Epitope Databases and Analyse Resources (IEDB) [8]. The scattering of the disease-related information in the literature and the Internet is a big obstacle for those interested in rare diseases. Users often have problems in finding relevant information and assessing the quality of information from the Internet. Biomedical information holds promises for developing informatics methods for postgenomic and personalised medicine. The new knowledge can be applied in the prevention, diagnosis and treatment of diseases. Computerised information sources have many challenges related, for example, to terminology and ontology building, information extraction from texts, knowledge discovery from collections of documents, sharing and integrating knowledge from factual and textual databases, and semantic annotation. There is a need for a standardised nomenclature and data form that can be easily handled by computers and presented on any platform.

The ImmunoDeficiency Resource (IDR) integrates biomedical information related to PIDs into a web accessible knowledge base. The fact files, which form the core of the system, integrate biomedical knowledge from several heterogeneous and autonomous sources.

This paper illustrates numerous new features and improvements, which have been implemented since previous IDR releases [9,10], and details about data collection and automated database integration. The IDR is developed to serve anybody interested in PIDs and to provide relevant, up-to-date and validated information in an easily understandable and usable format.

Construction and Content

The IDR has been designed and implemented using eXtensible Markup Language XML [11], a system comprising a native XML database and an XML server. Data within the IDR is structured into document-centred XML and SHTML files. The interface to the IDR has been completely redesigned. It consists of a dynamic layout that can adapt to different screen sizes, from wide desktop screens to small mobile devices.

Numerous new features, such as the classification of PIDs, genes related to immunodeficiency, reference sequences, protein structures and animal model pages, have been added. Links are also provided to other IDR-fact file databases [12], IDbases for PID-causing mutations [2], and IDdiagnostics for PID diagnostic laboratories [13].

The IDR aims to provide comprehensive integrated knowledge about immunodeficiencies in an easily accessible form, targeting different types of users (doctors, scientists, nurses and patients and their families). The resource includes clinical, biochemical, genetic, structural and computational data and analyses. The main headings of the IDR are General Information, Bioinformatics, Immunology, and Interest Groups (Fig. 1).

The General Information class

Immunodeficiencies in the IDR are classified according to the molecular defects criteria [1] with links to the Online Mendelian Inheritance in Man (OMIM) database [14]. Information about the affected genes and loci are provided and linked with corresponding services. The ESID and PAGID recommendations for diagnostic criteria [15], the American Academy of Allergy, Asthma and Immunology (AAAAI) parameters [16], and different diagnostic guidelines [17] are also included. There is also a list for PID abbreviations.

At the core of the system are fact files (Fig. 2), which store information regarding disorders, genes, mutations, protein sequences, online resources, organisations and associations [12]. At present there are fact files for 158 diseases. The user interface allows fast access to the information. International Classification of Diseases (ICD) codes [18] are provided for those diseases where the codes are available. Each fact file provides basic information about the disease and the affected gene. The fact files have hyperlinks to other reliable Internet resources. The fact file data model and the Inherited Disease Markup Language (IDML) [12] were developed to facilitate disease information integration, storage and exchange. The fact files make use of the following specifications, standards and databases: HUGO nomenclature [19], Swiss-Prot [20], GeneCard [21], and SOURCE [22].

The IDML fact files have been generated for each PID. The major concepts in the fact files are general information, clinical information, molecular biology and other resources – all of which are linked to related information services (Fig. 1). Each of these elements comprises one or more additional levels. Table 1 summarised the major concepts and descriptions of the elements in the IDR-fact files.

The IDML schema [23], IDML document type definition file [24], examples of an IDML-document, and documentation on the syntax can be read from [25]. The validation in IDML fact files is done with the IDML validator program, available online at [26].

The Bioinformatics class

The bioinformatics section integrates numerous Web based services (Ensemble [27], Source [22], EntrezGene [28], euGenes [29], GeneLynx [30], UniGene [31], GeneCard [21], GenAtlas [32]). Reference sequences for PID genes are available for DNA and RNA from the EMBL database, and for protein data from SwissProt [20]. When

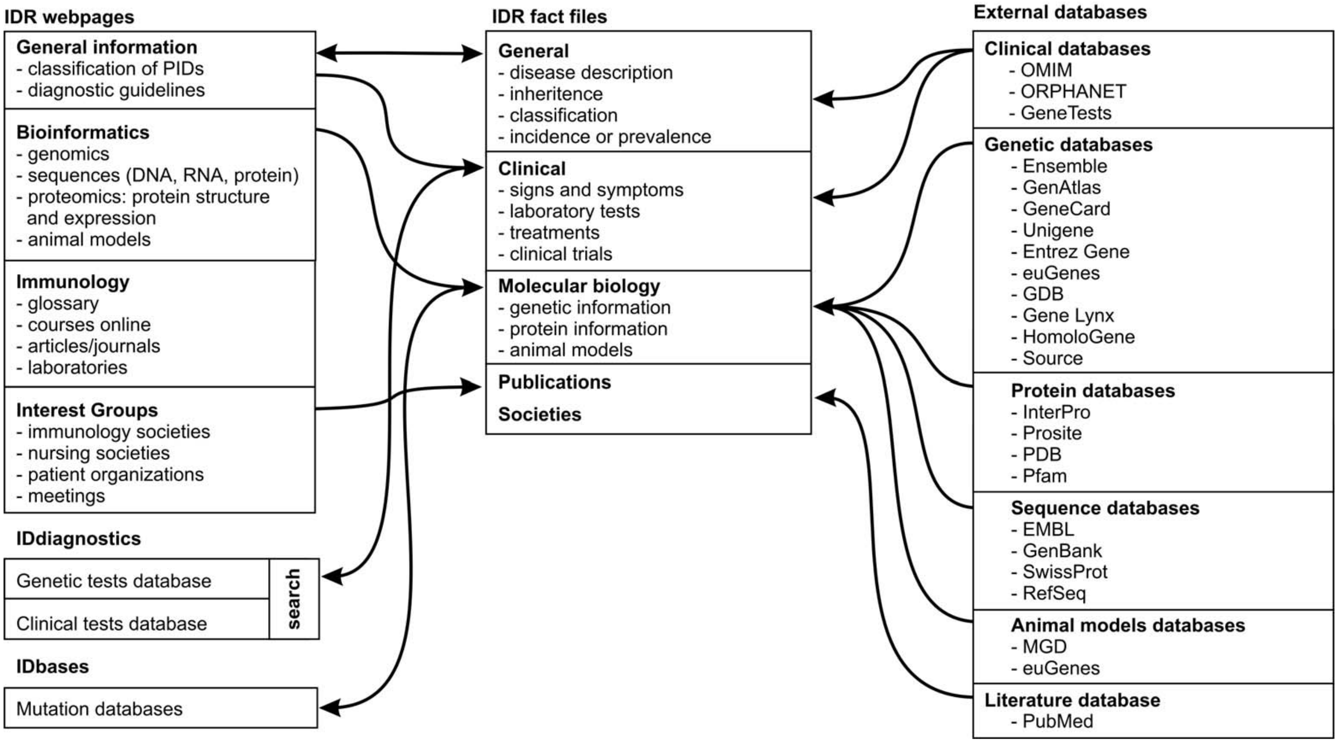


Figure 1

Concept map for the IDR knowledge service. IDR is composed of web pages grouped in to different class categories and a fact file database. The system is integrated with different internal and external databases to serve a wide category of users.

available, there are links to the protein structures and visualisation tools in the PDB [33]. The animal models page has been updated.

The IDbases [2] section provides, in addition to our own mutation registries, links to other IDbases. At the moment we have 115 databases with over 4,500 patient entries.

The Immunology and Interest Group classes

The immunology section lists collections of immunology related data sources including lectures on immunology and immunodeficiencies, and links to over 40 online immunology journals. A new feature is the glossary, which provides explanations for more than 800 immunology terms. Glossary terms are cross linked by each other, so by clicking one of these terms, such as 'antigen', not only is the explanation for the term provided, but also for a group of terms related to 'antigen'. This gives a broad overview of immunology terminology, which also makes it a useful tool for education.

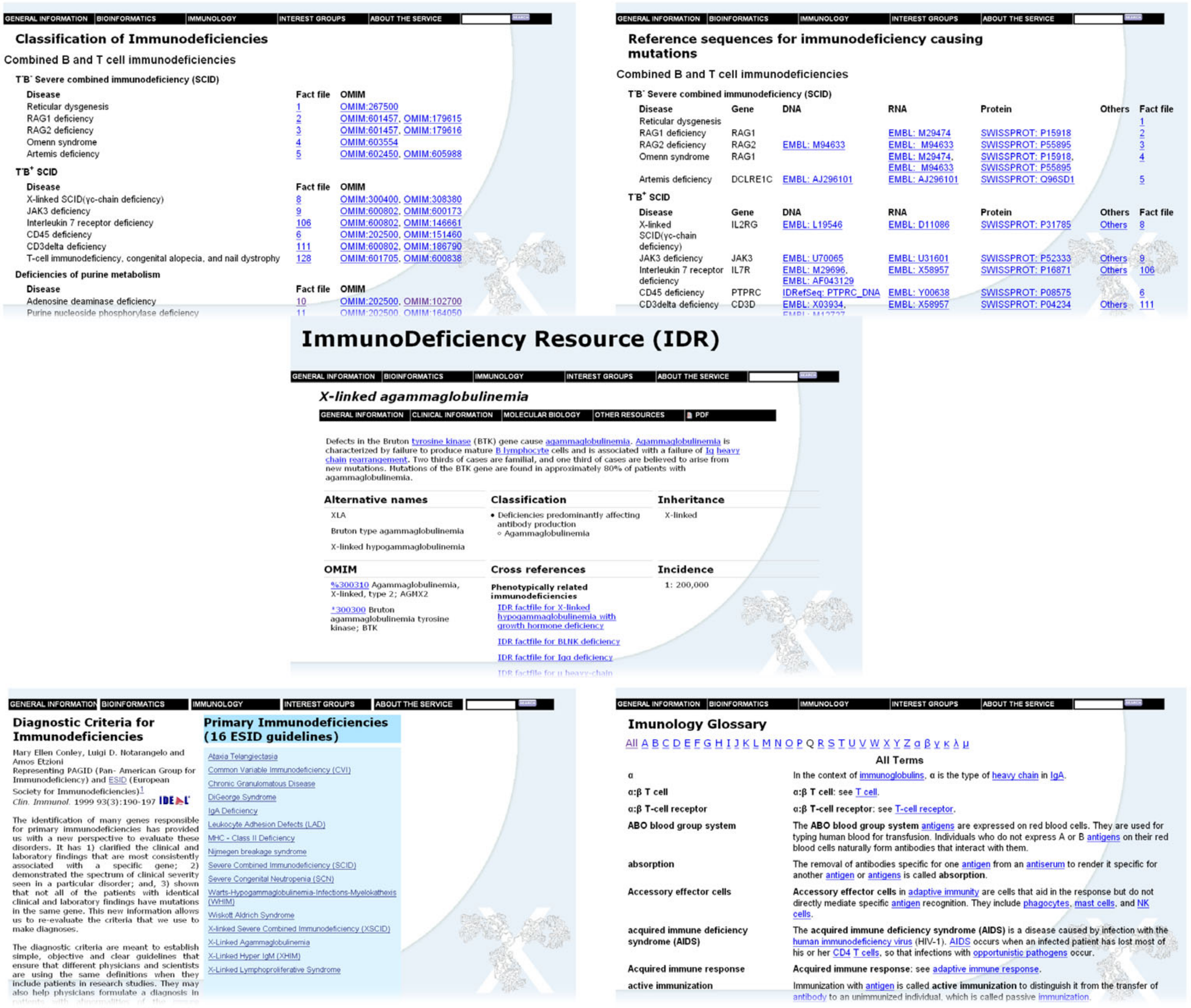
The interest group section contains links to immunology, immunodeficiency, and nursing and patient organisations. Several societies are related to immunodeficiency research, care and patients. The list of meetings and workshops is continuously updated.

Utility and Discussion

Accuracy and validation of data

The Internet contains a large number of pages. Search engines often give thousands of links but usually the most difficult task is to differentiate the useful and reliable data from other search results. In the IDR, the experts check all the data and approve only those sites with solid scientific and medical information. There will be at least one external expert for each immunodeficiency. Nursing and patient societies are also involved in the data validation process for their own interest groups.

The IDR is easy to navigate. The pages are colour coded for different interest groups: researchers, physicians, nurses, patients and families. By selecting the group of interest, the user can get specific pages produced and tailored for the particular group. This makes it easier for the user to find the most relevant and useful information. The IDR also provides an advanced text search facility, which can utilise Boolean logic searches with multiple keywords. Within a typical search, the user-entered search criteria are



Classification of Immunodeficiencies

Combined B and T cell immunodeficiencies

T⁻B⁻ Severe combined immunodeficiency (SCID)

Disease	Fact file	OMIM
Reticular dysgenesis	1	OMIM:267500
RAG1 deficiency	2	OMIM:601457, OMIM:179615
RAG2 deficiency	3	OMIM:601457, OMIM:179616
Omenn syndrome	4	OMIM:603554
Artemis deficiency	5	OMIM:602450, OMIM:605988

T⁻B⁺ SCID

Disease	Fact file	OMIM
X-linked SCID(γc-chain deficiency)	8	OMIM:300400, OMIM:308380
JAK3 deficiency	9	OMIM:600802, OMIM:600173
Interleukin 7 receptor deficiency	106	OMIM:600802, OMIM:146661
CD45 deficiency	6	OMIM:202500, OMIM:151460
CD3delta deficiency	111	OMIM:600802, OMIM:186790
T-cell immunodeficiency, congenital alopecia, and nail dystrophy	128	OMIM:601705, OMIM:600838

Deficiencies of purine metabolism

Disease	Fact file	OMIM
Adenosine deaminase deficiency	10	OMIM:202500, OMIM:102700
Purine nucleoside phosphorylase deficiency	11	OMIM:202500, OMIM:164050

Reference sequences for immunodeficiency causing mutations

Combined B and T cell immunodeficiencies

T⁻B⁻ Severe combined immunodeficiency (SCID)

Disease	Gene	DNA	RNA	Protein	Others	Fact file
Reticular dysgenesis						1
RAG1 deficiency	RAG1		EMBL: M29474	SWISSPROT: P15918		2
RAG2 deficiency	RAG2	EMBL: M94633	EMBL: M94633	SWISSPROT: P55895		3
Omenn syndrome	RAG1		EMBL: M29474, EMBL: M94633	SWISSPROT: P15918, SWISSPROT: P55895		4
Artemis deficiency	DCLRE1C	EMBL: AJ296101	EMBL: AJ296101	SWISSPROT: Q96SD1		5

T⁻B⁺ SCID

Disease	Gene	DNA	RNA	Protein	Others	Fact file
X-linked SCID(γc-chain deficiency)	IL2RG	EMBL: L19546	EMBL: D11086	SWISSPROT: P31785	Others	8
JAK3 deficiency	JAK3	EMBL: U70065	EMBL: U31601	SWISSPROT: P52333	Others	9
Interleukin 7 receptor deficiency	IL7R	EMBL: M29696, EMBL: AF043129	EMBL: X58957	SWISSPROT: P16871	Others	106
CD45 deficiency	PTPRC	IDRefSeq: PTPRC_DNA	EMBL: Y00638	SWISSPROT: P08575		6
CD3delta deficiency	CD3D	EMBL: X03934, EMBL: M12727	EMBL: X58957	SWISSPROT: P04234	Others	111

ImmunoDeficiency Resource (IDR)

X-linked agammaglobulinemia

Defects in the Bruton tyrosine kinase (BTK) gene cause agammaglobulinemia. Agammaglobulinemia is characterized by failure to produce mature B lymphocyte cells and is associated with a failure of Ig heavy chain rearrangement. Two thirds of cases are familial, and one third of cases are believed to arise from new mutations. Mutations of the BTK gene are found in approximately 80% of patients with agammaglobulinemia.

Alternative names	Classification	Inheritance
XLA Bruton type agammaglobulinemia X-linked hypogammaglobulinemia	• Deficiencies predominantly affecting antibody production ◦ Agammaglobulinemia	X-linked

OMIM	Cross references	Incidence
%300310 Agammaglobulinemia, X-linked, type 2; AGMX2 *300300 Bruton agammaglobulinemia tyrosine kinase; BTK	**Phenotypically related immunodeficiencies** IDR factfile for X-linked hypogammaglobulinemia with growth hormone deficiency IDR factfile for BLNK deficiency IDR factfile for Iga deficiency IDR factfile for μ heavy-chain	1: 200,000

Diagnostic Criteria for Immunodeficiencies

Mary Ellen Conley, Luigi D. Notarangelo and Amos Etzioni
Representing PAGID (Pan- American Group for Immunodeficiency) and ESID (European Society for Immunodeficiencies)[1]
Clin. Immunol. 1999 93(3):190-197 IDEAL

The identification of many genes responsible for primary immunodeficiencies has provided us with a new perspective to evaluate these disorders. It has 1) clarified the clinical and laboratory findings that are most consistently associated with a specific gene; 2) demonstrated the spectrum of clinical severity seen in a particular disorder; and, 3) shown that not all of the patients with identical clinical and laboratory findings have mutations in the same gene. This new information allows us to re-evaluate the criteria that we use to make diagnoses.

The diagnostic criteria are meant to establish simple, objective and clear guidelines that ensure that different physicians and scientists are using the same definitions when they include patients in research studies. They may also help physicians formulate a diagnosis in patients with abnormalities of the immune

Primary Immunodeficiencies (16 ESID guidelines)

- Ataxia Telangiectasia
- Common Variable Immunodeficiency (CVI)
- Chronic Granulomatous Disease
- DiGeorge Syndrome
- IgA Deficiency
- Leukocyte Adhesion Defects (LAD)
- MHC - Class II Deficiency
- Nijmegen breakage syndrome
- Severe Combined Immunodeficiency (SCID)
- Severe Congenital Neutropenia (SCN)
- Warts-Hypogammaglobulinemia-Infections-Myelokathexis (WHIM)
- Wiskott Aldrich Syndrome
- X-linked Severe Combined Immunodeficiency (XSCID)
- X-Linked Agammaglobulinemia
- X-Linked Hyper IgM (XHIM)
- X-Linked Lymphoproliferative Syndrome

Imunology Glossary

All A B C D E F G H I J K L M N O P Q R S T U V W X Y Z α β γ κ λ μ

All Terms

α	In the context of immunoglobulins, α is the type of heavy chain in IgA.
α:β T cell	**α:β T cell**: see T cell.
α:β T-cell receptor	**α:β T-cell receptor**: see T-cell receptor.
ABO blood group system	The **ABO blood group system** antigens are expressed on red blood cells. They are used for typing human blood for transfusion. Individuals who do not express A or B antigens on their red blood cells naturally form antibodies that interact with them.
absorption	The removal of antibodies specific for one antigen from an antiserum to render it specific for another antigen or antigens is called **absorption**.
Accessory effector cells	**Accessory effector cells** in adaptive immunity are cells that aid in the response but do not directly mediate specific antigen recognition. They include phagocytes, mast cells, and NK cells.
acquired immune deficiency syndrome (AIDS)	The **acquired immune deficiency syndrome (AIDS)** is a disease caused by infection with the human immunodeficiency virus (HIV-1). AIDS occurs when an infected patient has lost most of his or her CD4 T cells, so that infections with opportunistic pathogens occur.
Acquired immune response	**Acquired immune response**: see adaptive immune response.
active immunization	Immunization with antigen is called **active immunization** to distinguish it from the transfer of antibody to an unimmunized individual, which is called passive immunization.

Figure 2

IDR user interface. Screenshots of the main IDR pages. The new user interface provides faster access to the information and different new features, such as classification of diseases (top left), the gene related with the PIDs and their reference sequences (top right), glossary terms for immunology (bottom right), and diagnostic tools (bottom left). At the core of the system are fact files that provide clinical and molecular information for 158 primary immunodeficiency diseases (centre).

carried from an SHTML or XML form to a category specific PERL script, which performs the database queries.

The IDR can be used to discover many kinds of information as it is integrated with internal (IDbases and IDdiagnostics) and external databases (Fig. 1).

The IDbases have recently been integrated with the ESID patient registry [34], which collects clinical data for patients. This collaboration facilitates direct submission both to the ESID registry and IDbases. A similar arrangement will be made with the US Immunodeficiency Network (USIDNET).

A fact file is a user oriented user interface, which serves as a good starting point to explore information on hereditary diseases. The user can find not only information about the disease nomenclature and classification (OMIM, ICD10), but also a clinical description of the disease, inheritance and prevalence. The IDR-fact file facilitates finding information about laboratories which perform genetic tests for PIDs, using direct links to IDdiagnostics [13], GeneTest

Table 1: Major concepts and elements in IDR-fact files

Major concepts	Elements	Description
General Information	DiseaseName	Disease name
	Abbreviation	Abbreviation for disease name
	AlternativeNames	Alternatively used disease names
	Description	General description of disease
	Classification	Classifies disease in the fact files' hierarchy
	Omim	Link to the OMIM knowledge base
	ICD-10	WHO classification of diseases
	CrossReferences	References to the related fact files
	Incidence	Number of cases in population
Clinical Information	Clinical Description	Characteristic clinical features
	Diagnosis	Diagnostic guidelines, protocols and laboratories
	TherapeuticOptions	Treatment of disease
	ResearchPrograms	Clinical trials or research projects on-going
Molecular Biology	GeneInformation	Gene name, aliases, reference sequences, chromosomal location, maps, markers, variations and other gene related resources
	AnimalModels	Related transgenic animal data
	ProteinInformation	Protein features, structures, domains, motifs and other protein resources
	ExpressionPattern	Gene expression levels in a variety of cells and tissues
Other Resources	Publications	Related publications in PubMed
	Societies	General and disease specific societies
	OtherSites	Other related websites

[35] or ORPHANET [36]. Information for PID related genes contains the nomenclature, including aliases, sequences, chromosomal location and maps, variations, and mutations. The IDR-fact files also contain information about protein functions, structure, domains, motifs, subcellular location and post-translational modifications.

PID researchers will also find in the IDR-fact files lots of information for each disease to keep up-to-date with the literature, meetings and different associations in the field.

Future work

In the near future a new tool for faster and more accurate diagnosis, PIDexpert, will be added. PIDexpert is a medical expert system, designed to give the diagnostic picture of PIDs based on symptoms, signs, medical history, physical findings, and laboratory tests. Future tasks in the development of the IDR will focus on expanding its depth, breadth, and scope. The external database updates will be monitored so that any alterations are mirrored within the archive. We plan to further develop the IDR based on user feedback and our interactions with the PID community.

Conclusion

The IDR contains systematically organised, continuously updated and validated information that is valuable for clinicians and researchers and can improve the medical care of PIDs. IDR is the first knowledge service designed to capture both clinical and molecular information about primary immunodeficiencies and to address different types of users. It is validated and will be updated frequently. The IDR facilitates PID informatics and helps to parameterise *in silico* modelling of these diseases. The IDR has many potential users throughout the PID community, from doctors to patients and from immunoinformaticians to experimental immunologists and structural biologists.

Availability and requirements

The IDR database is freely available for academic use from the URL: http://bioinf.uta.fi/idr.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

CS carried out the data mining analysis, database development and drafted the manuscript. JV participated in database development and drafted the manuscript. MV conceived the study, participated in IDR design and coordination, and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The Academy of Finland, the Medical Research Fund of Tampere University Hospital, and the European Union, all supported the research and are gratefully acknowledged. We thank Jukka Lehtiniemi for help in the implementation of the IDR and Kathryn Rannikko for language correction.

References

1. Ochs HD, Smith CIE, Puck JM: **Primary immunodeficiency diseases: A molecular and genetic approach.** 2nd edition. New York, Oxford University Press; 2006.
2. Piirilä H, Väliaho J, Vihinen M: **Immunodeficiency mutation databases (IDbases).** *Hum Mutat* 2006, **27:**1200-1208.
3. Robinson J, Waller MJ, Fail SC, Marsh SG: **The IMGT/HLA and IPD databases.** *Hum Mutat* 2006, **27:**1192-1199.
4. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwagama CK, Flower DR: **AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data.** *Immunome Res* 2005, **1:**4.
5. Ortutay C, Siermala M, Vihinen M: **Molecular characterization of the immune system: emergence of proteins, processes, and domains.** *Immunogenetics* 2007, **59(5):**333-48.
6. Ortutay C, Siermala M, Vihinen M: **ImmTree: Database of evolutionary relationships of genes and proteins in the human immune system.** *Immunome Res* 2007, **3:**4.
7. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: **SYFPEITHI: database for MHC ligands and peptide motifs.** *Immunogenetics* 1999, **50:**213-219.
8. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A: **The immune epitope database and analysis resource: from vision to blueprint.** *PLoS Biol* 2005, **3:**e91.
9. Väliaho J, Riikonen P, Vihinen M: **Novel immunodeficiency data servers.** *Immunol Rev* 2000, **178:**177-185.
10. Väliaho J, Pusa M, Ylinen T, Vihinen M: **IDR: the ImmunoDeficiency Resource.** *Nucleic Acids Res* 2002, **30:**232-234.
11. **Extensible Markup Language (XML) 1.0** *World Wide Web Consortium* [http://www.w3.org/TR/REC-xml/].
12. Väliaho J, Riikonen P, Vihinen M: **Distribution of immunodeficiency fact files with XML-from Web to WAP.** *BMC Med Inform Decis Mak* 2005, **5:**21.
13. Samarghitean C, Väliaho J, Vihinen M: **Online registry of genetic and clinical immunodeficiency diagnostic laboratories, IDdiagnostics.** *J Clin Immunol* 2004, **24:**53-61.
14. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30:**52-55.
15. Conley ME, Notarangelo LD, Etzioni A: **Diagnostic criteria for primary immunodeficiencies. Representing PAGID (Pan-American Group for Immunodeficiency) and ESID (European Society for Immunodeficiencies).** *Clin Immunol* 1999, **93:**190-197.
16. Bonilla FA, Bernstein IL, Khan DA, Ballas ZK, Chinen J, Frank MM, Kobrynski LJ, Levinson AI, Mazer B, Nelson RP Jr., Orange JS, Routes JM, Shearer WT, Sorensen RU: **Practice parameter for the diagnosis and management of primary immunodeficiency.** *Ann Allergy Asthma Immunol* 2005, **94:**S1-63.
17. de Vries E, Clinical Working Party of the European Society for Immunodeficiencies, E S I D: **Patient-centred screening for primary immunodeficiency: a multi-stage diagnostic protocol designed for non-immunologists.** *Clin Exp Immunol* 2006, **145:**204-214.
18. **International Classification of Diseases (ICD-10)** [http://www.who.int/classifications/apps/icd/icd10online/]
19. Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S: **Genew: the Human Gene Nomenclature Database, 2004 updates.** *Nucleic Acids Res* 2004, **32:**D255-7.
20. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
21. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.** *Bioinformatics* 1998, **14:**656-664.
22. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31:**219-223.
23. **IDML schema** [http://bioinf.uta.fi/idml/idml.xsd.txt.shtml]
24. **DML document type definition file** [http://bioinf.uta.fi/idml/idml.dtd.txt.shtml]
25. **IDML document** [http://bioinf.uta.fi/idml/]
26. **IDML validator** [http://bioinf.uta.fi/cgi-bin/submit/IDMLvalidator.cgi]
27. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35:**D610-7.
28. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
29. Gilbert DG: **euGenes: a eukaryote genome information system.** *Nucleic Acids Res* 2002, **30:**145-148.
30. Lenhard B, Hayes WS, Wasserman WW: **GeneLynx: a gene-centric portal to the human genome.** *Genome Res* 2001, **11:**2151-2157.
31. Schuler GD: **Pieces of the puzzle: expressed sequence tags and the catalog of human genes.** *J Mol Med* 1997, **75:**694-698.
32. Frezal J: **Genatlas database, genes and development defects.** *C R Acad Sci III* 1998, **321:**805-817.
33. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures.** *Arch Biochem Biophys* 1978, **185:**584-591.
34. Eades-Perner AM, Gathmann B, Knerr V, Guzman D, Veit D, Kindle G, Grimbacher B: **The European internet-based patient and research database for primary immunodeficiencies: results 2004-06.** *Clin Exp Immunol* 2007, **147:**306-312.
35. Pagon RA, Tarczy-Hornoch P, Baskin PK, Edwards JE, Covington ML, Espeseth M, Beahler C, Bird TD, Popovich B, Nesbitt C, Dolan C, Marymee K, Hanson NB, Neufeld-Kaiser W, Grohs GM, Kicklighter T, Abair C, Malmin A, Barclay M, Palepu RD: **GeneTests-GeneClinics: genetic testing information for a growing audience.** *Hum Mutat* 2002, **19:**501-509.
36. **Online encyclopaedia of rare diseases** [http://www.orpha.net/orphacom/cahiers/reports-orphanet.htm]

Systematic Classification of Primary Immunodeficiencies Based on Clinical, Pathological, and Laboratory Parameters[1]

Crina Samarghitean,*† Csaba Ortutay,* and Mauno Vihinen[2]*†

The classification of diseases has several important applications ranging from diagnosis and choice of treatment to demographics. To date, classifications have been successfully created manually, often within international consortia. Some groups of diseases, such as primary immunodeficiencies (PIDs), are especially hard to nosologically cluster due, on one hand, to the presence of a wide variety of disorders and, in contrast, because of overlapping characteristics. More than 200 PIDs affecting components of the innate and adaptive immune systems have been described. Clinical, pathological, and laboratory characteristics were collected and used to group PIDs. A consensus of at least five independent methods provided a novel classification of 11 groups, which revealed previously unknown features and relationships of PIDs. Comparison of the classification to independent features, including the severity and therapy of the diseases, functional classification of proteins, and network vulnerability, indicated a strong statistical support. The method can be applied to any group of diseases. ***The Journal of Immunology,*** **2009, 183: 7569–7575.**

Primary immunodeficiencies (PIDs)[3] are a large and heterogenous group of disorders that have been organized manually into different categories (1–8) sometimes without a consensus. PIDs are mainly rare hereditary disorders of the immune system that often have serious consequences (1, 2). These diseases represent a challenge in their diagnosis and treatment due to overlapping symptoms and similarities between diseases. Infections are the hallmarks of PIDs (1, 9–11). A diagnosis will often be considered when infections are frequent or severe, resistant to standard therapies, or caused by unusual (opportunistic) organisms. Other manifestations include autoimmune (12, 13) and cancer diseases (14), granulomatosis (15), hemophagocytic syndrome (16, 17), angioedema (18), autoinflammation (19–21), thrombotic microangiopathy, or predisposition to allergy.

Clinical descriptions have already been made for more than 200 PIDs (4, 7, 22), for which 167 genetic etiologies have been described. PIDs have historically been defined and classified according to immunological phenotypes (3). Before molecular analyses were widely available, PIDs were classified according to the affected immune function, as follows: Ab production (B cells), cellular immunity (T cells) or both (combined immunodeficiencies), phagocyte function (neutrophils, monocytes), and complement activation, etc. This classification has been useful for certain practical purposes, but not from the mechanistic point of view, because many PIDs do not easily fit into the scheme. On clinical grounds, immunodeficiencies can be classified into two broad groups according to whether all features are the result of the immune defect (immunodeficiency syndromes) or whether many, even prominent ones, cannot be explained by the immune defect (syndromes with immunodeficiency).

The behavior of even the most complex of systems is based on the interaction of their components. These components can be reduced to a series of nodes that are connected to each other by links, which together form a network (23). Most real networks in technological, social, and biological systems share common designs that are simple and quantifiable. In medicine, network analysis has been used to characterize, e.g., the spread of epidemics (24), to determine ways to control them (25), and to identify novel target genes for prostate cancer (26).

When networks formed from diseases, the involved genes, and their phenotypes have been investigated (27–29), only a few PIDs have been included. Information about a protein interaction network for the immunome (30) has been used together with gene ontology terms (31) to predict novel PID candidate genes (32).

In this study, our goal was to develop a systematic, mathematical classification of PIDs. Previous PID classifications have been derived from observational correlations between pathological and clinical features. The foundation of the method is the description of the diseases based on 87 clinical and laboratory parameters.

Altogether, six methods belonging to two categories were used to organize the diseases based on the characteristics. Three clustering methods were applied to the multivariate problem to form disease groups in which members are most similar to each other (33).

The other three methods are from the emerging field of community analysis of networks. A community is a set of nodes with many edges (connections) inside the community and few edges outside it. Community analysis is a powerful tool for finding groups in interconnected entities (34–36); therefore, PIDs interpreted as a network can be analyzed this way, and diseases strongly associated with each other can be identified. Our systematic approach can be applied for classifying any other disease group.

Materials and Methods

To obtain a systematic classification for PIDs, a novel approach was developed. The method applies advanced computational tools for clustering and network analysis. We used altogether six methods to group PIDs based on characteristics that they share. Three of the methods were for clustering, and three for network community analysis.

*Institute of Medical Technology, University of Tampere, Tampere, Finland; and †Tampere University Hospital, Tampere, Finland

Received for publication June 9, 2009. Accepted for publication September 23, 2009.

[1] This work was supported by the Finnish Academy, the Sigrid Juselius Foundation, the Medical Research Fund of Tampere University Hospital, and Tampere City Hall.

[2] Address correspondence and reprint requests to Dr. Mauno Vihinen, Institute of Medical Technology, FI-33014 University of Tampere, Finland. E-mail address: mauno.vihinen@uta.fi

[3] Abbreviations used in this paper: PID, primary immunodeficiency; DC, dendritic cell; ICD, International Classification of Diseases; IDR, ImmunoDeficiency Resource.

www.jimmunol.org/cgi/doi/10.4049/jimmunol.0901837

Immunodeficiencies and selection of the parameters

Data for PIDs were collected from the ImmunoDeficiency Resource (IDR) (4), IDdiagnostics (37), IDbases (38), and literature (1, 2, 5–7, 39–43). Only detailed reports with statistical information including clinical symptoms and measured laboratory values characterizing PIDs were included. When diseases without specific symptoms were excluded, there were altogether 194 PIDs left.

For each disease, all signs, symptoms, and laboratory values mentioned in the literature were collected. The initial list contained 420 parameters. Parameters characterizing only one to four diseases were omitted or merged to others.

For example, IgM lymphoma, which is according to the literature associated only with hyper-IgM syndrome type 2, was merged to the more general term of other malignancies. Similarly, encephalitis, meningoencephalitis, meningitis, conjunctivitis, iritis, episcleritis, and brain abcess were grouped to a more general term of CNS infections. Cerebellar ataxia, pathognomonic for ataxia-telangieactasia, and ataxia-telangieactasia-like disease were merged to neurological or CNS abnormalities together with other signs, such as peripheral neuropathy, speech delay, and convulsions. Finally, after iterative process, we had 87 informative parameters (supplemental Table S1).[4] All of the parameters had an equal weight in the analysis.

Cluster and network community analysis

Cluster and network analyses were performed in the R statistical environment (44) using the igraph (45) and cluster program libraries. Three different variations of *K*-means clustering were used to analyze the dataset. The Clustering Large Applications (clara) method computes a list representing the clustering of the data into *k* clusters. Partitioning Around Medoids (pam) partitions (clusters) the data into *k* clusters around medoids, which are representative objects of a dataset from which the distances to the other points in the cluster are computed. The Fuzzy Analysis Clustering (fanny) method computes a partition grouping of the data into *k* clusters. The number of clusters was chosen by maximizing the average width of the clusters. In fuzzy clustering, data elements can belong to more than one cluster, and thus, each disease has a set of membership levels, which indicate the strength of association to each cluster.

Three methods were applied to find highly interconnected parts of the network. Community structure via short random walks is a walktrap community analysis, which searches for densely connected subgraphs, i.e., communities (34). When moving from one node to a connected one, short random walks tend to stay in the same community. The second method uses community structure detection based on the leading eigenvector of the community matrix (35). The method looks for densely connected subgraphs by calculating the leading nonnegative eigenvector of the modularity matrix of the graph. The third method tries to find communities in graphs via a spin-glass model and simulated annealing (36).

Combination of clustering and network results

The data for PIDs are incomplete because in many diseases just a few, even a single, patient was known, and therefore, the most prominent signs were difficult to define. In rare diseases, some symptoms may occur frequently just by chance. Therefore, to obtain the most reliable and robust grouping and a consistent and robust view of the disease grouping patterns, a consensus classification based on the co-occurrence of the diseases in four, five, or six methods was generated. Using this approach, the results are expected to be independent from the biases of the individual methods.

Statistics

To evaluate which of the binary parameters significantly supported the identified disease clusters, we tested whether a parameter had significantly different distribution in the individual clusters compared with the entire dataset by calculating *p* values using the hypergeometric distribution. The threshold for significance used was a *p* value 0.05. A similar evaluation was performed for the binary clinical and functional properties of the diseases and respective proteins.

Correlation of disease clusters to clinical, genetic, functional, and network properties

The consensus network graph was used to visualize a number of properties for the diseases, and involved genes and proteins. Information about the prevalence of the PIDs was obtained from the IDR (4) and other sources (1, 2, 40, 46). When the prevalence was not known and only a few cases were reported in literature, it was assumed to be $<1/10^7$. Data on inheritance were retrieved from the IDR and literature (1, 2, 6, 40). For treatment modalities, the most common treatments were listed for each disease (1, 2, 39). Functional classifications of the proteins in the immunome (47) were obtained from the Immunome Knowledge Base (48).

Results

PID network

Novel PID grouping was obtained by applying altogether six methods for clustering and network analysis. Our approach revealed associations that were not previously obvious and led to the identification of distinct novel groups. Fig. 1*A* shows the results when at least five methods agreed on the clustering. The diseases are indicated by the affected genes, when known. If all six methods are required to agree, the only difference is that the 11 major groups are divided further to smaller subgroups. There is one giant cluster that contains the majority of the PIDs and some separate clusters and singleton PIDs. Some details of the classification may change in the future when more information becomes available, yet still the major features will remain. In the consensus graph for the PIDs (Fig. 1*A*), 1,285 pairs of diseases are grouped together by at least five methods of a possible 18,721, and of the 12,721 that are grouped by at least one method. The results are also available in an interactive web page at http://bioinf.uta.fi/PID_classification.

Disease clusters

The analytical approach combining six different and independent methods provides a highly robust classification when at least five of the methods were required to agree on the grouping details. The dendogram reveals 11 well-defined disease clusters (DCs) of at least 4 diseases (Fig. 1*A*). First, we analyzed the nature of the PIDs in the clusters and then investigated the properties of the diseases and the corresponding genes and proteins in the clusters.

Most of the clusters are very homogeneous and contain related diseases. An overall view of the DCs shows that clusters III and VII contain (almost) exclusively SCIDs, whereas in group IX the diseases are related to the complement system and in DCs I and XI to phagocyte functions; in DC VIII are fever syndromes, and in DC X Fanconi anemias. All of the MHC II genes are in cluster III, and all of the known MHC I diseases are in cluster V. The classical complement pathway diseases are in clusters V and IX. Diseases in cluster II are related to DNA instability and DNA damage repair, except for G6PC3. DCs IV, V, and VI are the most heterogenous. Cluster IV contains numerous receptor and signaling molecule-related diseases. Some of the proteins behind these disorders form transmembrane channels. DC V contains mainly Ab and complement deficiencies together with some SCIDs. The majority of the group VI diseases are related to phagocytosis and apoptosis. All of the groups have strong statistical support (supplemental Table S2).

These groups, which have been exclusively generated based on disease characteristics, indicate the power of the method. Other information for the PID genes, proteins, and their functions further support the classification (Fig. 1, *B–E*). Also, these results are statistically significant (supplemental Table S2).

The previous PID classifications have relied heavily on the cell types in which the disease-related genes are normally expressed. Thus, one of the major differences to these is that Ab deficiencies, combined PIDs, and diseases related to phagocytosis are widely scattered in our graph (Fig. 1). These diseases are very heterogenous in their symptoms and signs, and affect numerous parts of the immune system. The highest concentration of SCIDs is in DCs II, IV, and VII, whereas Ab deficiencies mainly appear in clusters IV and V. Phagocyte diseases are exclusive to DCs I and XI, but also appear in DCs IV, V, and VI.

[4] The online version of this article contains supplemental material.

To test whether the etiologically rather homogeneous DCs shared any similarities, the distribution of characteristics describing independent clinical, functional, genetic, and network properties of the diseases and the corresponding genes and proteins were investigated. These features were not used for the original clustering of the PIDs.

The PIDs were divided into four groups according to their severity. Severity was not among the symptoms used in our classification because it is not routinely used in the clinical description of the diseases. SCID is a pediatric emergency situation because the condition is life threatening, whereas some of the other PIDs are just mild. In fact, the vast majority of PID cases have been thought to remain undiagnosed due to mild symptoms. The severity shows a very homogeneous pattern in some of the clusters (Fig. 1*B*). Diseases in DCs II, VII, and IX belong to two categories, whereas in almost all of the remaining clusters there are almost exclusively moderate and severe or severe and life-threatening diseases.

PIDs are treated in a number of ways that can be grouped to a small number of categories, including Ig treatment, the use of antibiotics, antifungals or antivirals, immunomodulators, and the reconstitution of the immune system by (haploidentical) bone marrow transplantation or hematopoietic stem cell transplantation. The data for the treatment of the diseases in Fig. 1*C* indicate that the majority of the clusters are very homogeneous in regard to treatments and there are clear differences between DCs. There are DCs in which all of the diseases are treated with the same battery of therapeutic modalities, and in almost all the clusters some of the therapeutic options can be used for all of the coclustered PIDs. Also, based on these results, the DCs reliably reflect the properties of diseases, and the therapy applied to diseases within DCs is usually similar.

The cellular functions have been determined for all the genes and proteins required in the immunome (entirety of immune system) (49). There are functional groups for, e.g., the surface receptors in clusters of differentiation classification, chemokines, and their receptors, humoral immunity proteins, and those involved in cellular immunity, Ag processing, and transcription factors. The distribution of the functional properties in the PID classification is shown in Fig. 1*D*.

The majority of the group I proteins are involved in inflammation and cellular immunity. Group II and VII proteins have functions as transcription factors involved in humoral and cellular immunity. Humoral immunity is most prevalent in DC IV, complement proteins in DC IX, and both humoral and complement functions in DC V. Inflammation is the function involved in DC VIII. Many of the proteins have several classifications because they are typically overlapping. Also, at this level, the grouped proteins and diseases share many common properties because the functions are very homogeneous within DCs and differ between them.

Vulnerability is a systems biology measure that indicates how crucial a certain node is for the network (30). Vulnerabilities of proteins in the immunome protein interaction network were color coded compared with the average vulnerability in the entire network (Fig. 1*E*). There are some clusters in which the vulnerabilities are related, especially those in DCs III, VI, and VII. The most vulnerable diseases are widely scattered throughout the network. Only some of the SCID proteins that are related with the most severe PIDs are highly vulnerable. Also, previously, the vast majority of disease genes were shown to be peripheral in the network (27).

Inheritance pattern (supplemental Fig. S1) and prevalence (supplemental Fig. S2) did not show any correlation within DCs at all. This was expected because these characteristics are not likely to affect the etiology of diseases.

Discussion

We developed a novel approach for nosology and produced a systematic classification of PIDs based on parameters across several clinical, pathological, and physiological dimensions. Our approach combines existing clustering and network partition methods to classify these diseases. The new classification shares certain features with previous groupings, yet is different in a number of details. For example, the new classification indicates that cell-type expression, which has previously been one of the major classification criteria, cannot be very reliably used for classification of PIDs.

Clustering methods are widely used in many fields, such as in microarray data analysis. In medicine, cluster analysis has been used in the nosological splitting of different phenotypes, for example in Marshall and Stickler syndromes (50, 51), and more recently to classify patients with chronic pain (52) and to develop a new taxonomy for airway diseases (53, 54). Because the disease data were not complete due to many of the PIDs being extremely rare, we combined both network and clustering methods and used their consensus to obtain a robust grouping for PIDs. Considering the consensus of four, five, or six independent methods makes the results robust and independent from the individual methods.

Comparison with previous classifications

Previous in silico disease classification methods have been based on shared genes in disorders (27), protein interactions (55), protein complexes (28), or tissue-specific gene expression (29). For example, Human Phenotype Ontology terms describe clinical features and can be used for disease grouping (56). However, because only a few PIDs were included in these classification studies, we unfortunately cannot compare our results with them.

The new PID classification differs in details from those published earlier. The most comprehensive classification with the largest number of PIDs is from the European Society for Immunodeficiencies (ESID) registry (7) based on the International Union of Immunological Societies classification of 150 diseases (6). This scheme contains seven defined disease groups. The IDR uses and expands the classification (from Ref. 2) in 11 classes. The American Academy of Allergy, Asthma, and Immunology, the American College of Allergy, Asthma, and Immunology, and the Joint Council of Allergy, Asthma, and Immunology have classified 97 PIDs in 5 groups (39). The International Classification of Diseases (ICD10) contains 100 PIDs in 10 categories (8). These classifications have been useful, although they have disagreed on a number of disorders. The task of classifying the widely variable PIDs is hardly any more possible to do manually due to the very high dimensionality of the data. A systematic, mathematical classification can be used as an alternative or complement approach to the existing methods.

The earlier classifications were color coded and visualized in supplemental Fig. 3. The previous studies contained only subsections of the PIDs. The color codes were also chosen so that interclassification comparisons are possible because related diseases in the different groupings have the same colors. The more homogeneous the color is within a cluster, the more similar the classifications are. The IDR and ESID classifications agree very well in DCs I, II, IV, V, VII, VIII, IX, X, and XI. ICD10 agrees well in DCs III, V, VII, IX, X, and XI. There are, however, only 100 PIDs in the ICD system. The American Academy of Allergy, Asthma, and Immunology grouping behaves similarly to ICD, except for DC X, diseases that are not included at all. In conclusion, the novel

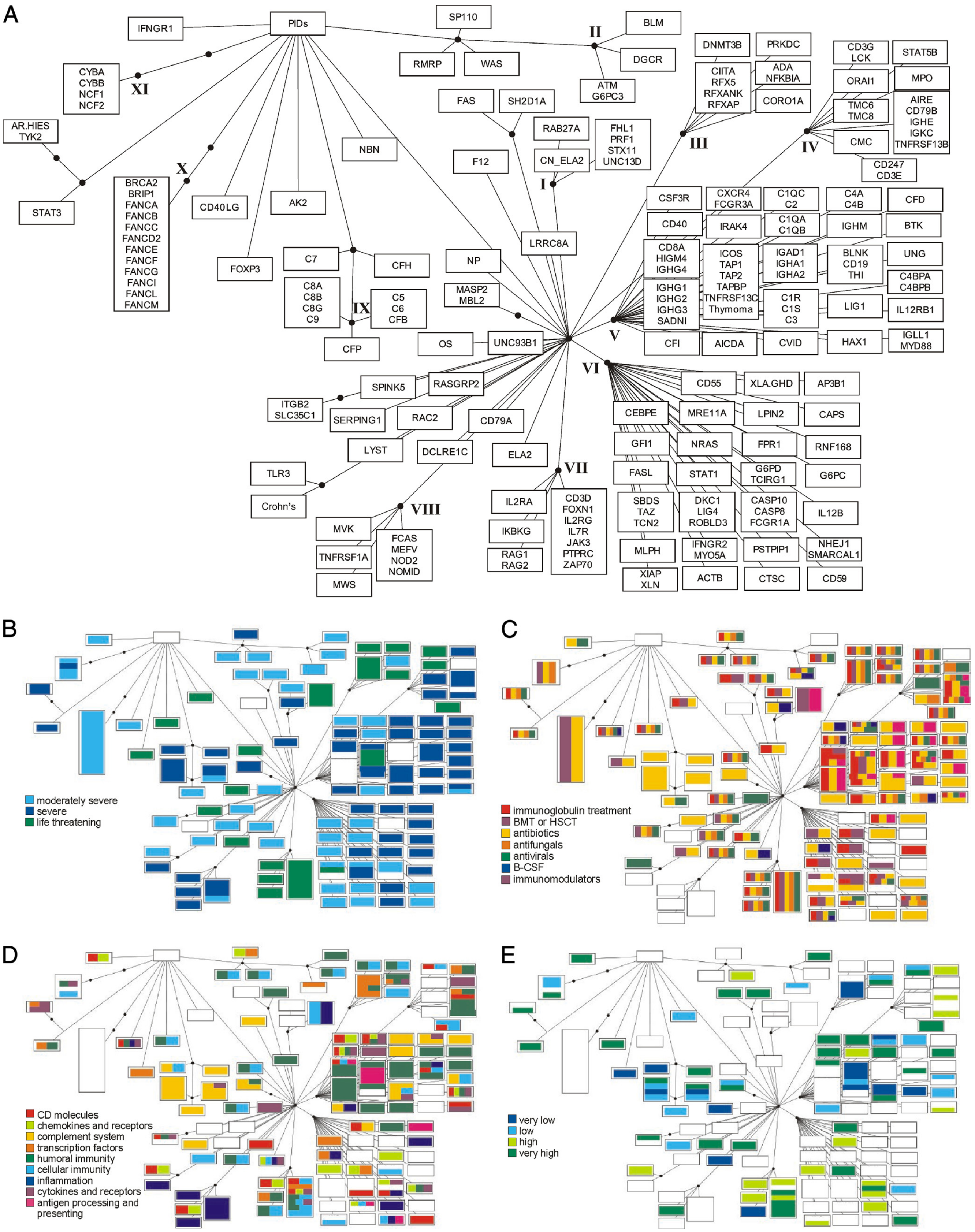


FIGURE 1. Consensus for the six methods used to group the PIDs. *A*, Relationship of the diseases. The white rectangles indicate the grouping of diseases by five (those attached to the first black bullet from the PIDs root) and six (the second bullet) methods. Diseases are indicated by the systematic names of affected genes, when known. Otherwise, the following names were used when no genes are identified in relation to the disease: AR-HIES, autosomal recessive hyperimmunoglobulin E recurrent infection syndrome; CMC, chronic mucocutaneous candidiasis; CVID, common variable immunodeficiency; FHL1, familial hemophagocytic lymphohistiocytosis type 1; HIGM4, hyper-IgM syndrome type 4; SADNI, specific Ab deficiency with normal Ig concentrations; THI, transient hypogammaglobulinemia of infancy; thymoma, thymoma with immunodeficiency (Good's syndrome); and XLA/GHD, X-linked hypogammaglobulinemia with growth hormone deficiency. When one gene is involved in more than one disease, the following abbreviations were used:

classification agrees with numerous features in the earlier groupings. All of the previous classifications have a smaller number of groups, which are typically divided in the new nosology. There are also a number of diseases that are not consistent with the novel classification, or with the other previous groupings. The systematic mathematical approach is capable of solving these cases and places the diseases in groups using solely the given parameters.

Parameters characterizing primary immunodeficiencies

In PIDs, the signs and symptoms of infections may be repetitive, severe, or refractory to therapy and caused by organisms with low virulence (39). We grouped the parameters for infections according to microbial taxonomy and site of infection.

Autoimmune diseases and malignancies are complications of many immunodeficiencies. Based on the frequency of associations with autoimmune diseases, PIDs can be grouped in three groups, as follows: systematic (>80% of the patients with the disorder have autoimmune disease symptoms), strong (20–80%), and mild (<20% of the patients) and absent (12, 57). Some PID patients appear susceptible also to atopy and lupus-like syndromes (58).

Malignancies occur with great frequency in certain immunodeficiencies. The types of malignancies depend on the PID, the age of the patient, and any possible viral infection(s). B cell malignancies, especially non-Hodgkin's lymphomas, are predominant. Other types of malignancies encountered in PIDs are T cell malignancies and leukemia (14).

Problems in lymphoproliferation (hepatomegaly, splenomegaly, lymphadenopathy) are typical for some PIDs. EBV infection is associated with many lymphoproliferation-linked immunodeficiencies (59–61).

Some PID patients have chronic respiratory problems, such as asthma, chronic obstructive disease, chronic inflammatory lung disease, emphysema/lung cysts, or interstitial pneumonia (62–65). Cardiovascular diseases such as congenital cardiac anomalies, cardiomyopathy, and hematologic abnormalities are found in certain PIDs (66, 67).

Some of the gastrointestinal diseases are associated with PIDs, including esophageal atresia, Crohn's disease, chronic inflammatory bowel disease, ulcerative colitis, granulomatous colitis, celiac disease, malabsorption, Hirschprung disease, and anal stenosis (15, 68–70). Hepatobiliary tract diseases in PID patients include storage liver disease, hepatic vascular occlusion disease, and sclerosing cholangitis (71, 72). Kidney diseases associated with PIDs include renal anomalies, renal dysfunction, amyloidosis, renal failure, IgA nephropathy, and glomerulonephritis (73).

Although physical findings are often absent, may be nonspecific, or very discreet, many PIDs have characteristic features. A common feature in PID patients is failure to thrive (child) or wasting (adult). Facial abnormalities, such as microcephaly or dysmorphism, are characteristic in some PIDs. Neurological abnormalities can include ataxia, peripheral neuropathy, speech delay, mental retardation, retinal lesions, photophobia, convulsions, or psychomotor retardation, whereas gastrointestinal abnormalities appear as severe gingivostomatitis, recurrent aphthae, periodontitis, delayed shedding of primary teeth, palatal weakness/cleft, gastric outlet obstruction, or diarrhea (1, 2, 39–41).

Ligamentous laxity/hyperextensive joints, limited extension of elbows, costocondral junction abnormality, rib abnormalities, metaphyseal chondrodysplasia/dysostosis, pectus carinatum, spondyloepiphyseal dysplasia, hip degeneration, or short limb dwarfism are among the skeletal abnormalities found in PIDs (74). The skin is frequently affected in immunodeficiencies. Erythroderma, eczema/atopic dermatitis, pyoderma, or granuloma is common. The presence of petechiae or bruises suggests a bleeding problem, as in phagocyte disorders (75–79).

The definitive diagnosis of PIDs depends on laboratory evidence, including assessments of humoral and cellular immunity and molecular analysis. The immunologic phenotype is based on laboratory tests of immune function, such as serum Ig levels, specific Ab titers, peripheral blood lymphocyte subpopulations, measures of T cell function, assays of phagocytes, and complement function or serum component level (1, 2, 37, 39–41).

Most of the parameters have binary values (yes, no), whereas, for example, the laboratory parameters are quantitative. The parameters were chosen so that they represent different important features of PIDs.

The new PID classification

The obtained 11 disease clusters are very robust due to them being the consensus of at least five methods. The *p* values show the significance of the observations. More detailed subgrouping is available by using the consensus of all the six methods. In Fig. 1*A*, diseases coclustered by all the six methods are within the boxes, whereas in the DCs at least five methods agree on the placement of PIDs within the graph. In addition to the actual classification, the PID parameters could offer guidelines for medical descriptions of PIDs. The classification allows a novel and fresh look at the relationships of PIDs, the genes behind them, and the encoded proteins. The network is far more complex than the previous mainly cell-type-based groupings might have led to imagine. Based on the classification and the parameters, it might be possible to develop novel diagnostic schemes for PIDs.

The correlation with other independent information not used for the original classification implies that the classification reliably reflects numerous properties of the diseases and the genes and proteins mutated in them. Data for the disease parameters were not complete, especially for the ultra-rare PIDs, and thus the homogeneity of the groups could even increase in the future when more reliable statistical information for the symptoms and laboratory characteristics will be available.

Diseases affecting genes involved in the same pathway do often cocluster, and they share a similar etiology, for example, JAK3 and IL2RG, and proteins in IFN-mediated immunity, including STAT1 and IFNGR2; MHC II diseases in DC III are all parts of the same protein complex as well as the MHC I components in DC V; complement components are in DC V and IX, and components of membrane bound oxidase in DC XI; BLNK, BTK, and MyD88, which have been suggested to be downstream of CD19 signaling, are all in DC V.

The new PID classification resulting from our approach can have several applications. Because it was generated independently from the existing classifications using solely mathematical analysis of clinical

CAPS, cryopyrin-associated periodic syndromes; Crohn's, Crohn's disease; CN_ELA2, cyclic neutropenia; ELA2, severe congenital neutropenia; FCAS, familial cold autoinflammatory syndrome; MWS, Muckle-Wells syndrome; NOMID, neonatal-onset multisystem inflammatory disease; WAS, Wiskott-Aldrich syndrome; and XLN, X-linked neutropenia. OS denotes Omenn syndrome, in which multiple disease genes were identified. Independent properties of the diseases and their respective proteins have a rather homogeneous distribution with the disease clusters for *B*, disease severity; *C*, therapeutic options; *D*, functional properties of proteins; and *E*, vulnerability of the nodes in the PID protein interaction network. For supporting statistical parameters, see supplemental Table S2. The color-coded properties are shown in each subgraph.

and laboratory parameters, it can be used in evaluation and development of other classifications. Disease groups defined by experts and found also by our independent approach have a strong indication of their existence. The new classification will be used in IDR (4) for organizing diseases and information about them. Another possible application is detailed demographics and mortality records.

The classification and the dataset also serve the development of diagnostic expert systems, which requires objective criteria for diagnosis. Expert systems are useful, especially in case of rare diseases like PIDs. The computer-based disease classification applied in this study can also identify the key symptoms and laboratory parameters that can help the experts to diagnose correctly these diseases.

The approach can be applied also to other groups of diseases.

Acknowledgments

We thank Jukka Lehtiniemi for help with the figures.

Disclosures

The authors have no financial conflict of interest.

References

1. Stiehm, E. R., H. D. Ochs, J. A. Winklestein, and E. Rich. 2004. *Immunologic Disorders in Infants and Children.* Elsevier, New York.
2. Ochs, H. D., C. I. E. Smith, and J. M. Puck. 2007. *Primary Immunodeficiency Diseases: A Molecular and Genetic Approach.* Oxford University Press, Oxford.
3. Notarangelo, L., J. L. Casanova, M. E. Conley, H. Chapel, A. Fischer, J. Puck, C. Roifman, R. Seger, and R. S. Geha. 2006. Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee Meeting in Budapest, 2005. *J. Allergy Clin. Immunol.* 117: 883–896.
4. Samarghitean, C., J. Väliaho, and M. Vihinen. 2007. IDR knowledge base for primary immunodeficiencies. *Immunome Res.* 3: 6.
5. Shearer, W. T., R. H. Buckley, R. J. Engler, A. F. Finn, Jr., T. A. Fleisher, T. M. Freeman, H. G. Herrod III, A. I. Levinson, M. Lopez, R. R. Rich, et al. 1996. Practice parameters for the diagnosis and management of immunodeficiency: The Clinical and Laboratory Immunology Committee of the American Academy of Allergy, Asthma, and Immunology (CLIC-AAAAI). *Ann. Allergy Asthma Immunol.* 76: 282–294.
6. Geha, R. S., L. D. Notarangelo, J. L. Casanova, H. Chapel, M. E. Conley, A. Fischer, L. Hammarstrom, S. Nonoyama, H. D. Ochs, J. M. Puck, et al. 2007. Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. *J. Allergy Clin. Immunol.* 120: 776–794.
7. Guzman, D., D. Veit, V. Knerr, G. Kindle, B. Gathmann, A. M. Eades-Perner, and B. Grimbacher. 2007. The ESID Online Database network. *Bioinformatics* 23: 654–655.
8. WHO. 2007. *International Statistical Classification of Diseases and Related Health Problems.* World Health Organization, Geneva.
9. Riches, P. G. 1992. Viral infections complicating primary immunodeficiencies. *Clin. Ter.* 140: 123–129.
10. Antachopoulos, C., T. J. Walsh, and E. Roilides. 2007. Fungal infections in primary immunodeficiencies. *Eur. J. Pediatr.* 166: 1099–1117.
11. Bustamante, J., S. Boisson-Dupuis, E. Jouanguy, C. Picard, A. Puel, L. Abel, and J. L. Casanova. 2008. Novel primary immunodeficiencies revealed by the investigation of pediatric infectious diseases. *Curr. Opin. Immunol.* 20: 39–48.
12. Carneiro-Sampaio, M., and A. Coutinho. 2007. Tolerance and autoimmunity: lessons at the bedside of primary immunodeficiencies. *Adv. Immunol.* 95: 51–82.
13. Arkwright, P. D., M. Abinun, and A. J. Cant. 2002. Autoimmunity in human primary immunodeficiency diseases. *Blood* 99: 2694–2702.
14. Salavoura, K., A. Kolialexi, G. Tsangaris, and A. Mavrou. 2008. Development of cancer in patients with primary immunodeficiencies. *Anticancer Res.* 28: 1263–1269.
15. Assari, T. 2006. Chronic granulomatous disease; fundamental stages in our understanding of CGD. *Med. Immunol.* 5: 4.
16. Janka, G. E. 2007. Hemophagocytic syndromes. *Blood Rev.* 21: 245–253.
17. Filipovich, A. H. 2008. Hemophagocytic lymphohistiocytosis and other hemophagocytic disorders. *Immunol. Allergy Clin. North Am.* 28: 293–313.
18. Frank, M. M. 2008. 8. Hereditary angioedema. *J. Allergy Clin. Immunol.* 121: S398–S401.
19. Shinkai, K., T. H. McCalmont, and K. S. Leslie. 2008. Cryopyrin-associated periodic syndromes and autoinflammation. *Clin. Exp. Dermatol.* 33: 1–9.
20. Stojanov, S., and M. F. McDermott. 2005. The tumor necrosis factor receptor-associated periodic syndrome: current concepts. *Expert Rev. Mol. Med.* 7: 1–18.
21. Brydges, S., and D. L. Kastner. 2006. The systemic autoinflammatory diseases: inborn errors of the innate immune system. *Curr. Top. Microbiol. Immunol.* 305: 127–160.
22. Marodi, L., and L. D. Notarangelo. 2007. Immunological and genetic bases of new primary immunodeficiencies. *Nat. Rev. Immunol.* 7: 851–861.
23. Barabasi, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5: 101–113.
24. Pastor-Satorras, R., and A. Vespignani. 2001. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86: 3200–3203.
25. Pastor-Satorras, R., and A. Vespignani. 2002. Immunization of complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65: 036104.
26. Savli, H., A. Szendroi, I. Romics, and B. Nagy. 2008. Gene network and canonical pathway analysis in prostate cancer: a microarray study. *Exp. Mol. Med.* 40: 176–185.
27. Goh, K. I., M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi. 2007. The human disease network. *Proc. Natl. Acad. Sci. USA* 104: 8685–8690.
28. Lage, K., E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, et al. 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* 25: 309–316.
29. Lage, K., N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak. 2008. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. USA* 105: 20870–20875.
30. Ortutay, C., and M. Vihinen. 2008. Efficiency of the immunome protein interaction network increases during evolution. *Immunome Res.* 4: 4.
31. Barrell, D., E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler. 2009. The GOA database in 2009: an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.* 37: D396–D403.
32. Ortutay, C., and M. Vihinen. 2009. Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res.* 37: 622–628.
33. Kaufman, L., and P. Rousseeuw. 1990. *Finding Groups in Data: an Introduction to Cluster Analysis.* Wiley, New York.
34. Pons, P., and M. Latapy. 2005. Computing communities in large networks using random walks. *arXiv:physics* 0512106.
35. Newman, M. E. J., and M. Girvan. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* 69: 026113.
36. Newman, M. E. J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74: 036104.
37. Samarghitean, C., J. Väliaho, and M. Vihinen. 2004. Online registry of genetic and clinical immunodeficiency diagnostic laboratories, IDdiagnostics. *J. Clin. Immunol.* 24: 53–61.
38. Piirilä, H., J. Väliaho, and M. Vihinen. 2006. Immunodeficiency mutation databases (IDbases). *Hum. Mutat.* 27: 1200–1208.
39. Bonilla, F. A., I. L. Bernstein, D. A. Khan, Z. K. Ballas, J. Chinen, M. M. Frank, L. J. Kobrynski, A. I. Levinson, B. Mazer, R. P. Nelson, Jr., et al. 2005. Practice parameter for the diagnosis and management of primary immunodeficiency. *Ann. Allergy Asthma Immunol.* 94: S1–S63.
40. Spickett, G. 2006. *Oxford Handbook of Clinical Immunology and Allergy.* Oxford University Press, Oxford.
41. De Vries, E. 2006. Patient-centered screening for primary immunodeficiency: a multi-stage diagnostic protocol designed for non-immunologists. *Clin. Exp. Immunol.* 145: 204–214.
42. Eades-Perner, A. M., B. Gathmann, V. Knerr, D. Guzman, D. Veit, G. Kindle, and B. Grimbacher. 2007. The European internet-based patient and research database for primary immunodeficiencies: results 2004–06. *Clin. Exp. Immunol.* 147: 306–312.
43. Wehr, C., T. Kivioja, C. Schmitt, B. Ferry, T. Witte, E. Eren, M. Vlkova, M. Hernandez, D. Detkova, P. R. Bos, et al. 2008. The EUROclass trial: defining subgroups in common variable immunodeficiency. *Blood* 111: 77–85.
44. Teo, Y. Y. 2008. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr. Opin. Lipidol.* 19: 133–143.
45. Csardi, G., and T. Nepusz. 2006. The igraph software package for complex network research. *InterJournal Complex Systems* 1695.
46. Prevalence of rare diseases: bibliographic data, Orphanet Report Series, May 2009:1. http://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf
47. Ortutay, C., M. Siermala, and M. Vihinen. 2007. Molecular characterization of the immune system: emergence of proteins, processes, and domains. *Immunogenetics* 59: 333–348.
48. Ortutay, C., and M. Vihinen. 2009. Immunome Knowledge Base (IKB): an integrated service for immunome research. *BMC Immunol.* 10: 3.
49. Ortutay, C., and M. Vihinen. 2006. Immunome: a reference set of genes and proteins for systems biology of the human immune system. *Cell. Immunol.* 244: 87–89.
50. Ayme, S., and M. Preus. 1984. The Marshall and Stickler syndromes: objective rejection of lumping. *J. Med. Genet.* 21: 34–38.
51. Verloes, A. 1995. Numerical syndromology: a mathematical approach to the nosology of complex phenotypes. *Am. J. Med. Genet.* 55: 433–443.
52. Sheffer, C. E., J. A. Deisinger, J. E. Cassisi, and K. Lofland. 2007. A revised taxonomy of patients with chronic pain. *Pain Med.* 8: 312–325.
53. Haldar, P., I. D. Pavord, D. E. Shaw, M. A. Berry, M. Thomas, C. E. Brightling, A. J. Wardlaw, and R. H. Green. 2008. Cluster analysis and clinical asthma phenotypes. *Am. J. Respir. Crit. Care Med.* 178: 218–224.
54. Wardlaw, A. J., M. Silverman, R. Siva, I. D. Pavord, and R. Green. 2005. Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clin. Exp. Allergy* 35: 1254–1262.
55. Feldman, I., A. Rzhetsky, and D. Vitkup. 2008. Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. USA* 105: 4323–4328.

56. Robinson, P. N., S. Kohler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* 83: 610–615.

57. Carneiro-Sampaio, M., and A. Coutinho. 2007. Immunity to microbes: lessons from primary immunodeficiencies. *Infect. Immun.* 75: 1545–1555.

58. Karim, M. Y. 2006. Immunodeficiency in the lupus clinic. *Lupus* 15: 127–131.

59. Tran, H., J. Nourse, S. Hall, M. Green, L. Griffiths, and M. K. Gandhi. 2008. Immunodeficiency-associated lymphomas. *Blood Rev.* 22: 261–281.

60. Wallet-Faber, N., C. Bodemer, S. Blanche, E. Delabesse, C. Eschard, N. Brousse, and S. Fraitag. 2008. Primary cutaneous Epstein-Barr virus-related lymphoproliferative disorders in 4 immunosuppressed children. *J. Am. Acad. Dermatol.* 58: 74–80.

61. Elenitoba-Johnson, K. S., and E. S. Jaffe. 1997. Lymphoproliferative disorders associated with congenital immunodeficiencies. *Semin. Diagn. Pathol.* 14: 35–47.

62. Garcia-Laorden, M. I., J. Sole-Violan, F. Rodriguez de Castro, J. Aspa, M. L. Briones, A. Garcia-Saavedra, O. Rajas, J. Blanquer, A. Caballero-Hidalgo, J. A. Marcos-Ramos, et al. 2008. Mannose-binding lectin and mannose-binding lectin-associated serine protease 2 in susceptibility, severity, and outcome of pneumonia in adults. *J. Allergy Clin. Immunol.* 122: 368–374, 374.e1–2.

63. Basile, N., S. Danielian, M. Oleastro, S. Rosenzweig, E. Prieto, J. Rossi, A. Roy, and M. Zelazko. 2009. Clinical and molecular analysis of 49 patients with X-linked agammaglobulinemia from a single center in Argentina. *J. Clin. Immunol.* 29: 123–129.

64. Bott, L., J. Lebreton, C. Thumerelle, J. Cuvellier, A. Deschildre, and A. Sardet. 2007. Lung disease in ataxia-telangiectasia. *Acta Paediatr.* 96: 1021–1024.

65. Quinti, I., A. Soresina, G. Spadaro, S. Martino, S. Donnanno, C. Agostini, P. Claudio, D. Franco, A. Maria Pesce, F. Borghese, et al. 2007. Long-term follow-up and outcome of a large cohort of patients with common variable immunodeficiency. *J. Clin. Immunol.* 27: 308–316.

66. Shprintzen, R. J. 2008. Velo-cardio-facial syndrome: 30 years of study. *Dev. Disabil. Res. Rev.* 14: 3–10.

67. Sweeney, R. T., G. J. Davis, and J. A. Noonan. 2008. Cardiomyopathy of unknown etiology: Barth syndrome unrecognized. *Congenit. Heart Dis.* 3: 443–448.

68. Marks, D. J., K. Miyagi, F. Z. Rahman, M. Novelli, S. L. Bloom, and A. W. Segal. 2009. Inflammatory bowel disease in CGD reproduces the clinicopathological features of Crohn's disease. *Am. J. Gastroenterol.* 104: 117–124.

69. Barton, L. L., S. L. Moussa, R. G. Villar, and R. L. Hulett. 1998. Gastrointestinal complications of chronic granulomatous disease: case report and literature review. *Clin. Pediatr.* 37: 231–236.

70. Daniels, J. A., H. M. Lederman, A. Maitra, and E. A. Montgomery. 2007. Gastrointestinal tract pathology in patients with common variable immunodeficiency (CVID): a clinicopathologic study and review. *Am. J. Surg. Pathol.* 31: 1800–1812.

71. Cliffe, S. T., M. Wong, P. J. Taylor, E. Ruga, B. Wilcken, R. Lindeman, M. F. Buckley, and T. Roscioli. 2007. The first prenatal diagnosis for veno-occlusive disease and immunodeficiency syndrome, an autosomal recessive condition associated with mutations in SP110. *Prenat. Diagn.* 27: 674–676.

72. Hayward, A. R., J. Levy, F. Facchetti, L. Notarangelo, H. D. Ochs, A. Etzioni, J. Y. Bonnefoy, M. Cosyns, and A. Weinberg. 1997. Cholangiopathy and tumors of the pancreas, liver, and biliary tree in boys with X-linked immunodeficiency with hyper-IgM. *J. Immunol.* 158: 977–983.

73. De Heer, E., and D. J. Peters. 2008. Innate immunity as a driving force in renal disease. *Kidney Int.* 73: 7–8.

74. Sordet, C., A. Cantagrel, T. Schaeverbeke, and J. Sibilia. 2005. Bone and joint disease associated with primary immune deficiencies. *Joint Bone Spine* 72: 503–514.

75. Heyworth, P. G., A. R. Cross, and J. T. Curnutte. 2003. Chronic granulomatous disease. *Curr. Opin. Immunol.* 15: 578–584.

76. De Raeve, L., M. Song, J. Levy, and F. Mascart-Lemone. 1992. Cutaneous lesions as a clue to severe combined immunodeficiency. *Pediatr. Dermatol.* 9: 49–51.

77. Saurat, J. H. 1985. Eczema in primary immune-deficiencies: clues to the pathogenesis of atopic dermatitis with special reference to the Wiskott-Aldrich syndrome. *Acta Derm. Venereol. Suppl.* 114: 125–128.

78. Moin, A., A. Farhoudi, M. Moin, Z. Pourpak, and N. Bazargan. 2006. Cutaneous manifestations of primary immunodeficiency diseases in children. *Iran J. Allergy Asthma Immunol.* 5: 121–126.

79. Chowdhury, M. M., A. Anstey, and C. N. Matthews. 2000. The dermatosis of chronic granulomatous disease. *Clin. Exp. Dermatol.* 25: 190–194.

PID CLASSIFICATION

Systematic classification of primary immunodeficiencies based on clinical, pathological and laboratory parameters

Crina Samarghitean, Csaba Ortutay, Mauno Vihinen

Supplemental legends:

Fig. S1. Inheritance of PIDs.

Fig. S2. Prevalence of PIDs.

Fig. S3. Comparison of PID classification schemes.

Table S1. Parameters used to characterize signs, symptoms and laboratory values in PIDs.

Table S2. Parameters statistically enriched in Disease Clusters.

Fig. S1. Inheritance of PIDs. The layout of the graph is the same as in Fig. 1*A*.

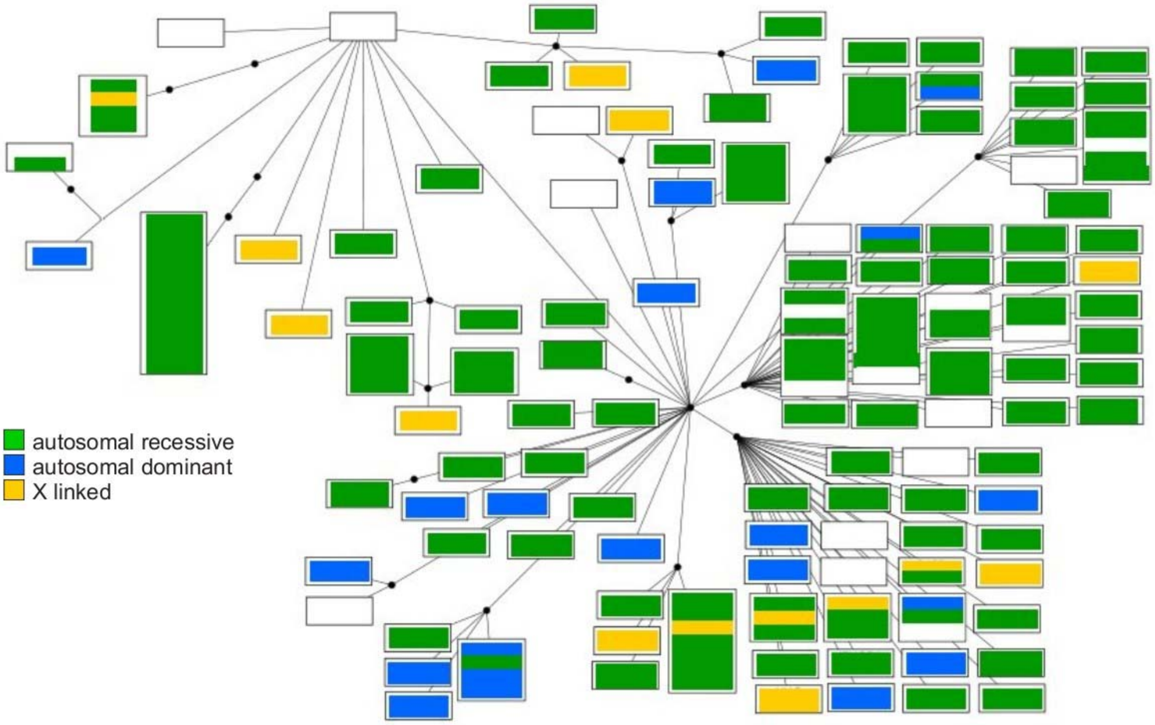


Fig. S2. Prevalence of PIDs. The layout of the graph is the same as in Fig. 1*A*.

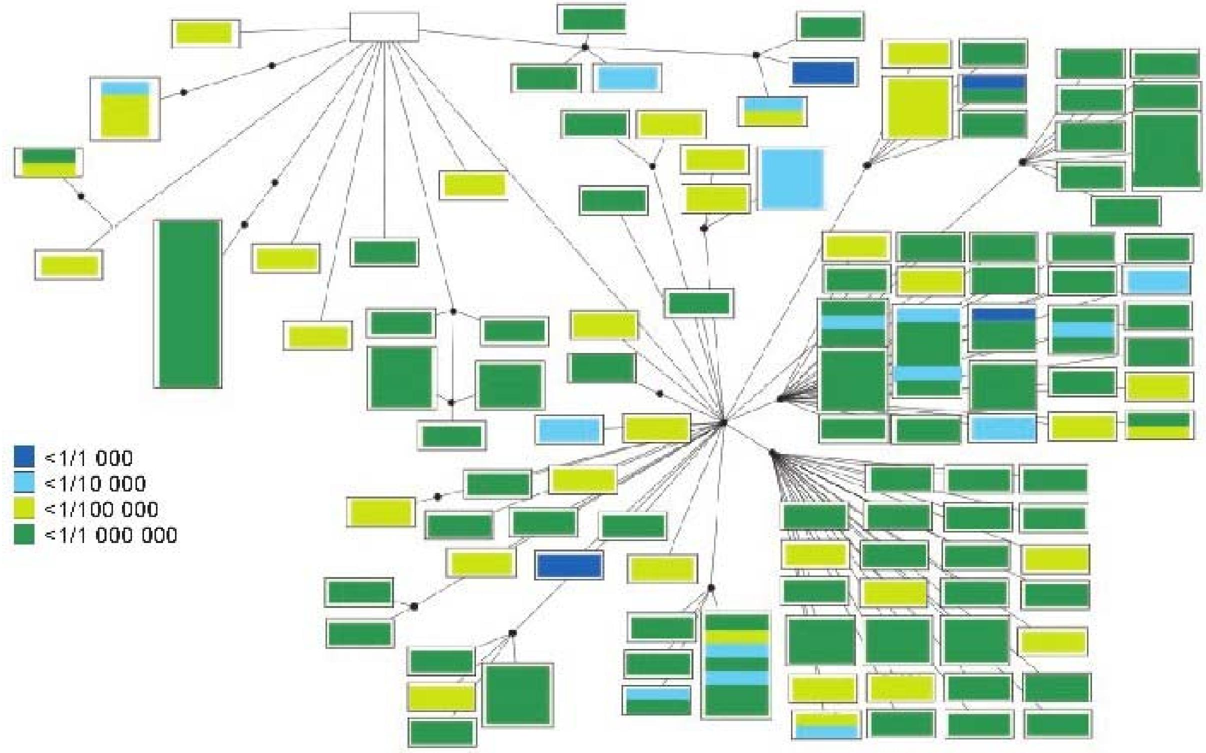


Fig. S3. Comparison of PID classification schemes. The layout of the graph is the same as in Fig. 1*A*.

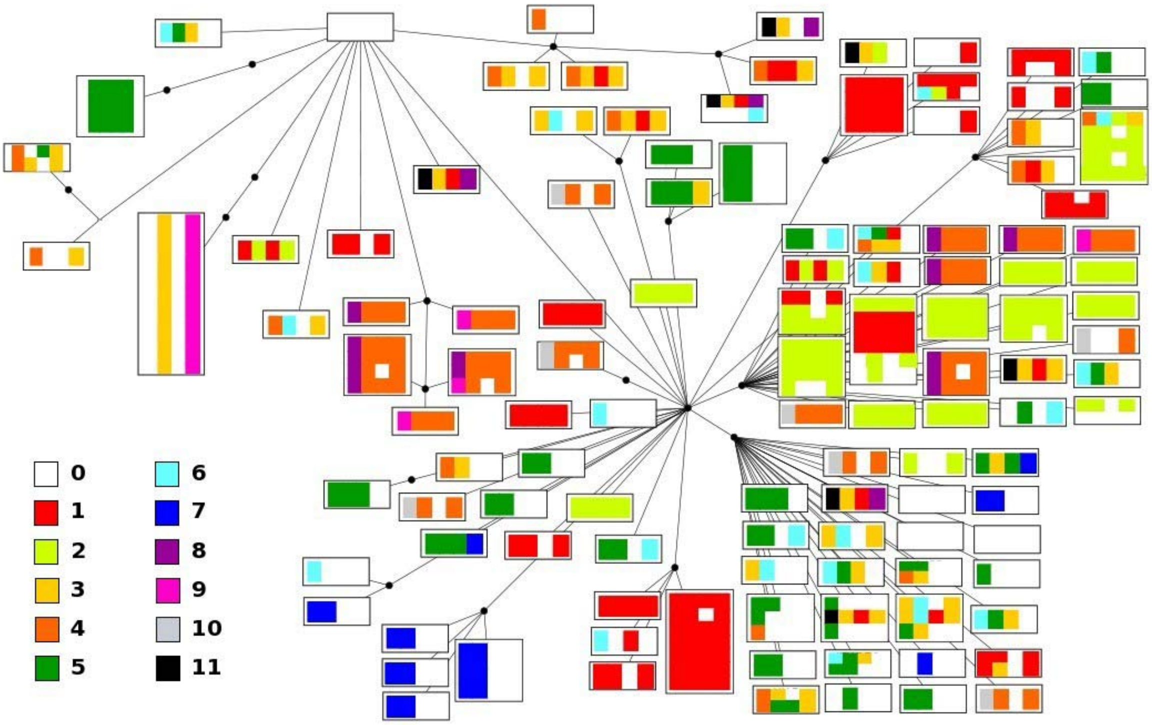
The novel PID classification was compared with four previous systems In each cells the four coloured columns represents the following disease classification systems: IDR, ESID, AAAAI, ICD-10 from left to right, respectively. The colour coding for the individual systems is as follows:

ImmunoDeficiency Resource (IDR) classification
0 Not available
1 Combined B and T cell immunodeficiencies
2 Deficiencies predominantly affecting antibody production
3 Defects in lymphocyte apoptosis
4 Other well-defined immunodeficiency syndromes
5 Defects of phagocyte function
6 Defects of innate immune system, receptors and signaling components
7 Periodic fever syndromes
8 Defects of the classical complement cascade proteins
9 Defects of the alternative complement pathway
10 Defects of complement regulatory proteins
11 DNA breakage associated syndromes and DNA epigenetic modification syndromes

The European Society for Immunodeficiencies (ESID) classification
0 Not available
1 Predominantly T cell deficiencies, T^-B^- Severe combined immunodeficiency (SCID)
2 Predominantly antibody disorder
3 Other well definied PIDs
4 Complement deficiencies
5 Phagocytic disorders
6 Autoimmune and immunedysregulation syndromes
7 Autoinflammatory syndromes

American Academy of Allergy Asthma and Immunology (AAAAI) classification
0 Not available
1 Combined immunodeficiency
2 Humoral immunodeficiency
3 Cellular immunodeficiency
4 Phagocytic cell disorders
5 Complement deficiencies

International Classification of Diseases 10 (ICD-10)
0 Not available
1 Combined immunodeficiencies (D81)
2 Immunodeficiency with predominantly antibody defects (D80)
3 Immunodeficiency associated with other major defects (D82)
4 Other immunodeficiencies (D84)
5 Functional disorders of polymorphonuclear neutrophils (D71)
6 Agranulocytosis (D70)
7 Albinism (E70.3)
8 Cerebellar ataxia with defective DNA repair (G11.3)
9 Constitutional aplastic anaemia (D61)
10 Crohn's disease [regional enteritis] (K50)



0
1
2
3
4
5
6
7
8
9
10
11

Table S1. Parameters used to characterize signs, symptoms and laboratory values in PIDs.

Parameter	Values[a]	Number of diseases
Infections		
locations		
central nervous system	yes/no	44
upper respiratory tract	yes/no	91
lower respiratory tract	yes/no	92
skin	yes/no	52
gastro-intestinal	yes/no	33
musculo-skeletal	yes/no	31
other	yes/no	55
taxonomy		
Susceptibility to		
- bacteria, Gram (+) (*Streptococcus pneumonie*, *Haemophilus influenzae*)	0: no	134
	1: intermediate	24
	2: high	30
- bacteria, Gram (+) (*Staphylococcus aureus*)	0: low/no	132
	1: intermediate	17
	2: high	29
- bacteria, Gram (-) (*Neisseria*)	yes/no	13
- mycobacteria	0: low/no	150
	1: intermediate	2
	2: high	27
- viruses	0: no	133
	1: intermediate	11
	2: high	34
Susceptibility to fungi (especially *P. carinii*)	0: no	188
	1: intermediate	12
	2: high	33
- *Candida*	yes/no	43

Parameter	**Values**[a]	**Number of diseases**
- *Aspergillus*	yes/no	11
- *Cryptococcus*	yes/no	5
- *Histoplasma*	yes/no	5
Susceptibility to Protozoa	1: *Giardia*	9
	2: *Giardia*+*Cryptosporidium*	5
	3: C*ryptosporidium*	1
Immune system dysregulation		
autoimmunity		
frequency	0: AID<20%	153
	1: 20<AID<80	14
	2: AID>80%	11
systemic lupus erythematosus (and SLE-like syndrome)	yes/no	26
vasculitis	yes/no	13
rheumatoid arthritis/juvenile rheumatoid arthritis /arthritis	yes/no	21
autoimmune haemolytic anaemia (AIHA)	yes/no	23
idiopathic thrombocytopenia (ITP)	yes/no	17
endocrine AID	yes/no	6
malignancies		
B cell malignancies	yes/no	24
T cell malignancies	yes/no	12
leukemia	yes/no	19
other	yes/no	17
allergy/atopy	yes/no	15
Associated diseases		
chronic respiratory problems	yes/no	25
cardio-vascular diseases	yes/no	11
gastro-intestinal diseases	yes/no	25
biliary tract and liver diseases	yes/no	10

Parameter	**Values**[a]	**Number of diseases**
kidney and urogenital diseases	yes/no	36
infertility	yes/no	17
Signs and symptoms		
failure to thrive (child) or wasting (adult)	yes/no	36
chronic diarrhea	yes/no	38
delayed cord separation < 2 weeks	yes/no	5
fever	yes/no	25
growth retardation	yes/no	45
pleural abnormalities	yes/no	5
gastrointestinal abnormalities	yes/no	29
hepatomegaly	yes/no	25
splenomegaly	yes/no	31
lymphadenopathy	yes/no	25
bleeding tendency	yes/no	35
skeletal or connective tissue abnormalities	yes/no	31
facial abnormalities		
- microcephaly	yes/no	21
- dental abnormalities/periodontitis	yes/no	10
- other facial anomalies/dysmorphism	yes/no	16
skin defects		
- photosensitivity	yes/no	5
- cutaneous telangiectasia	yes/no	5
- pigmentation defects	yes/no	14
- rash	yes/no	10
- eczema/atopic dermatitis	yes/no	17
- erythroderma	yes/no	7
- granuloma	yes/no	7
- other skin defects	yes/no	40
neurological or CNS abnormalities	yes/no	44
-mental retardation	yes/no	5

Parameter	Values[a]	Number of diseases
-ocular abnormalities	yes/no	7
Laboratory features		
low red blood cell count (anemia)	yes/no	55
white blood cell (leukocytes) count	-1: decreased	67
	0: normal/no	96
	1: increased	21
low neutrophil count (<500/ml)	yes/no	56
lymphocyte cell count affected	yes/no	82
thrombocytopenia	yes/no	28
polymorphonuclear cells affected	yes/no	32
monocytes/macrophages affected	yes/no	29
circulating T cells affected	-1: decreased	51
	0: normal/no	136
	1: increased	7
circulating B cells affected	-1: decreased:	32
	0: normal/no	155
	1: increased	7
NK cells affected	-1: decreased	25
	0: normal/no	168
	1: increased	1
IgG	-1: decreased	63
	0: normal/no	139
	1: increased	2
IgA	-1: decreased	65
	0: normal/no	124
	1: increased	5
IgM	-1: decreased	43
	0: normal/no	141
	1: increased	10

Parameter	Values[a]	Number of diseases
IgE	-1: decreased	38
	0: normal/no	144
	1: increased	12
antibody response to booster immunization impaired	yes/no	19
low lymphocyte proliferation to mitogens	yes/no	20
CH50 abnormal (low or absent)	yes/no	21
AH50 abnormal (low or absent)	yes/no	12
low C4 or C3 concentrations	yes/no	7
spontaneous activation of the complement pathway	yes/no	19
radiation sensitivity/ chromosomal instability	yes/no	22
missing lymph nodes	yes/no	9
absence of thymus	yes/no	11
apoptosis	-1: decreased	4
	0: normal/no	177
	1: increased	13
chemotaxis	yes/no	8
killing (faulty O2 production)	yes/no	7

[a] no – either the parameter is not related or data is missing

Table S2. Parameters statistically enriched in Disease Clusters. p-values are shown in brackets.

Cluster	Diseases and affected genes	Parameters significantly overrepresented in the disease clusters	Treatments and functional categories significant in the clusters
I	UNC13D STX11 RAB27A PRF1 FHL1 CN_ELA2	other sites of infections (2.17e-03) fever (1.24e-06) neurological or CNS abnormalities (9.75e-04) gastrointestinal abnormalities (2.79e-03) hepatomegaly (7.21e-05) splenomegaly(1.96e-04) lymphadenopathy (7.21e-05) bleeding tendency (3.36e-04) low neutrophil count (<500/ml) (1.11e-04) thrombocytopenia (1.23e-04) lymphocyte cell count affected (6.02e-04)	BMT or HSCT (1.01e-02) immunomodulators (2.93e-02) cellular immunity (4.4e-02) inflammation (9.49e-03)
II	G6PC3 DGCR ATM BLM	other sites of infection (3.26e-02) infections in upper respiratory tract (9.93e-03) infections in lower respiratory tract (1.02e-02) endocrine-AID (4.34e-03) B cell malignancies(4.54e-02) leukemia (3.84e-02) cardio-vascular diseases(1.44e-02) infertility (3.17e-02) growth retardation (2.18e-02) other facial anomalies/dysmorphism (2.85e-02) microcephaly (2.98e-03) photosensitivity(2.95e-03) cutaneous telangiectasia(2.95e-03) neurological or CNS abnormalities (2.08e-02) low lymphocyte proliferation to mitogens (4.19e-02) radiation sensitivity/chromosomal instability (4.9e-02) absence of thymus (1.44e-02)	-

Cluster	Diseases and affected genes	Parameters significantly overrepresented in the disease clusters	Treatments and functional categories significant in the clusters
III	PRKDC NFKBIA ADA CORO1A RFXAP RFXANK RFX5 DNMT3B CIITA	infections in upper respiratory tract (2.60e-05) infections in lower respiratory tract (2.80e-05) skin infections (3.19e-04) susceptibility to bacteria, Gram (+) (*Staphylococcus aureus*) (3.03e-05) susceptibility to *Candida* (5.05e-06) failure to thrive or wasting (2.22e-08) chronic diarrhea (5.15e-05) gastro-intestinal diseases (1.12e-03) growth retardation (1.60e-04) neurological or CNS abnormalities (1.41e-04) lymphocyte cell count affected (1.31e-05) low lymphocyte proliferation to mitogens (3.57e-02)	immunoglobulin treatment (2.02e-05) BMT or HSCT (1.61e-05) antibiotics (5.36e-04) antifungals (2.32e-05) antivirals(8.31e-05) transcription factors (1.8e-04) humoral immunity (2.11e-02)
IV	STAT5B TNFRSF13B TMC8 TMC6 ORAI1 MPO LCK IGKC IGHE CMC CD79B CD3G AIRE CD3E CD247	infections in central nervous system(4.2e-02) infections in upper respiratory tract (2.61e-03) infections in lower respiratory tract (1.89e-02) susceptibility to *Candida* (1.9e-04) other malignancies (1.95e-02) low neutrophil count (<500/ml) (1.85e-02)	immunoglobulin treatment (1.29e-02) antibiotics (1.45e-02) antifungals (8.61e-04) antivirals(6.96e-05)

Cluster	Diseases and affected genes	Parameters significantly overrepresented in the disease clusters	Treatments and functional categories significant in the clusters
V	IRAK4 UNG TNFRSF13C Thymoma THI TAPBP TAP2 TAP1 ICOS HAX1 FCGR3A CXCR4 CD40 AICDA CSF3R CD19 BLNK IL12RB1 MYD88 IGLL1 IGHA2 IGHA1 IGAD1 CVID BTK SADNI IGHG3 IGHG2 IGHG1 LIG1 IGHG4 HIGM4 CFI C4B C4A C3 C2 C1S C1R C1QC C1QB C1QA CFD CD8A C4BPB C4BPA IGHM	infections in central nervous system(6.37e-03) infections in upper respiratory tract (2.4e-26) infections in lower respiratory tract (4.01e-26) gastro-intestinal infections (3e-02) susceptibility to mycobacteria (4.59e-03) susceptibility to *Candida* (9.07e-03) failure to thrive or wasting (1.20e-04) chronic diarrhea (5.26e-03) systemic lupus erythematosus (1.76e-03) rheumatoid arthritis (3.3e-06) idiopathic thrombocytopenia (3.90e-02) leukemia (6.87e-03) chronic respiratory problems (3.70e-02) cardio-vascular diseases (1.61e-02) infertility (1.13e-02) fever (1.26e-02) growth retardation (1.89e-04) other facial anomalies/dysmorphism (1.46e-02) microcephaly (4.19e-03) skin pigmentation defects (2.42e-02) neurological or CNS abnormalities (1.56e-03) gastrointestinal abnormalities (2.43e-02) hepatomegaly (1.58e-03) splenomegaly(3.75e-03) bleeding tendency (2.91e-02) skeletal or connective tissue abnormalities (3.81e-04) thrombocytopenia (7.72e-04) polymorphonuclear cells affected (3.02e-04) monocytes/macrophages affected (4.59e-03) CH50 abnormal (low or absent) (9.58e-04) low C4 or C3 concentrations (4.5e-02) spontaneous activation of the complement pathway (1.76e-03) radiation sensitivity/chromosomal instability (2.29e-02)	immunoglobulin treatment (2.75e-05) BMT or HSCT (1.64e-03) antibiotics (5.64e-16) antifungals (4.15e-02) immunomodulators (1.73e-11) complement system (8.87e-04) transcription factors (1.16e-02) humoral immunity (5.02e-05) cytokines (not chemokins) and receptors (4.69e-02) antigen processing and presenting (4.43e-02)

Cluster	Diseases and affected genes	Parameters significantly overrepresented in the disease clusters	Treatments and functional categories significant in the clusters
VI	NRAS FPR1 CD59 XLA.GHD G6PC FCGR1A CTSC SMARCAL1 ROBLD3 RNF168 NHEJ1 LIG4 TCN2 TCIRG1 TAZ STAT1 SBDS PSTPIP1 MYO5A MRE11A MLPH IL12B IFNGR2 G6PD ACTB DKC1 CEBPE AP3B1 CASP8 CASP10 CAPS XLN XIAP GFI1 FASL CD55 LPIN2	infections in central nervous system(2.53e-04) infections in upper respiratory tract (4.12e-06) infections in lower respiratory tract (1.68e-07) skin infections (4.47e-02) gastro-intestinal infections (4.75e-02) susceptibility to bacteria, Gram (+) (*Staphylococcus aureus*) (1.14e-02) susceptibility to *Candida* (3.02e-03) chronic diarrhea (2.63e-02) systemic lupus erythematosus (3.66e-02) other malignancies (3.95e-02) growth retardation (3.83e-02) other facial anomalies/dysmorphism (3.1e-02) other skin defects (4.18e-02) eczema/atopic dermatitis (3.26e-02) splenomegaly(2.88e-02) low neutrophil count (<500/ml) (1.17e-02) thrombocytopenia (2.73e-02) lymphocyte cell count affected (2.46e-02) monocytes/macrophages affected (2.44e-02) chemotaxis (3.29e-02) low lymphocyte proliferation to mitogens (1.84e-02) spontaneous activation of the complement pathway (2.22e-02)	BMT or HSCT (2.94e-02) antibiotics (3.57e-02) antifungals (3.84e-03) antivirals (1.35e-02) immunomodulators (2.09e-02) chemokins and receptors (3.88e-02) complement system (2.80e-02) humoral immunity (3.84e-03)

Cluster	Diseases and affected genes	Parameters significantly overrepresented in the disease clusters	Treatments and functional categories significant in the clusters
VII	ZAP70 RAG2 RAG1 PTPRC JAK3 IL7R IL2RG IL2RA IKBKG FOXN1 CD3D	infections in upper respiratory tract (6e-05) infections in lower respiratory tract (6.48e-05) susceptibility to bacteria, Gram (+) (*Staphylococcus aureus*) (2.97e-05) susceptibility to mycobacteria (3.93e-11) susceptibility to *Candida* (4.03e-06) failure to thrive or wasting (3.21e-10) chronic diarrhea (4.44e-08) eczema/atopic dermatitis (6.17e-03) lymphocyte cell count affected (9.66e-07) low lymphocyte proliferation to mitogens (1.07e-02) absence of thymus (1.18e-03)	immunoglobulin treatment (1.67e-06) BMT or HSCT (3.58e-05) antibiotics (9.63e-05) antifungals (2.10e-08) antivirals (1.39e-07) humoral immunity (1.26e-02) cellular immunity (2.48e-04)
VIII	TNFRSF1A NOMID NOD2 MWS MVK MEFV FCAS	infections in central nervous system(2.07e-02) musculo-skeletal infections (4.99e-07) vasculitis (2.94e-04) kidney and urogenital diseases (1.15e-02) fever (1.10e-07) rash (4.56e-11) neurological or CNS abnormalities (1.95e-04) ocular abnormalities (5.74e-10) pleural abnormalities (8.5e-09) gastrointestinal abnormalities (1.86e-05) splenomegaly (7.15e-03) skeletal or connective tissue abnormalities (4.99e-07) polymorphonuclear cells affected (6.19e-07) monocytes/macrophages affected (2.26e-05)	antibiotics (1.65e-02) inflammation (1.43e-06)

Cluster	Diseases and affected genes	Parameters significantly overrepresented in the disease clusters	Treatments and functional categories significant in the clusters
IX	CFP CFB C9 C8G C8B C8A C6 C5	infections in central nervous system(7.82e-07) infections in upper respiratory tract (4.4e-02) infections in lower respiratory tract (4.27e-02) susceptibility to bacteria, Gram (-) (*Neisseria*) (1.77e-11) systemic lupus erythematosus (4.72e-02) CH50 abnormal (low or absent) (1.02e-05) AH50 abnormal (low or absent) (7.06e-12) spontaneous activation of the complement pathway (5.52e-06)	immunoglobulin treatment (4.85e-02) antibiotics (1.34e-02) complement system (2.33e-08)
X	FANCM FANCL FANCI FANCG FANCF FANCE FANCD2 FANCC FANCB FANCA BRIP1 BRCA2	other sites of infections (4.63e-02) infections in upper respiratory tract (8.84e-03) infections in lower respiratory tract (8.47e-03) susceptibility to bacteria, Gram (+) (*Staphylococcus aureus*) (4.41e-02) leukemia (3.79e-15) kidney and urogenital diseases (3.66e-11) infertility (5.23e-16) growth retardation (6.75e-10) microcephaly (1.97e-14) skin pigmentation defects (9.18e-18) bleeding tendency (2.58e-11) skeletal or connective tissue abnormalities (5.41e-12) low neutrophil count (<500/ml) (7.08e-09) thrombocytopenia (1.38e-12) lymphocyte cell count affected (1.31e-02) polymorfonuclear cells affected (8.2e-12) radiation sensitivity/chromosomal instability(4.1e-14)	immunoglobulin treatment (1.03e-02) BMT or HSCT (3.45e-07) antibiotics (4.06e-05) antifungals (4.93e-02) antivirals (2.77e-02) humoral immunity (4.93e-02)

Cluster	Diseases and affected genes	Parameters significantly overrepresented in the disease clusters	Treatments and functional categories significant in the clusters
XI	NCF2 NCF1 CYBB CYBA	other sites of infections (2.18e-03) infections in central nervous system(1.04e-03) infections in upper respiratory tract (9.93e-03) infections in lower respiratory tract (1.02e-02) skin infections (1.82e-03) gastro-intestinal infections (3.8e-04) musculo-skeletal infections (3.03e-04) susceptibility to bacteria, Gram (+) (*Staphylococcus aureus*) (2.31e-03) susceptibility to mycobacteria (2.68e-04) susceptibility to *Aspergillus* (4.62e-06) failure to thrive or wasting (5.19e-04) chronic diarrhea (6.28e-04) systemic lupus erythematosus (1.57e-04) chronic respiratory problems (1.36e-04) gastro-intestinal diseases (1.36e-04) kidney and urogenital diseases (5.19e-04) growth retardation (1.21e-03) eczema/atopic dermatitis (2.97e-05) granuloma (5.3e-07) gastrointestinal abnormalities (2.37e-04) hepatomegaly (1.36e-04) splenomegaly (3.03e-04) lymphadenopathy (1.36e-04) low neutrophil count (<500/ml) (2.45e-03) polymorfonuclear cells affected (3.4e-04) monocytes/macrophages affected (2.68e-04) killing (faulty O2 production) (5.3e-07)	BMT or HSCT (8.08e-03) antibiotics (3.65e-02) antifungals (2.02e-03)