Bioinformatics strategies for disease gene identification

Een wetenschappelijke proeve op het gebied van de
Natuurwetenschappen, Wiskunde en Informatica.

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de Rector Magnificus prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op maandag 28 november 2005
des namiddags om 3.30 uur precies

door

Markus Antoine van Driel

geboren op 12 maart 1974
te Arnhem.

Promotors:

prof. dr. G. Vriend
prof. dr. H.G. Brunner

Co-promotor:

prof. dr. J.A.M. Leunissen (WU)

Manuscriptcommissie:

prof. dr. H.G. Stunnenberg
prof. dr. G.J.B. van Ommen (LUMC)
dr. J.H.P. Hackstein

ISBN-10: 9090198334
ISBN-13: 9789090198330

Cover:

“Molecules of Life”
Compilation of 2089 protein structures reflecting a human embryo.
© 2005 Helmut Grubmüller (with permission)

Printed by:

Drukkerij Print Partners Ipskamp te Enschede

2DGE - *2-Dimensional Gel Electrophoresis*

3D - *3-Dimensional*

ABO - *ABO blood system*

ADULT - *Acro-Dermato-Ungual-Lacrimal-Tooth (syndrome)*

AIDS - *Acquired Immune Deficiency Syndrome*

arrayCGH - *array-based Comparative Genomic Hybridization*

ASP - *Affected Sib Pair*

BBS - *Bardet-Biedl Syndrome*

BIND - *Biomolecular Interaction Network Database*

BLAST - *Basic Local Alignment Search Tool*

BLOSUM - *BLOcks SUbstitution Matrix*

bp - *base pair(s)*

BRENDA - *BRaunschweiger ENzyme DAtabase*

CAP - *College of American Pathologists*

CCM - *Chemical Cleavage Mismatches*

cDNA - *copy DNA*

ChIP - *Chromatin Immuno-Precipitation*

cM - *centiMorgan*

CMBI - *Centre for Molecular and Biomolecular Informatics*

COG - *Clusters of Orthologous Groups*

CS - *OMIM Clinical Synopsis field*

CT - *SNOMED Clinical Terms*

DB - *DataBase*

dbSNP - *Single Nucleotide Polymorphism database*

DDBJ - *DNA DataBank of Japan*

DGGE - *Denaturing Gradient Gel Electrophoresis*

DHMHD - *Dysmorphic Human and Mouse Homology Database*

DHPLC - *Denaturing High Performance Liquid Chromatography*

DIP - *Database of Interacting Proteins*

DNA - *Deoxyribose Nucleic Acid*

DSSP - *Definition of Secondary Structure of Proteins*

ECM - *Enzymatic Cleavage Mismatches*

EEC - *Ectrodactyly-Ectodermal dysplasia-Clefting*

EMBL - *European Molecular Biology Laboratory (institute/nucleic acid database)*

eq. - *equation*

EST - *Expressed Sequence Tag*

FA - *Fanconi Anaemia*

FASTA - *Fast Alignment Search Tools All/Anything*

FISH - *Fluorescence In Situ Hybridization*

GDB - *Genome DataBase*

GO - *Gene Ontology*

GOA - *Gene Ontology Annotation*

GXD - *Gene eXpression Database*

HGMD - *Human Gene Mutation Database*

HGP - *Human Genome Project*

HMM - *Hidden Markov Model*

HPP - *Human Phenome Project*

HPRD - *Human Protein Reference Database*

HTML - *HyperText Markup Language*

HT - *High-Throughput*

IBD - *Identity-By-Descent*

ICD - *International Classification of Disease*

ICD-9-CM/ICD-10-CM - *International Classification of Disease version 9/10 with Clinical Modification*

IR - *Information Retrieval*

KEGG - *Kyoto Encyclopedia of Genes and Genomes*

KOG - *(euKaryotes) clusters of Orthologous Groups*

LDDB - *London Dysmorphology Database*

MALDI-TOF MS - *Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry*

MAPH - *Multiplex Amplifiable Probe*

Hybridization
Mb - *Megabase(s)*
MCA - *Multiple Congenital Anomalies*
MDS - *Multi-Dimensional Scaling*
MEBD - *Muscle–Eye–Brain Disease*
MeSH - *Medical Subject Headings*
MGD - *Mouse Genome Database*
MIM - *Mendelian Inheritance in Man (see also OMIM)*
MIMMAP - *Mendelian Inheritance in Man MAP*
MIPS - *Munich Information center for Protein Sequences*
MLC - *Mouse Locus Catalog*
MLPA - *Multiplex Ligation-dependent Probe Amplification*
MPSS - *Massively Parallel Signature Sequencing*
mRNA - *messenger RNA*
MS - *Mass Spectrometry*
NCBI - *National Center for Biotechnology Information*
NCHGR - *National Center for Human Genome Research*
NLM - *National Library of Medicine*
NMR - *Nuclear Magnetic Resonance*
OMIM - *Online Mendelian Inheritance in Man*
OPD - *Oto-Palato-Digital (syndrome)*
OSMED - *Oto-Spondylo-MegaEpiphyseal Dysplasia*
PAM - *Point Accepted Mutations*
PCR - *Polymerase Chain Reaction*
PDB - *Protein Data Bank*
PFAM - *Protein FAMilies (database)*
PH - *Pallister–Hall*
PIR - *Protein Information Resource*
PIV - *Polydactyly, Imperforate anus, and Vertebral anomalies*
POSSUM - *Pictures of Standard Syndromes and Undiagnosed Malformations*
QTL - *Quantitative Trait Locus*
RefSeq - *Reference Sequence (database)*
REMTREMBL - *Remaining TrEMBL*
RNA - *Ribonucleic Acid*
RNAi - *RNA interference*
RT-PCR - *Reverse Transcription-Polymerase Chain Reaction*
SAGE - *Serial Analysis of Gene Expression*
SLO - *Smith–Lemli–Opitz*
SMD - *Stanford Microarray Database*
SNOMED - *Systematized NOmenclature of MEDicine*
SNP - *Single Nucleotide Polymorphism*
SPTREMBL - *SwissProt TrEMBL*
SRS - *Sequence Retrieval System*
SSCP - *Single Strand Conformation Polymorphism*
STS - *Sequence Tagged Site*
SWISSNEW - *SWISSprot NEW (updates)*
SwissProt - *Protein knowledgebase*
TBASE - *Transgenic/Targeted mutation dataBASE*
TBC - *TuBerCulosis*
TDT - *Transmission Disequilibrium Test*
TPC - *Trismus-PseudoCamptodactyly (syndrome)*
TrEMBL - *Translated EMBL*
TREMBLNEW - *TrEMBL NEW (updates)*
TTD - *TrichoThioDystrophy*
TX - *OMIM full-TeXt field*
UCSC - *University of California Santa Cruz*
UMLS - *Unified Medical Language System*
UniProt - *United Protein (databases)*
UPGMA - *Unweighted Pair Group Method with Arithmetic Mean*
URL - *Uniform Resource Locator*
WHO - *World Health Organization*
WWW - *World Wide Web*
XP - *Xeroderma Pigmentosum*

Chapter 1

General introduction

1.1 Genetic disease

Genetic disorders are caused by abnormalities in the genetic material (or DNA[1]). These abnormalities can be located in a specific piece of DNA, or gene, which encodes instructions on how to make proteins. Essential life functions are performed by these proteins and they make up the majority of cellular structures in all cells. Changes in a gene are referred to as mutations. Genetic diseases can be caused by a mutation in a gene, but also by other abnormalities in the DNA. In general, four different types of genetic disorders can be distinguished:

1. Monogenetic

Monogenetic (also called Mendelian or single gene) disorders are caused by a mutation in a single gene. A mutated gene usually results in a mutated protein, which can no longer carry out its normal function. Single-gene disorders are manifest in about 1 out of every 400 newborns (in Northern-Netherlands, www.eurocatnederland.nl). However, as life progresses these numbers may increase considerably due to mutations that have effects at a later age.

2. Multifactorial

This type of disorders is due to mutations in multiple genes in combination with external factors, such as lifestyle and environment. Some of the most common chronic disorders are multifactorial disorders, among which are cardiovascular disease, diabetes and Alzheimer's disease. Complex patterns of inheritance and the involvement of (often unknown) environmental factors makes it more difficult than in monogenetic disorders to identify the genes, and to ascertain the risk of carrying or passing on the disorder.

3. Chromosomal

The complete DNA of an individual is known as the genome and can be found in almost every human cell in distinct structures called chromosomes. The chromosomes are located in a special compartment of the cell called the nucleus. A nucleus holds 23 chromosome pairs, of which 22 have one identical chromosome inherited from each parent. Two chromosomes determine the sex of the individual; the sex chromosomes X and Y. Females have two of the same kind of sex chromosome (XX), while males have two distinct sex chromosomes (XY). Abnormalities in the chromosomal number or structure e.g. (partial) deletion, extra copies, breakage and (partial) rearrangements, can result in disease.

4. Mitochondrial

Apart from the chromosomes, a tiny portion of the human DNA is present in small compartments in the cell known as mitochondria that functions as the power plant of the cell. Genetic disorders due to mutations in the mitochondrial DNA are rare. Mitochondrial DNA mutations can affect both male and female, but are transmitted only via the mother.

[1] See also the list of abbreviations section.

Overall, genetic disorders (monogenetic and multifactorial) affect two to three infants in every 100 births (www.eurocatnederland.nl). Although congenital abnormalities are less common than diseases like AIDS and tuberculosis (TBC), they are important. Genetic conditions are a major cause of neonatal and childhood diseases and of infant death in the western world[1]. Furthermore, studies of congenital developmental disorders have made major contributions to our understanding of the developmental process and the function of the genes and proteins involved[2].

1.2 Phenotype classifications

An adequate description of the clinical phenotype is essential for diagnosis and research. It can help to group and classify patients with the same disorder. Congenital anomalies can be divided in abnormalities and minor variants[3]. Embryonic development consists of several (more or less virtual) phases. Aberrations in these phases have a specific clinical outcome. A commonly used classification system of congenital anomalies distinguishes between malformations, deformations, disruptions, syndromes, sequences, and associations. An intrinsically abnormal development process resulting in morphological defects of an organ is called a malformation. Malformations are caused by genetic and/or environmental factors. Congenital anomalies account for approximately 2-3% of the (live births) newborns (www.eurocatnederland.nl). For multiple malformations this is 1%[4]. The complexity of a malformation depends on the time of onset, early defects lead to more serious consequences.

Deformations are anomalies caused by non-disruptive mechanical forces that distort normal developing structures. They usually originate in fetal life and can be the result of both maternal or fetal factors. The forces that lead to deformations can be extrinsic for example by an amniotic tear or by crowding in the case of twin foetuses. Intrinsic defects can also generate a deformation, for example when malformation of the urinary tract leads to insufficient amniotic fluid. Foetuses that grow in a uterine environment where not enough amniotic fluid is present (oligohydramnios) have a flattened face due to compression of the face against the uterine wall.

Disruptions are structural defects, which are caused by interference with a genetically normal development. They can result from events like infection or amniotic bands.

Minor anomalies are deviations of the process of developmental fine-tuning, called phenogenesis. Minor anomalies are structural variations that do not cause significant functional impairment. They divide into rare variations and common variants based on their prevalence and implication. Rare variants have a prevalence lower than 4%, whereas common variants are above 4%[3].

The clinical outcome of any of these events may result in multiple congenital anomalies (MCA). Relations between anomalies can be classified with the terms syndrome, sequence, and association. A sequence occurs when a single developmental defect results in a chain of secondary (tertiary) defects. The group of defects can be traced back to the original event. Syndromes are anomalies that contain multiple malformations due to a single known underlying cause. The expression of defects in syndromes is variable and diagnosis of a syndrome depends on recognizing the overall pattern of the anomalies. An association is not a sequence or syndrome, but a statistically defined non-random group of anomalies.

Various medical classification systems are in use and are being developed. Each system emphasizes different aspects and has a different purpose. The International Classification of Disease (ICD) is the oldest classification system and is maintained by the World Health Organization (WHO; http://www.who.int/classifications/icd/en/). This system was designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. The 10th version of ICD (ICD-10) is adopted worldwide to describe causes of death. A clinical modification of ICD (ICD-9-CM/ICD-10-CM) has become a diagnostic classification framework for all general epidemiological and many health management purposes. Systematized Nomenclature of Medicine (SNOMED) Clinical Terms (CT) is a reference terminology that is maintained by SNOMED International, a division of the College of American Pathologists (CAP) (http://www.snomed.org/). SNOMED-CT is designed to consistently describe medical files that contain diagnoses and treatment. Both ICD and SNOMED-CT are general systems for clinical coding of disease features of conditions. The London Dysmorphology Database (LDDB) provides a more specific overview of genetic diseases and their definitions/nomenclature[5]. Unfortunately, to date there are no internationally accepted terms for defining congenital abnormalities. Several attempts have been made to come to a universally accepted definition of global terms like malformation, deformation, and syndrome, but no consensus has been established. This also applies to definitions of individual defects.

In the absence of standardized nomenclature, clinical information is mainly available in a free text format in the literature (PubMed)[6]. Utilizing this information for e.g. extracting, comparing, or indexing, requires thesauri or medical language systems. The Medical Subject Headings (MeSH) Thesaurus was designed by the National Library of Medicine (NLM), and is a controlled vocabulary used for indexing articles for PubMed[7]. MeSH provides a standardized way to retrieve information that uses the same concepts, but different terminology. In 1986, NLM launched the Unified Medical Language System (UMLS)[8]. The purpose of UMLS is to aid in the development of computer systems, which ‘comprehend’ the meaning of the medical texts. Currently, UMLS contains around a million concepts, which map to MeSH, ICD-9-CM, SNOMED and other coding systems using the UMLS Metathesaurus. This Metathesaurus is a multi-purpose and multi-lingual vocabulary that contains medical concepts, their various aliases, and the relationships among them.

1.3 Gene identification/finding of inherited disease

For every gene there is a task in the cell. Therefore the identification of disease genes is not very different from finding the genes that are responsible for normal functions or normal attributes such as eye or hair colour. The variation or mutation may be within a gene/protein, or within a regulatory part of the genome that for example affects the amount of protein being produced. A mutation in a gene changes the protein, which alters the way the task is usually performed, and this results in a disease.

The identification of disease genes was tremendously accelerated by the invention of the Polymerase Chain Reaction (PCR) technique by Kary Mullis in 1983[9]. Mullis was awarded the Nobel Prize in Chemistry in 1993 for his invention. With PCR unlimited numbers of copies of a strand of DNA could be made. PCR offers the opportunity to specifically multiply a single DNA fragment out of a DNA haystack. Linkage studies and mutation screening became easier with PCR and the number of identified (disease) genes increased rapidly.

In 1989, the National Center for Human Genome Research (NCHGR) was created, first directed by James D. Watson, known as co-discoverer with Francis Crick of the double-helical structure of DNA. By October 1990 the Human Genome Project (HGP) officially started to map and sequence all human DNA. The main focus for the first five years was on creating a detailed genetic map of the human genome. Sequencing of the DNA started in April 1996, the same year the DNA sequence of the first eukaryotic genome (brewer's yeast, *Saccharomyces cerevisiae*) was completed[10]. In a parallel effort, the private company Celera Genomics founded by the Applera Corporation and J. Craig Venter started to sequence the human genome in 1998. Their aim to finish the sequencing effort within two years was met. The competition between the two efforts accelerated the process, and by early 2001, both the International Human Genome Sequencing Consortium and Celera Genomics published a draft version of the sequence[11, 12]. A first analysis of the human genome sequence remarkably showed that there are only 30.000 to 40.000 human genes, and not 100.000 as previously thought[13]. By 2003, the human genome sequence was completed and the number of genes was further adjusted to 20.000-25.000[14, 15]. Note, however, that although a so-called complete DNA sequence of the human species has been published, the sequence is still frequently updated with new and/or rearranged sequences, and some parts are still missing.

With all the genetics technology in place, identification of disease related mutations in Mendelian single gene disorders now mainly depends on having the right patients and families. However, the genetic analysis of complex diseases still remains a difficult task, and most genes for multifactorial disease remain to be discovered.

1.4 Candidate definition

A candidate gene is a gene that is suspected to be involved in a genetic disease. The reason for a gene to be a candidate can be based on its functional characteristics or on the chromosomal position. Various methods are employed to find a list of candidate genes (see paragraph 1.5). Once a list of candidates has been generated it is important to prioritise the candidates such that the lab work is minimized.

Now that a large portion of the human genes has been identified and the sequence is known, it has become easier to identify candidate genes. However, the number of candidate genes identified based on chromosomal position can still be more than one hundred. Therefore, disease gene identification from a list of candidate genes often remains laborious.

1.5 Candidate strategies

The strategies to identify disease genes evolved together with the technical possibilities in molecular biology. The two major strategies that were developed are position dependent cloning and position independent or functional cloning.

1.5.1 Position dependent

As opposed to position independent cloning, in position dependent cloning the gene can be found by genetic methods without knowledge of the disease pathology, even if the biochemical basis is unknown.

1.5.1.1 Positional cloning

Chromosomal allocation of the first human gene was accomplished in 1911 by Wilson. He was able assign the gene for colour blindness to the X-chromosome using phenotypic criteria[16]. In 1986, Nathans *et al.* were the first to characterise and isolate the red, green, and blue visual pigment genes involved in colour blindness[17]. The first autosomal gene mapped was in 1968 by Donahue *et al.* This was the gene for the Duffy blood group system whose function can be compared to other blood systems like ABO and Rhesus[18].

Classical positional cloning is based on pure (cyto)-genetic data to pinpoint the defective gene to a particular chromosome or sub-chromosomal location. Various steps are involved in the process: **a)** Assignment of a genetic disease gene to a small sub-chromosomal candidate region by linkage analysis. **b)** Generation of a physical map of this region based on clones. **c)** Identification of coding sequences within the region. **d)** Prioritise genes as candidates for mutation screening. **e)** Testing the candidate genes for mutations. In 1986, the first gene was identified with this approach[19]. By 1995, about fifty disease genes were identified using positional cloning. The positional cloning approach was expediting gene identification and was increasingly favoured over other functional strategies[20].

With the start of the human genome project in 1990, more detailed mapping information was released. This combination of positional cloning and the mapping information release resulted in a slow but sure shift from genes identified with classical positional cloning towards positional candidate gene identification. In 1997, the majority of the genes were identified with this positional candidate gene strategy[21]. Completion of the human genome draft sequence in 2001 allowed identification of genes *in silico*, speeding up the positional cloning process (see also 1.6.5).

1.5.1.2 Chromosomal aberrations

In rare, but important patients, a chromosomal aberration can point directly to the gene or genes involved in the genetic disease. Various aberrations can occur, such as translocations or inversions. In balanced chromosomal changes genetic material is evenly exchanged with no extra or missing information. Such aberrations are particularly useful in disease gene identification. Breakpoints are often located within the causative gene itself or in the gene's vicinity. Rarely, these functional segments can be located as far as 1 Mb from the gene[22, 23]. This is known as a position effect. Fortunately for geneticists, such long-distance effects are rare. Unbalanced chromosomal aberrations are also useful for gene identification, as they can indicate a chromosomal region of interest. They do not, however, point directly to the causative gene. Because a varying amount of genetic material is lost in unbalanced abnormalities, the causative gene is not the only one affected but frequently other genes in the neighbourhood as well. When large chromosomal sections are lost this commonly results in (severe) mental retardation[24]. Mental retardation is a frequent feature of syndromes and especially in X-linked disorders. Over 140 X-linked syndromic forms have been described so far[25].

Sometimes, patients suffer from different genetic disorders at the same time. This can be the result of a chromosomal deletion of a contiguous gene set[26-29]. Small chromosomal deletions can be easily missed using classic microscopy analysis. These microdeletions can be detected by techniques like array-based comparative genomic hybridization (arrayCGH) or FISH[30, 31].

1.5.2 Position independent

In some occasions there is some knowledge of the disease pathology, which can be used to identify the disease gene without positional information.

1.5.2.1 Functional cloning

This strategy relies on the presumed biological function of a disease gene, as predicted from the disease phenotype. For this strategy it is necessary to understand the biochemical and/or pathogenetic background of the disease. Efforts using this approach may start with a (partially) purified protein, which can be used to deduce a partial amino acid sequence. These sequences can then be translated back to possible DNA sequences that in turn can be used to raise antibodies or to design primers for cDNA library screening. Before 1980, the majority of the disease genes found, were identified by this approach, because none or very limited genetic mapping information was available. Diseases with a clear (biochemical) change, such as phenylketonuria and haemophilia, were elucidated first[32, 33]. Later on, newly developed techniques of positional cloning quickly became the preferred strategy (see paragraph 1.5.1.1). Still today, functional cloning is used and providing clues for the genes involved in disease[34, 35].

1.5.2.2 Homologous phenotypes in animal models.

Knowledge from animal models like mouse (*Mus musculus*) and the fruit fly (*Drosophila melanogaster*) has aided substantially in finding disease genes[36]. Information from model organisms to identify disease genes can be utilized in various ways. In some instances genes for a specific phenotype were localized in the mouse and then mapped back on the human genome via a so-called Oxford-grid (see figure 1). This grid shows what regions of human chromosomes correspond to mouse chromosomal regions. If natural or induced mutations in the mouse gene show phenotypic similarity with the human disorder under investigation, then these orthologous genes can become candidates because they map into the disease locus[37].

1.5.2.3 Gene expression.

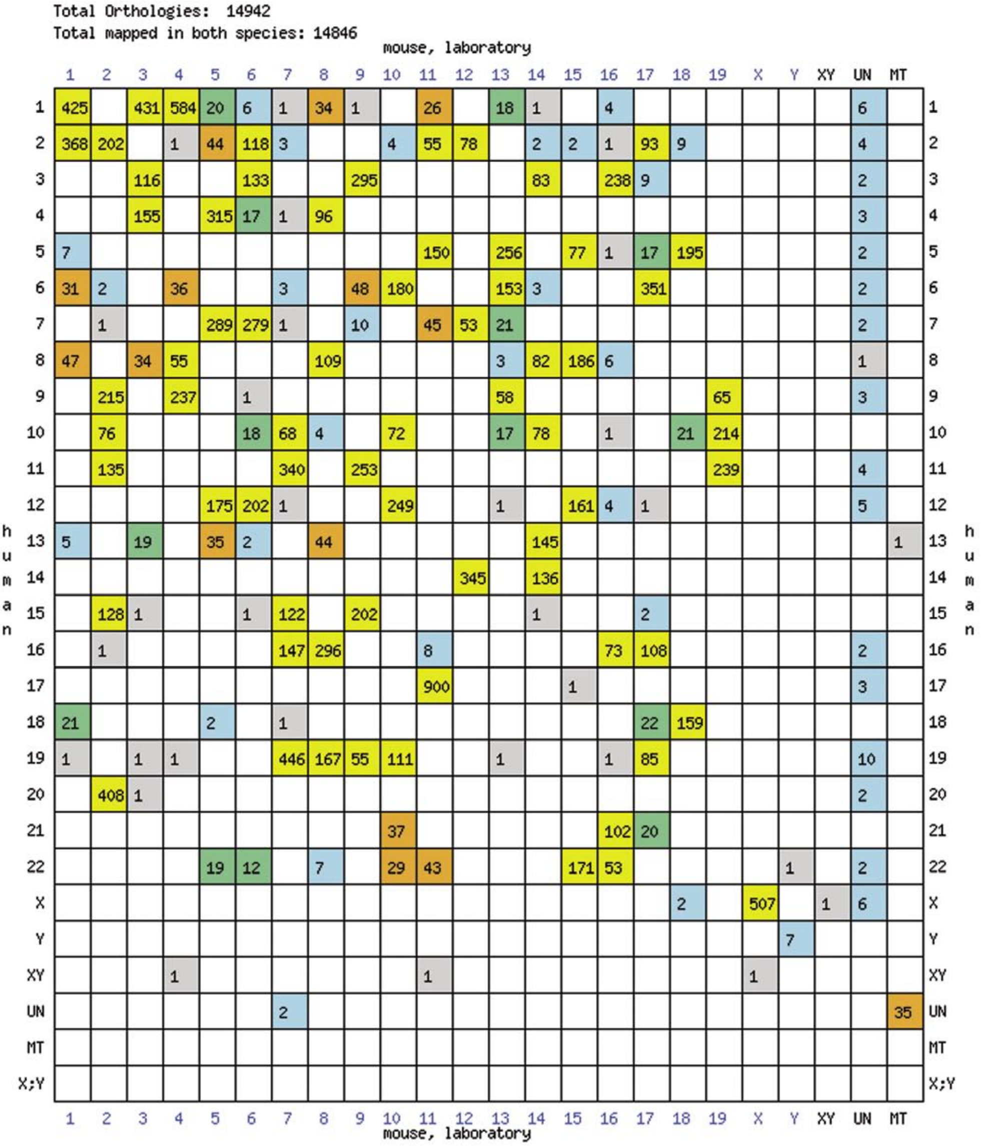
Methods that utilize a presupposed specific characteristic of the disease genes have proven to be successful. One approach for example tries to enrich for genes that have a specific function in the tissue(s) affected in the disease[38, 39]. Sequence characteristics can sometimes be predicted, for instance in the case of expanded tri-nucleotide repeats in several neurological diseases[40].

Aberrant levels of gene transcription can point to a disease gene. When a candidate gene encodes an mRNA which is quantitatively or qualitatively different in patients with the disease compared to a control group, one might identify this gene by a genome-wide expression screen. Various methods to screen for these gene expression differences were developed during the last decades (see also paragraph 1.7).

The microarray methods are increasingly used to measure the levels of expression. A microarray is a small two dimensional array of deposited or synthesized genes or gene fragments. The array carrier is typically glass, silicon wafer, or filter. Since the order of the samples is known it is possible to screen the array with an applied DNA or RNA sample, in a high-throughput and parallel manner. Initially, in microarray experiments only DNA/RNA samples were profiled, but today also antibodies or proteins can be tested.

1.5.3 Disease gene confirmation

To prove that the candidate is in fact the gene, demonstration of a genetic mutation is needed. Mutation analysis in small patients groups can be done by direct sequencing. For screening larger groups other methods are usually more cost efficient: Single Strand Confirmation Polymorphism (SSCP)[41], Denaturing Gradient Gel Electrophoresis (DGGE)[42],



Total Orthologies: 14942
Total mapped in both species: 14846

mouse, laboratory

human

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y	XY	UN	MT	
1	425		431	584	20	6	1	34	1		26		18	1		4							6		1
2	368	202		1	44	118	3			4	55	78		2	2	1	93	9					4		2
3			116			133			295					83		238	9						2		3
4			155		315	17	1	96															3		4
5	7										150		256		77	1	17	195					2		5
6	31	2		36			3		48	180			153	3			351						2		6
7		1			289	279	1		10		45	53	21										2		7
8	47		34	55				109					3	82	186	6							1		8
9		215		237		1							58						65				3		9
10		76				18	68	4		72			17	78		1		21	214						10
11		135					340		253										239				4		11
12					175	202	1			249			1		161	4	1						5		12
13	5		19		35	2		44						145										1	13
14												345		136											14
15		128	1			1	122		202					1			2								15
16		1					147	296			8					73	108						2		16
17											900				1								3		17
18	21				2		1										22	159							18
19	1		1	1			446	167	55	111			1			1	85						10		19
20		408	1																				2		20
21										37						102	20								21
22					19	12		7		29	43				171	53					1		2		22
X																		2		507		1	6		X
Y																					7				Y
XY				1							1									1					XY
UN							2																	35	UN
MT																									MT
X;Y																									X;Y
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y	XY	UN	MT	

mouse, laboratory

Figure 1 An Oxford-grid. The grid shows the relationship between human and mouse chromosomes. Chromosome location of either of the species often predicts the chromosome location in the other species. The colours indicate the number of orthologies: Grey (1), Blue (2-10), Green (11-25), Orange (26-50), Yellow (>50). From the Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine. (http://www.informatics.jax.org, April, 2005). (Colour version: see appendix 2)

Denaturing High Performance Liquid Chromatography (DHPLC)[43], Chemical Cleavage Mismatches (CCM)[44, 45], Enzymatic Cleavage Mismatches (ECM)[46]. Comparative sequencing strategy based on Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry (MALDI-TOF MS) analyses can be used for high-throughput discovery of genomic sequence polymorphisms[47]. Copy number changes in genomic DNA can be detected using e.g. Multiplex Amplifiable Probe Hybridization (MAPH) and Multiplex Ligation-dependent Probe Amplification (MLPA). Both techniques are suited to quantify up to fifty loci in a single reaction[48-50].

Changes in the splicing process of the gene may be missed when screening protein-coding DNA sequences only, but are detectable at the RNA level using RT-PCR. With RT-PCR and related methods it is possible to evaluate whether the spatio-temporal gene expression pattern is compatible with the phenotype of interest. Ultimately, final proof may require a functional test, i.e. examination of the effect of induced mutation in model organisms or the restoration of a normal phenotype by adding the normal gene.

1.5.4 Complex disorders

The focus of disease identification is shifting from Mendelian traits to complex disorders. Usually complex disorders are multifactorial and many such diseases, like heart and vascular disease are quite common. Variation in the phenotype among individuals is part of the evolutionary adaptation through (natural) selection, and determines (partially) the vulnerability of an individual to disease[51, 52].

In general the following steps are applicable to research of a complex disease. **a)** Establishing that the disease is indeed (partially) caused by genetic factors. Some traits can be influenced by e.g. the family environment (like behaviour or diet). **b)** Perform segregation analysis to determine the type of inheritance. Inheritance can vary from Mendelian to polygenic, depending on penetrance and environment. The mode of inheritance determines the linkage analysis methods that are applicable. **c)** Linkage analysis to map the susceptibility loci. In Mendelian diseases a parametric linkage model can be used, because the precise genetic model is known. In complex diseases this model is usually not clear and a non-parametric method will be preferred. These methods trace shared chromosomal regions among affected individuals. Examples are 'Identity-By-Descent' (IBD) or 'Affected Sib Pair' analysis (ASP). **d)** Population-association studies to fine map the susceptibility gene. Association methods, like the transmission disequilibrium test (TDT), create associations between unrelated individuals, whereas linkage analysis works only within families. **e)** Elucidate the DNA sequences/genes; confirm their molecular and biochemical action and involvement. All of this is not as straightforward as it is in Mendelian disorders, because a single change is not sufficient to cause the complex disease. Susceptibility is often modelled as a quantitative trait locus (QTL), rather than a single DNA change[53].

1.6 Bioinformatics approach to disease gene identification

The human genome project and the release of the genomic sequences of a continuously increasing number of other species provide the opportunity to analyse the organization of the human genome. Genomic sequences, as well as full-length cDNA sequences, expressed sequence tags (ESTs) and large-scale expression micro-array data of model organisms like *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Mus musculus*[54-57] offer invaluable resources for studying (human) genes. These encompass gene transcription and translation, which are referred to as the transcriptome and proteome when studied for the whole organism. In addition, understanding gene evolution offers prospects for functional assessments. Addressing these different aspects is far from trivial in view of the fact that very large amounts of data are stored in numerous different databases (see appendix 1), yet much of the data suffers from chronic incompleteness and errors. This makes the use of (high-performance) computing an essential tool for decoding the information that is contained in these databases.

1.6.1 Sequence databases

Primary sequence databases are the major repositories for nucleic acid sequences, and together form the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ)[58], the European Molecular Biology Laboratory database (EMBL)[59], and GenBank[60] at National Center for Biotechnology Information (NCBI, Bethesda, MD, USA). These three organizations exchange data on a daily basis. Similarly, protein sequences are stored in SwissProt and PIR[61, 62]. Connections between the nucleic acid and protein databases based on translations from annotated coding regions in GenBank and EMBL, are stored in UniProt[61] and RefSeq[63]. Other sequence directed data encompass protein family, domain and motif organization or classification as deposited in e.g. PFAM[64] or integrated in the InterPro database[65]. Structural information at different levels of detail can be found in e.g. the PDB and DSSP[66, 67].

1.6.2 Sequence related databases

Functional data linked to genes and proteins cover several molecular and cellular aspects. Annotation and/or experimentally derived data is deposited in the form of metabolic pathways (KEGG, Reactome)[68, 69], spatio and/or temporal expression patterns (Unigene, SMD)[6, 70], and protein-protein interactions (BIND, HPRD)[71, 72].

Human genetics traits and phenotypes are described in the literature (PubMed)[6], but the main resource is the Online Mendelian Inheritance in Man (OMIM) database[73], developed by Dr. Victor A. McKusick and his colleagues at Johns Hopkins University (Baltimore, Maryland, USA). This catalogue contains over 15.000 entries, describing both human genetic traits (~5.000) and human genes/proteins (>10.000). OMIM also holds genetic variation data linked to diseases, but in this respect its scope is limited. Mutation data is also deposited in specialist databases like the Human Gene Mutation Database (HGMD)[74] and in the Single Nucleotide Polymorphism database (dbSNP)[6].

1.6.3 Retrieval systems and browsers

In addition to the resources described in the previous paragraphs there are gene (group) specific, disease specific, species specific, and computational result databases. Although most databases have a web interface to access them, retrieval systems like NCBIs Entrez[75], the widely used Sequence Retrieval System[76], or the more recently developed MRS[77] provide easy and fast access to a collection of different databases. The main focus of these retrieval systems is fetching a set of database entries that meet the user query. Three genome browsers, NCBIs map viewer, ENSEMBL, and the UCSC browser now provide multi-species interfaces and integrate the major data repositories[6, 78, 79]. Furthermore, they continually analyse the genomes and keep track of annotations (figure 2).

1.6.4 Sequence analysis

1.6.4.1 Homology

The identification of a disease gene is not always straightforward. Sophisticated algorithms and tools are required to understand the data. One way to understand more about the sequence of interest is to compare it with other known sequences, using sequence analysis tools. The French molecular biologist François Jacob described Nature as a tinkerer and not an inventor[80]. New functions require new proteins, which require new genes. Rather than designing sequences *de novo*, new genes are adapted during evolution from existing sequences. These adaptations are a gradual process and often we can recognize similarity between new and already known sequences. Sequences that show significant sequence similarity are thus considered to be related and we can transfer information from the known to the new sequence. Two or more sequences are said to be homologous if they are alike because of shared ancestry.

1.6.4.2 Pairwise alignment

A most basic task in sequence analysis is the comparison of two sequences. But, evaluating the two sequences like two character strings is only possible if the sequences are very much alike. Crucial is the concept of alignment, because sequences can have insertions, deletions, duplications, inversions, substitutions, etc. Insertions and deletions are considered gaps and must be dealt with, usually by a penalty score for starting and extension of a gap. Furthermore,

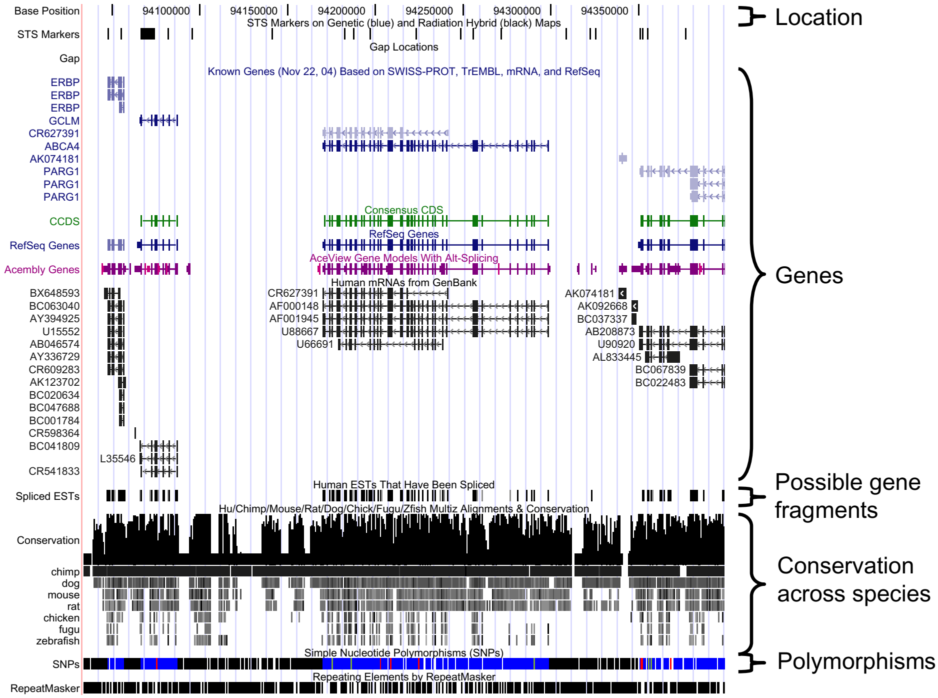


Figure 2 Partial screenshot of the UCSC genome browser[79]. The (horizontal) annotation tracks of a subchromosomal region are shown (Chr. 1: 94.033.925-94.399.032 Bp; Human May 2004 assembly). The vertical lines indicate the position on the chromosome. Many items are clickable, giving detailed information. More information on the tracks can be found at http://genome.ucsc.edu (April, 2005).

some amino acids in protein sequences can be replaced without significantly disrupting the protein structure and thereby its function while other changes are severe. Almost all sequence alignment algorithms use a scheme to score in such a way that the resulting alignment is optimally biologically relevant. Other aspects such as three dimensional protein structure and evolutionary relationship of the sequences are also addressed. A number of schemes with different characteristics and applicability have been published: Dayhoff/PAM and BLOSUM[81-83].

Finding the optimal alignment by making all possible alignments including gaps and scoring them is computationally very intensive, even for only a moderate number of sequences. Algorithms that find optimal alignment while reducing computation time were initially based on so-called dynamic programming. Dynamic programming is a general algorithm for solving certain optimization problems. “Programming” does not refer to a computer program, but is mathematical jargon for using a fixed set of rules to arrive at a solution. Dynamic programming was first applied for sequence alignments in 1970 by Needleman and Wunsch[84].

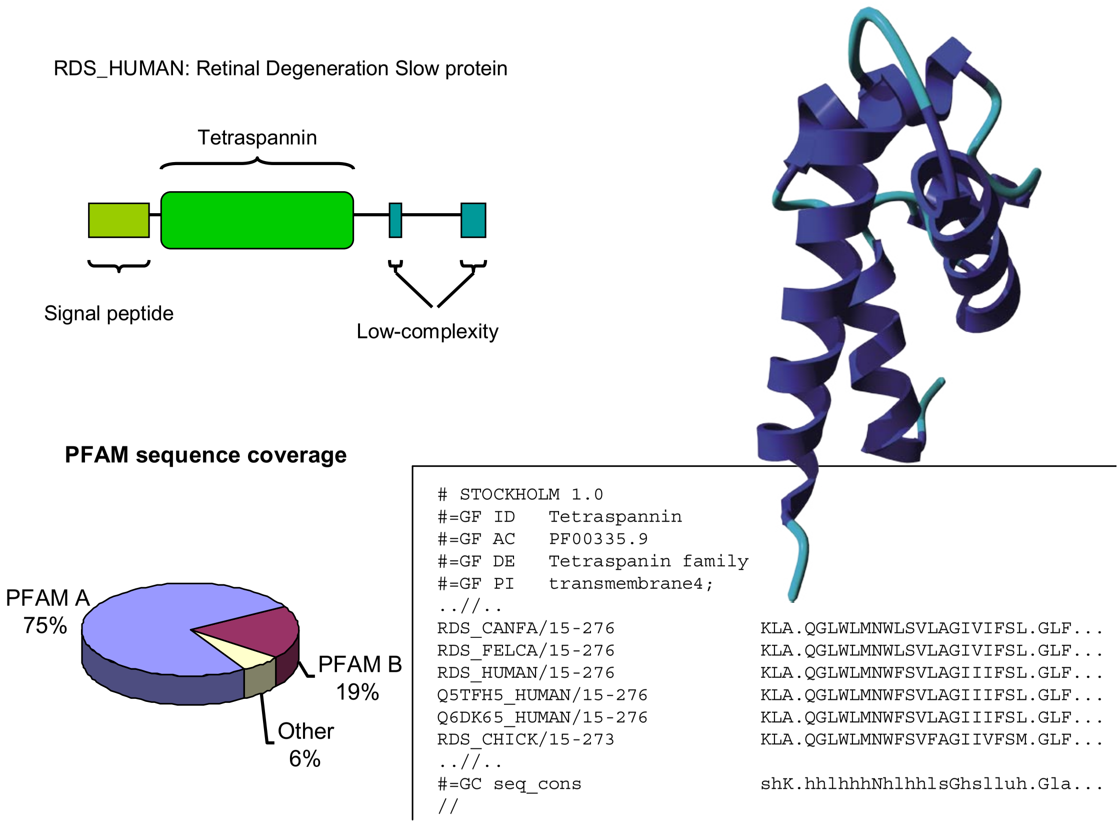
The Needleman-Wunsch algorithm is designed to align two sequences head-to-tail, a so-called global alignment, allowing gaps. Because the two sequences are treated as potentially equivalent, the main application for this algorithm is to identify conserved regions and differences, e.g. two genes with similar function but of different species.
Another application is to use alignments to identify conserved motifs/domains or for finding a smaller sequence in a genomic sequence. In this situation, two sequences are not necessarily related, and the best scoring alignment is called the local alignment. Smith and Waterman developed such a local alignment algorithm using dynamic programming[85].

1.6.4.3 Heuristic homology searches

To identify all sequences in a database that are broadly similar to the gene of interest using the Smith and Waterman algorithm requires a lot of computational resources and time. The Basic Local Alignment Search Tool (BLAST)[56] uses a different method to find candidates for alignment without making the actual full alignment during the search through the database. BLAST is a heuristic alignment algorithm, which is faster than Smith-Waterman but is not guaranteed to find the optimal alignment according to the specified scoring scheme. Heuristic alignment algorithms search for small identical sequences in the database to identify possible high scoring alignment matches. Another heuristic sequence search tool is FASTA[86] that can search a sequence against nucleotide and protein databases. FASTA can be very specific when identifying long regions of low similarity particularly for highly diverged sequences. BLAST is faster than FASTA and performs equally for highly similar sequences, but less well for very diverged sequences. Various adaptations to the BLAST algorithm have been made to increase the sensitivity for specific problems without losing too much speed[87, 88].

1.6.4.4 Probabilistic models

Computational sequence analysis may have difficulty determining the alignment significance and is often plagued by low signal-to-noise ratios. To circumvent this, probabilistic models like hidden Markov models (HMMs) are currently used. They provide a general approach to the types of statistical problems that often occur in sequence analysis[89]. HMMs are used to model a single sequence or a family of sequences like protein domains. Proteins may contain structural domains that are independent from the rest of the protein in terms of folding and stability. Such domains are found in multiple proteins, and are named according to their biological function. The PFAM database contains multiple protein sequence alignments (figure 3), functional annotation, and profile hidden Markov models for protein domain families[64, 90]. A profile HMM is a mathematical description of a domain, which can be used to search for that domain in new sequences or to identify those sequences in a database.

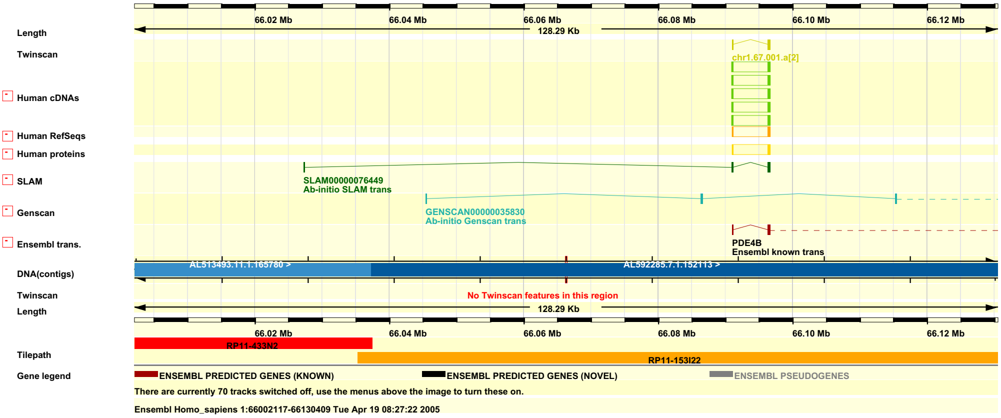


```
# STOCKHOLM 1.0
#=GF ID   Tetraspannin
#=GF AC   PF00335.9
#=GF DE   Tetraspanin family
#=GF PI   transmembrane4;
..//..
RDS_CANFA/15-276              KLA.QGLWLMNWLSVLAGIVIFSL.GLF...
RDS_FELCA/15-276              KLA.QGLWLMNWLSVLAGIVIFSL.GLF...
RDS_HUMAN/15-276              KLA.QGLWLMNWFSVLAGIIIFSL.GLF...
Q5TFH5_HUMAN/15-276           KLA.QGLWLMNWFSVLAGIIIFSL.GLF...
Q6DK65_HUMAN/15-276           KLA.QGLWLMNWFSVLAGIIIFSL.GLF...
RDS_CHICK/15-273              KLA.QGLWLMNWFSVFAGIIVFSM.GLF...
..//..
#=GC seq_cons                 shK.hhlhhhNhlhhlsGhslluh.Gla...
//
```

Figure 3 The tetraspannin protein domain covers a large portion of the RDS protein (UniProt: RDS_HUMAN). A selection of the PFAM database record of this domain is shown at the lower-right (PFAM: PF00335). The 3D structure of the domain has been partially elucidated (top-right; PDB: 1G8Q (large extra-cellular loop); Graphics made with YASARA: http://www.yasara.org/). Version 17 (March 2005) of PFAM contains 7868 protein domains/families and consists of PFAM A (curated), PFAM B (automatically generated) and other (bottom-left).

1.6.5 *In silico* positional cloning

Once the critical region for a genetic disease has been determined by linkage analysis or other positioning methods, the human genome sequence can be used to identify positional candidate disease genes. Genome browsers, biological databases, and other bioinformatics tools all contribute to the gene finding strategy. The first step is to search for all genes between two genetic markers on the chromosome under study. Numerous genes were positionally cloned in projects that depended on the genome sequence, even using only the draft version[13]. Essential for *in silico* positional cloning is a proper description of the location of the genes and of other annotations like regulatory elements. Computational algorithms and tools have been developed to identify all genes on the human genome sequence[91, 92]. None of these is perfect and genes may be missed, or false genes have been found and annotated. In practice, this means that for a predicted gene different sources have to be evaluated manually. Criteria are whether the gene overlaps with ESTs, whether homologous sequences from other species exist, and whether the intron-exon structure is correct for proper RNA to protein translation. Sequence analysis tools like BLAST help with this manual survey. The three major genome browsers perform multiple sequence analyses and prediction programs and present the integrated results to the user (figure 4).



No Twinscan features in this region

There are currently 70 tracks switched off, use the menus above the image to turn these on.

Ensembl Homo_sapiens 1:66002117-66130409 Tue Apr 19 08:27:22 2005

Figure 4 Partial screenshot of the Ensembl genome browser[78]. Shown is a part of chromosome 1: 1:66.002.117-66.130.409 bp. The view is similar to the UCSC browser (figure 2). Different transcripts are predicted by GenScan, SLAM, and TwinScan[91-93]. The exons (vertical blocks) predicted by SLAM can be confirmed by other sources (e.g. cDNA). From http://www.ensembl.org/ (April, 2005).

1.6.6 *In silico* functional cloning and candidate gene selection

In theory, every gene within the disease critical region can cause the disease. When the critical region is large or the gene density is high, positional candidates are many. Screening all these genes in the wet-lab is very labor intensive and thus expensive. Therefore it is more favourable to perform an *in silico* selection, combining positional information and function prediction. There may already be a suspicion on the biochemical and/or pathogenic background of the disease. This can be taken as a starting point for the selection of candidate genes.

The expression pattern of the gene is likely to overlap the main affected tissues in the disease. If a genetic eye disorder affecting the retina is studied, the gene should be expressed in retina. It is possible to search the database for (all) genes expressed in retina e.g. when no linkage data is available[94, 95]. This *in silico* strategy can be used to generate candidate disease genes[96, 97].

Domain, motif, and profile databases and tools permit a scan for other expected biochemical and cellular function characteristics. Possible examples are transporter, trans-membrane, enzyme, structural, or DNA binding function. Local genome organization and known regulatory elements can be useful in disease gene identification[98-101].

For known genes, the knowledge in literature (e.g. PubMed) and the OMIM database can be used to select the candidate genes. If there is already a gene causing a related disease in the critical region, it is possible that the disorders are allelic, that is they are caused by different mutations in the same gene. Further, genes located within the critical disease region that have a functional similarity and/or relation with genes causing related diseases, can be considered good candidates.

Knowledge of model organisms makes comparative candidate gene selection possible. This situation applies when a gene is known, which causes a similar phenotype in another species. This is a powerful indicator for selecting a candidate gene. Yet, a direct comparison between the phenotypes in human and the model organism can be complicated because of the different anatomy. Transfer of knowledge by phenotype is most straightforward in other mammalian species like *Mus musculus* that are evolutionarily close to humans. An example is the disease gene identification in Ectrodactyly-Ectodermal Dysplasia-Clefting syndrome (EEC)[37].

1.6.7 Candidate gene prediction

Some genes are more likely to be involved in a specific disease than others based on the knowledge of that gene. This information is not always available, and various methods can be used to predict candidate genes for diseases.

1.6.7.1 Sequence based prediction

Which gene underlies a disease may not be immediately obvious from the list of candidates. When none of the known genes has mutations, one may try to find new genes in the critical region. Methods for the identification of new genes rely on sequence signatures that distinguish coding from non-coding sequences. Among the most frequently used discriminating characteristics are splice site motifs, which are found in the majority of the human genes. At the 5’ end of an intron a GT is found, whilst AG is found at the 3’ end. Algorithms to find exons also evaluate other characteristics. A single genome predictor algorithm like GENSCAN uses a general probabilistic model of the gene structure including transcriptional, translational and splicing signals, length distributions, and compositional features of exons, introns and intergenic regions[91]. The algorithm used in the TWINSCAN program similarly predicts exons but improves the accuracy by taking advantage of genome comparisons, and is called a dual genome prediction algorithm[92]. Multiple genome prediction methods have only recently been developed and though not used to their full potential yet, they promise to be more accurate than single and dual models[102].

1.6.7.2 Comparative prediction

Genes, whether newly identified or not, without a clear annotation, can be involved in human genetic disorders. Comparative genome analysis of more distantly related species presents us with a wealth of opportunities for studying evolution and gene/protein function. Although still under intense development, comparative genome analysis is already offering gene/protein function prediction methods at various biological levels.

Homology-based function prediction transfers information from known genes/proteins to unknown sequences and remains the primary method to determine the function of a new gene. However, in most genomes about 30-40% of the genes lack a clear functional annotation[103]. This limits the predictive value of homology-based methods like BLAST, FASTA, and Smith-Waterman. Orthology-based protein function prediction uses information from multiple genomes, which renders it more specific. As orthologs are due to a speciation rather than to a duplication event in evolution, they are more likely to perform the same function in different species. Furthermore, comparison is gene-based instead of sequence based. Orthologs can be determined via various methods and have been stored in databases like Protein World[104], COG/KOG[105, 106], or OrthoDisease[107].

Functionally related genes may have protein-protein interactions that are subject to evolutionary pressure to stay and act together, because a change in one of the proteins impinges on the function of the second (or others). This concept has been used to predict protein function and hence transfer knowledge by functional relation. The tendency to evolve similarly (or complementarily) becomes apparent from the phylogenetic distribution of genes. This strategy has been applied successfully in the elucidation of the function of the frataxin gene involved in Friedreich's ataxia (MIM 229300), which is a neurodegenerative disorder[108, 109].

Prokaryotes use a regulatory structure that controls a number of related genes within a region of the genome, the so-called operon. In prokaryotes genomic organization, order, and proximity of genes are commonly used for function prediction[110]. Eukaryotes lack these operon structures, but genomic-context based prediction may still be possible to some extent[111, 112]. Other methods for function prediction use correlated gain or loss of genes, co-evolution, and conservation of co-expression are reviewed in [113].

1.7 Functional genomics

The completion of the human genome sequence is undoubtedly an important milestone in biology. Completion is a relative term, since an unknown number of genes remain to be identified. Small non-coding RNAs called microRNAs have been shown to play important roles in gene regulation. Only a limited number of microRNAs have been identified, but there are various indications that their current number is much higher[114]. Even, a complete list of all human genes using the methods described will not allow us to comprehend the function of all the genes or their interactions in the cell(s). This means that attention is shifting from identification to functional annotation. In addition, there is a shift from single gene/system to whole genome/cell analyses. In other words there is a shift from a reductionist to a holistic approach. This global analysis of the function of genes is the foundation of functional genomics. Core concepts of functional genomics are the genomic expression or transcriptome, and the resulting proteins or proteomics. Ultimately, functional genomics defined as the functional annotation of the genes, will lead to systems biology or a complete and integrated picture of cellular physiology at the molecular level.

Transcriptomics deals with gene expression in terms of RNA levels, regulation, processing, and turn-over/degradation. Observing expression changes in disease can reveal the role of the gene or its relations with other genes involved. Formerly, RNA expression was determined by transferring the RNA sample to a carrier membrane followed by screening with a radioactively labelled fragment of the gene of interest: Northern blotting[115]. This screening was generally on a gene-by-gene basis. RT-PCR then allowed the reverse transcription of RNAs from typically a specific tissue/cell type into copy DNA (cDNA) and the creation of cDNA libraries. This permitted global expression analysis as well as the identification of genes, which were tissue specific or at least highly expressed genes. Such genes might be involved in diseases affecting that tissue. A range of (differential) expression analysis methods comparing disease and control gene expression have been developed. These include databases analyses for ESTs, differential display PCR, micro array, serial analysis of gene expression (SAGE), and massively parallel signature sequencing (MPSS) [116-119]. Data obtained from such experiments using micro arrays; can reveal the links between the individual genes. Analysis involves clustering techniques and requires special or newly developed software. Transcriptomics has been useful in the identification of disease genes and potential drug targets[120-123].

Proteins are the products of genes and the study of their abundance and molecular function is the subject of proteomics. Often proteins are post-translationally modified by for example partial cleavage or phosphorylation. Furthermore, not every gene transcript results in a protein product[114]. Some genes encode functional RNAs. Both protein modifications result in differences with the transcriptome and this argues for a separate analysis of the proteome. Similar to genes, an analysis of protein expression can reveal interactions and may highlight changes that result in disease. Methods like 2-dimensional gel electrophoresis (2DGE) and mass spectrometry (MS) are used to study protein expression patterns of specific tissue or cellular samples[124, 125].

The 3-dimensional (3D) structure of proteins gives direct insight into their molecular mechanisms, which can be explored for designing drugs or developing antibodies. Structural genomics aims to solve the 3D structures of proteins. X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) are two major experimental methods by which protein structures are resolved[66]. Unfortunately, both methods are very labor intensive and consequently the number of structures solved and stored in the Protein Databank (PDB) is vastly outnumbered by the number of known proteins [126, 127] (figure 5). Fortunately, the number of distinct structures seems to be relatively small, less than 10.000 by most estimates[128]. *Ab initio* prediction of the 3D protein structure, also known as the protein folding problem, is very complicated. Numerous attempts have been undertaken, but so far none has succeeded. Developing faster method for resolving protein structures is now the main focus[129].

The function of a protein may not be (fully) clarified when analyses of its sequence, expression, and even its 3D structure have been done. Obviously, the molecular partners of the protein may provide essential clues. Among other experimental methods affinity chromatography, yeast-two-hybrid analysis, and co-immunoprecipitation might allow isolation of the associated protein(s).

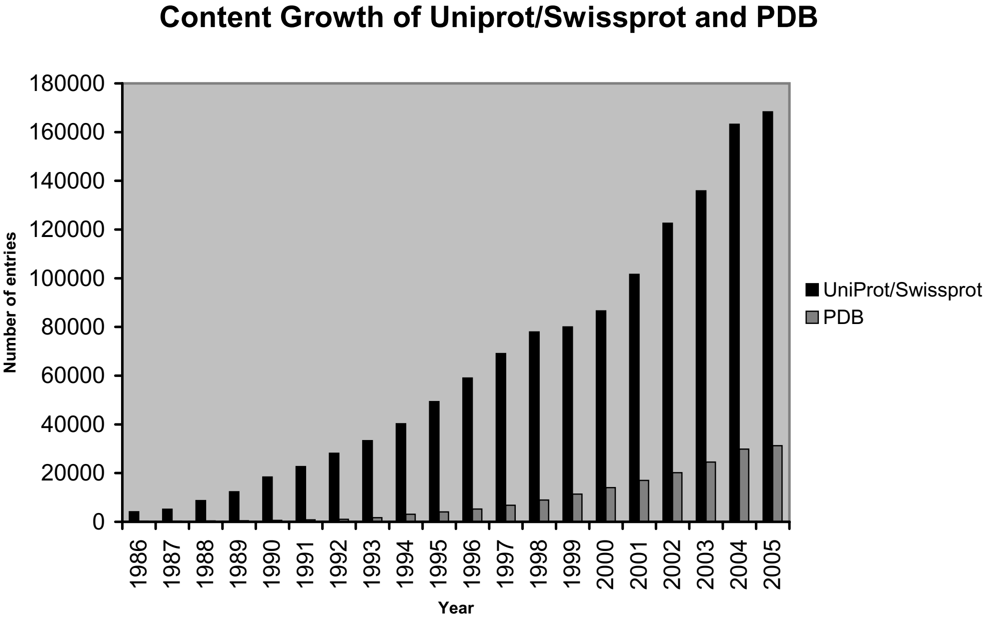


Figure 5 The growth in the number of structures in the PDB and the number of proteins in UniProt/ Swissprot[61, 66]. Note that structure models are included in the PDB. (From http://www.rcsb.org/pdb/ and http://www.uniprot.org/. April, 2005).

1.8 Biological networks

Over the last decade, high-throughput interaction screening methods have been developed and used to map different biological interactions. These maps provide a first glimpse of the overall organization of the molecular networks in biological systems. Biological networks are expected to change the general perception on biological systems such as the relation between genotype and phenotype. The networks model physical or functional relationships ranging from atomic to organism or even population level. For example, protein interaction data can be found in databases like BIND, HPRD, MIPS, DIP[71, 130-132]. Networks consist of nodes and edges, describing e.g. proteins and their interactions respectively (figure 6).

Networks are being generated from data generated by (functional) genomics studies. Such networks are still incomplete and contain multiple errors[133]. Large-scale experiments are conducted to establish the networks. A first glance into the protein-protein interaction/ functional networks for e.g. *Saccharomyces cerevisiae* is available from yeast-two-hybrid and MS studies[134-137]. Another technique, chromatin immunoprecipitation (ChIP), was used to identify binding sites for transcriptional regulators in yeast[138].

The networks indicate that there are general characteristics of biological networks such as the topology that are similar to other complex networks. Similar non-random patterns were identified in networks from the World Wide Web as were present in biological networks[139-141]. In a random network the distribution of the connectivity is even. Metabolic networks are not random. However, they are modular and scale-free, showing a very uneven distribution of connectivity between the nodes with a few very well connected nodes called hubs. Scale-free networks follow a Power law distribution (figure 7). Modularity in biological systems is not a new concept and was previously recognized in e.g. operons, embryonic development, and even proposed by the founders of cell-theory, the German botanist Matthias Schleiden and physiologist Theodor Schwann[142, 143].

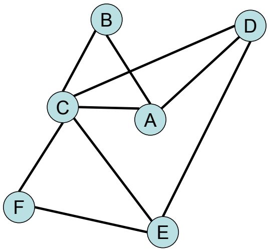


Figure 6 Protein interaction network example. The proteins A-F are called nodes. The interaction relations between the proteins(/nodes) are represented by the lines between them, the edges.

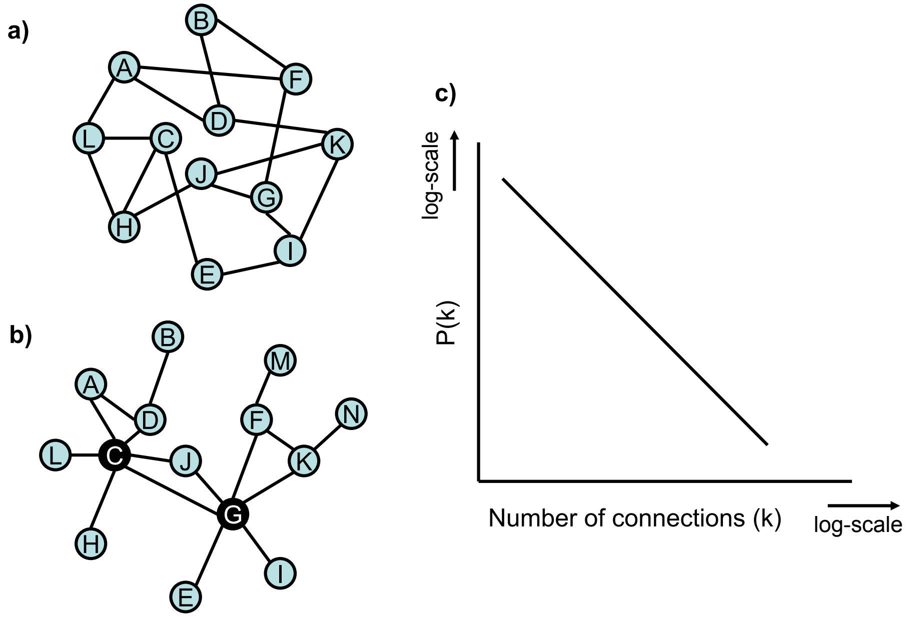


Figure 7 a) Random network: most nodes have approximately the same number of links. The connectivity follows a Gaussian distribution (not shown). b) Scale-free network: some nodes are highly connected (hubs, C and G), whilst other nodes have only a few links. c) Connectivity of a scale-free network follows a power-law distribution.

1.9 Phenotype networks/phenome

A major goal in genetic research is the identification of genotypes that are associated with human phenotypes. That overlapping (syndrome) phenotypes might reflect biological relationships has been recognized for many years[144]. Phenotypes that share major clinical features are called ‘phenotype communities’ or ‘syndrome families’, and can be considered like disease phenotype networks. These phenotypic relations have shown to be helpful to identify new genes or mutations[145].

For complex diseases, much effort is put in elucidating networks that describe genes, proteins, interactions, or function. Although these projects are complex and crucial to understand complex diseases, phenotypes are essential. Phenotypes are the ultimate representation of the genotypes. However, current approaches to define phenotypes are probably inadequate to fully explore the genotypic data. A global network of phenotypes would describe the physical properties of an organism, its physiology, morphology and behaviour. The physical totality of all traits of an organism or of one of its subsystems is called the phenome. Studying the variation of the phenotype among individuals is the subject of phenomics.

Freimer and Sabatti proposed the Human Phenome Project (HPP), an international effort to create phenotypic databases and new methods to analyse such phenotypic data[146].

The human phenome would be beneficial for the understanding of both Mendelian and complex diseases and (the identification of) their causes. Recognition of the importance and the potential of the phenotype also translated into phenome initiatives in a number of other species[147-151]. A primary task for the HPP is to define what constitutes a phenotype. A broad spectrum of definitions in the biological literature for the term phenotype shows that this is not trivial, because it can refer to morphologic, biochemical, physiological or behavioural features[152]. Since covering all aspects of the phenome is too broad, the HPP and similar projects are mainly focusing on (human) disease phenotypes. As highlighted in paragraph 1.2, standardization in terminology and classification is a slow process. New methodology for storage and analysis of phenotypic data awaits development, including (non-parametric) statistical tools for association studies, clustering, data mining, machine- and statistical learning.

1.10 Bioinformatics approach to bio-networks

(Functional) genomics data is the main source for the (re-)construction of bionetworks (see also paragraph 1.8). Analysis of these networks using bioinformatics includes determining the architecture of networks, function prediction, simulation analysis, and other computational approaches.

The several types of networks that emerge are not independent. Rather they form a network consisting of networks, which determine the cell characteristics. Unravelling networks requires bioinformatics. Essential is to model and understand the topological and dynamic properties of the networks. Networks consist of nodes and edges and are called graphs in mathematical jargon. Each network has measures, which allow comparison and characterization. The connectivity of the nodes (number of edges), connectivity distribution, shortest and average path length, and cluster coefficient (if there is a link between A and B, and B and C, how often is A linked to C) are basic network measures[153]. Important to the understanding of the cellular network was the discovery that most cellular networks follow a scale-free topology, including for example metabolic[153] and protein-protein interactions[154] networks. The network's ability to respond to changes like e.g. gene deletion is called the network robustness. Scale-free networks when compared to random networks are more robust, because of the many nodes with a few connections and the small number of hubs. However, the tolerance to failure comes with a high vulnerability to attacks on the few crucial nodes (hubs)[140]. This has been found in gene knock-out experiments[154].

Cellular functional organization is likely to have a modular structure[155]. Modularity refers to groups of nodes that have a rather distinct function. Various methods were developed to identify modules in networks either by topology[141, 156] or by integrated functional data[157]. Network approaches are used to enhance function prediction for example in protein interaction[158, 159].

Topology alone is not capable of describing the function of cellular networks. Important aspects for characterizing these networks are also the strength of the edges and the (spatio-) temporal features[153]. Metabolic networks are studied with respect to the flux of (the amount of) substrates and products. These analyses enable the formulation and verification of hypotheses on the importance of various reactions[160, 161].

The release of ever more (genomics) data requires new computational approaches to analyse the networks. This ranges from new theories for topology characterization to identification of modules, clusters and their dynamics in relation to biological function.

1.11 Phenotype networks and bio-networks

All bio-networks whether interactome, metabolome, or phenome will continue to grow as knowledge accumulates. The purpose of the study of these networks is a better understanding of the biological systems. Large-scale mutagenesis projects are conducted and are an important resource for functional annotation. However, the number of mutations that need to be induced and screened to be comprehensive for both functional annotations as well as for the phenotypic consequences is unclear, and probably gigantic. To close the gap between genotype and phenotype (networks) it is also necessary to work on the phenotype side of the problem.

For many years, it has been recognized that biological relationship can be found in overlapping disease phenotypes[144]. Furthermore, functionally similar genes, when mutated, have a characteristic pattern of human disease[162]. Naturally occurring mutations and the related phenotypes are a fruitful source and may provide a bridge to start to close the gap.

Mutations are stored in databases like HGMD, and genetic variations in dbSNP. Phenotype descriptions are found in expert databases like OMIM, the London Dysmorphology Database (LDDB), Dysmorphic Human and Mouse Homology Database (DHMHD) and the Pictures of Standard Syndromes and Undiagnosed Malformations database (POSSUM)[5, 163, 164]. Besides the genome browsers, there are few initiatives to integrate specifically the (human) phenotype and genotype data[165, 166]. Also, the phenome is not determined by disease phenotypes alone. All phenotype variations are a potential source for a better understanding of biology. The international HapMap project aims to determine the common patterns of DNA sequence variation using samples from populations with ancestry from parts of Africa, Asia and Europe[167]. Projects like HapMap will also give insight in the genetic basis of phenotype characteristics e.g. racial differences.

Phenotype data from literature has been explored to find biological relationships. But, the exploration of this information has only recently begun to be explored and all methods are aiming to identify genes for human genetic disorders[101, 168-170].

1.12 Outline of thesis

These studies concern the relationships between diseases and genes, and the relationships among diseases as indicators of the underlying molecular mechanisms. Over 5.000 human genetic traits have been described. In this work we focus on human genetics disorders in general, rather than on a particular class of diseases. The present study aims to reveal some general principles of human disease and the bioinformatics strategies to explore them for the identification of disease genes.

Chapter 2 describes a system called the GeneSeeker. The growing number of data collections is a direct consequence of the HGP. These databases make it possible to identify candidate genes for human genetic diseases, but all have specific target groups, contain particular information, and present this to the user in different formats. Furthermore, volume and complexity of these resources are a challenge to the researcher. The GeneSeeker aims to integrate various databases using a strategy that mimics the "normal" strategy used by the researcher. GeneSeeker is particularly well suited for syndromes in which the disease gene displays altered expression patterns in the affected tissue(s).

Chapter 3 discusses the technical background of the GeneSeeker. Because of the bulk of data, the GeneSeeker uses a remote querying method, that obviates data warehousing and guarantees that the most recent data are queried. The analyses performed reduce the time-consuming process of browsing manually to a few minutes. All options are presented to the user via a web-interface.

Chapter 4 discusses the use of phenotype information from syndrome families as a tool for functional genomics. Syndromes are constituted by a specific combination of clinical features. We discuss how this can be systematically explored and ask if this knowledge can be transferred to biological relationships. Furthermore, we discuss the possibilities of a phenotype map for the identification of disease genes.

Chapter 5 describes a text mining method and analysis of the human phenome. The systematic exploration of human disease phenotypes and molecular relationships of the underlying genes has only been attempted in a few studies. We describe a fully automated method to generate such a phenotype map. Additionally, we analyse this phenomap for biologically meaningful information, and discuss the applications that could be derived from this approach.

References

1. Aase, J.M., *Diagnostic Dysmorphology.* 1990, New York: Plenum Medical.
2. Donnai, D. and A.P. Read, *How clinicians add to knowledge of development.* Lancet, 2003. **362**(9382): p. 477-84.
3. Merks, J.H., et al., *Phenotypic abnormalities: terminology and classification.* Am J Med Genet A, 2003. **123**(3): p. 211-30.
4. Cohen, M.M.J., *The child with multiple birth defects.* 1982, New York: Raven Press.
5. Winter, R.M. and M. Baraitser, *The London Dysmorphology Database.* J Med Genet, 1987. **24**(8): p. 509-10.
6. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D39-45.
7. Lipscomb, C.E., *Medical Subject Headings (MeSH).* Bull Med Libr Assoc, 2000. **88**(3): p. 265-6.
8. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.
9. Saiki, R.K., et al., *Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia*. Science, 1985. **230**(4732): p. 1350-4.
10. Goffeau, A., et al., *Life with 6000 genes.* Science, 1996. **274**(5287): p. 546, 563-7.
11. McPherson, J.D., et al., *A physical map of the human genome.* Nature, 2001. **409**(6822): p. 934-41.
12. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
13. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
14. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
15. Larsson, T.P., et al., *Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery.* FEBS Lett, 2005. **579**(3): p. 690-8.
16. Wilson, E.B., *The sex chromosomes.* Arch. Mikrosk. Anat. Enwicklungsmech., 1911. **77**: p. 249-271.
17. Nathans, J., et al., *Molecular genetics of inherited variation in human color vision.* Science, 1986. **232**(4747): p. 203-10.
18. Donahue, R.P., et al., *Probable assignment of the Duffy blood group locus to chromosome 1 in man.* Proc Natl Acad Sci U S A, 1968. **61**(3): p. 949-55.
19. Royer-Pokora, B., et al., *Cloning the gene for an inherited human disorder--chronic granulomatous disease--on the basis of its chromosomal location.* Nature, 1986. **322**(6074): p. 32-8.
20. Collins, F.S., *Positional cloning moves from perditional to traditional.* Nat Genet, 1995. **9**(4): p. 347-50.
21. Monreal, A.W., et al., *Mutations in the human homologue of mouse dl cause autosomal recessive and dominant hypohidrotic ectodermal dysplasia.* Nat Genet, 1999. **22**(4): p. 366-9.
22. de Kok, Y.J., et al., *Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4.* Hum

Mol Genet, 1996. **5**(9): p. 1229-35.

23. Kleinjan, D.A. and V. van Heyningen, *Long-range control of gene expression: emerging mechanisms and disruption in disease.* Am J Hum Genet, 2005. **76**(1): p. 8-32.

24. Iwakoshi, M., et al., *9q34.3 deletion syndrome in three unrelated children.* Am J Med Genet A, 2004. **126**(3): p. 278-83.

25. Ropers, H.H. and B.C. Hamel, *X-linked mental retardation.* Nat Rev Genet, 2005. **6**(1): p. 46-57.

26. Bremond-Gignac, D., et al., *Combination of WAGR and Potocki-Shaffer contiguous deletion syndromes in a patient with an 11p11.2-p14 deletion.* Eur J Hum Genet, 2005.

27. Johnston, J.J., et al., *Clinical and molecular delineation of the Greig cephalopolysyndactyly contiguous gene deletion syndrome and its distinction from acrocallosal syndrome.* Am J Med Genet A, 2003. **123**(3): p. 236-42.

28. Yntema, H.G., et al., *A novel ribosomal S6-kinase (RSK4; RPS6KA6) is commonly deleted in patients with complex X-linked mental retardation.* Genomics, 1999. **62**(3): p. 332-43.

29. Ballabio, A., et al., *Contiguous gene syndromes due to deletions in the distal short arm of the human X chromosome.* Proc Natl Acad Sci U S A, 1989. **86**(24): p. 10001-5.

30. Veltman, J.A., et al., *Definition of a critical region on chromosome 18 for congenital aural atresia by arrayCGH.* Am J Hum Genet, 2003. **72**(6): p. 1578-84.

31. Vissers, L.E., et al., *Mutations in a new member of the chromodomain gene family cause CHARGE syndrome.* Nat Genet, 2004. **36**(9): p. 955-7.

32. Friedman, P.A., et al., *Detection of hepatic phenylalanine 4-hydroxylase in classical phenylketonuria.* Proc Natl Acad Sci U S A, 1973. **70**(2): p. 552-6.

33. Gitschier, J., et al., *Characterization of the human factor VIII gene.* Nature, 1984. **312**(5992): p. 326-30.

34. Nilssen, O., et al., *alpha-Mannosidosis: functional cloning of the lysosomal alpha-mannosidase cDNA and identification of a mutation in two affected siblings.* Hum Mol Genet, 1997. **6**(5): p. 717-26.

35. Matsuura, S., et al., *Genetic mapping using microcell-mediated chromosome transfer suggests a locus for Nijmegen breakage syndrome at chromosome 8q21-24.* Am J Hum Genet, 1997. **60**(6): p. 1487-94.

36. Banfi, S., et al., *Identification and mapping of human cDNAs homologous to Drosophila mutant genes through EST database searching.* Nat Genet, 1996. **13**(2): p. 167-74.

37. Celli, J., et al., *Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome.* Cell, 1999. **99**(2): p. 143-53.

38. den Hollander, A.I., et al., *Isolation and mapping of novel candidate genes for retinal disorders using suppression subtractive hybridization.* Genomics, 1999. **58**(3): p. 240-9.

39. den Hollander, A.I., et al., *Mutations in a human homologue of Drosophila crumbs cause retinitis pigmentosa (RP12).* Nat Genet, 1999. **23**(2): p. 217-21.

40. Everett, C.M. and N.W. Wood, *Trinucleotide repeats and neurodegenerative disease.* Brain, 2004. **127**(Pt 11): p. 2385-405.

41. Orita, M., et al., *Rapid and sensitive detection of point mutations and DNA*

polymorphisms using the polymerase chain reaction. Genomics, 1989. **5**(4): p. 874-9.

42. Fischer, S.G. and L.S. Lerman, *DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory.* Proc Natl Acad Sci U S A, 1983. **80**(6): p. 1579-83.
43. Xiao, W. and P.J. Oefner, *Denaturing high-performance liquid chromatography: A review.* Hum Mutat, 2001. **17**(6): p. 439-74.
44. Cotton, R.G., N.R. Rodrigues, and R.D. Campbell, *Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations.* Proc Natl Acad Sci U S A, 1988. **85**(12): p. 4397-401.
45. Bui, C.T., et al., *Chemical cleavage reactions of DNA on solid support: application in mutation detection.* BMC Chem Biol, 2003. **3**(1): p. 1.
46. Youil, R., B. Kemper, and R.G. Cotton, *Detection of 81 of 81 known mouse beta-globin promoter mutations with T4 endonuclease VII--the EMC method.* Genomics, 1996. **32**(3): p. 431-5.
47. Stanssens, P., et al., *High-throughput MALDI-TOF discovery of genomic sequence polymorphisms.* Genome Res, 2004. **14**(1): p. 126-33.
48. White, S.J., M.H. Breuning, and J.T. den Dunnen, *Detecting copy number changes in genomic DNA: MAPH and MLPA.* Methods Cell Biol, 2004. **75**: p. 751-68.
49. Vink, G.R., et al., *Mutation screening of EXT1 and EXT2 by direct sequence analysis and MLPA in patients with multiple osteochondromas: splice site mutations and exonic deletions account for more than half of the mutations.* Eur J Hum Genet, 2004.
50. Sellner, L.N. and G.R. Taylor, *MLPA and MAPH: new techniques for detection of gene deletions.* Hum Mutat, 2004. **23**(5): p. 413-9.
51. Glazier, A.M., J.H. Nadeau, and T.J. Aitman, *Finding genes that underlie complex traits.* Science, 2002. **298**(5602): p. 2345-9.
52. Dean, M., *Approaches to identify genes for complex human diseases: lessons from Mendelian disorders.* Hum Mutat, 2003. **22**(4): p. 261-74.
53. Strachan, T. and A.P. Read, *Human molecular genetics*. 2004, London: Garland Science.
54. *The yeast genome directory.* Nature, 1997. **387**(6632 Suppl): p. 5.
55. *Genome sequence of the nematode C. elegans: a platform for investigating biology.* Science, 1998. **282**(5396): p. 2012-8.
56. Adams, M.D., et al., *The genome sequence of Drosophila melanogaster.* Science, 2000. **287**(5461): p. 2185-95.
57. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
58. Miyazaki, S., et al., *DDBJ in the stream of various biological data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D31-4.
59. Kanz, C., et al., *The EMBL Nucleotide Sequence Database.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D29-33.
60. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D34-8.
61. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33 Database Issue**: p. D154-9.

62. Wu, C. and D.W. Nebert, *Update on genome completion and annotations: Protein Information Resource.* Hum Genomics, 2004. **1**(3): p. 229-33.

63. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D501-4.

64. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.

65. Mulder, N.J., et al., *InterPro, progress and status in 2005.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D201-5.

66. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

67. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

68. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-80.

69. Joshi-Tope, G., et al., *The Genome Knowledgebase: a resource for biologists and bioinformaticists.* Cold Spring Harb Symp Quant Biol, 2003. **68**: p. 237-43.

70. Ball, C.A., et al., *The Stanford Microarray Database accommodates additional microarray platforms and data formats.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D580-2.

71. Alfarano, C., et al., *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D418-24.

72. Peri, S., et al., *Human protein reference database as a discovery resource for proteomics.* Nucleic Acids Res, 2004. **32 Database issue**: p. D497-501.

73. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.

74. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.

75. Schuler, G.D., et al., *Entrez: molecular biology database and retrieval system.* Methods Enzymol, 1996. **266**: p. 141-62.

76. Etzold, T. and P. Argos, *SRS--an indexing and retrieval tool for flat file data libraries.* Comput Appl Biosci, 1993. **9**(1): p. 49-57.

77. Hekkelman, M.L. and G. Vriend, *MRS: A fast and compact retrieval system for biological data.* Nucleic Acids Res, 2005. **33 Webserver Issue**: p. W766-9.

78. Hubbard, T., et al., *Ensembl 2005.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D447-53.

79. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

80. Jacob, F., *Evolution and tinkering.* Science, 1977. **196**(4295): p. 1161-6.

81. Dayhoff, M.D., R.M. Schwartz, and B.C. Orcutt, *A model of evolutionary change in proteins.* In Dayhoff, M. O., ed., Atlas of protein sequence and structure. Vol. 5. 1978, Washington D.C.: National Biomedical Research Foundation. 345.

82. Dayhoff, M.O., R.V. Eck, and C.M. Park, *A model of evolutionary change in proteins.* In Dayhoff, M. O., ed., Atlas of protein sequence and structure. Vol. 5. 1972, Washington D.C.: National Biomedial Research Foundation. 89-89.

83. Henikoff, S. and J.G. Henikoff, *Automated assembly of protein blocks for database*

searching. Nucleic Acids Res, 1991. **19**(23): p. 6565-72.

84. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
85. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
86. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.
87. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
88. Zhang, Z., et al., *Protein sequence similarity searches using patterns as seeds.* Nucleic Acids Res, 1998. **26**(17): p. 3986-90.
89. Durbin, R., et al., *Biological sequence analysis.* 1998, Cambridge: Cambridge University Press.
90. Sonnhammer, E.L., S.R. Eddy, and R. Durbin, *Pfam: a comprehensive database of protein domain families based on seed alignments.* Proteins, 1997. **28**(3): p. 405-20.
91. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.
92. Korf, I., et al., *Integrating genomic homology into gene structure prediction.* Bioinformatics, 2001. **17 Suppl 1**: p. S140-8.
93. Alexandersson, M., S. Cawley, and L. Pachter, *SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model.* Genome Res, 2003. **13**(3): p. 496-502.
94. Barrett, T., et al., *NCBI GEO: mining millions of expression profiles--database and tools.* Nucleic Acids Res, 2005. **33**(Database issue): p. D562-6.
95. Liang, S., et al., *Novel retinal genes discovered by mining the mouse embryonic RetinalExpress database.* Mol Vis, 2004. **10**: p. 773-86.
96. Katsanis, N., et al., *A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14326-31.
97. Lord-Grignon, J., et al., *Characterization of new transcripts enriched in the mouse retina and identification of candidate retinal disease genes.* Invest Ophthalmol Vis Sci, 2004. **45**(9): p. 3313-9.
98. Lopez-Bigas, N. and C.A. Ouzounis, *Genome-wide identification of genes likely to be involved in human genetic disease.* Nucleic Acids Res, 2004. **32**(10): p. 3108-14.
99. Adie, E.A., et al., *Speeding disease gene discovery by sequence based candidate prioritization.* BMC Bioinformatics, 2005. **6**(1): p. 55.
100. Smith, N.G. and A. Eyre-Walker, *Human disease genes: patterns and predictions.* Gene, 2003. **318**: p. 169-75.
101. Turner, F.S., D.R. Clutterbuck, and C.A. Semple, *POCUS: mining genomic sequence annotation to predict disease genes.* Genome Biol, 2003. **4**(11): p. R75.
102. Brent, M.R. and R. Guigo, *Recent advances in gene structure prediction.* Curr Opin Struct Biol, 2004. **14**(3): p. 264-72.
103. Galperin, M.Y. and E.V. Koonin, *'Conserved hypothetical' proteins: prioritization of targets for experimental study.* Nucleic Acids Res, 2004. **32**(18): p. 5452-63.
104. Hulsen, T., J. de Vlieg, and P. Groenen, *Protein World, a comprehesive protein knowledge database based upon Smith-Waterman sequence comparisons.* in

preparation.

105. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.

106. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**(1): p. 41.

107. O'Brien, K.P., I. Westerlund, and E.L. Sonnhammer, *OrthoDisease: a database of human disease orthologs.* Hum Mutat, 2004. **24**(2): p. 112-9.

108. Huynen, M.A., et al., *The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly.* Hum Mol Genet, 2001. **10**(21): p. 2463-8.

109. Muhlenhoff, U., et al., *The yeast frataxin homolog Yfh1p plays a specific role in the maturation of cellular Fe/S proteins.* Hum Mol Genet, 2002. **11**(17): p. 2025-36.

110. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D433-7.

111. Fukuoka, Y., H. Inaoka, and I.S. Kohane, *Inter-species differences of co-expression of neighboring genes in eukaryotic genomes.* BMC Genomics, 2004. **5**(1): p. 4.

112. Lercher, M.J., A.O. Urrutia, and L.D. Hurst, *Clustering of housekeeping genes provides a unified model of gene order in the human genome.* Nat Genet, 2002. **31**(2): p. 180-3.

113. Gabaldon, T. and M.A. Huynen, *Prediction of protein function and pathways in the genome era.* Cell Mol Life Sci, 2004. **61**(7-8): p. 930-44.

114. Mattick, J.S. and I.V. Makunin, *Small regulatory RNAs in mammals.* Hum Mol Genet, 2005. **14 Suppl 1**: p. R121-32.

115. Alwine, J.C., D.J. Kemp, and G.R. Stark, *Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5350-4.

116. Liang, P. and A.B. Pardee, *Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction.* Science, 1992. **257**(5072): p. 967-71.

117. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.

118. Velculescu, V.E., et al., *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.

119. Brenner, S., et al., *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nat Biotechnol, 2000. **18**(6): p. 630-4.

120. Hertzano, R., et al., *Transcription profiling of inner ears from Pou4f3(ddl/ddl) identifies Gfi1 as a target of the Pou4f3 deafness gene.* Hum Mol Genet, 2004. **13**(18): p. 2143-53.

121. Weston, M.D., et al., *Mutations in the VLGR1 gene implicate G-protein signaling in the pathogenesis of Usher syndrome type II.* Am J Hum Genet, 2004. **74**(2): p. 357-66.

122. McCaffrey, T.A., et al., *High-level expression of Egr-1 and Egr-1-inducible genes in mouse and human atherosclerosis.* J Clin Invest, 2000. **105**(5): p. 653-62.

123. Clarke, P.A., R. te Poele, and P. Workman, *Gene expression microarray technologies in the development of new therapeutic agents.* Eur J Cancer, 2004. **40**(17): p. 2560-91.

124. Li, J.B., et al., *Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene.* Cell, 2004. **117**(4): p. 541-52.

125. Tang, K., et al., *Mining disease susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass spectrometry.* J Proteome Res, 2004. **3**(2): p. 218-27.

126. Deshpande, N., et al., *The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema.* Nucleic Acids Res, 2005. **33**(Database issue): p. D233-7.

127. Chance, M.R., et al., *High-throughput computational and experimental techniques in structural genomics.* Genome Res, 2004. **14**(10B): p. 2145-54.

128. Vendruscolo, M. and C.M. Dobson, *A glimpse at the organization of the protein universe.* Proc Natl Acad Sci U S A, 2005.

129. Burley, S.K., et al., *Structural genomics: beyond the human genome project.* Nat Genet, 1999. **23**(2): p. 151-7.

130. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.

131. Mewes, H.W., et al., *MIPS: analysis and annotation of proteins from whole genomes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D41-4.

132. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.

133. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.

134. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180-3.

135. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.

136. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.

137. Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-7.

138. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.

139. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.

140. Albert, R., H. Jeong, and A.L. Barabasi, *Error and attack tolerance of complex networks.* Nature, 2000. **406**(6794): p. 378-82.

141. Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks.* Science, 2002. **297**(5586): p. 1551-5.

142. Schleiden, M.J., *Beiträge zur Phytogenesis.* Müller's Arch. Anat. Physiol. Wiss. Med., 1838: p. 136-176.

143. Schwann, T., *Mikroskopische Untersuchungen über die Übereinstimmung in der Struktur und dem Wachsthum der Thiere und Pflanzen.* 1839, Berlin: Verlag der Sander'schen Buchhandlung.

144. Pinsky, L., *The polythetic (phenotypic community) system of classifying human malformation syndromes.* Birth Defects Orig Artic Ser, 1977. **13**(3A): p. 13-30.

145. Brunner, H.G., B.C. Hamel, and H. Van Bokhoven, *The p63 gene in EEC and other syndromes.* J Med Genet, 2002. **39**(6): p. 377-81.

146. Freimer, N. and C. Sabatti, *The human phenome project.* Nat Genet, 2003. **34**(1): p.

15-21.

147. Mashimo, T., et al., *Rat Phenome Project: the untapped potential of existing rat strains.* J Appl Physiol, 2005. **98**(1): p. 371-9.

148. Bogue, M.A. and S.C. Grubb, *The Mouse Phenome Project.* Genetica, 2004. **122**(1): p. 71-4.

149. de la Cruz, N., et al., *The Rat Genome Database (RGD): developments towards a phenome database.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D485-91.

150. Rual, J.F., et al., *Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library.* Genome Res, 2004. **14**(10B): p. 2162-8.

151. Varki, A., et al., *Great Ape Phenome Project?* Science, 1998. **282**(5387): p. 239-40.

152. Mahner, M. and M. Kary, *What exactly are genomes, genotypes and phenotypes? And what about phenomes?* J Theor Biol, 1997. **186**(1): p. 55-63.

153. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.

154. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.

155. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.

156. Snel, B., P. Bork, and M.A. Huynen, *The identification of functional modules from the genomic association of genes.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5890-5.

157. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.

158. Jansen, R., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data.* Science, 2003. **302**(5644): p. 449-53.

159. Huynen, M.A., et al., *Function prediction and protein networks.* Curr Opin Cell Biol, 2003. **15**(2): p. 191-8.

160. Edwards, J.S., R.U. Ibarra, and B.O. Palsson, *In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.* Nat Biotechnol, 2001. **19**(2): p. 125-30.

161. Ibarra, R.U., J.S. Edwards, and B.O. Palsson, *Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.* Nature, 2002. **420**(6912): p. 186-9.

162. Jimenez-Sanchez, G., B. Childs, and D. Valle, *Human disease genes.* Nature, 2001. **409**(6822): p. 853-5.

163. Evans, C.D., et al., *The dysmorphic human-mouse homology database (DHMHD): an interactive World-Wide Web resource for gene mapping.* J Med Genet, 1996. **33**(4): p. 289-94.

164. Evans, C.D., *Computer systems in dysmorphology.* Clin Dysmorphol, 1995. **4**(3): p. 185-201.

165. Kahraman, A., et al., *PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics.* Bioinformatics, 2005. **21**(3): p. 418-20.

166. Toyoda, T. and A. Wada, *Omic space: coordinate-based integration and analysis of genomic phenomic interactions.* Bioinformatics, 2004. **20**(11): p. 1759-65.

167. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

168. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.

169. Freudenberg, J. and P. Propping, *A similarity-based method for genome-wide prediction of disease-relevant human genes.* Bioinformatics, 2002. **18 Suppl 2**: p. S110-5.

170. Tao, Y.C. and R.L. Leibel, *Identifying functional relationships among human genes by systematic analysis of biological literature.* BMC Bioinformatics, 2002. **3**(1): p. 16.

Chapter 2

A new web-based data mining tool for the identification of candidate genes for human genetic disorders

Marc A. van Driel, Koen Cuelenaere, Patrick P.C.W. Kemmeren,
Jack A.M. Leunissen and Han G. Brunner

European Journal of Human Genetics (2003) 11, 57 – 63.
(*With permission*).

Abstract

To identify the gene underlying a human genetic disorder can be difficult and time consuming. Typically, positional data delimit a chromosomal region that contains between 20 and 200 genes. The choice then lies between sequencing large numbers of genes, or setting priorities by combining positional data with available expression and phenotype data, contained in different internet databases. This process of examining positional candidates for possible functional clues may be performed in many different ways, depending on the investigator's knowledge and experience. Here, we report on a new tool called the GeneSeeker, which gathers and combines positional data and expression/phenotypic data in an automated way from nine different web-based databases. This results in a quick overview of interesting candidate genes in the region of interest. The GeneSeeker system is built in a modular fashion allowing for easy addition or removal of databases if required. Databases are searched directly through the web, which obviates the need for data warehousing. In order to evaluate the GeneSeeker tool, we analysed syndromes with known genesis. For each of 10 syndromes the GeneSeeker program generated a shortlist that contained a significantly reduced number of candidate genes from the critical region, yet still contained the causative gene. On average, a list of 163 genes based on position alone was reduced to a more manageable list of 22 genes based on position and expression or phenotype information. We are currently expanding the tool by adding other databases. The GeneSeeker is available via the web-interface (http://www.cmbi.kun.nl/GeneSeeker/).

Introduction

Positional cloning and candidate gene analysis are commonly used as complementary strategies for the identification of genes involved in human genetic disorders. With the recent completion of the human genome draft sequence a comprehensive list of positional candidate genes can often be obtained. For many diseases the critical interval will be between 0.5 and 10 cM, with the number of genes anywhere between 5 and 300. Prioritising these genes for mutation analysis is the logical next step. This requires that the researcher collects information from various sources on expression patterns, biological function, animal models, related human diseases and other relevant data. Clearly, researchers differ widely in their ability to retrieve relevant information that is stored in a growing number of separate (and often unlinked) on-line databases. Moreover, this process tends to be very time-consuming, and many hours may go into collecting and sorting the relevant information. Integrating information from the databases in an automatic way would allow researchers to get a quick snapshot overview of their particular candidate region.

Here we report on a new bioinformatics tool, which gathers both positional as well as expression/phenotypic data in an automated way from nine different databases and then combines this information using Boolean operators. This results in a quick overview of candidate genes in the genetic region of interest. The GeneSeeker system is built in a modular fashion, making it easy to maintain and expand. A further advantage is that there is no need for data warehousing or updating because the databases are searched directly through the web.

In its present form, the GeneSeeker tool uses the Genome Database (GDB)[1] and the Online Mendelian Inheritance in Man (OMIM (URL: http://www.ncbi.nlm.nih.gov/omim/)) to obtain human mapping data. Genetic localisations specified by the user are also translated with the aid of an ‘Oxford-grid’, to search the appropriate mouse databases (e.g. the Mouse Genome Database (MGD)[2]. The key tissues affected by the genetic disorder are used to query phenotypic or expression related databases, including the OMIM phenotype fields, Swissprot[3], and Medline (National Library of Medicine, Bethesda, USA) for data on human phenotypes and the Gene Expression Database (GXD)[4], the Transgenic/Targeted Mutation Database (TBASE)[5], and the Mouse Locus Catalog (MLC)[2] for gene expression patterns and phenotypes in mice. A general overview of the data flow within the program is given in figure 1.

Materials and methods

The GeneSeeker interface

The homepage of the GeneSeeker (http://www.cmbi.kun.nl/GeneSeeker/) allows the user to specify the genetic mapping information. This can be a chromosome, a chromosome arm, or

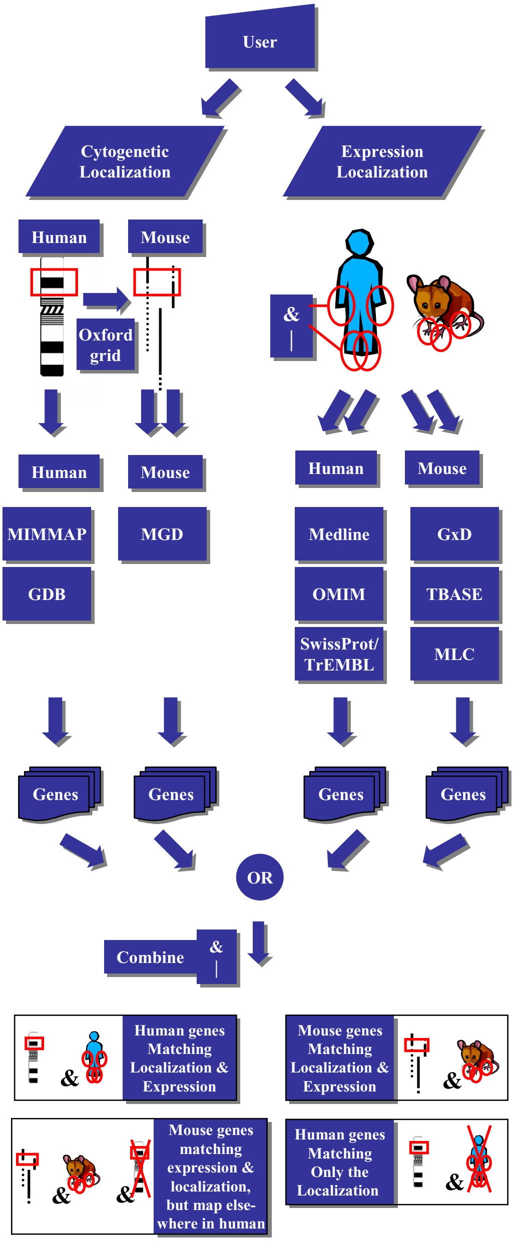

User
Cytogenetic Localization
Expression Localization
Human
Mouse
Oxford grid
&
|
Human
Mouse
Human
Mouse
MIMMAP
MGD
GDB
Medline
GxD
OMIM
TBASE
SwissProt/ TrEMBL
MLC
Genes
Genes
Genes
Genes
OR
Combine
&
|
&
Human genes Matching Localization & Expression
Mouse genes Matching Localization & Expression
&
&
&
Mouse genes matching expression & localization, but map elsewhere in human
Human genes Matching Only the Localization
&

Figure 1 A general overview of the GeneSeeker program. The query entered by the user is pre-processed for Human and Mouse databases and subsequently reformulated into the format appropriate for each database. The database queries result in lists of genes, which are combined by Boolean operators according to the query as formulated by the user. The results are presented in the four boxes at the foot of the figure.

a range (e.g. 7p15-7p14). If necessary, a combination of genetic localisations can be entered (e.g. 4p16-4p14 *or* 4q31-4q35). Gene expression or phenotypic information can be entered in a separate box, in which the user specifies the tissue names where either direct RNA expression or phenotypic expression of the candidate gene is expected. For example, the phenotype of Hand-Foot-Uterus Syndrome[6] can be translated into the expression terms 'limb *or* genital'. Advanced options include a thesaurus[7], which can be used to include alternatives and hence broaden the expression search term. In case of 'limb *or* genital' use of the thesaurus will result in 'upper limb' *or* arm *or* limb *or* joint *or* 'lower limb' *or* 'hip joint' *or* toes *or* digit *or* 'male genital' *or* testes *or* testis *or* 'Sertoli cells' *or* 'female genital' *or* ovaries *or* ovary *or* uterus *or* vagina'.
A number of refinement options have been implemented, such as the possibility to exclude databases, to exclude housekeeping or user-specified genes, to change the maximum distance for the Oxford-grid (used in Human-to-Mouse map translation, see below), and to put multiple searches in a batch list.

Databases used

The GeneSeeker searches three types of databases: genetic localisation, gene expression, and phenotypic databases. GDB and MIMMAP (a reformatted version of the OMIM gene mapping information) are searched for genes in a specified chromosome location in humans, while MGD is queried for mouse genes in the homologous regions. From the gene expression and phenotypic databases GXD, SWISSPROT, TrEMBL, MLC, OMIM, TBASE and Medline, all the genes are extracted which match the given expression terms. Database web addresses are given in table 1.

Gene naming

Different databases cause the data retrieved to be in different output formats. In contrast, communication between program processes, logical combination and analysis of the data obtained require a uniform nomenclature. To circumvent this problem a list of synonyms was created using the gene-name information stored in SWISSPROT in combination with GDB's 'alias' information. This synonym list is updated weekly, and its use should remove a number of potential naming problems. As an exception, the gene naming process in TBASE is highly variable and this could not always be neutralized by the use of these lists.

Extraction of genes from the databases

Human gene names are selected upon the fact that they must consist of one or more capital letters and/or numbers in the databases. Mouse genes begin with a capital letter, followed by one or more lowercase letters or numbers. The id-numbers of the genes in the databases are taken as a unique identifier. In cases where mouse databases are used (TBASE/MGD/GXD), the mouse gene names are translated into human gene names, using a list obtained from GXD. The obtained gene names are compared with the synonym list, obtained from SWISSPROT. If a synonym exists, this then replaces the gene name. The gene names are reported to the GeneSeeker program together with an URL-encoded link (Uniform Resource Locator) to the entry.

Table 1 Database URL's. The number of entries is based on the query formulation used by the GeneSeeker to extract human/mouse related information, and thus can differ from the total number of entries in the database.

Data bank	*No.entries*	*URL*
Localisation databases		
OXFORD	5652	http://www.informatics.jax.org/[a]
MIMMAP	7171	http://www.ncbi.nlm.nih.gov/omim/
MGD	24925	http://www.informatics.jax.org/
GDB	51917	http://www.gdb.org/gdb/
Expression and phenotype databases		
SWISSPROT	5908	http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
SWISSNEW	5875	http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
SPTREMBL	23567	http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
REMTREMBL	19036	http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
TREMBLNEW	10394	http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
Medline	58978	http://www.ncbi.nlm.nih.gov/[b]
TBASE	6768	http://tbase.jax.org/[b]
GXD	24925	http://www.informatics.jax.org/
OMIM	13099	http://www3.ncbi.nlm.gov/omim/
MLC	24925	http://www.informatics.jax.org/
Other database(s)		
GeneCards	20417	http://bioinformatics.weizmann.ac.il/cards/

[a]Accessed after processing from the local mirror site: http://www.cmbi.kun.nl/srs/. [b]Accessed directly using this address. Remaining sites were accessed using the local mirror site: http://www.cmbi.kun.nl/.

Databases containing locus information

The MIMMAP and GDB databases are searched for all genes between two human genetic locations, including any overlapping genes.

In order to be able to search mouse gene localisation databases, human genetic localisations are converted to mouse localisations by an 'Oxford-grid' as supplied by the MGD[2]. Mouse genes are searched positionally rather than on similarity, since the goal is to find extra genes. The input for the 'Oxford-grid' is a human chromosome number with a band range (e.g. 7p15-p21). This location is then translated into mouse chromosomes with a chromosomal range in cM. Two locations are taken as one range if they are closer to each other than specified in the maximum distance. If not, they are returned as two separate regions. Each region is returned with a standard extension of 5 cM, so as not to miss any genes located on the region boundaries. The output from the 'Oxford-grid' contains mouse chromosomes with their cM-range and is reported back to the GeneSeeker main program. This range is subsequently used to query MGD for homologous mouse genes.

Databases containing gene expression or phenotype information

For all genes expressed in a certain tissue type or associated with a phenotypic feature involving a specified tissue or organ the description and comment lines, matching 'human' are extracted from the SWISSPROT, SWISSNEW, TrEMBL, and TrEMBLNEW databases. The gene names are selected from the gene-name field.

The same procedure is followed for the Medline database, using an advanced Boolean search for all genes expressed in a certain tissue type or sharing a phenotypic feature of a syndrome in humans (query: 'tissue *and* human[orgn] *not* mouse[orgn] *not* rat[orgn]'). The human gene names are selected by the fact that these begin with two capital letters, followed by one or more capital letters or numbers. Common abbreviations such as DNA, RNA, PCR, and others are filtered out.

All genes expressed in a certain tissue type or with a phenotypic feature of a syndrome are extracted from the TBASE database using the 'phenotype' field and selecting the mouse as the organism, from GXD and OMIM using the 'abstract' field and 'text' field respectively, and from the MLC database using the 'phenotype' field. The obtained mouse gene names are translated into human gene names, using a list obtained from MLC itself.

Test selection of human genetic disorders

To test the ability of the GeneSeeker program to identify candidate genes, eight syndromes with known genesis where used. We also evaluated two syndromes whose genesis at the time of the query had not yet been published: Acro-Dermato-Ungual-Lacrimal-Tooth (ADULT) syndrome[8] and Noonan syndrome[9, 10] (table 2). These two examples presented an excellent opportunity to test the system without the noise from direct pointers to the gene in the databases used.

Table 2 Selected disorders

Syndrome (OMIM#)	*Expression terms*	*Genetic localisation*
Acro-Dermato-Ungual-Lacrimal-Tooth Syndrome [103285]	Limb/Hand/Teeth	3q27
Alagille Syndrome [118450]	Liver/Eye/Heart	20p12
Hand-Foot-Uterus Syndrome [140000]	Limb/Genital	7p15-7p14.2
Holt-Oram Syndrome [142900]	Limb/Heart	12q24.1
Multiple Synostoses Syndrome 1 [186500]	Ear/Joint	17q22
Noonan Syndrome [163950]	Skeletal/Heart	12q24.1
Renal-Coloboma Syndrome [120330]	Renal/Eye	10q24.3-10q25.1
Townes-Brocks Syndrome [107480]	Limb/Ear	16q12.1
Tricho-Dento-Osseous Syndrome [190320]	Ectoderm/Skeleton/Tooth	17q21.3-17q22
Ulnar-Mammary Syndrome [181450]	Limb/Mammary	12q24.1

Querying the GeneSeeker

The setup of the GeneSeeker makes it possible to submit queries in a number of ways. To benchmark the performance, accuracy, and the flexibility of the system, the same query was formulated in different ways. Each syndrome mentioned in table 2 was queried using primary expression terms (table 2) combined with the Boolean operators *and* or *or* for all terms. For example Alagille syndrome was formulated once as 'liver *and* eye *and* heart', and also as 'liver *or* eye *or* heart'. In addition, in some queries the thesaurus/embryological terms were used. Thus, 'eye' became (eyes *or* eye *or* conjunctiva *or* cornea *or* lens *or* optic nerve *or* retina *or* vitreous *or* 'conjunctival sac').

Evaluation

Each result is saved as a HTML file in a separate directory, containing the output from the different databases analysed by the GeneSeeker. The output of the analysis is presented in four tables. (1) A list of human genes in the correct genetic region and matching the specified expression profile, (2) a list of mouse genes matching the syntenic region(s) as well as the expression profile, but with no matching human gene name, (3) a list of mouse genes found in the syntenic region in mouse, for which the homologous human gene is found to map outside the critical interval, and (4) a list of all the remaining human genes that are present in the genetic interval, but which do not match the expression profile. The data in the HTML files was extracted and converted to a spreadsheet for further analysis.

Results

The evaluation queries were performed in batch in June and July 2001. The processing time of queries using both genetic localisation and expression/phenotypic information varied from 2 min for simple queries to 10 min for complex queries. The number of hits per database for the genetic localisation, expression/phenotypic and the combined queries are presented in figure 2a, b, c respectively. For the genetic localisation query, no large differences were found between the GDB and MIMMAP. The number of MGD hits is relatively large for several regions. First, because this list includes both mouse genes whose human homologues are present in GDB as well as a smaller list of mouse genes for which a human homologue could not be identified, either directly or by applying the synonym list. In addition, the conversion through the Oxford grids caused more genes to be retrieved because of wider segment limits. Both figure 2b and c show a very small contribution from TBASE compared to the other queried databases. This likely reflects the relatively small number of genes in TBASE as well as inconsistent gene naming.

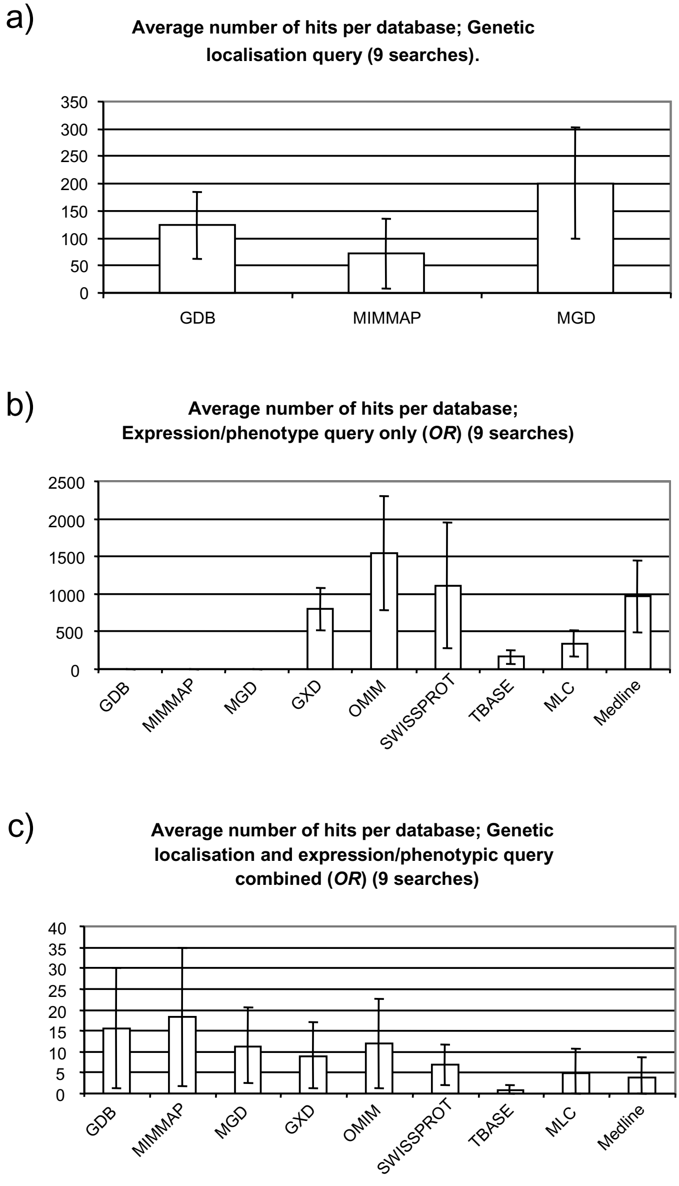
All the causative genes were found in the queries done with only the genetic localisation data (table 3). The average number of genes in a disease critical interval was 165. This number varied from 322 in the case of Tricho-Dento-Osseous syndrome in 17q21.3-17q22, to only 49 in Townes-Brocks syndrome located at 16q12.1.

Combining genetic localisation with expression/phenotype data was most successful if a Boolean *or* was used to combine expression sites. In all 10 such cases, the causative gene was retrieved. Starting from an average number of 165 positional candidate genes (range 49 – 322), the number of candidate genes that matched both location and expression pattern was reduced to 22 (range 2 – 63). A match was also obtained for both syndromes for which the gene had not previously been identified as causing the disease. For ADULT syndrome, a candidate gene list of 12 genes was generated, reflecting a 10-fold reduction from 116 positional candidate genes. *TP63*, which was subsequently been proven to be the causative gene for ADULT syndrome, was present among these 12 selected genes[11].

A similar result was obtained for Noonan syndrome. Using ‘skeletal *and* heart’ as search terms, the number of genes from chromosome band 12q24.1 was 174. This was reduced to 10 in the candidate gene shortlist. Among this final selection was the *PTPN11* gene, which indeed causes Noonan syndrome[10].

Discussion

Human disease genes can sometimes be rapidly identified by using information on RNA expression patterns or by studying knockout phenotypes in mice. For instance systematic screens for genes expressed in retina or inner ear are currently being applied successfully in labs around the world in order to identify genes for deafness or blindness respectively[12-14]. Similarly, direct comparison of human and mouse phenotypes allowed for the rapid



a) Average number of hits per database; Genetic localisation query (9 searches).

b) Average number of hits per database; Expression/phenotype query only (*OR*) (9 searches)

c) Average number of hits per database; Genetic localisation and expression/phenotypic query combined (*OR*) (9 searches)

Figure 2 The average number of hits per database in the candidate gene list of the GeneSeeker: (a) only the genetic localisation query; (b) only expression/ phenotypic query, combining the search terms with a Boolean or; (c) genetic localisation and expression/ phenotypic information combined with a Boolean. For the expression terms the thesaurus table is used. The ranges indicate the standard deviation of the mean.

Table 3 Selectivity and reduction. All queries were performed using expanded thesaurus terms

Syndrome	*Localisation selectivity*	*Expression selectivity*		*Candidate gene selectivity*	
	Total	*or*	*and*	*or*	*and*
Acro-Dermato-Ungual-Lacrimal-Tooth Syndrome	1/116	1/2664	0/109	1/12	0/2
Alagille Syndrome	1/102	1/6435	1/420	1/17	1/3
Hand-Foot-Uterus Syndrome	1/148	1/4550	1/78	1/29	1/2
Holt-Oram Syndrome	1/154	1/4523	1/136	1/18	1/2
Multiple Synostoses Syndrome 1	1/273	1/846	0/102	1/10	0/5
Noonan Syndrome[1]	1/174	1/1090	1/349	1/21	1/12
Renal-Coloboma Syndrome	1/154	1/3774	1/504	1/33	1/7
Townes-Brocks Syndrome	1/49	1/1874	0/74	1/2	0/1
Tricho-Dento-Osseous Syndrome	1/322	1/3046	1/189	1/63	1/4
Ulnar-Mammary Syndrome	1/154	1/4523	0/136	1/18	0/2

The localisation selectivity represents the number of genes in the genetic region, whereas the expression selectivity reflects the number of genes that match the expression terms specified by the user; the candidate selectivity is the combination of the two. The Boolean operators indicate how the primary expression terms (table 2) are combined. [1]: This analysis was performed in October 2001.

recognition of *ROR2* as the Robinow syndrome gene[15, 16].

A systematic approach to this conservation of phenotypes has already been attempted and is presented in the Dysmorphic Human-Mouse Homology Database (DHMHD)[17]. Others have attempted to use cross-species conservation with invertebrates to identify genes that underlie human developmental syndromes and diseases[18].

All this argues for a systematic bioinformatics approach that includes all available information from public databases to prioritise among positional candidate genes. In a pilot experiment it was previously shown that it is possible to use a bioinformatics approach to identify plausible candidate genes for human multiple congenital anomaly syndromes by systematically using data on murine gene expression patterns[7].

We have since developed this data mining approach further to create a web-based tool that combines data on genetic localisation from OMIM, GDB, and MGD with data on gene expression from GXD and SWISSPROT/TrEMBL and data on phenotypes in humans

(OMIM, Medline) and mice (MLC, TBASE). This approach mimics the steps currently undertaken in most human genetics labs around the world once a critical region for a genetic disease is identified. The biggest advantage of the current automated approach is that it provides combined data from nine databases in a matter of minutes, rather than hours or days if individual databases are queried one gene at a time. Genetic localisation and expression databases were used in almost equal proportion. The number of hits per database varied greatly, but all contributed to the final selection of candidate genes. Of all databases, TBASE contributed the smallest number of genes. This is partly due to the fact that gene names in TBASE often do not conform to the nomenclature used in OMIM or GDB. Moreover, TBASE presently contains information on only a small number of genes.

The most successful search strategy with the GeneSeeker was by using the thesaurus in combination with *or* Boolean operators. More restrictive strategies failed in a significant proportion of cases, suggesting that the data in the databases is still incomplete, and that inappropriate search terms may have been used that failed to detect the presence of the gene in one or more databases. Some of these failures are to be expected as no single system has been adopted for scoring expression patterns and phenotype across the various databases. This situation is likely to improve considerably over the next few years. First, the genomic databases presently contain only draft versions of the genome with many genes yet to be identified, and properly annotated[19, 20]. In addition, efforts are currently underway to set up more complete databases on gene expression and on knockout phenotypes. As one example, a comprehensive inventory of expressed mouse genes during development is in progress[21]. Adding such databases to combined data mining strategies as presented here for the GeneSeeker may further improve their performance. Given the simplicity of the approach that is incorporated in the GeneSeeker tool, one might have expected that other similar applications might already exist. To the best of our knowledge this is not the case. Specifically, no program appears to be available that evaluates gene expression or phenotype information to aid with selection of positional candidate genes.

It is encouraging that all 10 causative genes were found for the human malformation syndromes with known genesis, and that this was accompanied by an on average 10-fold reduction compared to using localisation data only. We acknowledge that only prospective studies of syndromes that have not yet been defined molecularly can establish the true value of the bioinformatics tool described here. However, a considerable number of human disease genes have already been identified wholly or partly by virtue of comparing their mutant murine phenotypes and expression patterns. This by itself suggests that this approach can only become more effective as more information becomes available for each human gene. The modular setup employed in this first version of the GeneSeeker should allow easy expansion by adding further databases to improve the detection rate of disease genes. We are currently adding the Unigene database[22] and other EST database sources in order to expand the available information on expression patterns. SAGE (Serial Analysis of Gene Expression) data can also be added in the future, thereby further improving the sensitivity of the tool. Some text-based modules such as Medline may become more effective by using MeSH (Medical Subject Heading) terms and context sensitive searches.

Additional features like a comparison of old and new results, automatic selection of expression terms based on the OMIM clinical synopsis, and the ability to use STS marker data and physical coordinates rather than chromosome bands to specify genetic localisation are currently under development.

In its current form, the GeneSeeker is mainly suited for malformation syndromes in which the assumption can be made that the disease gene has an aberrant or absent gene expression in the affected tissues. For metabolic diseases other strategies can be applied, for example incorporating biochemical pathways such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)[23]. This can also be added to future versions of the program.

In spite of the obvious limitations of the system, even in its present early stage of development the GeneSeeker (version 2.0) offers researchers a useful tool to generate a starting list of candidate genes involved in human genetic disorders. The GeneSeeker site will be continuously updated and curated by the Centre for Molecular and Biomolecular Informatics at Nijmegen University. The current setup of the GeneSeeker relies on external databases. This means that regular checks of web addresses and database structures will be necessary to avoid losing individual databases. We intend to provide such regular follow-up, and note that the setting within a centre that provides support for more than 70 databases already would seem to be ideal for this. (Average downtime for these databases over the past 3 years has been less than 0.5%.) Also, the current GeneSeeker system has a number of advantages. Using external databases means that we avoid data warehousing. Therefore, all data are up to date and we would argue that in practice WWW front-ends are more stable than their underlying relational tables. In fact when changing the internals of the database, database developers often try to keep the WWW front-end unchanged. In conclusion, current developments in the availability of genomics data as well as improving bioinformatics strategies support the notion that data mining approaches as applied in the GeneSeeker may become a useful adjunct to wet lab experiments in human genetics.

Note added in proof

The program described here has previously been presented at scientific conferences and in abstracts as the ‘GeneMachine’[24]. In order to avoid confusion with a recently published program[25], we henceforth shall use the name GeneSeeker.

Acknowledgements

Partly supported by grants from N.W.O./Unilever (grant number 326756 to JAM Leunissen) and from the Irene kinderziekenhuis Foundation (to HG Brunner).

References

1. Letovsky, S.I., et al., *GDB: the Human Genome Database.* Nucleic Acids Res, 1998. **26**(1): p. 94-9.
2. Blake, J.A., et al., *The Mouse Genome Database (MGD): integration nexus for the laboratory mouse.* Nucleic Acids Res, 2001. **29**(1): p. 91-4.
3. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 45-8.
4. Ringwald, M., et al., *The Mouse Gene Expression Database (GXD).* Nucleic Acids Res, 2001. **29**(1): p. 98-101.
5. Woychik, R.P., et al., *TBASE: a computerized database for transgenic animals and targeted mutations.* Nature, 1993. **363**(6427): p. 375-6.
6. Stern, A.M., et al., *The hand-food-uterus syndrome: a new hereditary disorder characterized by hand and foot dysplasia, dermatoglyphic abnormalities, and partial duplication of the female genital tract.* J Pediatr, 1970. **77**(1): p. 109-16.
7. van Steensel, M.A., et al., *Probing the gene expression database for candidate genes.* Eur J Hum Genet, 1999. **7**(8): p. 910-9.
8. Propping, P. and K. Zerres, *ADULT-syndrome: an autosomal-dominant disorder with pigment anomalies, ectrodactyly, nail dysplasia, and hypodontia.* Am J Med Genet, 1993. **45**(5): p. 642-8.
9. Jamieson, C.R., et al., *Mapping a gene for Noonan syndrome to the long arm of chromosome 12.* Nat Genet, 1994. **8**(4): p. 357-60.
10. Tartaglia, M., et al., *Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome.* Nat Genet, 2001. **29**(4): p. 465-8.
11. Amiel, J., et al., *TP63 gene mutation in ADULT syndrome.* Eur J Hum Genet, 2001. **9**(8): p. 642-5.
12. Dryja, T.P., *Gene-based approach to human gene-phenotype correlations.* Proc Natl Acad Sci U S A, 1997. **94**(22): p. 12117-21.
13. den Hollander, A.I., et al., *Isolation and mapping of novel candidate genes for retinal disorders using suppression subtractive hybridization.* Genomics, 1999. **58**(3): p. 240-9.
14. Blackshaw, S., et al., *Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes.* Cell, 2001. **107**(5): p. 579-89.
15. van Bokhoven, H., et al., *Mutation of the gene encoding the ROR2 tyrosine kinase causes autosomal recessive Robinow syndrome.* Nat Genet, 2000. **25**(4): p. 423-6.
16. Afzal, A.R., et al., *Recessive Robinow syndrome, allelic to dominant brachydactyly type B, is caused by mutation of ROR2.* Nat Genet, 2000. **25**(4): p. 419-22.
17. van Steensel, M.A. and R.M. Winter, *Internet databases for clinical geneticists--an overview.* Clin Genet, 1998. **53**(5): p. 323-30.
18. Banfi, S., et al., *Identification and mapping of human cDNAs homologous to Drosophila mutant genes through EST database searching.* Nat Genet, 1996. **13**(2): p. 167-74.
19. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
20. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
21. Davidson, D., et al., *The Mouse Atlas Database: a community resource for mouse*

development. Trends Genet, 2001. **17**(1): p. 49-51.

22. Schuler, G.D., et al., *A gene map of the human genome.* Science, 1996. **274**(5287): p. 540-6.
23. Kanehisa, M., et al., *The KEGG databases at GenomeNet.* Nucleic Acids Res, 2002. **30**(1): p. 42-6.
24. Brunner, H.G., et al., *The Genemachine: A tool for the extraction and integration of information from web-based genetic databases.* Eur J Hum Genet, 2000. **8**: p. 130.
25. Makalowska, I., J.F. Ryan, and A.D. Baxevanis, *GeneMachine: gene prediction and sequence annotation.* Bioinformatics, 2001. **17**(9): p. 843-4.

Chapter 3

GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases

Marc A. van Driel, Koen Cuelenaere, Patrick P.C.W. Kemmeren,
Jack A.M. Leunissen, Han G. Brunner, Gert Vriend.

Nucleic Acids Research (2005) 33 (Web Server issue), W758-61.
(*With permission*)

Abstract

The identification of genes underlying human genetic disorders requires the combination of data related to cytogenetic localisation, phenotypes and expression patterns, to generate a list of candidate genes. In the field of human genetics, it is normal to perform this combination analysis by hand. We report on GeneSeeker (http://www.cmbi.ru.nl/GeneSeeker/), a web server that gathers and combines data from a series of databases. All database searches are performed via the web interfaces provided with the original databases, guaranteeing that the most recent data are queried, and obviating data warehousing. GeneSeeker makes the same selection of candidate genes as the human geneticist would have performed, and thus reducing the time-consuming process to a few minutes. GeneSeeker is particularly well suited for syndromes in which the disease gene displays altered expression patterns in the affected tissue(s).

Introduction

Identification of causative genes in human genetic disorders will be accelerated by the wealth of “omics” information being generated. Geneticists consult a number of databases to search for these genes. Each database concentrates on a different (molecular) aspect. In addition, databases have their own user interface, different formats to present the data, and sometimes even their own ontologies. Data such as gene localisation and expression patterns may be distributed over multiple databases.

Geneticists normally collect phenotypic and/or expression data and the genes in the chromosomal region(s) of interest, and combine these to get a list of candidate genes. The rationale for this is that the gene that causes a disease is likely to be expressed in the tissues affected by that disease[1-3]. Using model organisms, such as the mouse, it is often possible to obtain information on genes, proteins, protein interactions and other functional attributes that can be transferred to *Homo sapiens* by means of synteny and protein homology relations. The use of data from other species (such as mouse) often proves helpful in identifying the location or function of the equivalent human gene[4]. GeneSeeker mimics this multi-species identification strategy[5].

Material and Methods

Databases used.

Table 1 list the databases that GeneSeeker queries. These are divided over database groups (DB-groups). All databases are accessed through their standard WWW interfaces except MIMMAP and OXFORD. MIMMAP is a reformatted version of the OMIM[7] gene mapping information. OXFORD is used to translate human to mouse chromosomal locations, and is described in more detail in the pre-processing section. We use SRS (Sequence Retrieval System, Lion Biosciences, Cambridge, UK) to access these two databases[16]. The SRS parser was modified to allow searches for chromosomal ranges.

Data processing

The layout of the GeneSeeker web server is shown in figure 1. The user query consists of a chromosomal band range using standard nomenclature (e.g. “7p15–p21”). This cytogenetic localisation is passed through DB-group 1. Syntenic regions in the mouse are sought in DB-group 2 using an Oxford-grid. Tissues of interest or phenotypic features of a syndrome can be specified by the user as a Boolean expression that is split up and processed by DB-group 3. This modular set-up makes it easy to add extra DB-groups in the future. For every database, a plug-in was designed to perform all tasks from user-query pre-processing to query-result post-processing. These plug-ins deal with a series of technical topics, such as query reformatting, generating the correct URL, filling in the form on that database’s web interface, requesting all hits rather than in chunks, parsing the database HTML output, and so on.

Table 1 Databases accessed by the GeneSeeker

Database	**URL**
DB-group 1: localisation databases (human)	
OXFORD[6]	srs.bioasp.nl:4080
MIMMAP[7]	srs.bioasp.nl:4080
GDB[8]	www.gdb.org
DB group 2: localisation databases (Mouse)	
MGD[6]	www.informatics.jax.org
Datasets used in the interface	
GXD thesaurus	Van Steensel *et al.*[9]
Zuerich dataset	Brewer *et al.*[10, 11]
DB group 3: Expression/Phenotype databases	
PubMed (National Library of Medicine, Bethesda, Maryland, USA)	www.ncbi.nlm.nih.gov/pubmed
OMIM[7]	srs.bioasp.nl:4080
UniProt[12](SwissProt, TrEMBL, etc)	srs.bioasp.nl:4080
GXD[13]	www.informatics.jax.org
MLC[6]	www.informatics.jax.org
TBASE[14]	www.informatics.jax.org (was tbase, merged January 2005)
"Link out" database	
GeneCards[15]	bioinfo.weizmann.ac.il/cards/

The name of a gene can vary from database to database. The gene for the multi-drug resistance-associated protein 1, for example, is stored as *ABCC1*, *MRP,* or *MRP1,* depending on the database used. These gene nomenclature problems have to be solved because GeneSeeker depends on the gene names in the combination steps. For each DB-group the results are integrated with a Boolean OR. The resulting gene lists of the three DB-groups are combined according to the Boolean logic specified in the user query.

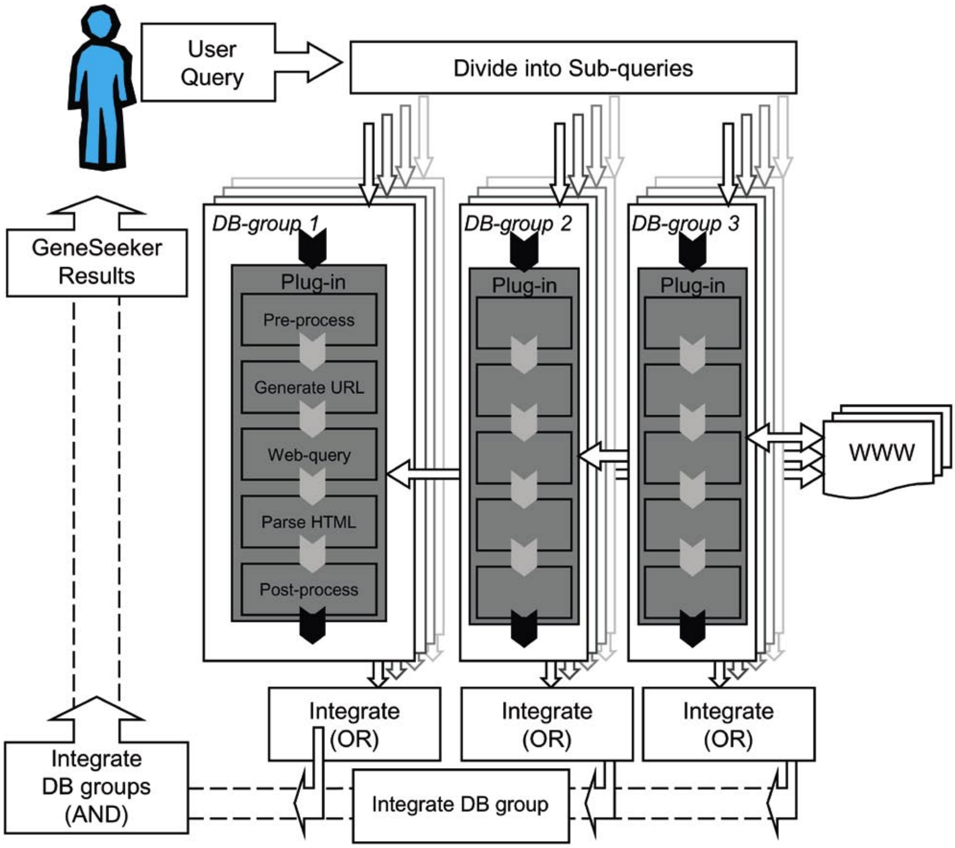
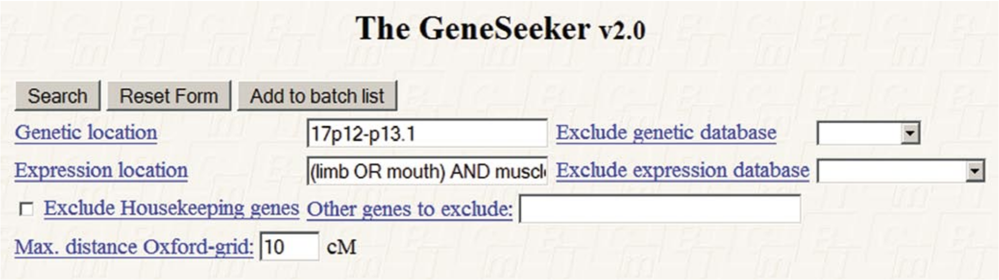


Figure 1 Overview of GeneSeeker. The query, which consists of a cytogenetic localisation, a phenotypic description and expression data, is divided over the three DB-groups that use the database-specific plug-ins to deal with all topics ranging from user-query pre-processing to post-processing of the query output. Results from each DB-group are merged with a Boolean OR. The results of the three DB-groups are combined as specified in the user query.

Implementation issues

Parallelisation.

The database plug-ins run in parallel to minimise the waiting time. A queuing system prevents excessive loads on remote servers. The plug-ins return the results of the queries to GeneSeeker as a list containing the gene names and corresponding database hyperlinks.



The GeneSeeker v2.0

Genetic location: 17p12-p13.1
Exclude genetic database:
Expression location: (limb OR mouth) AND muscl
Exclude expression database:
□ Exclude Housekeeping genes
Other genes to exclude:
Max. distance Oxford-grid: 10 cM

Figure 2 An example of a GeneSeeker query. Analyses of Trismus-Pseudocamptodactyly syndrome (TPC; MIM 158300) has been linked to 17p12-p13.1[18]. TPC is characterised by defects in muscle tissue mainly in limb and/or mouth. The options form is data not shown.

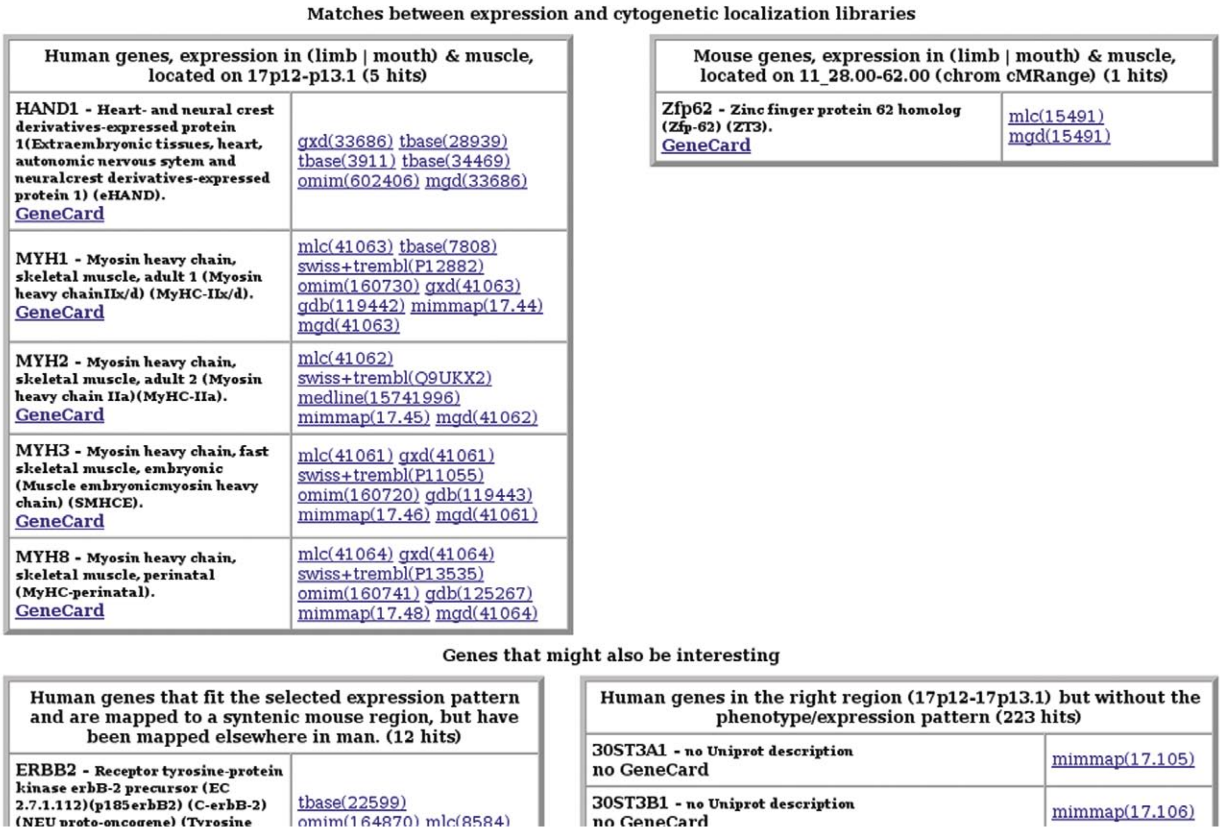
Mouse-Human synteny.

An Oxford grid[17] is used to find the homologous genes and gene regions in the mouse genome for all human chromosome locations entered by the user. A human chromosomal band range is translated into the corresponding mouse chromosome locations. Two mouse locations are combined if the genetic distance is shorter than a user-specified value (defaults to 10 cM). We regenerate this Oxford grid weekly to ensure that the latest synteny information is used in each query.

Gene Nomenclature. Inconsistent gene nomenclature is resolved using gene synonym information from UniProt database[12]. We use the MGD human homologues information to interconvert mouse and human gene names. We maintain local copies of these conversion tables because nearly all queries require that gene nomenclature problems be solved.

User interface

The GeneSeeker interface consists of the query form shown in figure 2 and an options form that usually requires no user input. A genetic localisation and the phenotypic/expression terms should be entered for a meaningful search. Databases that generate more noise than signal can be removed from the query. The user can also suppress the display of housekeeping genes or a specified list of genes. The options form contains a thesaurus[9] that can help the user to select the correct expression terms: for example, when the user is interested in a genetic trait that results in abnormalities in the brain, selection of the “brain” category returns the hints “brain *or* hindbrain *or* forebrain…”. Hints for the genetic localisation data can be found in a table containing frequently aberrant chromosomal bands in specific disorders taken from literature[10, 11]. The user can be notified on request about the completion of GeneSeeker searches by email. All parameters are linked to help screens. The results are presented in four tables (figure 3).



Matches between expression and cytogenetic localization libraries

Human genes, expression in (limb | mouth) & muscle, located on 17p12-p13.1 (5 hits)	

HAND1 - Heart- and neural crest derivatives-expressed protein 1(Extraembryonic tissues, heart, autonomic nervous sytem and neuralcrest derivatives-expressed protein 1) (eHAND).
GeneCard
	gxd(33686) tbase(28939)
	tbase(3911) tbase(34469)
	omim(602406) mgd(33686)

MYH1 - Myosin heavy chain, skeletal muscle, adult 1 (Myosin heavy chainIIx/d) (MyHC-IIx/d).
GeneCard
	mlc(41063) tbase(7808)
	swiss+trembl(P12882)
	omim(160730) gxd(41063)
	gdb(119442) mimmap(17.44)
	mgd(41063)

MYH2 - Myosin heavy chain, skeletal muscle, adult 2 (Myosin heavy chain IIa)(MyHC-IIa).
GeneCard
	mlc(41062)
	swiss+trembl(Q9UKX2)
	medline(15741996)
	mimmap(17.45) mgd(41062)

MYH3 - Myosin heavy chain, fast skeletal muscle, embryonic (Muscle embryonicmyosin heavy chain) (SMHCE).
GeneCard
	mlc(41061) gxd(41061)
	swiss+trembl(P11055)
	omim(160720) gdb(119443)
	mimmap(17.46) mgd(41061)

MYH8 - Myosin heavy chain, skeletal muscle, perinatal (MyHC-perinatal).
GeneCard
	mlc(41064) gxd(41064)
	swiss+trembl(P13535)
	omim(160741) gdb(125267)
	mimmap(17.48) mgd(41064)

Mouse genes, expression in (limb | mouth) & muscle, located on 11_28.00-62.00 (chrom cMRange) (1 hits)	

Zfp62 - Zinc finger protein 62 homolog (Zfp-62) (ZT3).
GeneCard
	mlc(15491)
	mgd(15491)

Genes that might also be interesting

Human genes that fit the selected expression pattern and are mapped to a syntenic mouse region, but have been mapped elsewhere in man. (12 hits)	

ERBB2 - Receptor tyrosine-protein kinase erbB-2 precursor (EC 2.7.1.112)(p185erbB2) (C-erbB-2) (NEU proto-oncogene) (Tyrosine	tbase(22599)
	omim(164870) mlc(8584)

Human genes in the right region (17p12-17p13.1) but without the phenotype/expression pattern (223 hits)	
30ST3A1 - no Uniprot description no GeneCard	mimmap(17.105)
30ST3B1 - no Uniprot description no GeneCard	mimmap(17.106)

Figure 3 The output of GeneSeeker for the Trismus-Pseudocamptodactyly syndrome query (see figure 2). It has been shown that mutations in the MYH8 gene can cause TPC[18]. Top left table: Genes that agree perfectly with the user query. Top right table: genes found in mouse syntenic regions that cannot be mapped automatically on the human genome, but match the expression pattern. Bottom left table: genes found in mouse syntenic regions that match the expression pattern, but map on the human genome outside the candidate cytogenetic region. Bottom right table: human genes in the candidate cytogenetic region that do not match the phenotype/expression pattern. All genes are hyperlinked to the underlying database, and, when possible, to GeneCards[15].

Results and Discussion

The GeneSeeker offers a user-friendly quick scan of several databases that are commonly used by geneticists to identify candidate genes for specific Mendelian diseases. As such, GeneSeeker uses those databases that are most appropriate for the questions asked. Several aspects are likely to change in the near future as genomics and genetics develop. For example, our usage of an Oxford grid can be improved or replaced as soon as consensus is reached about the localisation of genes on the mouse and human genomes among the various databases. Expression pattern information (e.g. microarray data) is growing rapidly, and is expected to become useful for GeneSeeker in the near future. At the moment, publicly available expression information is still sparse, scattered and not yet standardised.

In its present form, GeneSeeker is best suited for syndromes in which one can assume aberrant or absent gene expression in the affected tissues. GeneSeeker allows the user to query heterogeneous databases and obtain good candidate genes for the disease of interest based on positional, expression and model data[5]. With the present hardware set-up GeneSeeker can perform about 1000 searches per day.

Acknowledgements

We thank David Thomas for helpful corrections to the manuscript. This work was supported by N.W.O./Unilever, the Irene Kinderziekenhuis Foundation, and the EU FP6 Programme (LHSG-CT-2003-503265). Funding to pay the Open Access publication charges for this article was provided by Radboud University Nijmegen.

References

1. Blackshaw, S., et al., *Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes.* Cell, 2001. **107**(5): p. 579-89.
2. den Hollander, A.I., et al., *Isolation and mapping of novel candidate genes for retinal disorders using suppression subtractive hybridization.* Genomics, 1999. **58**(3): p. 240-9.
3. Dryja, T.P., *Gene-based approach to human gene-phenotype correlations.* Proc Natl Acad Sci U S A, 1997. **94**(22): p. 12117-21.
4. Chiang, A.P., et al., *Comparative genomic analysis identifies an ADP-ribosylation factor-like gene as the cause of Bardet-Biedl syndrome (BBS3).* Am J Hum Genet, 2004. **75**(3): p. 475-84.
5. van Driel, M.A., et al., *A new web-based data mining tool for the identification of candidate genes for human genetic disorders.* Eur J Hum Genet, 2003. **11**(1): p. 57-63.
6. Blake, J.A., et al., *MGD: the Mouse Genome Database.* Nucleic Acids Res, 2003. **31**(1): p. 193-5.
7. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.
8. Letovsky, S.I., et al., *GDB: the Human Genome Database.* Nucleic Acids Res, 1998. **26**(1): p. 94-9.
9. van Steensel, M.A., et al., *Probing the gene expression database for candidate genes.* Eur J Hum Genet, 1999. **7**(8): p. 910-9.
10. Brewer, C., et al., *A chromosomal deletion map of human malformations.* Am J Hum Genet, 1998. **63**(4): p. 1153-9.
11. Brewer, C., et al., *A chromosomal duplication map of malformations: regions of suspected haplo- and triplolethality--and tolerance of segmental aneuploidy--in humans.* Am J Hum Genet, 1999. **64**(6): p. 1702-8.
12. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32 Database issue**: p. D115-9.
13. Ringwald, M., et al., *The Mouse Gene Expression Database (GXD).* Nucleic Acids Res, 2001. **29**(1): p. 98-101.
14. Woychik, R.P., et al., *TBASE: a computerized database for transgenic animals and targeted mutations.* Nature, 1993. **363**(6427): p. 375-6.
15. Safran, M., et al., *Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE.* Nucleic Acids Res, 2003. **31**(1): p. 142-6.
16. Etzold, T. and P. Argos, *SRS--an indexing and retrieval tool for flat file data libraries.* Comput Appl Biosci, 1993. **9**(1): p. 49-57.
17. Edwards, J.H., *The Oxford Grid.* Ann Hum Genet, 1991. **55 (Pt 1)**: p. 17-31.
18. Veugelers, M., et al., *Mutation of perinatal myosin heavy chain associated with a Carney complex variant.* N Engl J Med, 2004. **351**(5): p. 460-9.

Chapter 4

From syndrome families to functional genomics

Han G. Brunner and Marc A. van Driel

Nat Rev Genet. 2004 Jul;5(7):545-51.
(*With permission*)

There are more than 2.000 monogenic syndromes in man. Each syndrome has a specific combination of phenotypic features, and each differs from other syndromes by only one or a few of those features. Could the ordering of phenotypes into syndrome families tell us about the relationships of the underlying genes? If so, such phenotype relationships could be systematically exploited to find new disease genes and provide clues to gene interactions, pathways and functions.

One way to define the function of the genes in a genome is to scrutinize mutant phenotypes in a systematic manner. Indeed, large-scale mutagenesis programmes that aim to do this have been completed for yeast and are underway for model systems such as *Caenorhabditis elegans*, the mouse and zebrafish. It is not clear how many different mutants will be required for such screens to be comprehensive. What is clear, however, is that creating a single knockout in a mouse model will rarely be sufficient to probe a gene's functions in development and homeostasis. In fact, to obtain a comprehensive view of gene function, the number of random mutations that need to be examined in detail at the phenotype level will probably be very large. Starting from interesting phenotypic differences and then comparing the mutations that underlie them might well be more effective. Spontaneous mutations are frequent in nature, and the study of their associated phenotypes can contribute to our understanding of gene function. This underlines the necessity for a phenotype-driven approach that saturates the genome with mutations, either with induced mutations (in animal models) or, in the case of humans, in disease states.

The study of naturally occurring mutations in humans has several advantages. First, differences in phenotypes reflect an alteration in a functional module, whether it is developmental or homeostatic. Second, the number of known monogenic phenotypes in humans is large, and many have been described in considerable clinical detail. Medical systems worldwide effectively operate as an enormous mutation screen of unequalled power. Because we are more likely to notice a subtle change in a human being than in a worm, we can detect the effects in more detail and at many more loci in humans than we can in any other species[1].

The vast amount of phenotypic information that is stored in a few accessible databases, such as Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/Omim; [2])[3], is another advantage of studying human disease phenotypes. Nonetheless, the systematic analysis of phenotype relationships has only recently begun to be explored[4-7].

Syndromes and syndrome families

To fully exploit the scientific potential of mutant phenotypes, we need to define them precisely[8]. The process of syndrome delineation starts at the clinical level, often from a few patients, and involves the continuous addition of new cases that help to establish and refine the phenotype[9]. Initially, cases of the syndrome that are identical or nearly identical to the original description are reported, and this might create the false impression that a syndrome is clinically homogeneous[9]. One possible way to avoid this would be to only include affected family members, and to exclude the index cases from the analyses[10]. Reports on patients with partly overlapping phenotypes frequently lead to debates between 'splitters and lumpers'[11]. As syndrome definition is initially intuitive and analytic, clinical arguments alone rarely resolve such debates. Only a few reports describe mathematical approaches to syndrome definition[12, 13], and syndrome diagnosis remains largely a matter of comparison with the 'ideal' aspect of the given syndrome. Once the gene for a syndrome has been found, a clearer picture emerges of what constitutes the core phenotype and its variants. However, having a molecular definition of a syndrome does not completely solve the problem of understanding its phenotypic variability: allelic mutations can be associated with considerable phenotypic diversity, and the action of modifier genes further adds to this variability[14]. Furthermore, as mutations in different genes can cause the same or related phenotypes, a strictly molecular classification would obscure the relationship between molecularly distinct syndromes at the phenotype level[15].

It has been recognized for many years that overlapping syndrome phenotypes might reflect biological relationships[16]. Designated 'phenotype communities' or 'syndrome families', these terms refer to groups of syndromes that share a large proportion of their key features. Because the process of syndrome delineation is iterative, syndrome phenotypes tend to evolve as more information becomes available. For this reason, and because of advances in molecular genetics, syndromes can merge or split over time, and can even disappear[17]. Consequently, syndrome families can only be arbitrarily defined as groups of syndromes that are coded as independent entities in expert databases, such as OMIM, the London Dysmorphology Database (LDDB; http://www.lmdatabases.com) or the Pictures of Standardized Syndromes and Undiagnosed Malformations (POSSUM; http://www.possum.net.au) database[2, 3]. In spite of these limitations, the syndrome-family concept has been remarkably successful in predicting allelic mutations, as in several skeletal dysplasias (BOX 1).

Biology of syndrome families

It is not immediately obvious that phenotypic overlap should be a reliable indicator of shared function. Perhaps perceived phenotypic similarities between syndromes are simply the result of our inability to appreciate the relevance of various discriminating features. For instance, we can ask whether the assignment of what seems to be a single human genetic disease to different genes represents diagnostic failure or shared biological mechanisms. From the examples given below, the latter would seem to be much more common.

The fact that mutations at different loci could lead to apparently the same human genetic disease was first recognized almost 50 years ago, when N. Morton showed that although some families with elliptocytosis were genetically linked to the Rhesus bloodgroup locus, other families were not[21]. It has since become clear that non-allelic genetic heterogeneity is frequent in Mendelian diseases, and that it can be extensive[22-26]. Genetic heterogeneity is commonly regarded as an obstacle to understanding human genetic disease, because it results

Box 1 An example of the syndrome-family approach: skeletal dysplasias

The syndrome-family approach was first systematically applied to the skeletal dysplasias. The number of these syndromes grew steadily from a single entity — achondroplasia — to more than 200 at the start of this century[18]. It was in 1985 that the German paediatrician J. Spranger proposed the syndrome-family system for grouping several clinically distinct skeletal dysplasias into a smaller number of categories, which he termed chondrodysplasia families[19], on the basis of key shared features. Almost 20 years on, molecular analysis has vindicated Spranger's classification. His concept of a Stickler–Kniest family has been shown to involve mutations in interacting collagen genes (see BOX 2), whereas the three members of the achondroplasia (weblink)[1] family of bone dysplasias all have mutations in fibroblast growth factor receptor 3 (FGFR3)[18]. Interestingly, recent data indicate that the Oto-Palato-Digital syndrome (OPD) and Larsen syndrome, which Spranger grouped together in a further bone-dysplasia family, arise through mutations in filamin A and B, respectively. Filamin A and B are two closely related members of a family of cytoplasmic proteins that regulate the structure and activity of the cytoskeleton by crosslinking actin into three-dimensional networks[20].

What is true for skeletal dysplasias is also true for other diseases. Human disease-phenotype groups might reflect the different functional domains of a single protein, the interaction between different proteins such as ligand and receptor, the interaction between proteins in a multiprotein complex and different steps in a cellular pathway.

in more ambiguous linkage assignments and slows down the gene-finding process. However, genetic heterogeneity can also be viewed in a different light: it might reflect interactions at the protein level, such as ligand–receptor interactions, the different subunits of a multiprotein complex or proteins that function at different steps of a metabolic pathway.

Such interactions have two implications: not only could we find the other genes more effectively once the first gene is found, but we can also speculate that apparently unrelated genes that are involved in the same phenotype will ultimately be shown to have a functional relationship. In this way, clinical classification could precede molecular verification, and syndromology could become a functional genomics tool.

From genetic heterogeneity to molecular interactions and pathways.

Fanconi anaemia (FA) is an example of a heterogeneous syndrome that has provided important information on the pathways that underlie it. This syndrome is an autosomal recessive disorder that is associated with cardiac, renal and limb malformations, as well as dermal pigmentary changes. Progressive bone-marrow disease can lead to bone-marrow failure and leukaemia. Somatic-cell hybrid studies that defined multiple complementation groups first showed that FA was genetically heterogeneous[27]; since then, at least nine FA loci have been identified and seven genes have been cloned. The function of each of these genes was unknown at the time that they were identified, although given that at the cellular level, FA was characterized by chromosome breakage, it was probable that these genes were involved in DNA repair.

[1] Weblinks: These terms are linked online to databases that are listed at the end of this paper.

Subsequently, it was shown that the proteins that the FA genes A, C, F, E and G encode indeed form a complex that functions in DNA repair. The FA D2 protein (FANCD2) functions as a downstream effector of this complex, and the latest FA gene to be identified (FANCL) encodes a putative ubiquitin ligase that might be responsible for the mono-ubiquitylation of FANCD2[22, 28]. So, the molecular genetics of a single syndrome can define a previously unknown multiprotein complex, which in this example has an important role in DNA repair.

The converse process is equally frequent in human molecular genetics. If a molecular pathway is already known, this knowledge can be used to accelerate the gene-finding process for a genetically heterogeneous disease. This candidate-gene approach was used to show that recessive mutations in each of the five subunits of translation initiation factor eIF2B cause the same recessively inherited leukoencephalopathic brain disease, which is characterized by vanishing white matter[29].

There are also several examples for which the same phenotype is caused by the mutation of either the receptor or the ligand gene for a signalling step. A striking example of this is the combined phenotype of pre-senile dementia and bone cysts. Mutations in a gene that encodes a membrane receptor on natural-killer and myeloid cells, and in the gene that encodes a ligand for this receptor, each resulted in this disease phenotype[30].

Syndrome families and pathways.

What is true for a single genetically heterogeneous disease might equally apply to distinct syndromes with phenotypic overlap. One such syndrome family comprises the Walker–Warburg syndrome, Fukuyama Muscular Dystrophy and muscle–eye–brain disease (MEBD (weblink)). These diseases are characterized by abnormal neuronal migration, variable eye involvement and congenital muscular dystrophy. Although these syndromes can usually be distinguished on clinical grounds alone, there is also a clear overlap in the features that define them, and it has previously been suggested that they form a clinical spectrum of diseases[31]. Biochemical staining of muscle specimens has shown that all three conditions share a defect in the glycosylation of proteins such as α-Dystroglycan. The predicted biochemical function for the three genes that are involved in Walker–Warburg syndrome (weblink), Fukuyama muscular dystrophy (weblink) and MEBD is consistent with a shared role in O-Mannosylation[32]. So, these three independent but clinically similar syndromes represent the disruption of different steps in the same biochemical pathway.

Similar interactions among gene products occur in Usher syndrome types 1B, 1C, 1D and 1G[26] and among Waardenburg syndrome types 1, 2 and 4[33, 34]. The clinical distinction of Waardenburg syndrome types 1, 2 and 4 pre-dated the identification of the causative genes by many years. As is common in these situations, both the clinical differences as well as the similarities turned out to be relevant for the various forms of Waardenburg syndrome. When the genes were found to be distinct, the clinical similarities immediately stimulated research to identify the functional module that links them together at the level of transcription.

Given that mutations that affect different steps in a pathway can cause the same monogenic phenotype, in some instances, the combined effect of mutations in different genes might interact to produce digenic or polygenic inheritance of human-malformation syndromes. Indeed, several examples of digenic or triallelic inheritance in humans have been recorded[35]. The most striking is Bardet–Biedl Syndrome (BBS (weblink)). In this syndrome, there is genetic interaction between several of the loci. Although the functions of the BBS genes remain largely unknown, the additive effect of mutations at *BBS1*, *BBS2*, *BBS4*, *BBS6* (weblink) (*MKKS*) and *BBS7* (weblink) indicates that the products of the different *BBS* genes share at least part of their functions[24, 36]. Consistent with this prediction is the recent discovery that all known *BBS* genes for which there are homologues in *C. elegans* are expressed exclusively in ciliated neurons[37].

Exceptions to the rule

If different phenotypes result from mutations in a single gene, the probable explanation is that the mutations disrupt different functions that are encoded by that gene. Detailed genotype–phenotype analysis might indicate the localisation of such functions within a gene and the protein that it encodes.

For example, mutations in the *XPD* (weblink) (*ERCC2*) gene can cause either Xeroderma Pigmentosum (XP) or TrichoThioDystrophy (TTD (weblink))[38]. Sensitivity to sunlight with the development of cancer at an early age is a leading feature of XP. By contrast, patients with TTD have brittle hair and nails, ichthyotic skin, and physical and mental retardation. Many, but not all, patients with TTD are sensitive to sunlight, but they do not have any unusual pigmentation changes, and there are no reports of cancer in TTD patients. The *XPD* gene encodes a subunit of the basal transcription factor TFIIH. The TFIIH complex has two principal independent functions, one in the initiation of basal transcription and the other in DNA repair. It has been proposed that if the defect in the *XPD* gene affects the DNA-repair function of TFIIH without affecting its transcriptional role, XP will result. On the other hand, if the transcriptional role of TFIIH is affected, the consequence will be the developmental defects that are found in TTD[38]. The hypothesis that XP is a repair syndrome and that TTD is a transcription syndrome predicts that the mutations that are associated with the two disorders could be located at different sites in the gene. Indeed, the mutation spectrum is consistent with the hypothesis that the site of the mutation in the *XPD* gene determines the clinical phenotype[39].

Considerable phenotypic heterogeneity has also been demonstrated for the *p63* gene. At least six clinically different syndromes are caused by mutations in this gene. Here too, different classes of mutation cause different syndromes. The most parsimonious explanation for this pattern is that each phenotype reflects the disruption of a specific function or set of functions[40].

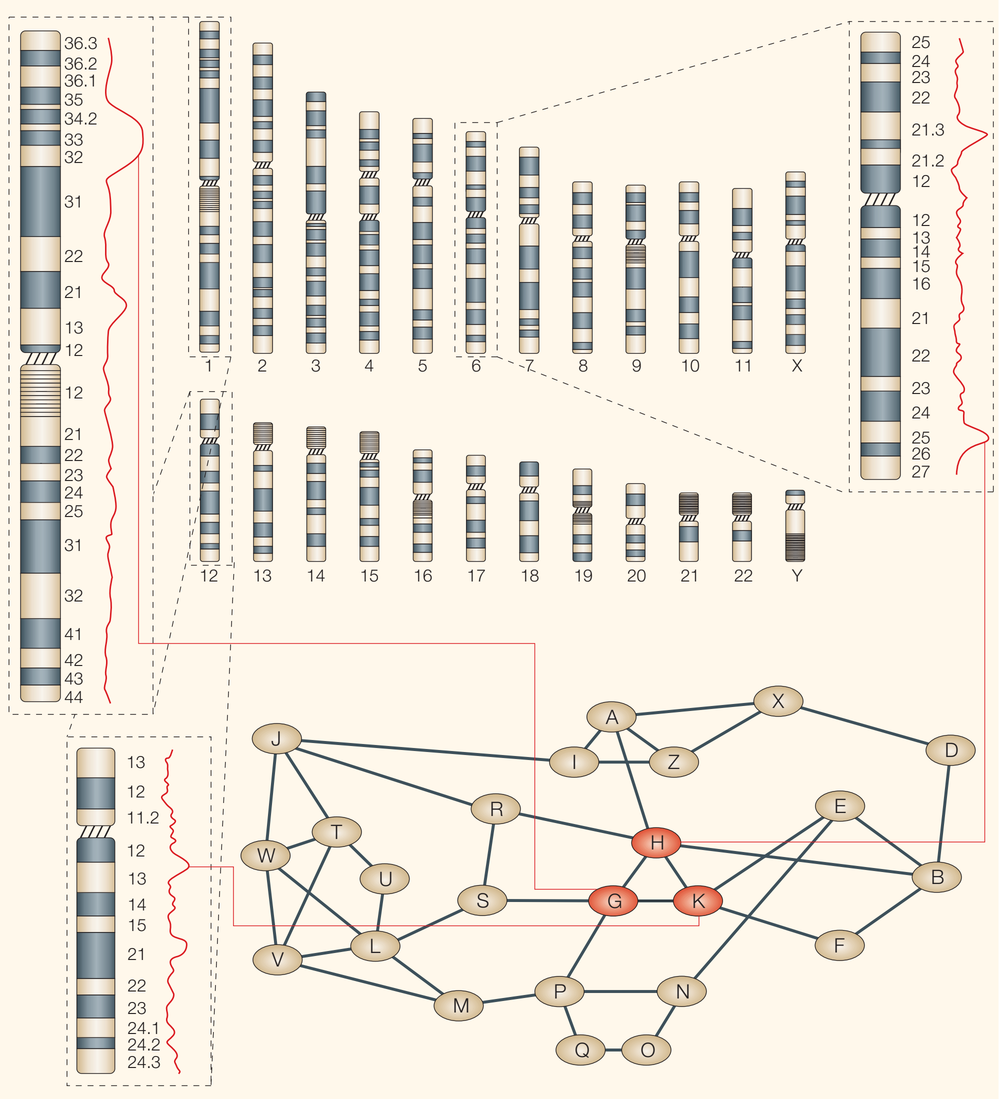


Figure 1 Integrating functional relationships and linkage analysis. Three separate genome scans for a hypothetical multifactorial disease result in multiple peaks of possible linkage (red traces) on 3 different human chromosomes (1, 6 and 12). Integrating linkage results with data from other sources, such as gene/protein networks (A … Z), points to a set of functionally related candidate genes (H, G and K) that together might explain the linkage patterns. Numbers along the highlighted chromosomes indicate the corresponding chromosome bands. (Colour version: see appendix 2)

Multifactorial disease

Phenotype clustering might equally be applied to multifactorial diseases. Although one or several distinct processes might be involved in disease causation for a multifactorial disease, we can predict that at least some of the individual genes will converge on a single biological process. A pertinent example might be Alzheimer disease (weblink), for which four genes are known. These Alzheimer genes are all involved in the proteolytic cleavage of amyloid precursor protein[41].

Current studies of multifactorial diseases do not take pathway information into account when analysing linkage data. We propose that bioinformatics analysis of the various interaction and pathway data sets (see, for example, refs [42, 43]) can be used to find the mechanism that links different genetic loci for a multifactorial disease. In this way, if a pattern emerges, it will immediately indicate the involvement of specific candidate genes (figure 1). These candidate genes can then be tested in more detail by examining SNPs that cover each of this set of genes at the different loci.

Pathway information can also be used to construct testable hypotheses about the additive or interactive effects of these loci. This strategy might aid the interpretation of linkage data that show that different loci are involved in a single disease. It would not be surprising to find that although the relative importance of a locus might differ between populations because of differences in the frequencies of mutant alleles, the loci themselves still form a recognizable pattern from a biological point of view. Whereas this combined linkage and data-mining approach remains to be tested for multifactorial diseases, a conceptually similar bioinformatics strategy showed promising results for several oligogenic diseases[44].

Bioinformatics and syndrome families

If we cease to view phenotypes as independent entities and apply the syndrome-family approach, we might discover new relationships between genes, proteins or cellular pathways. By clustering the known human phenotypes into groups, we can examine whether relationships at the phenotype level reflect shared functions at other biological levels, such as the proteome, genome or interactome. In principle, the biological order that underlies phenotype classification can now easily be explored in a systematic fashion. So far, only a few attempts at using phenotype clustering for the prediction of disease genes have been undertaken[7, 44].

OMIM contains 15.000 full-text records[2]. Of these, 5.000 describe a human phenotype, including some 2.000 syndromes. For approximately 1.200 human phenotypes, the corresponding genes are known. OMIM therefore holds data that can be used to validate ideas about phenotype-to-genotype relationships. However, although OMIM provides a wealth of information, its full-text character makes it difficult to analyse systematically. OMIM also lacks a standardized vocabulary and knows few rules for the organization and representation of its data. Other more specialized syndrome databases are not yet accessible through the Internet, and also lack a clear system for assigning the relative importance of the various syndrome features[3]. Finding similar phenotypic descriptions in such databases therefore requires text-analysis techniques.

Freudenberg and Propping manually extracted and analysed nearly 1.000 OMIM phenotype entries of diseases with known genetic origin using episodic occurrence, primary aetiology, primary tissue, mode of inheritance and age of onset as classification indices[7]. They collected approximately 10.000 Gene Ontology (GO)-annotated genes, and showed a good correlation between phenotypic clusters and GO-annotation clusters. Indeed, with stringent cut-offs, the correct gene was among the top 1.5% of the list in one-third of all cases.

We have used fully automated text mining to analyse all OMIM disease records. We used the medical subject headings (MeSH) as a repository of keywords. The MeSH system is organized as a hierarchical tree, in which 'nail' is part of 'finger', finger is part of 'hand' and hand is part of 'extremity'. In this way, it is possible to correct for differences in the level of detail of a description. To allow this process to be automated, the keyword frequencies were represented as vectors, with one vector per OMIM record (figure 2). Similar phenotypic descriptions have similar keyword frequencies, and therefore similar keyword vectors. We corrected for the length of the record, and applied the inverse-document-frequency technique to compensate, at least partly, for keyword frequency differences[45]. We then determined text-vector similarities.

Data from several test cases confirm the hypothesis that phenotype clustering — seen as high similarity between MeSH terms — is the result of underlying mutations that are located in genes that are involved in a similar function (figure 2). The approach was successfully tested for Stickler syndrome (BOX 2) and for Pallister–Hall syndrome (BOX 3). These are just two selected examples, but an analysis of all data (H.G.B. and M.D., unpublished observations (weblink)) shows a clear relationship between the phenotypic similarity scores and genetic similarity as measured by the Protein Families Database (Pfam: http://www.sanger.ac.uk/Software/Pfam), GO-annotations and even sequence alignments. These data strongly indicate that the principles that underlie phenotype clusters and syndrome families are relevant to all human genetic diseases, and should be explored further.

Limitations of the approach

One limitation of current methods for phenotype clustering relates to imprecisions in our methods to accurately and objectively define the phenotypes. Notably, OMIM is not a specialized database for syndromes, and other more comprehensive syndrome databases, such as POSSUM or LDDB, still do not provide estimates of the frequencies of the various syndrome features[3]. So, phenotype descriptions still contain considerable imprecision, as well as a significant subjective component.

In addition, the biological level at which the phenotype is determined might vary so that, for some syndromes, the unit of similarity is an entire organ or embryonic structure. For example, the syndromes that share split hand/foot malformation as one component feature are probably united simply because the underlying genes are involved in maintaining the proper function of the apical ectodermal ridge during embryogenesis[51].

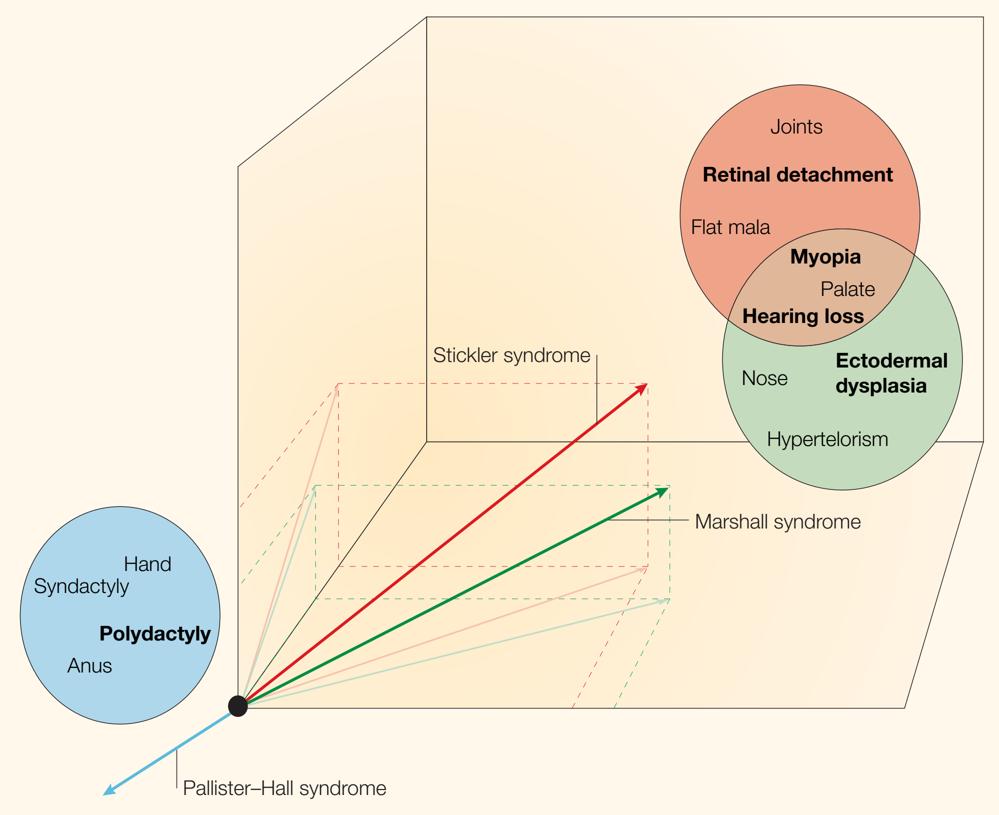
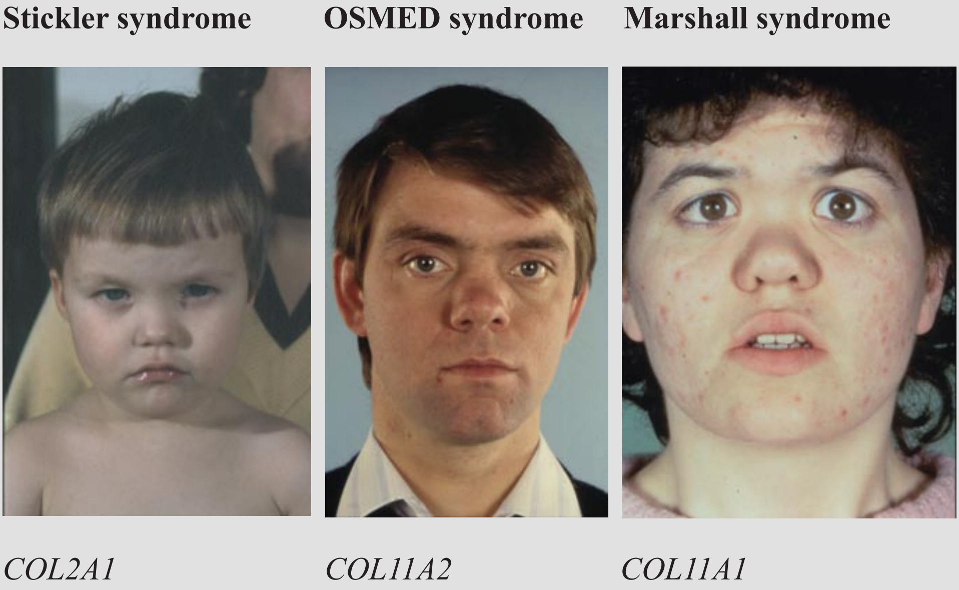


Figure 2 Phenotype clustering. Each arrow represents a keyword vector. The components in a keyword vector correspond to terms in the document (for example, 'myopia', 'joints', and so on). Vectors that point in the same direction are more alike. Three documents that describe three syndromes are shown (red: Stickler syndrome; green: Marshall syndrome; and blue: Pallister–Hall syndrome). Stickler and Marshall syndromes have similar phenotypes and share common terms ('myopia', 'palate', 'hearing loss') in their keyword vectors. Pallister–Hall syndrome (weblink) is clearly different, and shares few terms with Stickler and Marshall, and therefore, the vector points in a different direction. Some terms are more important than other terms. Terms are assigned different weights by computing the inverse document frequency, which gives a weight factor to each term in the collection. Terms in bold contribute the most to vector direction. Flat mala, low cheekbones; hypertelorism, widely-spaced eyes. (Colour version: see appendix 2)

Although the examples given here support the idea that genes that are involved in similar phenotypes will probably share some functional attribute, this will not always be true. For instance, whereas two genes that are involved in various types of osteopetrosis function in the acidification of the bone matrix[52, 53], another gene for osteopetrosis seems to work through a different mechanism[54].

Similarly, in some instances, a multifactorial disease such as atherosclerosis will result from genes that are involved in distinct process, such as lipid metabolism and hypertension,

Box 2 The Stickler syndrome family



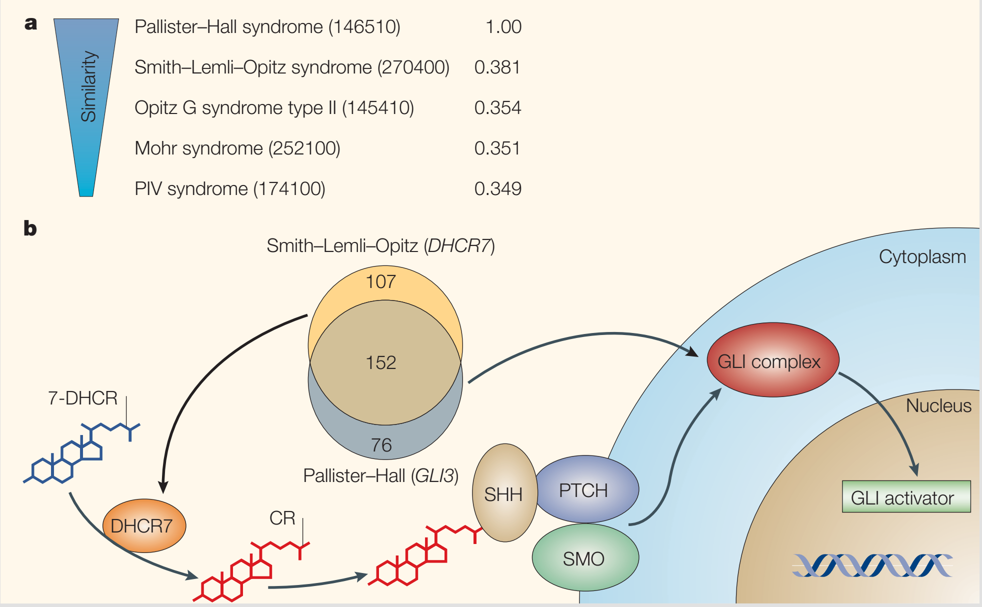
Stickler syndrome **OSMED syndrome** **Marshall syndrome**

COL2A1 *COL11A2* *COL11A1*

Stickler Syndrome is characterized by myopia, retinal detachment, hearing loss, cleft palate and arthropathy. Using the keyword-vector method for Stickler syndrome, several related syndromes were recovered: notably the Marshall and oto-spondylo-megaepiphyseal dysplasia (OSMED (weblink)) syndromes have similar facial and skeletal features (see figure) as Stickler syndrome, but differ in other characteristics such as deafness or blindness. Marshall and Stickler syndromes have been the subject of a long-standing debate about their status as independent entities[46]. Molecular investigations have since identified the mutations that cause these three disorders. Mutations in COL2A1 (weblink), COL11A2 (weblink) and COL11A1 (weblink) are responsible for the Stickler, OSMED and Marshall syndromes, respectively. All three genes encode fibrillar collagens that associate to form a single trimeric collagen-11 protein molecule[47]. Interestingly, Spranger grouped the Stickler, Marshall and OSMED syndromes together in a single chondrodysplasia family (see BOX 1). Figure modified with permission from ref. [47] © (1999) University of Chicago Press and ref. [48] © (1997) Wiley.

without clear molecular relationships. Nonetheless, for a concept to work, it does not need to be right all the time. If the syndrome family concept generates testable hypotheses, that in itself is sufficient justification for its application at the interface between clinical genetics and molecular biology. In fact, reports that two inherited forms of rickets, as well as a tumour-associated form of the disease, can be explained by abnormalities of the same module (involving either fibroblast growth factor 23 (FGF23 (weblink)) or its interacting molecule PHEX (weblink)) show that the implications of this concept might sometimes include non-genetic disease[55, 56].

Box 3 A family-syndrome approach to Pallister–Hall syndrome



Pallister–Hall syndrome (146510)	1.00
Smith–Lemli–Opitz syndrome (270400)	0.381
Opitz G syndrome type II (145410)	0.354
Mohr syndrome (252100)	0.351
PIV syndrome (174100)	0.349

For Pallister–Hall (PH) syndrome, a query with the keyword vector returns Smith–Lemli–Opitz (SLO (weblink)) syndrome as its nearest phenotypic neighbour (see figure part a; note that numbers in brackets after syndrome name refer to OMIM numbers; numbers in column to the right of the syndrome indicate proportion of keywords shared with Pallister–Hall syndrome) — these phenotypes share 152 keywords out of the 259 that describe SLO and the 228 that describe PH syndrome (see figure part b). This phenotypic similarity has been noted before[49]. It is now clear that the SLO mutation in DHCR7 (weblink) decreases cholesterol synthesis, and that this impairs sonic hedgehog (SHH) function. One of the downstream effectors of SHH is GLI3 (weblink), the gene that is mutated in Pallister–Hall syndrome (see figure part b). It would therefore seem that the sharing of phenotypic features between these syndromes might result from the fact that the causative mutations each disrupt SHH signalling through GLI3. Interestingly, a mutation of GLI3 also causes PIV (weblink) (polydactyly, imperforate anus and vertebral anomalies)[50]. Note also that other syndromes in the list given in part a either involve an apparently unrelated molecular mechanism (Opitz G syndrome (weblink)), or have not yet been determined at the molecular level (Mohr syndrome (weblink)). 7-DHCR, 7-dehydrocholesterol; CR, cholesterol; DHCR7, 7-dehydrocholesterol reductase; GLI3, zinc finger protein GLI3; PTCH, patched protein homologue 1; SMO, smoothened homologue precursor.

Conclusions

Of the ways in which we predict the function of human genes, mutant phenotypes are among the most reliable. Similar phenotypes can be assumed to result from the mutation of genes that are involved in the same biological process, until proved otherwise. By establishing the relationships among genes at the phenotype level, we can make predictions about interactions at the protein level. Viewed in this way, the human-phenotype collection is a powerful functional genomics tool.

To fully exploit the power of human phenotypes, the first requirement is that we develop more objective ways to define and quantify them. This is an important undertaking that requires that the presence or absence of features is rigorously documented in large numbers of patients[8]. Having a more objective assessment of the relative weights of features for a specific syndrome diagnosis is a further requirement. In addition, we have yet to develop new ways of performing multilocus linkage and association analysis that incorporate information on gene functions and gene relationships to pinpoint functional modules that could be causally involved in the disease under study. These reservations notwithstanding, we might expect that phenotype comparisons will continue to hint at molecular interactions and pathways that await discovery. In this sense, the future looks bright for phenotype clustering as a functional genomics tool.

Databases

(weblinks; see also appendix 1)

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nih.gov/Entrez
BBS1 | BBS2 | BBS4 | BBS6 (MKKS) | BBS7 | COL11A1 | COL11A2 | COL2A1 | DHCR7 | FANCD2 | FANCL | FGF23 | GLI3 | PHEX | XPD (ERCC2)

OMIM: http://www.ncbi.nlm.nih.gov/Omim
achondroplasia | Alzheimer disease | BBS | Fanconi anaemia | Fukuyama muscular dystrophy | Mohr syndrome | MEBD | Opitz G syndrome | OSMED | Pallister–Hall syndrome | PIV | SLO | TTD | Walker–Warburg syndrome

Further information

H.G.B. and M.D., unpublished observations: http://www.cmbi.kun.nl/articles/

References

1. Donnai, D. and A.P. Read, *How clinicians add to knowledge of development.* Lancet, 2003. **362**(9382): p. 477-84.
2. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.
3. Evans, C.D., *Computer systems in dysmorphology.* Clin Dysmorphol, 1995. **4**(3): p. 185-201.
4. Tao, Y.C. and R.L. Leibel, *Identifying functional relationships among human genes by systematic analysis of biological literature.* BMC Bioinformatics, 2002. **3**(1): p. 16.
5. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
6. van Driel, M.A., et al., *A new web-based data mining tool for the identification of candidate genes for human genetic disorders.* Eur J Hum Genet, 2003. **11**(1): p. 57-63.
7. Freudenberg, J. and P. Propping, *A similarity-based method for genome-wide prediction of disease-relevant human genes.* Bioinformatics, 2002. **18 Suppl 2**: p. S110-5.
8. Freimer, N. and C. Sabatti, *The human phenome project.* Nat Genet, 2003. **34**(1): p. 15-21.
9. Cohen, M.M., Jr., *Syndromology: an updated conceptual overview. III. Syndrome delineation.* Int J Oral Maxillofac Surg, 1989. **18**(5): p. 281-5.
10. Fraser, F.C. and A. Lytwyn, *Spectrum of anomalies in the Meckel syndrome, or: "Maybe there is a malformation syndrome with at least one constant anomaly".* Am J Med Genet, 1981. **9**(1): p. 67-73.
11. McKusick, V.A., *On lumpers and splitters, or the nosology of genetic disease.* Perspect Biol Med, 1969. **12**(2): p. 298-312.
12. Verloes, A., *Numerical syndromology: a mathematical approach to the nosology of complex phenotypes.* Am J Med Genet, 1995. **55**(4): p. 433-43.
13. Preus, M., *Numerical classification of syndromes.* Hosp Pract (Off Ed), 1985. **20**(6): p. 111-8, 127-9.
14. Romeo, G. and V.A. McKusick, *Phenotypic diversity, allelic series and modifier genes.* Nat Genet, 1994. **7**(4): p. 451-3.
15. Biesecker, L.G., *Lumping and splitting: molecular biology in the genetics clinic.* Clin Genet, 1998. **53**(1): p. 3-7.
16. Pinsky, L., *The polythetic (phenotypic community) system of classifying human malformation syndromes.* Birth Defects Orig Artic Ser, 1977. **13**(3A): p. 13-30.
17. Lindeman-Kusse, M.C., et al., *Cytogenetic abnormalities in two new patients with Pitt-Rogers-Danks phenotype.* Am J Med Genet, 1996. **66**(1): p. 104-12.
18. Superti-Furga, A., L. Bonafe, and D.L. Rimoin, *Molecular-pathogenetic classification of genetic disorders of the skeleton.* Am J Med Genet, 2001. **106**(4): p. 282-93.
19. Spranger, J., *Pattern recognition in bone dysplasias.* Prog Clin Biol Res, 1985. **200**: p. 315-42.
20. Krakow, D., et al., *Mutations in the gene encoding filamin B disrupt vertebral segmentation, joint formation and skeletogenesis.* Nat Genet, 2004. **36**(4): p. 405-10.

21. Morton, N.E., *The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type.* Am J Hum Genet, 1956. **8**(2): p. 80-96.
22. D'Andrea, A.D. and M. Grompe, *The Fanconi anaemia/BRCA pathway.* Nat Rev Cancer, 2003. **3**(1): p. 23-34.
23. Roberts, E., et al., *Autosomal recessive primary microcephaly: an analysis of locus heterogeneity and phenotypic variation.* J Med Genet, 2002. **39**(10): p. 718-21.
24. Beales, P.L., et al., *Genetic interaction of BBS1 mutations with alleles at other BBS loci can result in non-Mendelian Bardet-Biedl syndrome.* Am J Hum Genet, 2003. **72**(5): p. 1187-99.
25. Chiurazzi, P., B.C. Hamel, and G. Neri, *XLMR genes: update 2000.* Eur J Hum Genet, 2001. **9**(2): p. 71-81.
26. Weil, D., et al., *Usher syndrome type I G (USH1G) is caused by mutations in the gene encoding SANS, a protein that associates with the USH1C protein, harmonin.* Hum Mol Genet, 2003. **12**(5): p. 463-71.
27. Zakrzewski, S. and K. Sperling, *Genetic heterogeneity of Fanconi's anemia demonstrated by somatic cell hybrids.* Hum Genet, 1980. **56**(1): p. 81-4.
28. Meetei, A.R., et al., *A novel ubiquitin ligase is deficient in Fanconi anemia.* Nat Genet, 2003. **35**(2): p. 165-70.
29. van der Knaap, M.S., et al., *Mutations in each of the five subunits of translation initiation factor eIF2B can cause leukoencephalopathy with vanishing white matter.* Ann Neurol, 2002. **51**(2): p. 264-70.
30. Paloneva, J., et al., *Mutations in two genes encoding different subunits of a receptor signaling complex result in an identical disease phenotype.* Am J Hum Genet, 2002. **71**(3): p. 656-62.
31. Cormand, B., et al., *Clinical and genetic distinction between Walker-Warburg syndrome and muscle-eye-brain disease.* Neurology, 2001. **56**(8): p. 1059-69.
32. Beltran-Valero de Bernabe, D., et al., *Mutations in the O-mannosyltransferase gene POMT1 give rise to the severe neuronal migration disorder Walker-Warburg syndrome.* Am J Hum Genet, 2002. **71**(5): p. 1033-43.
33. Potterf, S.B., et al., *Transcription factor hierarchy in Waardenburg syndrome: regulation of MITF expression by SOX10 and PAX3.* Hum Genet, 2000. **107**(1): p. 1-6.
34. Bondurand, N., et al., *Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome.* Hum Mol Genet, 2000. **9**(13): p. 1907-17.
35. Ming, J.E. and M. Muenke, *Multiple hits during early embryonic development: digenic diseases and holoprosencephaly.* Am J Hum Genet, 2002. **71**(5): p. 1017-32.
36. Badano, J.L., et al., *Heterozygous mutations in BBS1, BBS2 and BBS6 have a potential epistatic effect on Bardet-Biedl patients with two mutations at a second BBS locus.* Hum Mol Genet, 2003. **12**(14): p. 1651-9.
37. Ansley, S.J., et al., *Basal body dysfunction is a likely cause of pleiotropic Bardet-Biedl syndrome.* Nature, 2003. **425**(6958): p. 628-33.
38. Bootsma, D. and J.H. Hoeijmakers, *DNA repair. Engagement with transcription.* Nature, 1993. **363**(6425): p. 114-5.
39. Lehmann, A.R., *The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases.* Genes Dev, 2001. **15**(1): p. 15-23.
40. Brunner, H.G., B.C. Hamel, and H. Van Bokhoven, *The p63 gene in EEC and other*

syndromes. J Med Genet, 2002. **39**(6): p. 377-81.

41. Kennedy, J.L., et al., *The genetics of adult-onset neuropsychiatric disease: complexities and conundra?* Science, 2003. **302**(5646): p. 822-6.
42. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
43. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.
44. Turner, F.S., D.R. Clutterbuck, and C.A. Semple, *POCUS: mining genomic sequence annotation to predict disease genes.* Genome Biol, 2003. **4**(11): p. R75.
45. Wilbur, W.J. and Y. Yang, *An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts.* Comput Biol Med, 1996. **26**(3): p. 209-22.
46. Ayme, S. and M. Preus, *The Marshall and Stickler syndromes: objective rejection of lumping.* J Med Genet, 1984. **21**(1): p. 34-8.
47. Annunen, S., et al., *Splicing mutations of 54-bp exons in the COL11A1 gene cause Marshall syndrome, but other mutations cause overlapping Marshall/Stickler phenotypes.* Am J Hum Genet, 1999. **65**(4): p. 974-83.
48. van Steensel, M.A., et al., *Oto- spondylo-megaepiphyseal dysplasia (OSMED): clinical description of three patients homozygous for a missense mutation in the COL11A2 gene.* Am J Med Genet, 1997. **70**(3): p. 315-23.
49. Donnai, D., J. Burn, and H. Hughes, *Smith-Lemli-Opitz syndromes: do they include the Pallister-Hall syndrome?* Am J Med Genet, 1987. **28**(3): p. 741-3.
50. Killoran, C.E., et al., *Overlap of PIV syndrome, VACTERL and Pallister-Hall syndrome: clinical and molecular analysis.* Clin Genet, 2000. **58**(1): p. 28-30.
51. Duijf, P.H., H. van Bokhoven, and H.G. Brunner, *Pathogenesis of split-hand/split-foot malformation.* Hum Mol Genet, 2003. **12 Spec No 1**: p. R51-60.
52. Kornak, U., et al., *Mutations in the a3 subunit of the vacuolar H(+)-ATPase cause infantile malignant osteopetrosis.* Hum Mol Genet, 2000. **9**(13): p. 2059-63.
53. Kornak, U., et al., *Loss of the ClC-7 chloride channel leads to osteopetrosis in mice and man.* Cell, 2001. **104**(2): p. 205-15.
54. Chalhoub, N., et al., *Grey-lethal mutation induces severe malignant autosomal recessive osteopetrosis in mouse and human.* Nat Med, 2003. **9**(4): p. 399-406.
55. Bowe, A.E., et al., *FGF-23 inhibits renal tubular phosphate transport and is a PHEX substrate.* Biochem Biophys Res Commun, 2001. **284**(4): p. 977-81.
56. Shimada, T., et al., *Cloning and characterization of FGF23 as a causative factor of tumor-induced osteomalacia.* Proc Natl Acad Sci U S A, 2001. **98**(11):p. 6500-5.

Chapter 5

A text-mining analysis of the human phenome

Marc A. van Driel, Jorn Bruggeman, Gert Vriend,
Han G. Brunner, Jack A.M. Leunissen

submitted

Abstract

Background
A number of large-scale efforts are underway to define the relationships between genes and proteins in various species. But few attempts have been made to systematically classify all such relationships at the phenotype level. Also, it is unknown whether such a phenotype map would carry biologically meaningful information.

Results
We have used text-mining to classify over 5.000 human phenotypes contained in the Online Mendelian Inheritance in Man (OMIM) database. We find that there is a continuum of phenotypes. Also, phenotype clusters reflect biological modules of interacting functionally related genes. Phenotype similarity is positively correlated with a number of measures of gene function, including relatedness at the level of protein sequence, protein motifs, functional annotation, and direct protein-protein interaction.

Conclusions
Phenotype clustering reflects the modular nature of human disease genetics. Thus, the phenomap may be used to predict candidate genes for diseases as well as functional relations between genes and proteins. Such predictions will further improve if a unified system of phenotype descriptors is developed.
The phenotype similarity data is accessible through a web interface at http://www.cmbi.ru.nl/MimMiner/.

Background

Functional annotation of genes an important challenge once the sequence of a genome has been completed. Gene annotation encompasses a variety of functional attributes, from structural motifs, through cellular function, to associations with specific functions and processes at the level of the organism.

Apart from descriptors at the gene and protein level, the phenotype effect of a mutated or deleted gene forms part of its functional annotation. Systematic mutation and RNA interference (RNAi) screens have been performed for selected phenotypes in *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*[1-4]. Also for *Mus musculus*, an ambitious project to mutagenize most or all genes has been conceived[5].
Previous studies have correlated various attributes of genes, such as predicted function or amino acid sequence length with the chance of encoding a human disease gene[6-8]. But only limited attention has been awarded to the grouping of phenotypes into a matrix as a means of predicting biological relations between genes and proteins. Phenotype analysis may suggest that genes act together if they cause a similar range of phenotypes when mutated. Systematic grouping of genes by their associated phenotypes may be referred to as phenomics.

Studies of specific phenotype groups in humans suggest that phenomics is possible in humans based on the large numbers of naturally occurring mutations and our detailed knowledge of the phenotypes that are associated[9, 10]. Qualitatively, the human mutation dataset surpasses that of most model organisms, because we can detect and describe human anomalies in more detail than in other species[11]. Specific examples illustrate that individual genes that cause a given phenotype tend to be linked at the biological levels as interacting proteins, as components of a multi-protein complex, or as steps in a biochemical pathway (figure 1).

We have classified over 5.000 phenotypes in humans on the basis of their phenotype similarities, into a single human phenome system. We have further devised and used a system for assigning similarity scores, which allows all genes with known phenotypes to be compared. This approach is very different from that which uses an artificial division into predetermined entities[12]. Given that the human phenome reflects the biology of the system, any phenotype classification should at least to some extent reflect other measures of gene function. We have therefore compared the organization of genes based on this human phenome map to their know interactions, and similarities at multiple levels including sequence, protein motifs, and assigned Gene Ontology functions. The analysis shows that for similar human disease phenotypes there is a consistent association at multiple levels of gene annotation. We propose that given more precise and standardized measures of the phenotype in humans the human phenome map could be a powerful tool for the prediction of the function of human genes.
The phenotype similarity data is accessible through a web interface. This interface, called the “MimMiner”, enables the user to retrieve the similarity ranking for a specific OMIM phenotype. Additionally, a sub tree for this disorder can be generated based on the UPGMA

algorithm. The MimMiner is available at http://www.cmbi.ru.nl/MimMiner/.

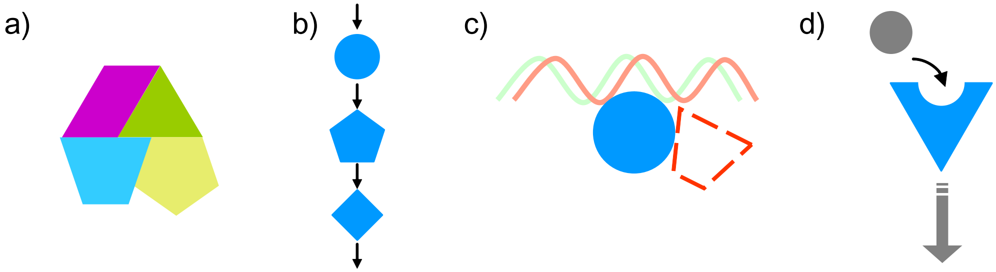


Figure 1 Types of functional relations. Phenotypes that are similar often reflect a close relationship at the genotype level. Different types of functional relations can occur. E.g. a) Subunits form one functional complex; b) Proteins acting in a pathway; c) DNA binding and regulation; d) Signal transduction via a receptor.

Results

Feature vectors

5.132 of the 16.357 OMIM records describe a phenotype, and their TX and CS fields were analysed for the presence of concepts from the anatomy (A) and disease (C) sections in the MeSH thesaurus. For 5.080 OMIM records we could match one or more MeSH terms. 3.778 of the possible 5.436 MeSH terms were found in the OMIM records. The observed concepts are stored in feature vectors; one feature vector per OMIM record. The number of concepts per record varies from 1 to 242 and the average number of concepts per vector is 16.4. The use of hypernyms (eq. 1) increases the average number of concepts per vector to 45.0 (min: 1; max: 477). This broadens the phenotype description and, more importantly, the number of common concepts between pairs of vectors increases from 0.85 to 5.88, allowing for a larger number of meaningful comparisons.

Normalization of the feature vectors by the inverse document frequency (eq. 2) and the correction for the record length (eq. 3) does not influence the number of concepts per vector, but this weighting influences the distances between feature vectors as determined with equation 4. The normalization step of equation 3 scales all concept frequencies to values between 0 and 1. This reduction is non-linear. For example, a specific concept like “Hair Follicle” becomes 1.7 times more important relative to the less specific concept “Skin”.

Comparing OMIM records

The 5.080x5.080 pair-wise feature vector similarities form the phenomap. The distribution of these similarity scores is shown in figure 3a. We used different methods to test whether the phenotypes form distinct groups. These methods map high-dimensional data to low dimensional space and are generally referred to as multi-dimensional scaling (MDS) techniques. General drawbacks of the MDS techniques are their computational requirements.

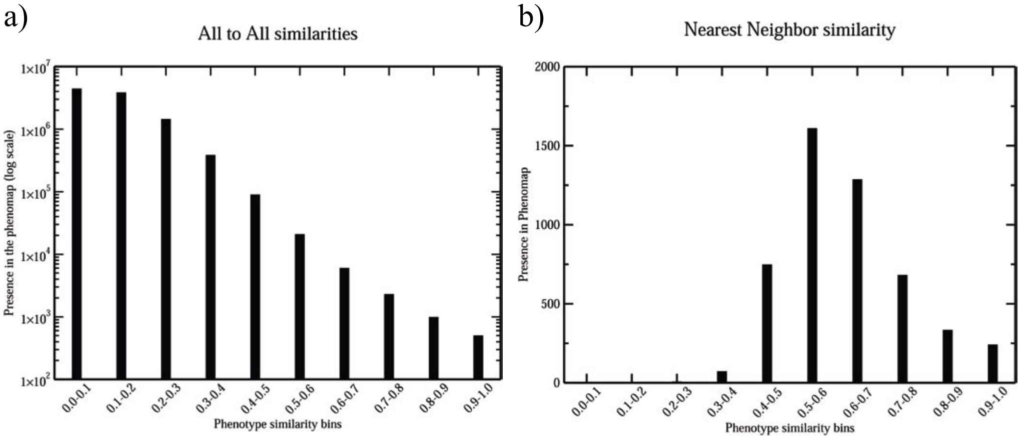


Figure 3 Histograms of phenotype similarity scores. a) Histogram of all pair-wise phenotype similarity scores of the 5.082 phenotype records. The vertical axis is logarithmic; most phenotype-phenotype pairs have a low similarity score. b) The best scores for all phenotypes in the disease phenotype dataset (nearest neighbour similarity).

We therefore used a sample that contains only the OMIM records with a CS field. The phenotype vectors were mapped in a two-dimensional space using Principal Component Analysis, Classic Multidimensional Scaling, and Independent Component Analysis[13] implementations from the "R" software package[14]. None of the methods showed evident clusters, suggesting that the human phenotypes are a continuum rather than distinct classes (data not shown).

We next examined in more detail those phenotype pairs that were characterized by a low average phenotype similarity. We found that phenotypes with a low average similarity score, hence a lower "connectivity" to the rest of the phenomap, corresponded to cases such as "Tobacco Addiction" (OMIM:188890) and "Atrial Tachyarrhythmia with Short PR Interval" (OMIM:108950). These OMIM records are either very small or contain MeSH terms which are infrequently found in other records. In contrast, phenotypes such as "Zellweger Syndrome" (OMIM:214100) or "Isolated Familial Hypoparathyroidism" (OMIM:146200) show significant similarity to large numbers of other OMIM records. Among the 500 phenotypes that had the lowest average similarity scores, we found that some MeSH sub-categories show over- or under representation compared to the average MeSH use in all phenotypes: "Body Regions" (A01, -6.5%), "Neoplasms" (C04, +6.3%), "Hemic and Lymphatic Diseases" (C15, +4.2%), "Neonatal Diseases and Abnormalities" (C16, -4.9%), and "Immunologic Diseases" (C20, +4.7%). Overall, more specific phenotypes were weakly connected, and more broadly defined phenotypes were strongly connected to the rest of the phenomap.

For each OMIM record the most similar of the other 5.079 records was identified. Figure 3b shows the distribution of these 5.080 highest similarity values. We asked whether moderately similar phenotype pairs might still yield reasonable hypotheses. This appeared to be the case in a number of instances. Even in the 0.3-0.4 bin individual cases displayed potentially relevant phenotypic similarities. For example, “Fibromuscular Dysplasia of Arteries” (OMIM:135580) and “Cardiomyopathy, Familial Hypertrophic” (OMIM 192600) have a phenotypic similarity score of 0.31, the 9th most similar score. The observation that these conditions are clinically interrelated is supported by two case reports[15, 16].

Next, we asked if clinically different conditions caused by mutation of the same gene were likely to be detected as having significant phenotype overlap by our system. Indeed, approximately 40% of phenotypes caused by allelic mutations were more similar to each other than to any other phenotype in the dataset (data not shown).

The conclusion is that we cannot define a general cut-off for similarity scores because even low-scoring OMIM records can occasionally contribute to our understanding of the diseases.

Phenotype - sequence similarity correlations

We asked whether the similarity at the phenotype level predicts similarity in gene/protein function. The causative gene/protein is known for 1.653 of the 5.080 OMIM records that describe a phenotype. These 1.653 phenotypes linked to 2.168 corresponding protein sequences (1.401 unique sequences). Sequences were extracted from the UniProt database and used to calculate all-against-all Smith-Waterman alignments. The corresponding 1.653x1.653 phenomap was extracted from the total phenomap. Figure 4a shows the fraction of significant sequence alignment similarities as a function of the phenotype similarity scores. The percentage of phenotype-pairs for which the causative proteins are similar increases with increasing phenotype similarity score from 0.6% to a maximum of 26.6%. Approximately half of these are due to different mutations in the same gene causing similar phenotypes. For example, “Robinow-Sorauf syndrome” (OMIM:180750) and the related “Saethre-Chotzen syndrome” (OMIM:101400) are both caused by a mutation in the TWIST1 protein (UniProt: Q15672). The other relations are due to mutations in different genes that share sequence similarity. For example, the “Rufous Oculocutaneous Albinism” (OMIM:278400) and “Albinism, Oculocutaneous, Type IB” (OMIM:606952) phenotypes show 0.68 phenotypic similarity. These diseases are caused by mutations in the TYRP1 (UniProt:P17643) and TYR (UniProt:P14679) proteins, respectively, that are 43% identical at the sequence level. Mutations in TYRP1 also cause “Oculocutaneous Albinism Type 3” (OMIM 203290).

Many proteins have multiple functional domains. These domains are not unique but appear in different combinations in various proteins. The PFAM system identifies domains through multiple sequence analysis and annotates them with their biological function. If two proteins share a functional domain, then mutations that influence the function of that domain may cause the disruption of the same or a similar process and thereby lead to similar phenotypes. This can be true, even if the full protein sequences do not share significant sequence identity. Figure 4b shows the percentage sequence pairs that share a PFAM domain as function of the phenotype similarity scores. The majority of pairs that share a PFAM domain (67% or 119/181, disallowing the same gene) in the 0.5-0.6 bin) also share significant overall sequence

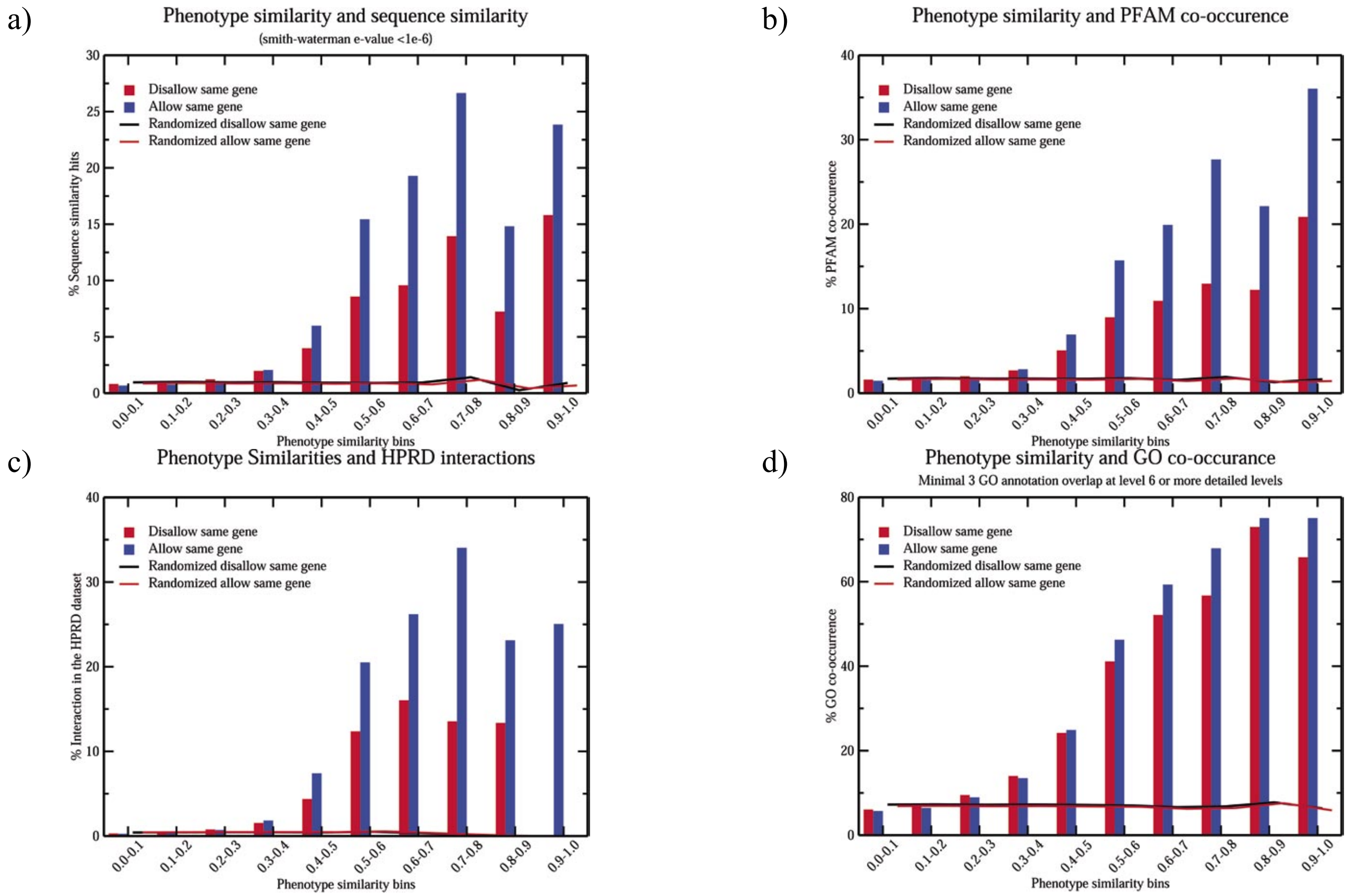


Figure 4 Phenotype similarity versus genotype relation categories. a) Sequence similarity (Smith-Waterman, threshold e-value 1e-6); b) Sharing of at least one PFAM domain; c) Protein-Protein interactions according to the HPRD database; d) Sharing of 3 or more GO annotations at the 6th or more detailed level. The 0.8-0.9 and 0.9-1.0 phenotype similarity bins suffer from low counts. The average signal of ten randomized phenomaps is at the level of the two lowest bins (data not shown). (Colour version: see appendix 2)

similarity. Only in a minority did sequence comparison based on domain sharing add new information compared to overall sequence similarity. For some phenotype pairs the proteins lack significant overall sequence identity, but do share one or more common structural features. For instance, “Long Qt Syndrome 3” (OMIM:603830) caused by a mutation in the Sodium Channel Protein Type V Alpha Subunit (SCN5A) (UniProt:Q14524) shares phenotypic characteristics with “Jervell And Lange-Nielsen Syndrome” (OMIM:220400) that can be caused by a mutation in the Potassium Voltage-Gated Channel Subfamily KQT Member 1 (KCNQ1) (UniProt:P51787). These 2 proteins have an “Ion Transporter” domain in common (PFAM:PF00520).

Phenotype - protein interaction correlations

The Human Protein Reference Database (HPRD) provides information relevant to the function of human proteins, including protein-protein interactions. We checked which of the 1653x1653 pairs have an interaction described in the HPRD (figure 4c). Although the HPRD dataset is sparser than the other datasets, and thus reveals fewer relations, 54% of these HPRD relations were not yet detected by the sequence alignments or the PFAM analyses. We performed a detailed inspection on the 212 protein interactions that are listed in the HPRD for pairs that have a phenotypic similarity between 0.3 and 0.4. Over 50% of the relations suggested a plausible reason for the phenotypic similarities. For example the clinical overlap between “Multiple Epiphyseal Dysplasia type 2” (OMIM:600204) and “Congenital Spondyloepiphyseal Dysplasia” (OMIM:183900) is not unexpected since they are caused by mutations in COL9A2 (UniProt:Q14055) and COL2A1 (UniProt:P02458) respectively. Another example in the 0.4-0.5 bin is the comparison of “Wiskott-Aldrich Syndrome” (OMIM:301000) with “Fleisher Syndrome” (OMIM:307200). These syndromes are caused by mutations in WAS (UniProt: P42768) and BTK (UniProt:Q06187), respectively. WAS and BTK do not share significant sequence similarity, nor do they share a PFAM domain. However, both are involved in cell growth regulation and cytoskeleton processes such as filopodium formation, podosome assembly, chemotaxis, receptor capping and phagocytosis in haematopoietic cells[17]. WAS is phosphorylated at Y291 by BTK leading to activation of the actin nucleating assembly complex Arp2/3 (actin related protein)[18-21]. In general, despite its sparse nature, the HPRD provided biologically plausible information.

Phenotype – functional process correlations

Various excellent databases describing pathway information are available, e.g. KEGG[22], BRENDA[23], Reactome[24], etc. Most of these databases focus on metabolic pathways, whereas less than 10% of the OMIM phenotype records relate to metabolic disorders. To get an impression of possible functional relations between genes/proteins, we compared their GO annotations. We defined GO similarity by the sharing of at least three GO annotations the sixth or more detailed GO level. The signal we find is well above the average of ten randomized matrices with a background percentage of ~7% over all bins (in the GO set, disallowing the same gene). The percentage of pairs that share three or more GO annotations increased as a function of the phenotypic similarity (figure 4d). Using the GO set 76% relations were new compared to the three other sets. A random set of 50 of the 786 proteins that share 3 or more GO annotations and that have a phenotypic similarity between 0.5 and 0.6 were inspected in more detail. A plausible reason for the phenotypic similarities was found in 37 of these 50 cases. Unlike in figure 4a-c less than half of the relations are due to a defect in the same gene, which suggest the criterion of sharing 3 GO annotations at the 6th level is less stringent than in the other data sets.

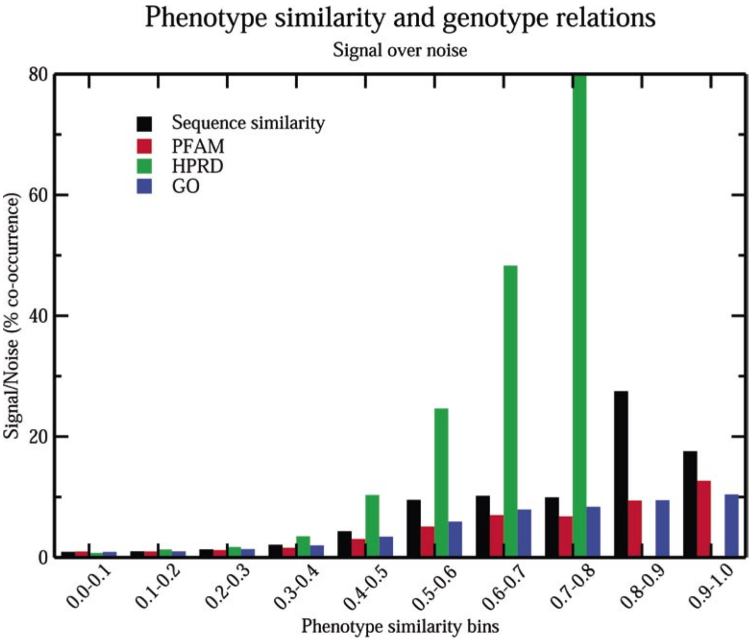


Figure 5 - Histogram of normalized genotype relations as a function of phenotype-phenotype similarity. All values are corrected for random information levels. Although the HPRD data set contains fewer relations than the other sets, the normalized signal is more than two times as strong, reflecting the quality of the HPRD dataset. The 0.8-0.9 and 0.9-1.0 phenotype similarity bins suffer from low counts. (Colour version: see appendix 2)

Nevertheless, sharing of GO annotations can be regarded as a relatively non-specific characterization of gene function. This type of relationship has a high noise level (fig. 4d). Overall, when the genotype by phenotype relations were normalized using random phenomap signals, the HPRD dataset was most efficient in providing non-random gene-gene relationships (fig. 5).

Discussion

We have developed a text-mining approach to map relationships between more than 5.000 human genetic phenotypes from the OMIM database. The resulting phenotype matrix has a number of characteristics that suggest that it might be a useful addition to other functional genomics tools such as the HPRD and KEGG. As expected, we find that different phenotypes associated with mutations of a single gene show considerable overlap. Such allelic conditions are each others best phenotypic hit in 40% of the cases. Conversely, nearest phenotypic neighbours shared at least one functional relationship in 50% of the cases. Also, the phenotype map reflects biologically relevant relationships with other genes. After exclusion of allelic conditions, there remained a significant positive correlation between phenotypic similarity on the one hand, and gene sequence, protein motifs, functional annotation, and known protein interactions on the other (figure 4). This underscores that human phenotypes reflect disturbance of functional modules, more than of individual genes. Further, the lack of obvious clusters in the OMIM phenotype matrix suggests that human disease phenotypes form a continuum. This in itself argues for a genome-wide view of phenotypes since any classification into predetermined classes would lead to a loss of information. One striking finding was that biologically meaningful relationships were mostly detected in the small subfraction of the phenotype relations with a similarity score greater than 0.4 (figure 4). The combined data suggests that we may indeed use phenotypic relationships as general indicators of biological and functional interactions at the gene and protein levels.

Several applications can be envisaged for the phenomap. First, our analysis suggests that the phenotype matrix may aid in the prediction of candidate genes for the 3.400 traits listed in OMIM whose molecular basis remains to be defined. Second, it is conceivable that one would take phenotypic relationships as the starting point for biochemical and cell biological experiments in order to prove a suspected link at the gene and protein levels. Experiments of this type have been shown to be successful (e.g. polycystic kidney disease (PKD1, PKD2), tuberous sclerosis (TSC1, TSC2), breast and ovarian cancer (BRCA1, BRCA2), and Fanconi anemia (FANCA-G))[25-28].

Finally, there may well be a point in pursuing large-scale phenotype analyses using more precise measures of the phenotypes themselves. OMIM was not designed as a structured database for phenotype analysis. Indeed, it does not contain rules for feature assignment and most of the phenotype information collected by our text-mining approach derives from free text fields. A more standardized method for phenotype description including frequency estimates for each feature would greatly increase the yield of the analyses of genotype-phenotype correlations.

Conclusions

Phenotype clustering reflects the modular nature if human disease genetics. Thus, the phenomap may be used to predict candidate genes for diseases as well as functional relations between genes and proteins. Such predictions will further improve if a unified system of phenotype descriptors is developed.

Methods

The OMIM database

The Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim/, November 25 2004) database is a catalogue of human genes and genetic disorders[29]. Each record in the OMIM database contains textual information, about one gene or one genetic disorder, literature references and links to other databases. The data are divided over different fields. We have used the full-text (TX) and the clinical synopsis (CS) field of all records that contain genetic disorders. We will refer to this combination of the TX and CS fields as a "record".

The OMIM database is a rich dataset containing 16.357 full-text records of which 5.132 describe an actual phenotype. The remaining 11.225 records contain variation, mutation, gene/protein, or other information. Approximately 33% of these 5.132 disease phenotypes are characterized at the gene level. OMIM was originally designed as a resource to be used manually. Reading OMIM with a computer program or database is therefore not trivial. We have automatically extracted the phenotypic features from each OMIM record using a series of text analysis techniques.

Creation of "feature vectors"

Retrieval of text-based information can be done on a keyword basis, or e.g. through natural language parsing[30], commonly referred to as "information retrieval" (IR). IR considers text to consist of documents and terms. In this analysis each OMIM record is a document. All words in the OMIM records were considered as terms. We did not use all words, but only those found in the anatomy (A) and the disease (C) sections of the Medical Subject Headings vocabulary (MeSH, http://www.nlm.nih.gov/mesh/meshhome.html). Each MeSH entry is a collection of phenotype-related terms with synonyms and plurals, which together we call a concept. Each concept is uniquely identified by a descriptor. For example, the concept "Neuron" also contains the synonym "Nerve Cell" and the plurals "Neurons" and "Nerve Cells", and is identified by the descriptor D009474. The MeSH concepts, rather than single keywords (like in keyword vectors) as used usually in IR, served as features characterizing OMIM records: every entry in the feature vectors represent a MeSH concept.

Each OMIM record was screened for concepts by matching the words in the records with MeSH terms. The number of times the terms of a given concept are found in an OMIM record reflects the concept's relevance to the phenotype. Some concepts cannot be described by a single term, e.g. "cleft palate". In such cases only the longest, most specific term was counted. So, "cleft palate" is used but not the single words "cleft" or "palate". A term was not allowed to span two sentences separated by a full stop. Non-specific concepts like "syndrome" or "disease" were excluded. This list of descriptor frequencies per OMIM record constitutes the initial feature vector.

Equation 1: For any concept c, its relevance r_c becomes the actual count of the concept in a document $r_{c,counted}$ plus the relevance sum of the concept's hyponyms $r_{hypo's}$. This sum is divided by the number of hyponyms $n_{hypo,c}$. This equation is applied iteratively from the most detailed level in the MeSH tree, till the highest hypernym level is reached. For example, "Sense organ" is a hypernym of the concepts "Eye", "Ear", "Nose", and "Taste buds". Therefore "Sense Organs" receives ¼ of the counts of each of those four hyponyms.

$$r_c := r_{c,counted} + \frac{\sum r_{hypo's}}{n_{hypo,c}}$$

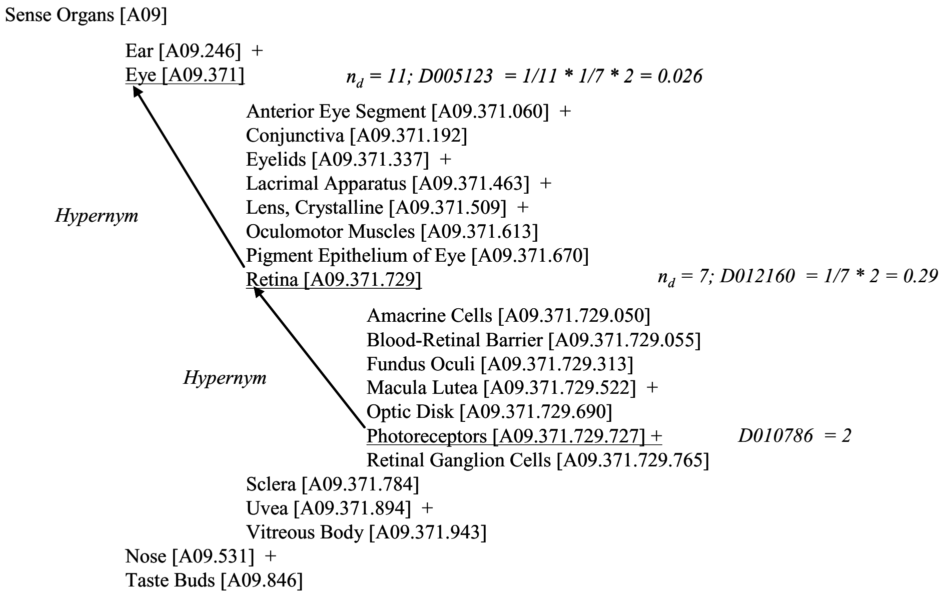

Sense Organs [A09]
Ear [A09.246] +
Eye [A09.371]
n_d = 11; D005123 = 1/11 * 1/7 * 2 = 0.026
Anterior Eye Segment [A09.371.060] +
Conjunctiva [A09.371.192]
Eyelids [A09.371.337] +
Lacrimal Apparatus [A09.371.463] +
Lens, Crystalline [A09.371.509] +
Hypernym
Oculomotor Muscles [A09.371.613]
Pigment Epithelium of Eye [A09.371.670]
Retina [A09.371.729]
n_d = 7; D012160 = 1/7 * 2 = 0.29
Amacrine Cells [A09.371.729.050]
Blood-Retinal Barrier [A09.371.729.055]
Fundus Oculi [A09.371.729.313]
Hypernym
Macula Lutea [A09.371.729.522] +
Optic Disk [A09.371.729.690]
Photoreceptors [A09.371.729.727] +
D010786 = 2
Retinal Ganglion Cells [A09.371.729.765]
Sclera [A09.371.784]
Uvea [A09.371.894] +
Vitreous Body [A09.371.943]
Nose [A09.531] +
Taste Buds [A09.846]

Figure 2 Example of concept expansion using the MeSH hierarchical structure. The concept "Photoreceptors" (with MeSH descriptor: D010786) is found twice in an OMIM record. Expansion of this concept gives the hypernym "Retina" (D012160). The relevance of the concept "Retina" is derived from "Photoreceptors" according to equation 1. "Retina" has 7 descendants or hyponyms, thus its relevance becomes 1/7 * 2 (for 2 times "Photoreceptors"). Similarly, Retina's hypernym "Eye" (D005123) has 11 hyponyms, so that "Photoreceptors" contributes 1/11 * 1/7 * 2 (= 0.026) to the relevance of "Eye". (Squared brackets indicate the MeSH tree position and a "+" sign that there are more specific concepts underneath).

Refinement of the feature vectors

MeSH concepts can be very broad like “Eye” or more specific like “Retina”. MeSH includes a concept hierarchy that describes relationships such as “Eye”-“Retina”-“Photoreceptors”. “Eye” is called a hypernym of “Retina”, which in turn is a hypernym of “Photoreceptors”, etc. Conversely, “Retina” is called a hyponym of “Eye”, etc. To ensure that, for example, the concepts “Eye” and “Retina” are recognized as similar, we use the MeSH hierarchy to encode this similarity in the feature vectors by increasing the value of all hypernyms as described in equation 1 (figure 2). Obviously, the value of hypernyms can be increased due to the presence of hyponyms, but not the other way around. Note that we use ‘hypernym’ and ‘hyponym’ not only to indicate true linguistically hypernyms/hyponyms, but also to indicate “part of” relationships.

Not all concepts in the OMIM records are equally informative. For example, “Retina Pigment Epithelium” occurs rarely, and thus provides more specific information than very frequently occurring terms like “Brain”. In practice this means that the overall importance of a concept depends on its total frequency of occurrence summed over all OMIM records; the more often it occurs, the less important it is. These differences in importance of concept frequencies in feature vectors were incorporated using the inverse document frequency measure of Wilbur and Yang[31] as explained in equation 2.

Equation 2: The inverse document frequency or global weight of concept c (gw_c) is the logarithm of the total number of records analysed (N; N=5.080) divided by the number of records that contain concept c, n_c.

$$r_c = r_c \cdot gw_c = r_c \cdot log_2 \frac{N}{n_c}$$

Not all OMIM records contain equally extensive descriptions. These differences will make a comparison between records difficult, because in the large records the diversity and the frequency of concepts will be higher than in the small records. The total number of concepts per feature vector theoretically can vary between 1 and 5436 MeSH entries. Equation 3[31] was used to (partly) correct for these record size differences.

The three feature vector corrections give a different result when applied in a different order. We first expanded all concepts via the MeSH hierarchical structure (eq. 1). Subsequently, all feature vectors were corrected using the inverse document frequency measure (eq. 2) followed by the local weight correction (eq. 3).

Equation 3: The local weight of concept c in a record is a function of the concept’s frequency r_c divided by the frequency of the most frequent concept in that record, r_{mf}.

$$r_c = 0.5 + 0.5 \cdot \frac{r_c}{r_{mf}}$$

Comparing OMIM records

The similarity between OMIM records can be quantified by comparing the feature vectors that are expanded and corrected by the two correction measures. Similarities between feature vectors were determined by the cosines of their angles (eq. 4)[30].

Equation 4: The similarity between the feature vectors X and Y (s(X,Y)) is a function of their respective concept frequencies x_i and y_i. This equation is also known as the cosine rule and $\sum_{i=1}^{l} x_i y_i$ is commonly known as the inner product of the vectors X and Y. The index i runs from 1 to the number of MeSH concepts l.

$$s(x, y) = \frac{\sum_{i=1}^{l} x_i y_i}{\sqrt{\sum_{i=1}^{l} x_i^2} \sqrt{\sum_{i=1}^{l} y_i^2}}$$

Phenotype-Genotype correlations

The matrix of all pair-wise vector similarities was denoted the phenomap. A subset of this phenomap contains all OMIM records for which the causative gene and protein are known, was used as a starting point for determining the relation between phenotypic similarities on the one hand and genotypic similarities on the other. All 1.653 phenotypes associated with a protein in the UniProt database[32] (http://www.uniprot.org) were then compared to four genotype-related datasets. The average of ten randomized phenomaps was used as a control for background signal.

The PFAM database[33] (http://www.sanger.ac.uk/Software/Pfam/) is a collection of multiple sequence alignments and hidden Markov models, typically used to study the domain organization of proteins. We used PFAM to determine whether pairs of genes share similar domains.

We compared the proteins associated with the 1.653 phenotypes from the UniProt database in an all-against-all Smith-Waterman analysis[34]. We used a Paracel computer (Blosum-90, version 5.03-88, Paracel Inc., Pasadena CA, USA) to check if genes are similar at the sequence level. Sequence pairs with an alignment e-value better than 10^{-6} were considered similar[35].

Protein-protein interactions were extracted from the interaction section of the HPRD database[36] (http://www.hprd.org/) and used to check whether the proteins are part of the same complex or interact in any other way.

The Gene Ontology (GO) database[37] (http://www.geneontology.org/) and the GO annotations (GOA)[38] (http://www.ebi.ac.uk/GOA/) were used to determine if two genes are part of the same functional category. The GO database provides three categories of terms to describe gene products. The "molecular function" category describes the tasks performed by gene products (e.g. ATPase activity); the "biological process" category describes biological mechanisms (e.g. mitosis); and the "cellular component" category describes sub-cellular structures, locations, and macromolecular complexes (e.g. nucleus, haemoglobin complex). The GO terms are hierarchically organized.

Two genes/proteins were considered related when they shared at least three GO terms at the sixth annotation level. Annotations at the more detailed levels (level 7, 8, etc) were converted to the corresponding annotation at the 6th level. For example, the RDS protein (UniProt: P23942) has the detailed annotation "visual perception" at level 7 (GO:0007601), which gets converted to the more general "sensory perception of light" (GO:0050953) at level 6.

Clustering

The conclusions drawn from clustering studies tend to depend on the level of detail of the clusters and thus on the number of clusters generated. A cluster study of the phenomap reveals that the phenotypes seem more continuously related, which makes every clustering attempt intrinsically subjective. In order to objectively visualize the relations between the phenotypes, we performed a hierarchical clustering that results in a tree-structure. Clustering was performed with the Unweighted Pair Group Method with Arithmetic Mean[39]. UPGMA is a simple and fast method that allows the user to cut the tree at any desired value after which the corresponding clustering is automatically generated.

Acknowledgements

The authors thank M. Huynen, B. Snel, H. van Bokhoven and R. van der Heijden for fruitful discussions. M. Oti, M. Hekkelman, T. Hulsen and J. van Rooij are acknowledged for providing expert technical assistance. This work was supported by the Netherlands Organization for Scientific Research (NWO), SON/Unilever (326756), and the EU FP6 Programme (LHSG-CT-2003-503265).

References

1. Boutros, M., et al., *Genome-wide RNAi analysis of growth and viability in Drosophila cells.* Science, 2004. **303**(5659): p. 832-5.
2. Giaever, G., et al., *Functional profiling of the Saccharomyces cerevisiae genome.* Nature, 2002. **418**(6896): p. 387-91.
3. Kamath, R.S., et al., *Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.* Nature, 2003. **421**(6920): p. 231-7.
4. Rual, J.F., et al., *Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library.* Genome Res, 2004. **14**(10B): p. 2162-8.
5. Auwerx, J., et al., *The European dimension for the mouse genome mutagenesis program.* Nat Genet, 2004. **36**(9): p. 925-7.
6. Jimenez-Sanchez, G., B. Childs, and D. Valle, *Human disease genes.* Nature, 2001. **409**(6822): p. 853-5.
7. Lopez-Bigas, N. and C.A. Ouzounis, *Genome-wide identification of genes likely to be involved in human genetic disease.* Nucleic Acids Res, 2004. **32**(10): p. 3108-14.
8. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
9. Brunner, H.G. and M.A. van Driel, *From syndrome families to functional genomics.* Nat Rev Genet, 2004. **5**(7): p. 545-51.
10. Freimer, N. and C. Sabatti, *The human phenome project.* Nat Genet, 2003. **34**(1): p. 15-21.
11. Donnai, D. and A.P. Read, *How clinicians add to knowledge of development.* Lancet, 2003. **362**(9382): p. 477-84.
12. Katsanis, N., et al., *A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14326-31.
13. Hyvarinen, A. and E. Oja, *Independent component analysis: algorithms and applications.* Neural Netw, 2000. **13**(4-5): p. 411-30.
14. R-Development-Core-Team, *R: A language and environment for statistical computing*. 2004: Vienna, Austria.
15. Safioleas, M., J. Kakisis, and C. Manti, *Coexistence of hypertrophic cardiomyopathy and fibromuscular dysplasia of the superior mesenteric artery.* N Engl J Med, 2001. **344**(17): p. 1333-4.
16. Scully, R.E., et al., *Case Records of the Massachusetts General Hospital: Case 9-1995.* N Engl J Med, 1995. **332**: p. 804-10.
17. Welch, M.D. and R.D. Mullins, *Cellular control of actin nucleation.* Annu Rev Cell Dev Biol, 2002. **18**: p. 247-88.
18. Baba, Y., et al., *Involvement of wiskott-aldrich syndrome protein in B-cell cytoplasmic tyrosine kinase pathway.* Blood, 1999. **93**(6): p. 2003-12.
19. Cory, G.O., et al., *Phosphorylation of tyrosine 291 enhances the ability of WASp to stimulate actin polymerization and filopodium formation. Wiskott-Aldrich Syndrome protein.* J Biol Chem, 2002. **277**(47): p. 45115-21.
20. Guinamard, R., et al., *Tyrosine phosphorylation of the Wiskott-Aldrich syndrome protein by Lyn and Btk is regulated by CDC42.* FEBS Lett, 1998. **434**(3): p. 431-6.
21. Torres, E. and M.K. Rosen, *Contingent phosphorylation/dephosphorylation provides a mechanism of molecular memory in WASP.* Mol Cell, 2003. **11**(5): p. 1215-27.

22. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-80.
23. Schomburg, I., et al., *BRENDA, the enzyme database: updates and major new developments.* Nucleic Acids Res, 2004. **32 Database issue**: p. D431-3.
24. Joshi-Tope, G., et al., *The Genome Knowledgebase: a resource for biologists and bioinformaticists.* Cold Spring Harb Symp Quant Biol, 2003. **68**: p. 237-43.
25. Chen, J., et al., *Stable interaction between the products of the BRCA1 and BRCA2 tumor suppressor genes in mitotic and meiotic cells.* Mol Cell, 1998. **2**(3): p. 317-28.
26. Qian, F., et al., *PKD1 interacts with PKD2 through a probable coiled-coil domain.* Nat Genet, 1997. **16**(2): p. 179-83.
27. van Slegtenhorst, M., et al., *Interaction between hamartin and tuberin, the TSC1 and TSC2 gene products.* Hum Mol Genet, 1998. **7**(6): p. 1053-7.
28. D'Andrea, A.D. and M. Grompe, *The Fanconi anaemia/BRCA pathway.* Nat Rev Cancer, 2003. **3**(1): p. 23-34.
29. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.
30. Hand, D.J., H. Mannila, and P. Smyth, *Principles of data mining.* Adaptive computation and machine learning. 2001, Cambridge, Mass.: MIT Press. xxxii, 546.
31. Wilbur, W.J. and Y. Yang, *An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts.* Comput Biol Med, 1996. **26**(3): p. 209-22.
32. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32 Database issue**: p. D115-9.
33. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32 Database issue**: p. D138-41.
34. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
35. Brenner, S.E., C. Chothia, and T.J. Hubbard, *Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 6073-8.
36. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.
37. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
38. Camon, E., et al., *The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro.* Genome Res, 2003. **13**(4): p. 662-72.
39. Sokal, R.R. and C.D. Mitchener, *A statistical method for evaluation of systematic relationships.* University of Kansas Scientific Bulletin, 1958. **28**: p. 1409-1438

Chapter 6

General discussion and conclusions

The aim of this research was to reveal some general principles underlying human genetic disease and to design bioinformatics strategies for the identification of disease genes. We combined information from gene specific databases and phenotype descriptions to achieve our goal. Two routes were followed. First, we selected genes within disease critical region. We selected genes that are expressed in the affected tissue(s) or where mutations in homologous genes in model organisms cause a similar phenotype. The technical implementation does not require extensive data warehousing, but uses a distributed model (chapter 2 and 3). Second, we explored the potential use of natural occurring disease phenotypes for function annotation. In chapter 4 we explore how phenotypes relate to one another and whether a comparison between phenotypic features may tell us what genes are involved. Also to what extent do phenotypic similarities reflect known and new molecular mechanisms. We devised a method for comparing phenotypes derived from the OMIM database that uses a textual similarity measure by an automated full text-analysis technique and analysed the phenotype-genotype relations (chapter 5).

Candidate gene prioritisation in syndromes by combining a pattern of organ involvement and positional information and utilization of mouse phenotypic data has been shown to be effective[1, 2]. This approach has been successfully implemented *in silico*. While the results of GeneSeeker on average lead to a tenfold reduction of the positional candidate genes, there are several limitations. These are the incompleteness of the databases, a lack of standardization for gene expression data and for phenotype and tissue descriptions. Also, the success of the method depends on a relatively strong relation between genotype and phenotype.

The phenotype is a representation of both genotype and environment. Phenotype relationships are a powerful method for function prediction[3-5]. We can therefore use the human phenotype collection and the underlying gene-phenotype relations as a tool for functional genomics. Several improvements can made to make full use of the phenotype potential as a functional genomics tool. These include clear and standardized definitions and (weighted) feature descriptions of human genetic disorders, other phenotype variations, and phenotype characteristics of gene knock-outs in model species.

So, how do our approaches to candidate gene identification and phenotype classification relate to other efforts within the field of human genetics and bioinformatics involving candidate gene prioritisation and phenotype classification? Identification of candidate gene is most likely to be successful when positional and functional routes are integrated. Different efforts have been made to incorporate both. Integration of data based on genomic context like in the UCSC genome browser and Ensembl[6, 7] resulted in step-by-step interfaces (e.g. EnsMart[8]) to extract data based on e.g. chromosomal position, expression[9], and Gene Ontology[10]. The enrichment for disease candidate genes using these database interfaces depends on the skills of the operating researcher. Only recently methods have been developed to systematically explore datasets for candidate disease genes. Four different approaches can be distinguished.

First, annotation data have been used to group genes with the same functional characteristics. This approach, developed by Perez-Iratxeta *et al.*[11], links Medline abstracts via MeSH and GO to RefSeq proteins[10, 12-14], and there by allows prioritisation of genes in disease loci. They tested 450 diseases were previously mapped to a specific locus but without a particular gene assigned. The resulting scores were compared to 100 diseases where the gene was known[11]. On average Perez-Iratxeta *et al.* tested 30Mb candidate regions. Assuming 20.000-25.000 human genes[15, 16], and an average gene density of 1 gene per 120kb, an 8-31 fold enrichment was calculated for this method.

Turner *et al.* prioritised candidate disease genes based on over-representation of functional annotation (GO) between loci for the same disease[17]. They tested 29 diseases and achieved an enrichment between 12 and 42-fold.

As a second approach, gene specific characteristics have been used in candidate disease gene identification. Sequence analysis of human/eukaryotic genes showed that human proteins with multiple long amino acid runs are more often linked with genetic disease than are shorter proteins[18]. Also, proteins involved in genetic diseases tend to be long, conserved, and without close paralogues[19]. Disease genes are more frequently found to be conserved in other species, but this can be due the preferential sequencing of known (disease) genes. Adie *et al.* tested sequence property analysis using alternating decision trees[20]. They found differences between random genes and disease genes based on a number of features, including: gene/cDNA/protein/3' UTR length, number of exons, distance to adjacent gene, higher level of conservation in mouse, signal peptide encoding, and 5' CpG islands. Their tests for candidate gene identification showed 2-25 fold enrichment. Smith *et al.* found similar differences between disease and non-disease genes. Using discriminant analysis they showed that these differences may help to predict human disease genes[21].

An approach to disease gene identification makes use of the multitude of gene and protein expression data that is produced by methods like RNA expression micro array analysis and SAGE. For example, Tiffen *et al.* developed a method which uses an anatomical ontology (eVOC)[9] to integrate biomedical literature and human gene expression data[22]. Using a controlled vocabulary, their method can be used without prior clinical knowledge. The enrichment reached is 1.5-3-fold and the correct gene was found in more than 85% of the cases.

A more clinically oriented fourth approach correlates clinical features of genetic disease with the functional classification of their underlying disease genes[23]. Freudenberg and Propping developed a method to cluster genetic diseases based on their phenotype similarity[24]. They manually attributed the disease phenotypic manifestations. In total, 878 diseases were tested for 10.672 candidate disease genes, achieving enrichment between 7 and 33-fold.

Similarly, Cantor *et al.* clustered OMIM[25] records based on the clinical synopsis section[26]. They reduced the disease characteristics to fifty categories. In a test of two diseases they found relations at the genotype level. Since the authors only intended a proof of principle on using OMIM for phenotype clustering they did not systematically analyse phenotype-

genotype relations.
Evaluation of the different methods does not allow direct comparison between them. Except for the method of Tiffen *et al.* [22] they all perform similarly, giving 7-10 fold enrichment in most cases. GeneSeeker (chapter 2 and 3) can be positioned in the third group. The GeneSeeker uses human as well as mouse expression/phenotypic data that is stored in various databases. This information is combined with positional data of both species. The GeneSeeker approach differs from the other candidate prioritisation approaches because of utilizing cross-species data. We achieved a 7-25 fold enrichment of candidate disease genes. However, GeneSeeker requires considerable prior clinical knowledge. In light of the comparable enrichment levels achieved with the different methods, it is likely that they can complement each other.
One question that we attempted to answer in these studies was whether the use of phenotype descriptions for automated candidate disease gene identification/enrichment is possible. Our studies and those of others suggest that this is the case (chapter 2 and 5). We found that phenotype similarity based on automatic quantification, correlates positively with a number of measures of gene function, including protein sequence, similarity shared protein motifs, functional annotation, and direct protein-protein interaction. The data suggests that phenotypic relationships may be used as indicators of biological and functional interactions at gene and protein levels (chapter 5).

To make the most of candidate disease prioritisation and phenotype information, a number of improvements are possible:

1. Develop new methods for sequence analysis
Gene identification is still not complete at the moment and a significant number of human genes remains to be elucidated[16]. Even more non-coding RNA genes (ncRNAs) await identification. Comparative genomics studies have shown highly conserved sequences in mammalian genomes[27]. The functional role of these conserved nongenic sequences (CNG) remains unknown and it is not unlikely that variations in CNGs are associated with phenotype variability and disease. If CNGs are important for human genetic disease, then clearly the current methods for disease gene identification will fail to identify them. Involvement of ncRNAs in human genetic disorders has been shown[28] and new methods to identify these genes have to be developed in order to broaden the view on disease mechanisms.

2. Expand experimental data, unify and integrate data sources
Much of the data that can be used for candidate gene prioritisation is scattered over various databases and suffers from incompleteness and errors. Standardization and protocols are needed to complete new and existing datasets. Initiatives such as MIAME (Minimum Information About a Microarray Experiment)[29] are a start, but still need a broader implementation and acceptance. Furthermore, more, structured, and detailed data on gene/protein expression, protein-protein interactions, regulation and other biological systems will

be essential for computational analysis of bionetworks, including DNA variations and their phenotypical consequences. If we are to achieve a comprehensive view of (human) biology these different (omics) dataset must be integrated. Data storage solutions and schemes are currently under development[30-32].

The different omics can be placed in layers creating an omics-space[31]. As the different types of omics evolve, links will be created between these layers. Ultimately, this results in a model of the organism, which is referred to as systems biology. Currently, omics data is mostly analysed by clustering methods. It is likely that new methodologies will be developed to study and make predictions across omics-space. Methods that may be used for the analysis of omics data include various statistical methods, network analysis, sequence analysis, machine learning, data mining, and visualization systems.

3. Improving statistical methods for linkage and association studies

The current shift from monogenetic to complex multifactorial genetic diseases emphasizes the need for genome-wide association studies for mapping of disease genes. The practical problems of data generation are expected to be solved in the not too distant future[33, 34]. However, study design and data analysis for these genome-wide association studies is not straight-forward and requires care[33]. Furthermore, concepts such as the likelihood of interacting loci, also known as epistasis, are frequently ignored in complex genetic trait studies[35]. New methods for multilocus linkage and association should take these concepts and mechanisms into account.

4. Elaboration and standardization of phenotype descriptions.

Current approaches to define (disease) phenotypes are inadequate for the generation of a reliable phenotype map. To begin with, there are different definitions of what a phenotype is[36]. For human diseases, ICD, SNOMED, and UMLS (http://www.who.int/classifications/icd/en/; http://www.snomed.org/; [37]) are three of the classification systems that are used to describe phenotypes. These systems are for the most part founded on clinical observations rather than on biological phenotype characteristics, which results in a gap between clinical and biological data. A central and standardized phenotype classification system awaits development and acceptance. Such a system should encompass definitions of phenotype features. Quantitative comparison of aberrant versus normal phenotypes is needed for weighting of those features, which calls for studies on ‘normal’ phenotypic (morphological) variation. Additionally, phenotype characteristics of model species/knock-out models can be used for comparative phenomics[38].

Phenotypic characters are not simple and they might change during development, ageing, and other environmental change. However, a detailed anatomical spatio-temporal atlas[39, 40] is feasible at this moment with medical imaging techniques. Integration and development of phenotype ontologies, such as Phenotype Attribute Ontology (PAtO) (http://obo.sourceforge.net/) or Mammalian Phenotype Ontology [41] will simplify phenotypic analysis. Moreover, free public access to both biological and clinical phenotype related databases is essential for further development of this field. Without doubt, our perception and description of phenotypic characters will change with our understanding of the biological basis of phenotypes.

5. Improve and integrate methods for candidate disease gene identification and phenotype classification.

The various methods to identify candidate disease genes in humans cover different concepts (figure 1). They either use functional and literature data, gene specific characteristics, anatomy based gene/protein expression data, or phenotype comparison analyses. As discussed earlier the current performance of the individual methods is comparable. These methods can complement each other and such combinations can improve the predictive performance. It will be essential to establish cross database and cross species ontologies, and to develop heuristics for candidate gene prioritisation that evaluates data and phenotype knowledge.

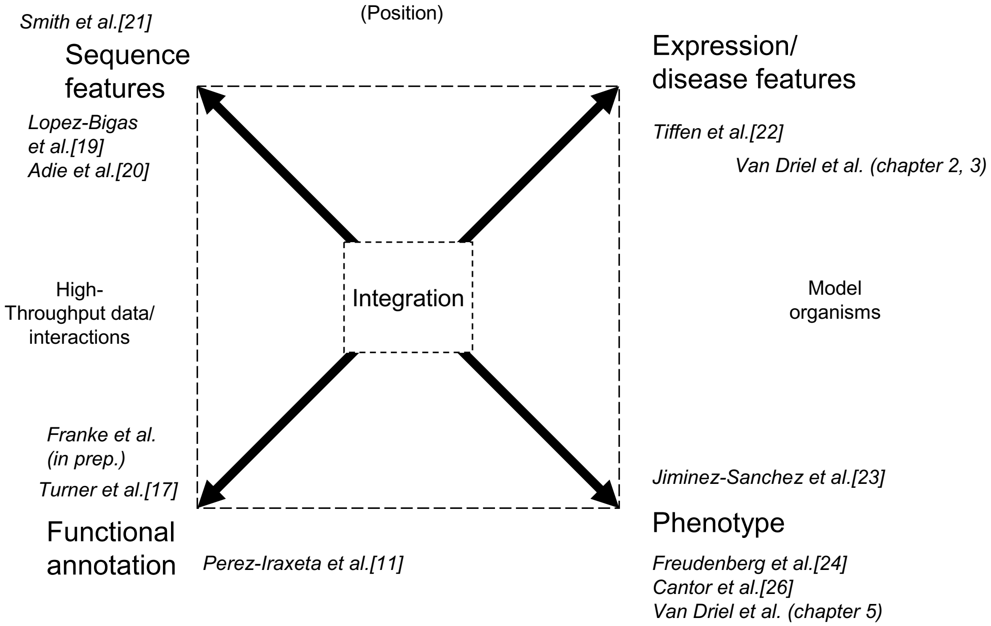


Figure 1 The various methods to identify candidate disease genes in humans are located relative to the different concepts (sequence features, expression/disease features, functional annotation, and phenotype).

In conclusion, we developed an approach that with the current set of databases, can be used for candidate disease gene prioritisation (chapters 2 and 3). *In silico* prioritisation methods are evolving and improvements can be made by integrating these methods (figure 1), by completion of datasets, and by development of standardized ontologies across databases and species. We believe that phenotype relationships are powerful predictors for biological function (chapters 4 and 5). These relationships reflect the modular nature of biology and may be used for candidate disease gene predictions. A unified system is needed and should recognize distinct phenotypic characters and classify them in a spatio-temporal model. New ways to perform and analyse multilocus linkage and association data, in combination with such a phenome system may further improve the performance of computational methods for candidate disease gene identification.

References

1. Celli, J., et al., *Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome.* Cell, 1999. **99**(2): p. 143-53.
2. van Bokhoven, H., et al., *Mutation of the gene encoding the ROR2 tyrosine kinase causes autosomal recessive Robinow syndrome.* Nat Genet, 2000. **25**(4): p. 423-6.
3. Spranger, J., *Pattern recognition in bone dysplasias.* Prog Clin Biol Res, 1985. **200**: p. 315-42.
4. Annunen, S., et al., *Splicing mutations of 54-bp exons in the COL11A1 gene cause Marshall syndrome, but other mutations cause overlapping Marshall/Stickler phenotypes.* Am J Hum Genet, 1999. **65**(4): p. 974-83.
5. van Steensel, M.A., et al., *Oto- spondylo-megaepiphyseal dysplasia (OSMED): clinical description of three patients homozygous for a missense mutation in the COL11A2 gene.* Am J Med Genet, 1997. **70**(3): p. 315-23.
6. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
7. Hubbard, T., et al., *Ensembl 2005.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D447-53.
8. Kasprzyk, A., et al., *EnsMart: a generic system for fast and flexible access to biological data.* Genome Res, 2004. **14**(1): p. 160-9.
9. Kelso, J., et al., *eVOC: a controlled vocabulary for unifying gene expression data.* Genome Res, 2003. **13**(6A): p. 1222-30.
10. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
11. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
12. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D39-45.
13. Lipscomb, C.E., *Medical Subject Headings (MeSH).* Bull Med Libr Assoc, 2000. **88**(3): p. 265-6.
14. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D501-4.
15. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
16. Larsson, T.P., et al., *Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery.* FEBS Lett, 2005. **579**(3): p. 690-8.
17. Turner, F.S., D.R. Clutterbuck, and C.A. Semple, *POCUS: mining genomic sequence annotation to predict disease genes.* Genome Biol, 2003. **4**(11): p. R75.
18. Karlin, S., et al., *Amino acid runs in eukaryotic proteomes and disease associations.* Proc Natl Acad Sci U S A, 2002. **99**(1): p. 333-8.
19. Lopez-Bigas, N. and C.A. Ouzounis, *Genome-wide identification of genes likely to be involved in human genetic disease.* Nucleic Acids Res, 2004. **32**(10): p. 3108-14.
20. Adie, E.A., et al., *Speeding disease gene discovery by sequence based candidate prioritization.* BMC Bioinformatics, 2005. **6**(1): p. 55.
21. Smith, N.G. and A. Eyre-Walker, *Human disease genes: patterns and predictions.* Gene, 2003. **318**: p. 169-75.

22. Tiffin, N., et al., *Integration of text- and data-mining using ontologies successfully selects disease gene candidates.* Nucleic Acids Res, 2005. **33**(5): p. 1544-52.
23. Jimenez-Sanchez, G., B. Childs, and D. Valle, *Human disease genes.* Nature, 2001. **409**(6822): p. 853-5.
24. Freudenberg, J. and P. Propping, *A similarity-based method for genome-wide prediction of disease-relevant human genes.* Bioinformatics, 2002. **18 Suppl 2**: p. S110-5.
25. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.
26. Cantor, M.N. and Y.A. Lussier, *Mining OMIM for Insight into Complex Diseases.* Medinfo, 2004. **2004**: p. 753-7.
27. Dermitzakis, E.T., A. Reymond, and S.E. Antonarakis, *Conserved non-genic sequences - an unexpected feature of mammalian genomes.* Nat Rev Genet, 2005. **6**(2): p. 151-7.
28. Ridanpaa, M., et al., *Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia.* Cell, 2001. **104**(2): p. 195-203.
29. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 2001. **29**(4): p. 365-71.
30. Ge, H., A.J. Walhout, and M. Vidal, *Integrating 'omic' information: a bridge between genomics and systems biology.* Trends Genet, 2003. **19**(10): p. 551-60.
31. Toyoda, T. and A. Wada, *Omic space: coordinate-based integration and analysis of genomic phenomic interactions.* Bioinformatics, 2004. **20**(11): p. 1759-65.
32. Hucka, M., et al., *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics, 2003. **19**(4): p. 524-31.
33. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nat Rev Genet, 2005. **6**(2): p. 95-108.
34. Wang, W.Y., et al., *Genome-wide association studies: theoretical and practical concerns.* Nat Rev Genet, 2005. **6**(2): p. 109-18.
35. Carlborg, O. and C.S. Haley, *Epistasis: too often neglected in complex trait studies?* Nat Rev Genet, 2004. **5**(8): p. 618-25.
36. Mahner, M. and M. Kary, *What exactly are genomes, genotypes and phenotypes? And what about phenomes?* J Theor Biol, 1997. **186**(1): p. 55-63.
37. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.
38. Lussier, Y.A. and J. Li, *Terminological mapping for high throughput comparative biology of phenotypes.* Pac Symp Biocomput, 2004: p. 202-13.
39. Burger, A., D. Davidson, and R. Baldock, *Formalization of mouse embryo anatomy.* Bioinformatics, 2004. **20**(2): p. 259-67.
40. Hunter, A., et al., *An ontology of human developmental anatomy.* J Anat, 2003. **203**(4): p. 347-55.
41. Smith, C.L., C.A. Goldsmith, and J.T. Eppig, *The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.* Genome Biol, 2005. **6**(1): p. R7.

Summary

Disease gene identification based on chromosomal localisation is sometimes difficult and often time-consuming. It requires collecting as much information on the disease as possible. Combining positional information with disease characteristics might give hints by which candidate disease genes can be selected. We combined data from various computer databases and developed bioinformatics strategies to explore the systematic identification of disease genes.

Genetic diseases are caused by changes called mutations in our hereditary material (DNA). In the western world genetic conditions are a major cause of neonatal and childhood diseases, and of infant death. Disease causing mutations play an important role in our understanding of human genetic diseases, and the molecular roles of genes and proteins in development. This knowledge can be utilized in genetic counselling, in genetic diagnosis, and in further research. Knowledge on disease genes may ultimately open possibilities for treatment.

Important indications for the identification of causes and mechanisms of disease come from clinical disease characteristics (the disease phenotype). Definition and classification of diseases are essential, both for clinical as well as for research purposes. Various classification systems have been developed and are used for this purpose. However, none of these systems is universally accepted and consequently a lot of clinical information is still stored in free-text literature databases.

Here, we present two bioinformatics strategies to identify human candidate disease genes (chapter 2, 3) and classify human disease phenotypes (chapter 4, 5). Our hypothesis was that internet databases, which contain gene specific data as well as databases that contain phenotype descriptions, can be utilized systematically to achieve our goals.

We developed a system (chapter 2 and 3; GeneSeeker) aiming to assist in the candidate gene identification process. A commonly applied strategy is to prioritise genes in the chromosomal interval obtained via pedigree studies. Prioritisation is based on gene characteristics such as expression in the affected tissue(s) and mutant phenotypes in model organisms. The GeneSeeker mimics this process in silico. In general, linkage analysis data restricts the disease gene to a chromosomal region of 20-200 genes. The system combines positional and expression/phenotype data from human and mouse. Chapter 2 describes the analysis of ten syndromes. For all syndromes tested including two which were novel at that moment, the candidate disease gene lists were reduced 7-25 fold while preserving the disease genes.

A more technical discussion on the GeneSeeker can be found in chapter 3. The program queries information from various databases directly on the World Wide Web. It retrieves information from nine different databases and the modular setup allows addition of other databases if needed. Direct searches via the original database web interfaces guarantee that the most recent data are queried, and this obviates the need for data warehousing. GeneSeeker makes candidate gene selections, thereby reducing the number of genes to be screened for mutation analysis.

GeneSeeker will only work for those syndromes in which the disease gene displays altered expression patterns in the affected tissue(s). In such cases, it provides a significant reduction in time spent on wet-lab experiments by giving helpful clues which genes to examine first for mutations. The field of candidate gene prediction and prioritisation is new and only a few efforts have been published using different data sources. Compared to these methods GeneSeeker performs similar. GeneSeeker requires prior clinical knowledge. However, it is the only method so far that incorporates localisation, expression, and phenotypic data from the mouse.

Chapter 4 discusses relationships at the phenotype level, and how phenotype information defines syndrome families which can then be explored as a tool for functional genomics. In this chapter we explore, using a number of examples form the literature, how similar syndromes/phenotypes relate to each other, and whether they can indicate what genes are involved. We argue that these similarities tend to reflect the biological relationships in the cell. Additionally, we argue about the applicability of a phenotype map for the identification of disease genes, including multifactorial/complex disorders.

In chapter 5 we describe a method devised to extract disease phenotypes from the full text descriptions in the OMIM database. Text mining techniques were used to identify phenotype concepts. In contrast to manually predefined categories, we used an automated method for systematic phenotypic feature identification and comparison. We have analysed the phenomap for the presence of biologically meaningful information and found that phenotype similarity is positively correlated with gene function at the level of protein sequence, motifs, functional annotation, and protein-protein interactions derived from the literature. The analysis suggests that the automatically derived phenotype map can serve as an indicator of biological relationships.

The aim of this research was to reveal some general principles of human genetic disease and the bioinformatics strategies to explore them for the identification of disease genes. Phenotype analysis is equally relevant to disease gene identification as it is to the functional annotation of the human genome: Over the course of the coming years major improvements are to be expected with regard to gene identification methods for monogenic, and multifactorial diseases, mutation detection, and various methods for cross-species comparisons. Bioinformatics approaches such as those described here, should be a useful addition to sequence and gene-based analyses. While much remains to be discovered from systematic phenotype-genotype analyses, an essential prerequisite will be the development of a standardized and internationally agreed nomenclature for phenotype definition that is applicable to humans as well as to the major model organisms that are in use today.

Glossary

α-DYSTROGLYCAN - A glycoprotein that binds to dystrophin, and helps to provide linkage between the sarcolemma and extracellular matrix in muscle.

ACHONDROPLASIA - The most common of the many types of short-limbed dwarfism. Achondroplasia is characterized by abnormal bone growth that results in short stature with disproportionately short arms and legs, a large head and characteristic facial features.

ACRO-DERMATO-UNGUAL-LACRIMAL-TOOTH (ADULT) SYNDROME - A syndrome with a variable expression that is similar to EEC syndrome. ADULT syndrome is characterized by skeletal, nail, breast, teeth, and other anomalies (see also OMIM: 103285).

ALAGILLE SYNDROME - Alagille syndrome is a multi-system hereditary disorder. Common findings in patients with this syndrome are reduced bile flow, congenital heart disease, bone defects, eye findings, and other (typical facial) features (see also OMIM: 118450).

ALBINISM, OCULOCUTANEOUS, TYPE IB - An autosomal recessive disorder characterized by absence of pigment in hair, skin, and eyes. Patients suffer from various eye problems such as reduced vision and photophobia (see also OMIM: 606952).

ATHEROSCLEROSIS - Age, lifestyle, diet and gene-related degeneration of arteries owing to deposition of lipoid plaques (atheromas) on inner arterial walls; it is the main cause of coronary artery disease and a leading cause of death.

ATRIAL TACHYARRHYTHMIA WITH SHORT PR INTERVAL - A condition characterized by heart rhythm abnormalities and a typical electrocardiographic pattern (see also OMIM: 108950).

BARDET–BIEDL SYNDROME - A genetic disorder that is linked to chromosomes 3, 15 and 16 that causes progressive blindness, obesity, extra fingers and toes, and mental retardation.

CARDIOMYOPATHY, FAMILIAL HYPERTROPHIC - A genetic disease of the heart muscle, which involves a thickening of the heart muscle (see also OMIM: 192600).

CHONDRODYSPLASIA - A disturbance in the development of cartilage, primarily the long bones. This can result in arrested growth and dwarfism.

CLASSIC MULTIDIMENSIONAL SCALING - Statistical technique reducing the number of dimensions of the data set.

CONGENITAL SPONDYLOEPIPHYSEAL DYSPLASIA - A genetic disease that features abnormal growth of the growing ends of bones in the spine, resulting in short trunk dwarfism and other problems (see also OMIM: 183900).

DYSMORPHOLOGY - The systematic examination and classification of abnormal external features.

ECTRODACTYLY-ECTODERMAL DYSPLASIA-CLEFTING SYNDROME (EEC) - A congenital disease characterized by split hand/feet, cleft lip, cleft palate, decreased hair growth, and other abnormalities of nail, teeth, and skin (see also OMIM: 604292).

ELLIPTOCYTOSIS - A hereditary abnormality of red blood-cell shape.

FANCONI ANEMIA - A genetically inherited bone marrow failure syndrome (see also OMIM: 227650).

FEATURE VECTOR - A numerical representation of the frequencies and weights of the different (predefined) features in a document. Features are the items which documents (e.g. phenotype descriptions) share or those in which they differ.

FIBROMUSCULAR DYSPLASIA OF ARTERIES - An arterial disease that produces stroke, hypertension, poor supply of oxygen to the muscles or myocardial infarction (see also OMIM: 135580).

FILOPODIUM - A thin protrusion from a cell, filled with a bundle of actin filaments that function in sensing environmental triggers to guide cell migration or axon extension.

FLEISHER SYNDROME - A growth hormone deficiency with a low level of antibodies, which results in frequent infections and other problems (see also OMIM: 307200).

FRIEDREICH'S ATAXIA - Friedreich's Ataxia is a slowly progressive disorder of the nervous system and muscles (see also OMIM: 229300).

FUKUYAMA TYPE MUSCULAR DYSTROPHY - A rare form of autosomal recessive muscular dystrophy, the symptoms of which begin before the age of 9 months and include mental retardation, loss of muscle tone or tension and weakness of the muscles (see also OMIM: 253800).

FUNCTIONAL GENOMICS - Global analysis of the function of genes in isolation and in concert with one another is the foundation of functional genomics. This includes analysis of genomic expression or transcriptome, and the resulting proteins or proteomics.

GENE ONTOLOGY - (GO). A hierarchical organization of concepts (ontology) with three organizing principles: molecular functions (the tasks done by individual gene products), biological processes (for example, mitosis) and cellular components (examples include the nucleus and the telomere).

GENETIC HETEROGENEITY - Clinically similar phenotypes caused by mutations in different genes or a combination of genes.

HAND-FOOT-UTERUS SYNDROME -

Besides hand, foot, and uterus abnormalities patients with this genetic disease also have genital tract and eye problems (see also OMIM: 140000).

HOLT-ORAM SYNDROME - This inherited disorder is also called heart-hand syndrome, and causes abnormalities of the upper limbs and heart (see also OMIM: 142900).

INDEPENDENT COMPONENT ANALYSIS - Technique that finds statistical independent components in the input data set.

INDEX CASE - The first diagnosed case in a family.

INTERACTOME - A complete set of macromolecular interactions (physical and genetic). Current use of the word tends to refer to a comprehensive set of protein–protein interactions.

INVERSE DOCUMENT FREQUENCY - A weight factor that is defined as the logarithm of the number of documents divided by the number of documents that contain that term.

ISOLATED FAMILIAL HYPOPARATHYROIDISM - An inheritable form of Hypoparathyroidism (reduction or absence of secretions of the parathyroid gland), which is characterized by low calcium levels in the blood and elevated phosphate levels (see also OMIM: 146200).

JERVELL AND LANGE-NIELSEN SYNDROME - Genetic functional heart disease with an abnormal electrocardiographic pattern, which can result in sudden death. Additionally, patients suffer from hearing loss (see also OMIM: 220400).

KEYWORD VECTOR - A numerical representation of the frequencies and weights of the different terms/keywords in a document.

LARSEN SYNDROME - A syndrome of multiple congenital dislocations and characteristic faces, notably with a prominent forehead, depressed nasal bridge and widely spaced eyes (see also OMIM: 150250/245600).

LEUKOENCEPHALOPATHIC BRAIN DISEASE - Leukoencephalopathy is the destruction of the myelin sheaths (white matter), which cover nerve fibers (see also OMIM: 603896).

LONG QT SYNDROME 3 - Genetic functional heart disease with an abnormal electrocardiographic pattern (see also OMIM: 603830).

MARSHALL SYNDROME - Rare genetic disorder that is characterized by a flattened nasal bridge, nostrils that tilt upwards, widely-spaced eyes, myopia, cataracts and hearing loss (see also OMIM: 154780).

METABOLOME - The collection of all metabolic units and pathway motifs in a cell, tissue, organ, etc.

MICRORNA - RNA (single-stranded), which is thought to play a regulatory role in gene expression. MicroRNAs (miRNAs) are typically ~20-25 nucleotides and transcribed normally from DNA, but they are not translated into protein.

MOHR SYNDROME - This syndrome shows several birth defects such as cleft palate, and other facial abnormalities, shortened limbs, and hands and feet malformations (see also OMIM: 252100).

MULTIPLE EPIPHYSEAL DYSPLASIA TYPE 2 - Genetic disease that features abnormal skeletal growth resulting in short stature and various limb problems (see also OMIM: 600204).

MULTIPLE SYNOSTOSES SYNDROME 1 - Patients with this syndrome suffer from abnormal bone fusions that lead to body width malformations in skeleton, head, neck, skin, and nails (see also OMIM: 186500).

MUSCLE-EYE-BRAIN DISEASE - Genetic disease that features weakness and dysfunction of the muscles, atypical neuronal migration, and various eye defects (see also OMIM: 253280).

NOONAN SYNDROME - The disease features include inborn heart defects, short stature, learning problems, indentation of the chest, blood clotting issues, and a distinctive appearance (see also OMIM: 163950).

OCULOCUTANEOUS ALBINISM TYPE 3 - A genetic trait with partial absence of pigmentation in skin, but normal retinal pigmentation (see also OMIM: 203290).

OLIGOGENIC - A trait is considered to be oligogenic if two or more genes work together to produce the phenotype. An oligogenic trait, which implies that few genes are involved, should be contrasted with a polygenic trait, which implies that many genes are involved in phenotype expression.

O-MANNOSYLATION - A form of glycosylation of proteins that begins by adding a mannose at serine and threonine residues.

OMICS - Omes is derived for the Greek for *all/complete*. In biology the suffix -omics is often used to describe biological subfields that involve large-scale data collection/analysis. Examples are genomics (DNA), proteomics (proteins), metabolomics (small molecules).

OTO-PALATO-DIGITAL SYNDROME - An X-linked condition with deafness, cleft palate, characteristic faces and a generalized bone dysplasia (see also OMIM: 311300/304120).

OPITZ G SYNDROME - OPITZ G syndrome is characterized by abnormality of eyes, penis, testes, and anus. Patients suffer from facial clefting, swallowing defects, and often there is a brain malformation (see also OMIM: 300000).

OSMED - OtoSpondyloMegaEpiphyseal Dysplasia is a very rare, more severe variant of the Stickler syndrome (see also OMIM: 215150).

OSTEOPETROSIS - Part of a range of diseases that are characterized by a generalized increase in skeletal density.

PALLISTER–HALL SYNDROME - An extremely rare genetic disorder that can be apparent at birth (congenital), the symptoms of which vary greatly in range and severity from case to case and can include a benign tumour of the hypothalamus, decreased pituitary function and the presence of extra fingers and/or toes (see also OMIM: 146510).

PHENOME - The physical totality of all traits of an organism or of one of its subsystems is called the phenome.

PHENYLKETONURIA - A genetic disorder in which a liver enzyme (phenylalanine hydroxylase) is defective, leads to mental retardation unless a special diet is followed (see also OMIM: 261600).

PIV SYNDROME - PIV (Polydactyly, Imperforate anus, and Vertebral anomalies) syndrome is characterized by the presence of more than the normal number of fingers or toes, partial or complete obstruction of the anus and vertebral anomalies (see also OMIM: 174100).

PODOSOME - Cell membrane structures that are involved in the adhesion process of various cells to a solid substrate.

POLYCYSTIC KIDNEY DISEASE - A genetic disease in which patients suffer from the growth of numerous cysts in the kidneys (see also OMIM: 601313).

PRINCIPAL COMPONENT ANALYSIS - Statistical technique that can be used to simplify a dataset by reducing the number of dimensions while retaining those characteristics of the dataset that contribute most to its variance.

PROTEOME - The complete set of proteins present in a cell, organ, etc.

RENAL-COLOBOMA SYNDROME - Genetic condition with kidney abnormalities and a cleft or defect in the eye (see also OMIM: 120330).

ROBINOW-SORAUF SYNDROME - An inherited disorder causing abnormalities of the skull and face and the hands and feet (see also OMIM: 180750).

ROBINOW SYNDROME - A syndrome characterized by abnormal face (resembling an early foetus), short forearms, and underdeveloped genitals, but no achondroplasia. This leads to dwarfism without mental retardation (see also OMIM: 268310).

RUFOUS OCULOCUTANEOUS ALBINISM - This trait is most common autosomal recessive disorder among native southern Africans and is characterized by bright copper-red coloration of the skin and hair and dilution of the colour of the iris (see also OMIM: 278400).

SAETHRE-CHOTZEN SYNDROME - Saethre-Chotzen syndrome is characterized by craniofacial and limb anomalies (see also OMIM: 101400).

SCALE-FREE NETWORKS - Networks consist of nodes and the connections between them. In a scale-free network, some nodes exhibit extremely high number of connections (called hubs). The vast majority of nodes are relatively poorly connected.

SERTOLI CELLS - Cells in the testis that are nurturing the developing sperm cells during the process of spermatogenesis.

SKELETAL DYSPLASIAS - Genetic disorders of the skeleton.

STICKLER–KNIEST FAMILY - The overlap between these two inherited disorders of the skeletal system.

STICKLER SYNDROME - This group of hereditary syndromes involves a characteristic facial appearance, eye abnormalities, hearing loss and joint problems. Many individuals are born with cleft palates (an opening in the roof of the mouth) (see also OMIM: 108300/604841/ 184840).

SYNDROME - The group or recognizable pattern of symptoms or abnormalities that indicate a particular trait or disease.

SYNDROMOLOGY - The recognition and classification of patterns of multiple congenital anomalies.

TOWNES-BROCKS SYNDROME - Clinical features of this syndrome are abnormal anus, and hand, foot and ear anomalies (see also OMIM: 107480).

TRANSCRIPTOMICS - The complete set of transcripts (RNAs) present in a cell, organ, etc.

TRICHO-DENTO-OSSEOUS SYNDROME - A hereditary condition that mainly involves the hair, teeth, and bones. Individuals are born with a full head of kinky hair. Nails are thin and likely to peel or fracture, but bones are denser (see also OMIM: 190320).

TRICHOTHIODYSTROPHY - Patients have brittle hair and nails, and their skin is dry, thickened and darker. They suffer from physical and mental retardation and approximately half of the patients are sensitive to light (see also OMIM: 601675).

TRISMUS-PSEUDOCAMPTODACTYLY SYNDROME - A genetic disorder characterized by the inability to completely open the mouth. Abnormal short muscle-tendons in the fingers cause the fingers to curve or bend (not permanent) when the hand is bent back (see also OMIM: 158300).

TUBEROUS SCLEROSIS - A genetic disorder with non-cancerous (benign) tumours in the kidneys, brain, eyes, heart, lungs, and skin. It is associated with mental retardation and seizures (see also OMIM: 191100).

ULNAR-MAMMARY SYNDROME - This genetic disorder is characterized by abnormalities affecting the bones of the forearms and hands and/or underdevelopment and dysfunction of certain sweat glands and/or the breasts (see also OMIM: 181450).

UPGMA - Unweighted Pair Group Method with Arithmetic mean (UPGMA) is a data clustering method. This method is often used to construct phylogenetic trees.

USHER SYNDROME - Recessively inherited deafness and retinitis pigmentosa.

WAARDENBURG SYNDROME - Dominantly inherited white forelock, unequal or reduced pigmentation of the iris and deafness.

WALKER–WARBURG SYNDROME - A rare autosomal recessive genetic disorder, the most consistent features of which are a lack of normal folds in the brain (lissencephaly), malformations of the back portion of the brain (cerebellum), abnormalities of the retina of the eye, and progressive degeneration and weakness of the voluntary muscles (congenital muscular dystrophy) (see also OMIM: 236670).

WISKOTT-ALDRICH SYNDROME - An X-linked recessive disease characterized by eczema, low platelet counts, immune deficiency, and bloody diarrhoea (see also OMIM: 301000).

XERODERMA PIGMENTOSUM - An inherited childhood skin eruption that is characterized by multiple pigmented spots (that resemble freckles) and larger atrophic lesions, eventually resulting in a glossy white thinning of the skin.

ZELLWEGER SYNDROME - A genetic disease characterized by the reduced ability to process toxic substances in cells of the liver, kidneys, and brain. Abnormal brain and nerve insulator development result and lead to mental retardation and other features (see also OMIM: 214100).

Appendix 1

Databases

Databases and other web addresses

Genome Browsers			
Name	URL	Description	Reference
Ensembl	http://www.ensembl.org/	The project provides a comprehensive and integrated source of annotation of large genome sequences	[1]
UCSC genome browser	http://genome.ucsc.edu/	Contains reference sequence and working draft assemblies for a large collection of genomes.	[2]
NCBI Map viewer	http://www.ncbi.nlm.nih.gov/mapview/	Map Viewer supports search and display of genomic information by chromosomal position.	[3]
Nucleic acid sequences			
EMBL	http://www.ebi.ac.uk/embl/	Europe's primary nucleotide sequence resource. Containing DNA and RNA sequences from direct submissions from individual researchers, genome sequencing projects and patent applications.	[4]
Genbank	http://www.ncbi.nlm.nih.gov/Genbank/	GenBank contains publicly available DNA and RNA sequences for more than 140.000 named organisms, submitted by individual researchers and batch submissions from large-scale sequencing projects.	[5]
DDBJ	http://www.ddbj.nig.ac.jp/	DNA Database of Japan	[6]
Protein sequences			
SwissProt/UniProt	http://www.uniprot.org	UniProt is a comprehensive, fully classified, and annotated protein sequence knowledgebase with cross-references.	[7]
PIR	http://pir.georgetown.edu/home.shtml	The Protein Information Resource (PIR) is an integrated public resource of protein informatics.	[8]
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	Provides a non-redundant collection of sequences representing genomic data, transcripts and proteins.	[9]

Protein families, domains and functional sites			
InterPro	http://www.ebi.ac.uk/interpro	An integrated documentation resource of protein families, domains and functional sites, was created to integrate the major protein signature databases.	[10]
PROSITE	http://www.expasy.org/prosite/	Consists of a large collection of biologically meaningful signatures that are described as patterns or profiles.	[11]
Pfam	http://www.sanger.ac.uk/Software/Pfam/	Large collection of protein families and domains.	[12]
PRINTS	http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/	A compendium of protein fingerprints (a group of conserved motifs used to characterise a protein family)	[13]
ProDom	http://prodes.toulouse.inra.fr/prodom/current/html/home.php	A comprehensive set of protein domain families automatically generated from the SwissProt and TrEMBL sequence databases.	[14]
SMART	http://smart.embl-heidelberg.de/	Simple Modular Architecture Research Tool, allows the identification and annotation of genetically mobile domains and the analysis of domain architectures.	[15]
TIGRFAMs	http://www.tigr.org/TIGRFAMs/	TIGRFAMs is a collection of manually curated protein families consisting of hidden Markov models, multiple sequence alignments, commentary, Gene Ontology assignments, literature references and pointers to related TIGRFAMs, Pfam and InterPro models	[16]
PIRSF	http://pir.georgetown.edu/pirsf/	PIR has extended its superfamily concept and developed the SuperFamily (PIRSF) classification system.	[17]
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/	Database of protein domain structures containing domains classified into superfamilies and sequence families.	[18]
PANTHER	http://panther.appliedbiosystems.com	Large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise.	[19]

Structural databases			
PDB	http://www.rcsb.org/pdb/	The single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.	[20]
DSSP	http://swift.cmbi.ru.nl/gv/dssp/	Containing secondary structure assignments for all protein entries in the Protein Data Bank (PDB).	[21]
HSSP	http://swift.cmbi.ru.nl/gv/hssp/	Multiple sequence alignments for all proteins in the PDB.	[22]
Organism specific databases			
SGD	http://www.yeastgenome.org/	A database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae*.	[23]
FlyBase	http://flybase.org/	FlyBase is a database of genetic and molecular data for Drosophila. FlyBase includes data on all species from the family Drosophilidae; the primary species represented is *Drosophila melanogaster*.	[24]
WormBase	http://www.wormbase.org	Wormbase is the model organism database for information about *Caenorhabditis elegans* and related nematodes.	[25]
MGI(/MGD)	http://www.informatics.jax.org	The Mouse Genome Informatics Database provides integrated access to data on the genetics, genomics and biology of the laboratory mouse.	[26]
GDB	http://gdbwww.gdb.org/	The Genome Database is a public repository of data on human genes, clones, STSs, polymorphisms and maps.	[27]
Classification systems and related resources			
ICD	http://www.who.int/classifications/icd/en/	This system was designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics	see URL
SNOMED	http://www.snomed.org/	Systematized Nomenclature of Medicine is a reference terminology for describing a complete medical file, including diagnosis and treatment.	see URL

ULMS	http://www.nlm.nih.gov/research/umls/	The purpose of the Unified Medical Language System is to aid the development of computer systems, which 'comprehend' the meaning of the medical texts	[28]
MeSH	http://www.nlm.nih.gov/mesh/	The Medical Subject Headings is a controlled vocabulary used for indexing articles for MEDLINE/PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts.	[29]
PubMed	http://www.ncbi.nlm.nih.gov/pubmed/	PubMed is the main life science literature source, including over 15 million citations for biomedical articles back to the 1950's.	[3]
Disease/variation databases			
HGMD	http://www.hgmd.org/	The Human Gene Mutation Database constitutes a comprehensive core collection of data on germ-line mutations in nuclear genes underlying or associated with human inherited disease.	[30]
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/	A repository for single base nucleotide substitutions and short deletion and insertion polymorphisms that contains 9.8 million human SNPs as well as about 5 million from a variety of other organisms.	[3]
OMIM	http://www.ncbi.nlm.nih.gov/Omim/	A catalogue of human genes and genetic disorders.	[31]
LDDB	http://www.hgmp.mrc.ac.uk/DHMHD/lddb.html	The London Dysmorphology Database contains over 3.000 non-chromosomal, multiple congenital anomaly syndromes that can be used both as an aid to diagnosis for the clinician and as a reference source.	[5]
DHMHD	http://www.hgmp.mrc.ac.uk/DHMHD/dysmorph.html	The Dysmorphic Human-Mouse Homology Database contains mouse/human malformation and mapping data.	[32]

POSSUM	http://www.possum.net.au/	Pictures of Standard Syndromes and Undiagnosed Malformations is a computer-based system that helps clinicians to diagnose syndromes in their patients.	[33]
Pathway databases			
KEGG	http://www.genome.jp/kegg/	Kyoto Encyclopedia of Genes and Genomes aiming to represent the cell and the organism in the computer	[34]
Reactome	http://www.reactome.org/	A curated resource of core pathways and reactions in human biology.	[35]
BRENDA	http://www.brenda.uni-koeln.de/	BRaunschweig ENzyme Database represents a comprehensive collection of enzyme and metabolic information, based on primary literature.	[36]
Expression databases			
Unigene	http://www.ncbi.nlm.nih.gov/UniGene/	UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters.	[3, 37, 38]
SMD	http://smd.stanford.edu/	Stanford Microarray Database functions as a public providing microarray data published by SMD users.	[39]
SAGEmap	http://www.ncbi.nlm.nih.gov/SAGE/	SAGEmap is a SAGE data resource for the query and retrieval and analysis of SAGE data from any organism.	[40]
GEO	http://www.ncbi.nlm.nih.gov/geo/	The Gene Expression Omnibus is a high-throughput gene expression data repository, as well as a curated, online resource for gene expression data browsing, query and retrieval.	[41]
Interaction databases			
BIND	http://bind.ca	Biomolecular Interaction Network Database archives biomolecular interaction, reaction, complex and pathway information.	[42]
HPRD	http://www.hprd.org/	Human Protein Reference Database integrates and visualizes information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for human proteins.	[43, 44]

MIPS	http://mips.gsf.de/	Munich Information Center for Protein Sequences provides protein sequence-related information based on whole-genome analysis.	[45]
DIP	http://dip.doe-mbi.ucla.edu/	Catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions.	[46]
Orthology databases			
COG	http://www.ncbi.nlm.nih.gov/COG/	Clusters of Orthologous Groups are a phylogenetic classification of proteins encoded in complete genomes	[47]
KOG	http://www.ncbi.nlm.nih.gov/COG/new/shokog.cgi	Similar to COG, but for eukaryotic genomes	[48]
OrthoDisease	http://orthodisease.cgb.ki.se	OrthoDisease is a comprehensive database of model organism genes that are orthologous to human disease genes.	[49]
Miscellaneous			
GO	http://www.geneontology.org/	Gene Ontology is a controlled vocabulary and is a collaborative effort to address the need for consistent descriptions of gene products in different databases.	[50]
GOA	http://www.ebi.ac.uk/GOA/	Aims to provide high-quality electronic and manual annotations to UniProt (Swiss-Prot, TrEMBL and PIR-PSD) using the standardized vocabulary of the Gene Ontology (GO).	[51]
GeneCards	http://bioinfo.weizmann.ac.il/cards/	GeneCards is a database of human genes, their products and their involvement in diseases.	[52]
Linkage	http://linkage.rockefeller.edu/	Web resources of genetic linkage analysis	see URL

References

1. Hubbard, T., et al., *Ensembl 2005.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D447-53.
2. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
3. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D39-45.
4. Kanz, C., et al., *The EMBL Nucleotide Sequence Database.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D29-33.
5. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D34-8.
6. Miyazaki, S., et al., *DDBJ in the stream of various biological data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D31-4.
7. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33 Database Issue**: p. D154-9.
8. Wu, C. and D.W. Nebert, *Update on genome completion and annotations: Protein Information Resource.* Hum Genomics, 2004. **1**(3): p. 229-33.
9. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D501-4.
10. Mulder, N.J., et al., *InterPro, progress and status in 2005.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D201-5.
11. Hulo, N., et al., *Recent improvements to the PROSITE database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D134-7.
12. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.
13. Attwood, T.K., *The PRINTS database: a resource for identification of protein families.* Brief Bioinform, 2002. **3**(3): p. 252-63.
14. Servant, F., et al., *ProDom: automated clustering of homologous domains.* Brief Bioinform, 2002. **3**(3): p. 246-51.
15. Letunic, I., et al., *SMART 4.0: towards genomic data integration.* Nucleic Acids Res, 2004. **32**(Database issue): p. D142-4.
16. Haft, D.H., J.D. Selengut, and O. White, *The TIGRFAMs database of protein families.* Nucleic Acids Res, 2003. **31**(1): p. 371-3.
17. Wu, C.H., et al., *PIRSF: family classification system at the Protein Information Resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D112-4.
18. Pearl, F., et al., *The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D247-51.
19. Mi, H., et al., *The PANTHER database of protein families, subfamilies, functions and pathways.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D284-8.
20. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
21. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

22. Sander, C. and R. Schneider, *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins, 1991. **9**(1): p. 56-68.

23. Dwight, S.S., et al., *Saccharomyces genome database: underlying principles and organisation.* Brief Bioinform, 2004. **5**(1): p. 9-22.

24. Drysdale, R.A., et al., *FlyBase: genes and gene models.* Nucleic Acids Res, 2005. **33**(Database issue): p. D390-5.

25. Chen, N., et al., *WormBase: a comprehensive data resource for Caenorhabditis biology and genomics.* Nucleic Acids Res, 2005. **33**(Database issue): p. D383-9.

26. Eppig, J.T., et al., *The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology.* Nucleic Acids Res, 2005. **33**(Database issue): p. D471-5.

27. Letovsky, S.I., et al., *GDB: the Human Genome Database.* Nucleic Acids Res, 1998. **26**(1): p. 94-9.

28. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.

29. Lipscomb, C.E., *Medical Subject Headings (MeSH).* Bull Med Libr Assoc, 2000. **88**(3): p. 265-6.

30. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.

31. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.

32. Evans, C.D., et al., *The dysmorphic human-mouse homology database (DHMHD): an interactive World-Wide Web resource for gene mapping.* J Med Genet, 1996. **33**(4): p. 289-94.

33. Evans, C.D., *Computer systems in dysmorphology.* Clin Dysmorphol, 1995. **4**(3): p. 185-201.

34. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-80.

35. Joshi-Tope, G., et al., *The Genome Knowledgebase: a resource for biologists and bioinformaticists.* Cold Spring Harb Symp Quant Biol, 2003. **68**: p. 237-43.

36. Schomburg, I., et al., *BRENDA, the enzyme database: updates and major new developments.* Nucleic Acids Res, 2004. **32 Database issue**: p. D431-3.

37. Schuler, G.D., et al., *A gene map of the human genome.* Science, 1996. **274**(5287): p. 540-6.

38. Boguski, M.S. and G.D. Schuler, *ESTablishing a human transcript map.* Nat Genet, 1995. **10**(4): p. 369-71.

39. Ball, C.A., et al., *The Stanford Microarray Database accommodates additional microarray platforms and data formats.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D580-2.

40. Lash, A.E., et al., *SAGEmap: a public gene expression resource.* Genome Res, 2000. **10**(7): p. 1051-60.

41. Barrett, T., et al., *NCBI GEO: mining millions of expression profiles--database and tools.* Nucleic Acids Res, 2005. **33**(Database issue): p. D562-6.

42. Alfarano, C., et al., *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D418-24.

43. Peri, S., et al., *Human protein reference database as a discovery resource for proteomics.* Nucleic Acids Res, 2004. **32 Database issue**: p. D497-501.

44. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.

45. Mewes, H.W., et al., *MIPS: analysis and annotation of proteins from whole genomes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D41-4.

46. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.

47. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.

48. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**(1): p. 41.

49. O'Brien, K.P., I. Westerlund, and E.L. Sonnhammer, *OrthoDisease: a database of human disease orthologs.* Hum Mutat, 2004. **24**(2): p. 112-9.

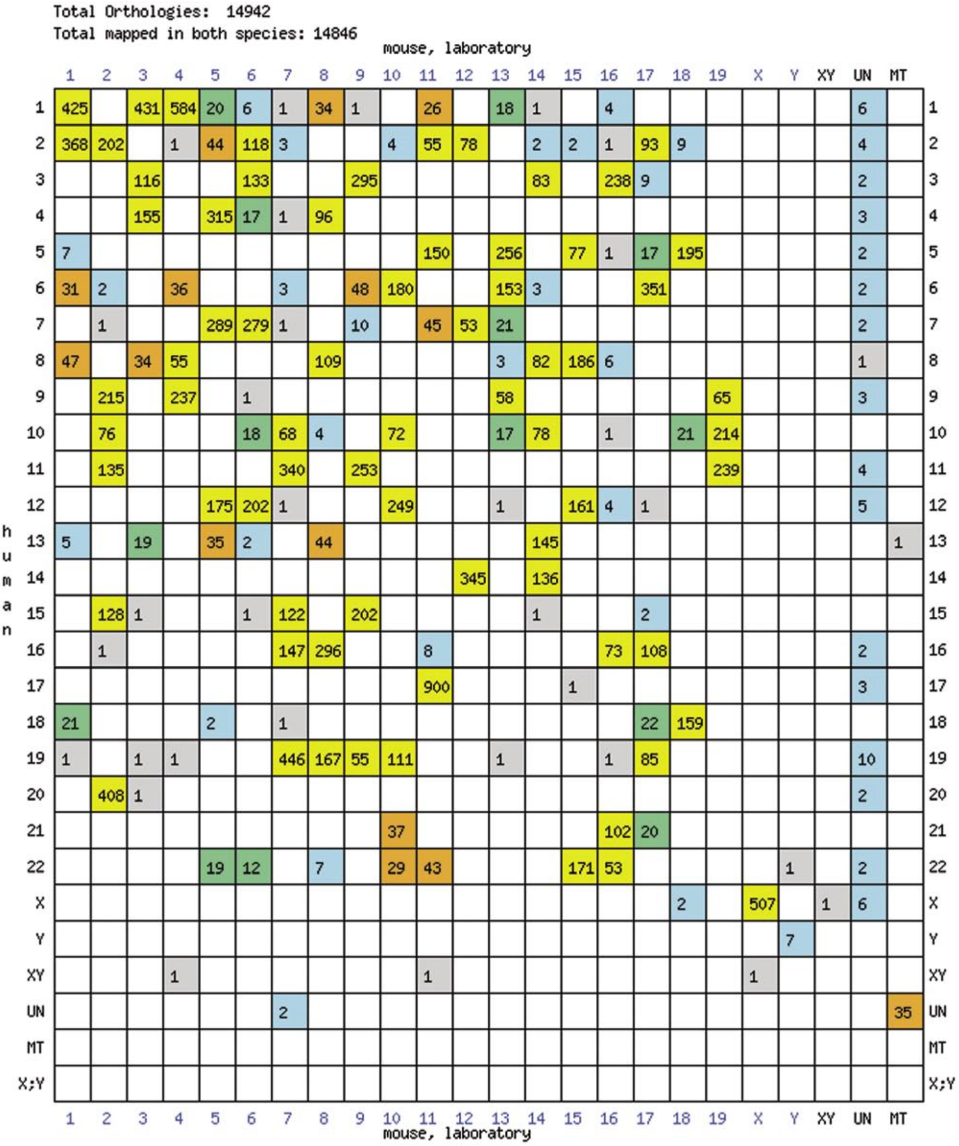
50. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.

51. Camon, E., et al., *The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D262-6.

52. Safran, M., et al., *Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE.* Nucleic Acids Res, 2003. **31**(1): p. 142-6.

Appendix 2

Colour figures



Total Orthologies: 14942

Total mapped in both species: 14846

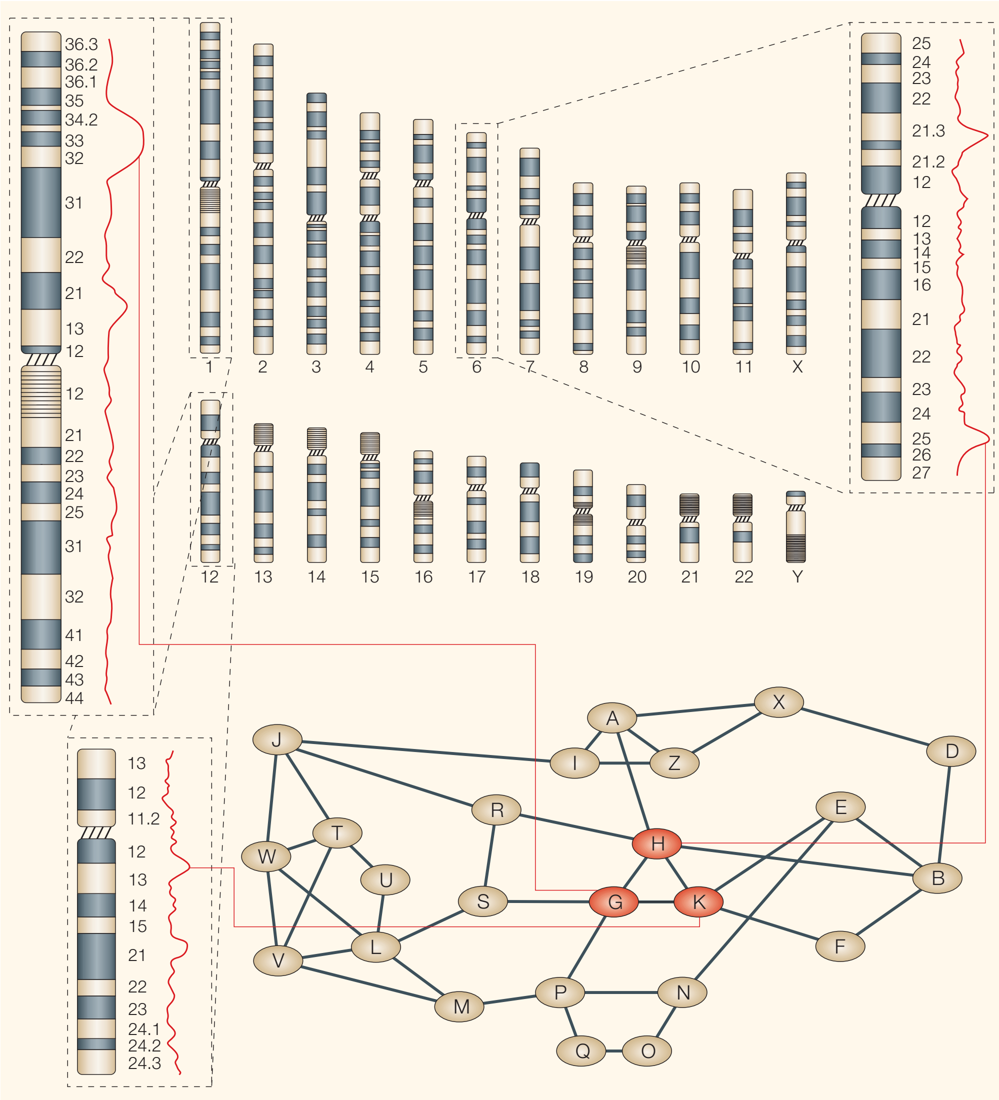
mouse, laboratory

human

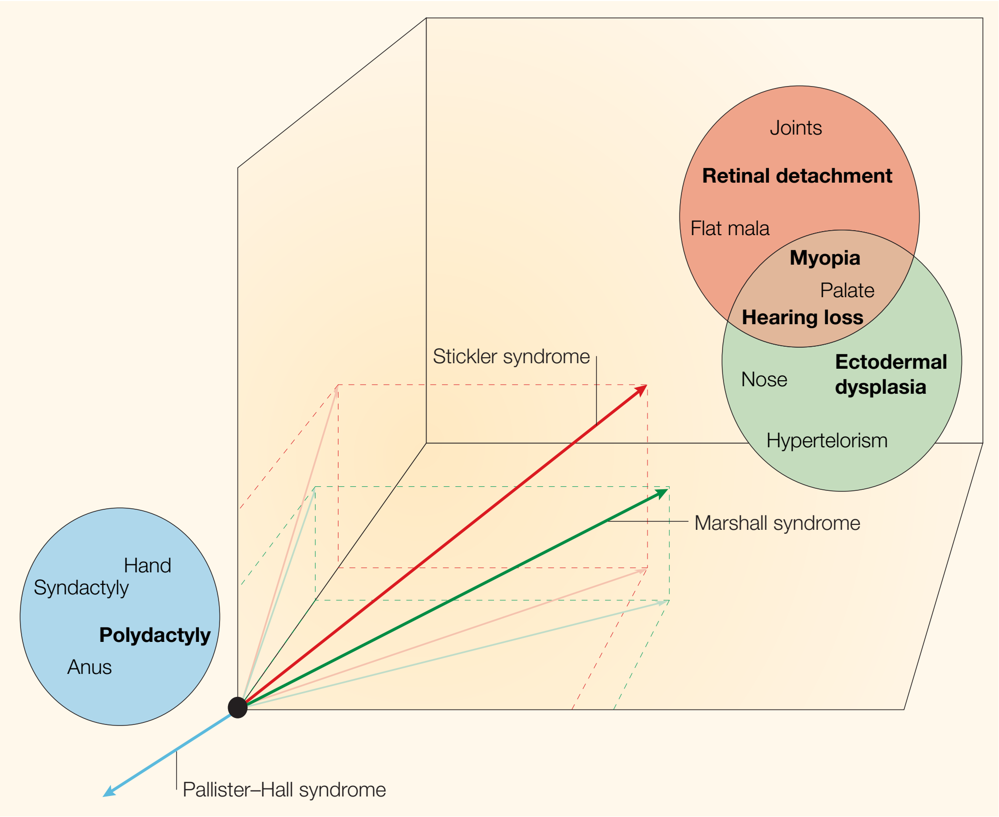
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y	XY	UN	MT	
1	425		431	584	20	6	1	34	1		26		18	1		4							6		1
2	368	202		1	44	118	3			4	55	78		2	2	1	93	9					4		2
3			116			133			295					83		238	9						2		3
4			155		315	17	1	96															3		4
5	7										150		256		77	1	17	195					2		5
6	31	2		36			3		48	180			153	3			351						2		6
7		1			289	279	1		10		45	53	21										2		7
8	47		34	55				109					3	82	186	6							1		8
9		215		237		1							58						65				3		9
10		76				18	68	4		72			17	78		1		21	214						10
11		135					340		253										239				4		11
12					175	202	1			249			1		161	4	1						5		12
13	5		19		35	2		44						145										1	13
14												345		136											14
15		128	1			1	122		202					1			2								15
16		1					147	296			8					73	108						2		16
17											900				1								3		17
18	21				2		1										22	159							18
19	1		1	1			446	167	55	111			1			1	85						10		19
20		408	1																				2		20
21										37						102	20								21
22					19	12		7		29	43				171	53					1		2		22
X																		2		507		1	6		X
Y																					7				Y
XY				1							1									1					XY
UN						2																		35	UN
MT																									MT
X;Y																									X;Y
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y	XY	UN	MT	

mouse, laboratory

Chapter 1 - Figure 1 An Oxford-grid. The grid shows the relationship between human and mouse chromosomes. Chromosome location of either of the species often predicts the chromosome location in the other species. The colours indicate the number of orthologies: Grey (1), Blue (2-10), Green (11-25), Orange (26-50), Yellow (>50). From the Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine. (http://www.informatics.jax.org, April, 2005).

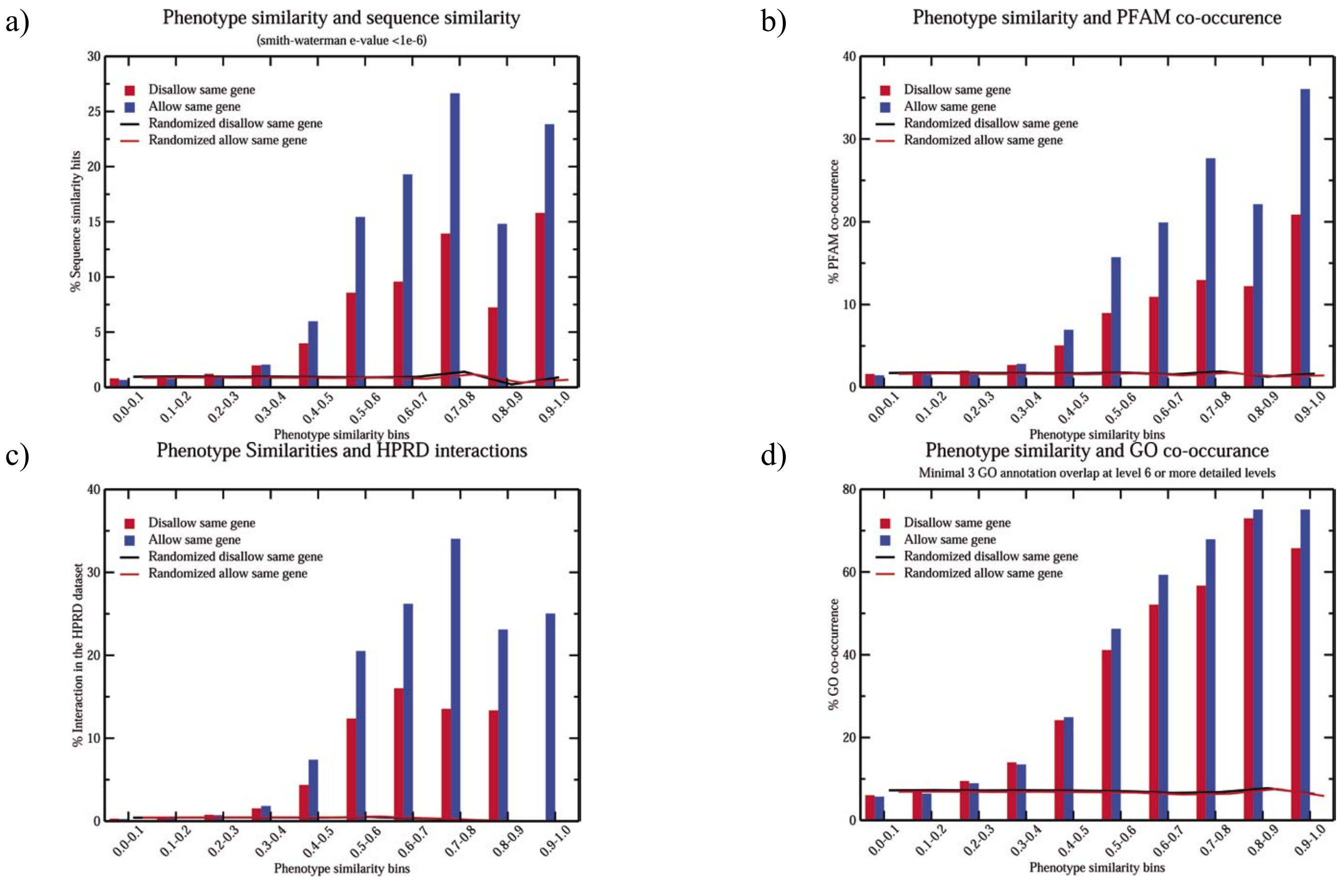


Chapter 4 - Figure 1 Integrating functional relationships and linkage analysis. Three separate genome scans for a hypothetical multifactorial disease result in multiple peaks of possible linkage (red traces) on 3 different human chromosomes (1, 6 and 12). Integrating linkage results with data from other sources, such as gene/protein networks (A … Z), points to a set of functionally related candidate genes (H, G and K) that together might explain the linkage patterns. Numbers along the highlighted chromosomes indicate the corresponding chromosome bands.

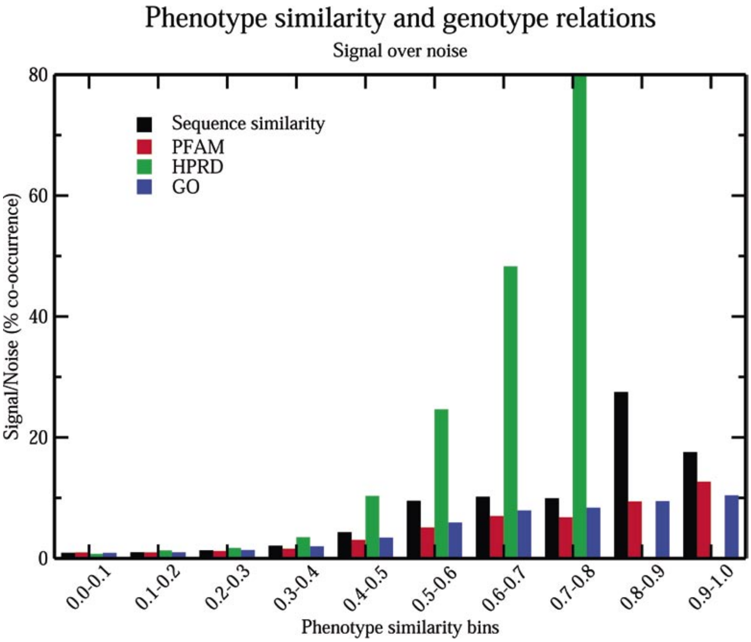


Chapter 4 - Figure 2 Phenotype clustering.

Each arrow represents a keyword vector. The components in a keyword vector correspond to terms in the document (for example, ‘myopia’, ‘joints’, and so on). Vectors that point in the same direction are more alike. Three documents that describe three syndromes are shown (red: Stickler syndrome; green: Marshall syndrome; and blue: Pallister–Hall syndrome). Stickler and Marshall syndromes have similar phenotypes and share common terms (‘myopia’, ‘palate’, ‘hearing loss’) in their keyword vectors. Pallister–Hall syndrome (weblink) is clearly different, and shares few terms with Stickler and Marshall, and therefore, the vector points in a different direction. Some terms are more important than other terms. Terms are assigned different weights by computing the inverse document frequency, which gives a weight factor to each term in the collection. Terms in bold contribute the most to vector direction. Flat mala, low cheekbones; hypertelorism, widely-spaced eyes.



Chapter 5 - Figure 4 Phenotype similarity versus genotype relation categories. a) Sequence similarity (Smith-Waterman, threshold e-value 1e-6); b) Sharing of at least one PFAM domain; c) Protein-Protein interactions according to the HPRD database; d) Sharing of 3 or more GO annotations at the 6th or more detailed level. The 0.8-0.9 and 0.9-1.0 phenotype similarity bins suffer from low counts. The average signal of ten randomized phenomaps is at the level of the two lowest bins (data not shown).



Chapter 5 - Figure 5 - Histogram of normalized genotype relations as a function of phenotype-phenotype similarity.All values are corrected for random information levels. Although the HPRD data set contains fewer relations than the other sets, the normalized signal is more than two times as strong, reflecting the quality of the HPRD dataset. The 0.8-0.9 and 0.9-1.0 phenotype similarity bins suffer from low counts.

Samenvatting

Ziektegen identificatie op basis van chromosomale lokalisatie is moeilijk en vaak tijdrovend. Het vereist het verzamelen van zoveel mogelijk informatie over de ziekte. Mogelijke hints voor het selecteren van kandidaat genen kunnen komen door het combineren van positie informatie en andere karakteristieken van de ziekte. We hebben gegevens van verschillende computer databases gecombineerd en bioinformatica strategieën ontwikkeld om zo systematisch ziektegenen te identificeren.

Genetische aandoeningen worden veroorzaakt door veranderingen in ons erfelijk materiaal (DNA). In de westerse wereld zijn genetische ziekten de hoofdoorzaak van neonatale-, kinderziekten en kindersterfte. Mutaties die leiden tot ziekten spelen een belangrijke rol in het doorgronden van erfelijke ziekten bij mensen, en de moleculaire rol van de genen en eiwitten in de ontwikkeling. Deze kennis kan worden gebruikt in erfelijkheidsadvisering, het stellen van diagnoses en in onderzoek. Kennis over ziektegenen kan uiteindelijk mogelijkheden creëren voor behandeling.

Belangrijke indicaties voor het opsporen van de oorzaken en mechanismen van ziekten, komen van de klinische karakteristieken van de ziekte (het ziekte fenotype). Zowel voor klinische- als onderzoeksdoeleinden zijn ziektedefinities en classificaties essentieel. Voor dit doel zijn diverse classificatiesystemen ontwikkeld en in gebruik. Geen van deze systemen is algemeen geaccepteerd met tot gevolg dat veel klinische informatie als vrije tekst ligt opgeslagen in literatuur databases.

We hebben twee bioinformaticastrategieën beschreven voor het identificeren van kandidaat ziektegenen bij de mens (hoofdstukken 2 en 3) en voor het classificeren van humane ziekten fenotypen (hoofdstukken 4 en 5). Onze hypothese was dat internet databases met genspecifieke informatie en databases met fenotypische beschrijvingen, systematisch kunnen worden gebruikt voor het vinden en karakteriseren van ziektegenen en processen.

We hebben een systeem ontwikkeld (hoofdstukken 2 en 3; GeneSeeker) dat als doel heeft ondersteuning te bieden in het proces voor het vinden van kandidaat genen. Een vaak gebruikte strategie is het prioritiseren van genen in het chromosomale interval dat is verkregen via stamboom onderzoek. Deze prioritisering wordt gebaseerd op gen karakteristieken zoals expressie in de aangedane weefsels en fenotypen in modelorganismen. Dit proces wordt door GeneSeeker in silico nagebootst. Over het algemeen kan de chromosomale regio d.m.v. koppelingsonderzoek worden gereduceerd tot 20-200 genen. Het ontwikkelde systeem combineert positionele informatie met expressie en fenotype data van mens en muis. In hoofdstuk 2 worden tien syndromen geanalyseerd. Voor deze syndromen, inclusief twee syndromen die ten tijde van de analyse nieuw en nog niet in de databases beschreven waren, werden de kandidaat ziekten genen lijsten gereduceerd met een factor 7-25 terwijl de ziektegenen behouden bleven.

Hoofdstuk 3 beschrijft de technische achtergrond van de GeneSeeker in meer detail. Het programma vraagt informatie op van diverse databases direct via het world wide web (WWW). Negen databases leveren informatie en de modulaire opzet van het systeem maakt het indien nodig mogelijk databases toe te voegen. Door het direct raadplegen van de originele databases via de web interfaces wordt het gebruik van de meest recente data gegarandeerd. Hierdoor kan data warehousing worden vermeden. GeneSeeker maakt kandidaatgen selecties zodat het aantal genen dat geanalyseerd moet worden kan worden beperkt.

GeneSeeker is over het algemeen het meest effectief in syndromen waarbij het ziektegen een veranderd expressie patroon heeft in de aangedane weefsels. In deze gevallen kan door de verkregen aanwijzingen tijd worden bespaard in de labexperimenten. Het veld voor kandidaatgen voorspelling en prioritisering is nieuw en slechts enkele methoden zijn gepubliceerd, ieder gebaseerd op verschillende databronnen. De prestaties van GeneSeeker zijn vergelijkbaar met deze methoden. Gebruik van GeneSeeker vereist klinische kennis vooraf. Echter, het is tot op heden de enige methode die lokalisatie, expressie, en fenotypische data van de muis exploreert.

In hoofdstuk 4 worden de relaties op fenotypisch niveau besproken. In dit hoofdstuk verkennen we, gebruik makend van een aantal voorbeelden uit de literatuur, hoe vergelijkbare syndromen/ fenotypen aan elkaar zijn gerelateerd, en of zij een indicatie kunnen zijn voor de betrokken genen. We beargumenteren dat deze gelijkenissen een afspiegeling zijn van de biologische relaties in de cel. Bovendien bespreken we de toepasbaarheid van een fenotypische kaart voor het identificeren van ziektegenen, ook bij multifactoriële/complexe aandoeningen.

In hoofdstuk 5 beschrijven we een methode voor het extraheren van fenotypische beschrijvingen vanuit de vrije tekst in de OMIM database. De fenotypische kenmerken werden m.b.v. tekstanalyse technieken geëxtraheerd. In tegenstelling tot het gebruik van vooraf (handmatig) gedefinieerde categorieën, hebben we een automatische methode gebruikt voor de fenotypische kenmerken identificatie en vergelijkingen. De fenotypische kaart werd geanalyseerd op biologisch relevante informatie. We vonden dat fenotypische gelijkenis positief gecorreleerd is met genfunctie op het niveau van eiwitsequentie, motieven, functionele annotatie, en eiwit-eiwit interacties afgeleid uit de literatuur. De analyse suggereert dat de automatisch gegenereerde fenotypische kaart inderdaad biologische relaties aan kan geven.

Het doel van dit onderzoek was om enkele algemene principes van erfelijke ziekten bij de mens te vinden en bioinformatica strategieën te ontwikkelen voor het identificeren van ziektegenen. Fenotype analyse is net zo relevant voor de identificatie van ziektegenen als het is voor de functionele annotatie van het menselijke genoom: Voor de komende jaren worden grote verbeteringen verwacht ten aanzien van genidentificatie-methoden voor monogenetische en multifactoriële ziekten, mutatie detectie, en methoden voor het vergelijken tussen organismen. Bioinformatica benaderingen zoals hier beschreven kunnen een aanvulling zijn voor analyses gebaseerd op sequentie en genen. Hoewel systematische fenotype-genotype analyses nog veel te ontdekken laten, is het essentieel om een gestandaardiseerde en internationaal aanvaarde nomenclatuur te ontwikkelen voor fenotype definitie, die zowel toepasbaar is op mensen als wel op de hedenten dage veel gebruikte modelorganismen.

Curriculum vitae

Marc van Driel werd op 12 maart 1974 geboren te Arnhem. Zijn HAVO-opleiding volgde hij aan de scholengemeenschap Presikhaaf te Arnhem, waar hij in 1992 zijn eindexamen behaalde. Vervolgens startte hij de Hogere Laboratorium opleiding in Nijmegen in de richting Biochemie/Biotechnologie. Gedurende deze opleiding deed hij de extra vakken toxicologie en management. In het kader van een afstudeerstage in 1995 bij de afdeling Antropogenetica van het Radboud ziekenhuis te Nijmegen, werkte hij aan de positionele klonering van het gen voor een erfelijke vorm van geslachtsgebonden gespleten gehemelte onder leiding van Dr. Hans van Bokhoven. In 1996 rondde hij de HLO opleiding af, waarna hij begon als research analist op de afdeling Antropogenetica. In de groep van Prof. Dr. Frans Cremers werkte hij aan de genetische karakterisatie van diverse erfelijke oogziekten (ziekte van Stargardt, ouderdoms macula degeneratie, retinitis pigmentosa, kegel-staaf dystrofie, choroideremie en cystoïde macula dystrofie). Vanaf 1998 zette hij onder leiding van Prof. Dr. Frans Cremers dit werk in de vorm van een promotie onderzoek voort in samenwerking met de afdeling Oogheelkunde (Prof. Dr. August Deutman en Dr. Carel Hoyng). In 2001 maakte hij de overstap van het laboratorium werk naar de bioinformatica en begon bij het Centrum voor Moleculaire en Biomoleculaire Informatica (CMBI). Onder leiding van Prof. Dr. Jack Leunissen (CMBI/WUR), Prof. Dr. Gert Vriend (CMBI) en Prof. Dr. Han Brunner (Antropogenetica, UMCN St. Radboud) werkte hij aan het in dit proefschrift beschreven onderzoek. Sinds juni 2005 werkt hij als bioinformaticus in het lab van Prof. Dr. Henk Stunnenberg (Moleculaire Biologie, Radboud Universiteit Nijmegen).

Dankwoord

Het proefschrift is klaar en dan is het ook tijd om terug te kijken. Veel mensen hebben een bijdrage geleverd en die wil ik graag hier bedanken. Als eerste mijn (co-)promotoren Gert, Han, en Jack. Jack, met de overstap van Antropogenetica begon ik bij jou op het CMBI met de GeneMachine, dat later tot ons beider ongenoegen helaas GeneSeeker moest gaan heten. Met je overstap naar Wageningen, om daar hoogleraar te worden, werd de afstand wel groter maar met het MimMiner project bleef het contact. Nu het proefschrift klaar is moeten we daar zeker een goede whisky op drinken. Gert, jouw betrokkenheid en inzet waren essentieel voor het tot stand komen van het proefschrift. Er was geen dag dat je niet even tijd had. Ook was er altijd tijd voor een praatje over het werk of onze andere interesses. Bedankt voor wat je me hebt geleerd en voor het vertrouwen in een goede afloop. Han, jouw bijdrage gaat terug tot de tijd dat ik nog bij Antropogenetica werkte. Daar is ons contact met het CMBI begonnen. Na mijn overstap was het dan ook logisch dat er contact bleef met jou. Dit contact groeide verder uit tot een intensieve samenwerking met dit boekje als resultaat. Bedankt voor alles wat je hebt gedaan.

Mijn paranimfen Maarten en Jacopo. Maarten, we hebben veel gelachen en ik heb veel geleerd van je programmeerkunsten. Gelukkig heb ik ook nog wat bij kunnen dragen: data::dumper.. Jacopo, vanaf het moment dat je bij Antropo kwam werken hebben we veel gepraat en nog meer pasta gegeten. Bedankt voor jullie steun.

Martijn, Berend, de discussies met jullie waren meer dan welkom en ik heb er zowel direct als indirect veel aan gehad. Bedankt. Barbara, Esther, zonder jullie organisatorische ondersteuning zouden er veel zijn blijven liggen. Cursusplanning, materiaal, maar ook talloze andere zaken liepen gesmeerd dankzij jullie. Koen, menigmaal heb ik je aan de telefoon gehad voor de Antropogenetica bioinformatica vragen. Je besloot over te stappen naar Lion Bioscience en ik nam je project over. Bedankt voor je bijdrage en dat Belgische pintke pakken we straks. Patrick en de studenten Eric, Pieter en Geerten wil ik bedanken voor hun bijdrage aan de GeneSeeker. Binnen het MimMiner project ben ik bijzondere dank verschuldigd aan Jorn en later Martin. De meeste van jullie doen nu zelf promotie-onderzoek, veel succes daarmee. Op het CMBI stonden alle deuren open en ik kon vrijwel altijd binnenlopen met een vraag of voor een praatje. De afdeling is erg gegroeid de afgelopen jaren en er zijn mensen gekomen en gegaan. Er zijn meer dan 60 namen te noemen, maar dat durf ik niet aan. Ik wil jullie allemaal bedanken voor de bijdrage aan dit proefschrift, de leuke werksfeer en de gezellige koffie/thee/lunchpauzes. Een tijd die me nog lang zal heugen.

Hoewel ik dit werk heb gedaan bij het CMBI zijn er een aantal mensen van Antropogenetica die een belangrijke rol hebben gespeeld. Hans, in 1995 begon ik mijn stage bij jou en dat was een erg leuke tijd. Je zangkwaliteiten zijn geloof ik nog niet echt verbeterd, maar die van mij ook niet. Bedankt voor je inzet en tijd. Frans, toen je hoorde dat ik in Amsterdam kon beginnen heb je snel een leuke plek voor me op het lab geregeld. Eerst als analist, maar later gaf je me de kans om promotie-onderzoek te gaan doen. We hebben veel gelachen maar ook in andere tijden had je altijd een luisterend oor. Bedankt. Anneke, Dorien, Jacopo en Alessandra, jullie waren in werk en privé altijd daar. Bedankt. Ook alle andere mensen met wie ik bij Antropogenetica en ook Oogheelkunde heb gewerkt wil ik hier graag bedanken voor de leuke en leerzame tijd. Vrienden buiten het werk zijn onmisbaar gebleken. Lieke, de squash-uurtjes waren altijd een welkome uitlaatklep of we nu speelden of niet. Maar ook voor een helpende hand was je altijd te bellen. Norbert, fotografie was en is vaak het onderwerp. Daarnaast passeerde van alles de revue met natuurlijk een glas wijn. Nu is het tijd voor een nieuw, wellicht gezamenlijk, foto project. Nathalie, je motivatie en vriendschap zijn al die tijd een steun in de rug geweest. Cappuccino kaneel? Dennis, Remco en Hedy. Nacht, dag, druk, vakantie of thuis en altijd interesse. Jullie sleurden me nukkig of niet overal mee naar toe, soms op de meest onverwachte momenten. Bedankt dat jullie altijd klaar stonden. De familie Oudakker, jullie vroegen altijd hoe het ging en waren immer gastvrij. Bedankt voor jullie steun.

Astrid, zonder jou was dit boekje er niet geweest. Als geen ander bracht je een lach, rust, talloze andere dingen en was je een steun en toeverlaat. Duizendmaal dank!

Mijn broers Arno en Lars en mijn ouders. Betere had ik me niet kunnen wensen. Dit proefschrift is ook van jullie. Het is het resultaat van jullie steun, opvoeding, en onvoorwaardelijke liefde.

Marc

List of publications

Allikmets, R., A. Hutchinson, R.A. Lewis, N.F. Shroyer, K. Dalakishvili, J.R. Lupski, K. Steiner, D. Pauleikhoff, F. Holz, B.H.F. Weber, P.S. Bernstein, N. Singh, N. Zabriskie, A. Peiffer, M. Leppert, J.M. Seddon, K. Zhang, J.S. Sunness, N.S. Udar, S. Yelchits, R. Silva-Garcia, K.W. Small, F. Simonelli, F. Testa, M. D'Urso, R. Brancato, E. Rinaldi, S. Ingvast, A. Taube, C. Wadelius, E. Souied, D. Ducroq, J. Kaplan, J.J.M. Assink, J.B. Brink, P.T.V.M. de Jong, A.A.B. Bergen, A. Maugeri, **M.A. van Driel**, C.B. Hoyng, F.P.M. Cremers, E. Paloma, R. Coco, S. Balcells, R. Gonzàlez-Duarte, S. Kermani, P. Stanga, A.C. Bird, S.S. Bhattacharya, and t.i.A.S. Consortium. 2000. Further evidence for an association of ABCR alleles with age-related macular degeneration. The International ABCR Screening Consortium. Am J Hum Genet 67: 487-491.

Brunner, H.G. and **M.A. van Driel**. 2004. From syndrome families to functional genomics. Nat Rev Genet 5: 545-551.

Cremers, F.P., D.J. van de Pol, **M. van Driel**, A.I. den Hollander, F.J. van Haren, N.V. Knoers, N. Tijmes, A.A. Bergen, K. Rohrschneider, A. Blankenagel, A.J. Pinckers, A.F. Deutman, and C.B. Hoyng. 1998. Autosomal recessive retinitis pigmentosa and cone-rod dystrophy caused by splice site mutations in the Stargardt's disease gene ABCR. Hum Mol Genet 7: 355-362.

den Hollander, A.I., J.B. ten Brink, Y.J. de Kok, S. van Soest, L.I. van den Born, **M.A. van Driel**, D.J. van de Pol, A.M. Payne, S.S. Bhattacharya, U. Kellner, C.B. Hoyng, A. Westerveld, H.G. Brunner, E.M. Bleeker-Wagemakers, A.F. Deutman, J.R. Heckenlively, F.P. Cremers, and A.A. Bergen. 1999a. Mutations in a human homologue of Drosophila crumbs cause retinitis pigmentosa (RP12). Nat Genet 23: 217-221.

den Hollander, A.I., **M.A. van Driel**, Y.J. de Kok, D.J. van de Pol, C.B. Hoyng, H.G. Brunner, A.F. Deutman, and F.P. Cremers. 1999b. Isolation and mapping of novel candidate genes for retinal disorders using suppression subtractive hybridization. Genomics 58: 240-249.

Go, S.L., A. Maugeri, J.J. Mulder, **M.A. van Driel**, F.P. Cremers, and C.B. Hoyng. 2003. Autosomal dominant rhegmatogenous retinal detachment associated with an Arg453Ter mutation in the COL2A1 gene. Invest Ophthalmol Vis Sci 44: 4035-4043.

Klevering, B.J., **M. van Driel**, D.J. van de Pol, A.J. Pinckers, F.P. Cremers, and C.B. Hoyng. 1999. Phenotypic variations in a family with retinal dystrophy as result of different mutations in the ABCR gene. Br J Ophthalmol 83: 914-918.

Klevering, B.J., **M. van Driel**, A.J. van Hogerwou, D.J. van De Pol, A.F. Deutman, A.J. Pinckers, F.P. Cremers, and C.B. Hoyng. 2002. Central areolar choroidal dystrophy associated with dominantly inherited drusen. Br J Ophthalmol 86: 91-96.

Maugeri, A., **M.A. van Driel**, D.J. van de Pol, B.J. Klevering, F.J. van Haren, N. Tijmes, A.A. Bergen, K. Rohrschneider, A. Blankenagel, A.J. Pinckers, N. Dahl, H.G. Brunner, A.F. Deutman, C.B. Hoyng, and F.P. Cremers. 1999. The 2588G-->C mutation in the ABCR gene is a mild frequent founder mutation in the Western European population and allows the classification of ABCR mutations in patients with Stargardt disease. Am J Hum Genet 64: 1024-1035.

Schulz, H.L., H. Stoehr, K. White, **M.A. van Driel**, C.B. Hoyng, F. Cremers, and B.H. Weber. 2002. Genomic structure and assessment of the retinally expressed RFamide-related peptide gene in dominant cystoid macular dystrophy. Mol Vis 8: 67-71.

van den Hurk, J.A., D.J. van de Pol, B. Wissinger, **M.A. van Driel**, L.H. Hoefsloot, I.J. de Wijs, L.I. van den Born, J.R. Heckenlively, H.G. Brunner, E. Zrenner, H.H. Ropers, and

F.P. Cremers. 2003. Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. Hum Genet 113: 268-275.

van Driel, M.A., J. Bruggeman, G. Vriend, H.G. Brunner, and J.A. Leunissen. 2005a. A text-mining analysis of the human phenome. submitted.

van Driel, M.A., K. Cuelenaere, P.P. Kemmeren, J.A. Leunissen, and H.G. Brunner. 2003. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. Eur J Hum Genet 11: 57-63.

van Driel, M.A., K. Cuelenaere, P.P. Kemmeren, J.A. Leunissen, H.G. Brunner, and G. Vriend. 2005b. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res 33: W758-761.

van Driel, M.A., A. Maugeri, B.J. Klevering, C.B. Hoyng, and F.P. Cremers. 1998. ABCR unites what ophthalmologists divide(s). Ophthalmic Genet 19: 117-122.

Van Lith-Verhoeven, J.J., **M.A. van Driel**, I.C. Meij, L. van Laer, A.J. Pinckers, H. Kremer, A.F. Deutman, H.G. Brunner, F.P. Cremers, and C.B. Hoyng. 2003. Clinical Classification of Autosomal Dominant Cystoid Macular Edema and Genetic Fine Mapping of the Underlying Defect. Invest. Ophthalmol. Vis. Sci. 44: 1496-.