Chapter 21:
Human disease

Learning objectives

After studying this chapter, you should be able to:

■ describe major categories of human disease;
■ explain different approaches to identifying disease-associated genes;
■ compare and contrast the main disease databases;
■ describe how studies of model organisms elucidate disease-related variation.

Outline

Human genetic disease: a consequence of DNA variation

Categories of disease

Disease databases

Approaches to identifying disease-associated genes and loci

Human disease genes in model organisms

Functional classification of disease genes

Perspective

Human disease: a consequence of variation

Genetic variation is responsible for the adaptive changes that underlie evolution.

Some changes improve the fitness of a species.
Other changes are maladaptive.
A maladaptation is a trait that is more harmful than helpful
For the individual in a species, these maladaptive changes represent disease.

Molecular perspective: mutation and variation

Medical perspective: pathological condition

Why is there such a diversity of diseases?

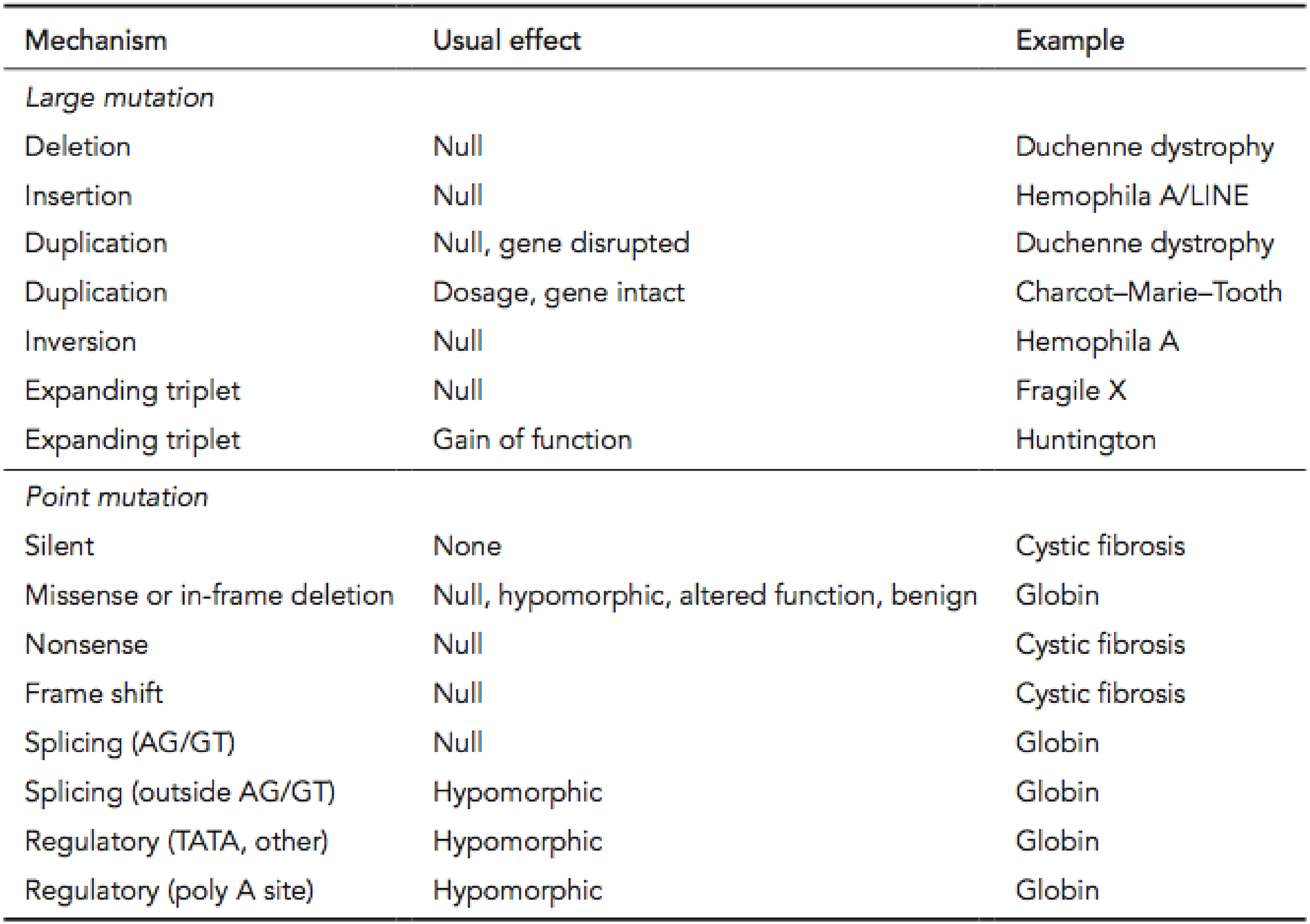
-- many regions of the genome may be affected

-- there are many mechanisms of mutation

-- genes and gene products interact with their molecular environments

-- an individual interacts with the environment in ways that may promote disease

Mechanisms of genetic mutation



Mechanism	Usual effect	Example
Large mutation		
Deletion	Null	Duchenne dystrophy
Insertion	Null	Hemophila A/LINE
Duplication	Null, gene disrupted	Duchenne dystrophy
Duplication	Dosage, gene intact	Charcot–Marie–Tooth
Inversion	Null	Hemophila A
Expanding triplet	Null	Fragile X
Expanding triplet	Gain of function	Huntington
Point mutation		
Silent	None	Cystic fibrosis
Missense or in-frame deletion	Null, hypomorphic, altered function, benign	Globin
Nonsense	Null	Cystic fibrosis
Frame shift	Null	Cystic fibrosis
Splicing (AG/GT)	Null	Globin
Splicing (outside AG/GT)	Hypomorphic	Globin
Regulatory (TATA, other)	Hypomorphic	Globin
Regulatory (poly A site)	Hypomorphic	Globin

AG/GT indicates mutations in the canonical first two and last two base pairs of an intron. From Beaudet *et al.* (2001).

- **Duchenne** muscular **dystrophy** (DMD) is a severe type of muscular **dystrophy**. The symptom of muscle weakness usually begins around the age of four in boys and worsens quickly. Typically muscle loss occurs first in the thighs and pelvis followed by those of the arms.

- **Haemophilia** A (or **hemophilia** A) is a genetic deficiency in clotting factor VIII, which causes increased bleeding and usually affects males. In the majority of cases it is inherited as an X-linked recessive trait, though there are cases which arise from spontaneous mutations.

- **Charcot–Marie–Tooth disease** (**CMT**) is one of the hereditary motor and sensory neuropathies, a group of varied inherited disorders of the peripheral nervous system characterized by progressive loss of muscle tissue and touch sensation across various parts of the body.

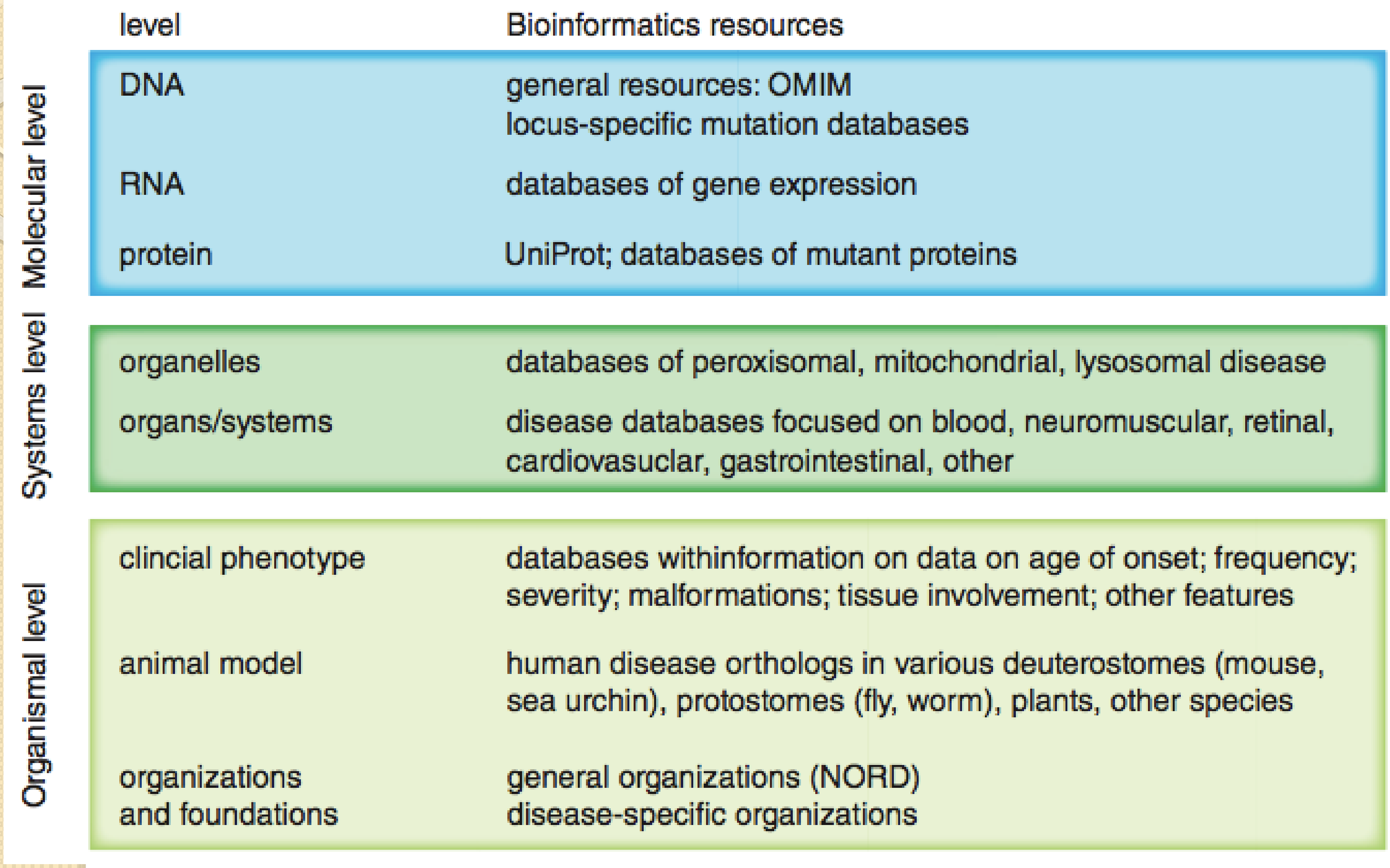
- **Cystic fibrosis** is a hereditary disease that affects the lungs and digestive system. The body produces thick and sticky mucus that can clog the lungs and obstruct the pancreas.

Bioinformatics perspectives on disease

The field of bioinformatics involves the use of computer algorithms and databases to study genes, genomes, and proteins.

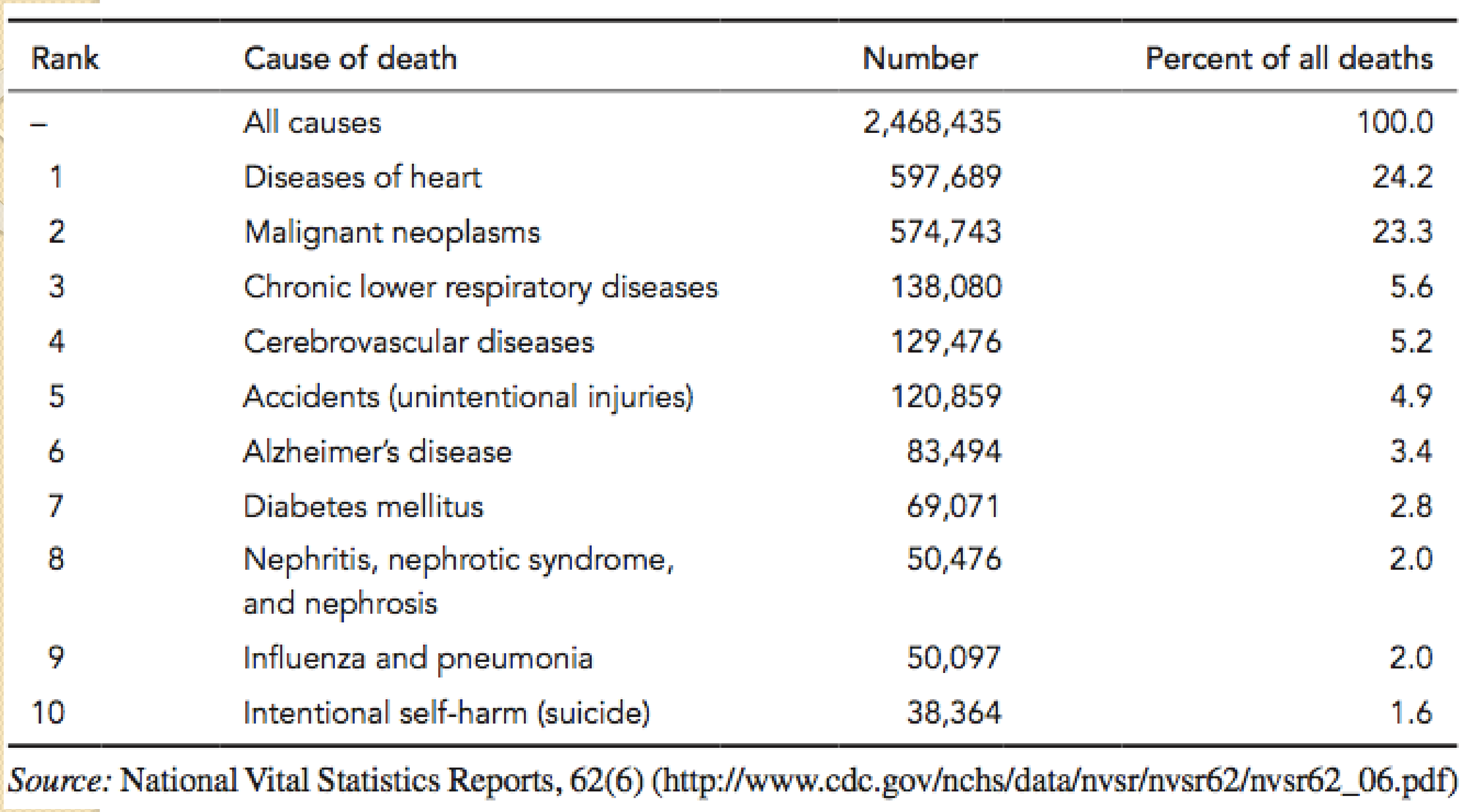
- DNA databases offer reference sequences to compare normal and disease-associated sequences
- Physical and genetic maps are used in gene-finding
- Protein structure studies allow study of effects of mutation
- Many functional genomics approaches applied to genes
- Insight into human disease genes is provided through the study of orthologs and their function

Bioinformatics resources for the study of human disease



	level	Bioinformatics resources
Molecular level	DNA	general resources: OMIM locus-specific mutation databases
	RNA	databases of gene expression
	protein	UniProt; databases of mutant proteins
Systems level	organelles	databases of peroxisomal, mitochondrial, lysosomal disease
	organs/systems	disease databases focused on blood, neuromuscular, retinal, cardiovasuclar, gastrointestinal, other
Organismal level	clincial phenotype	databases withinformation on data on age of onset; frequency; severity; malformations; tissue involvement; other features
	animal model	human disease orthologs in various deuterostomes (mouse, sea urchin), protostomes (fly, worm), plants, other species
	organizations and foundations	general organizations (NORD) disease-specific organizations

Leading causes of death



Rank	Cause of death	Number	Percent of all deaths
–	All causes	2,468,435	100.0
1	Diseases of heart	597,689	24.2
2	Malignant neoplasms	574,743	23.3
3	Chronic lower respiratory diseases	138,080	5.6
4	Cerebrovascular diseases	129,476	5.2
5	Accidents (unintentional injuries)	120,859	4.9
6	Alzheimer's disease	83,494	3.4
7	Diabetes mellitus	69,071	2.8
8	Nephritis, nephrotic syndrome, and nephrosis	50,476	2.0
9	Influenza and pneumonia	50,097	2.0
10	Intentional self-harm (suicide)	38,364	1.6

Source: National Vital Statistics Reports, 62(6) (http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_06.pdf)

Cause of death is based on the international Classification of Diseases, tenth revision.

Projected global deaths for selected causes of death, 2002–2030

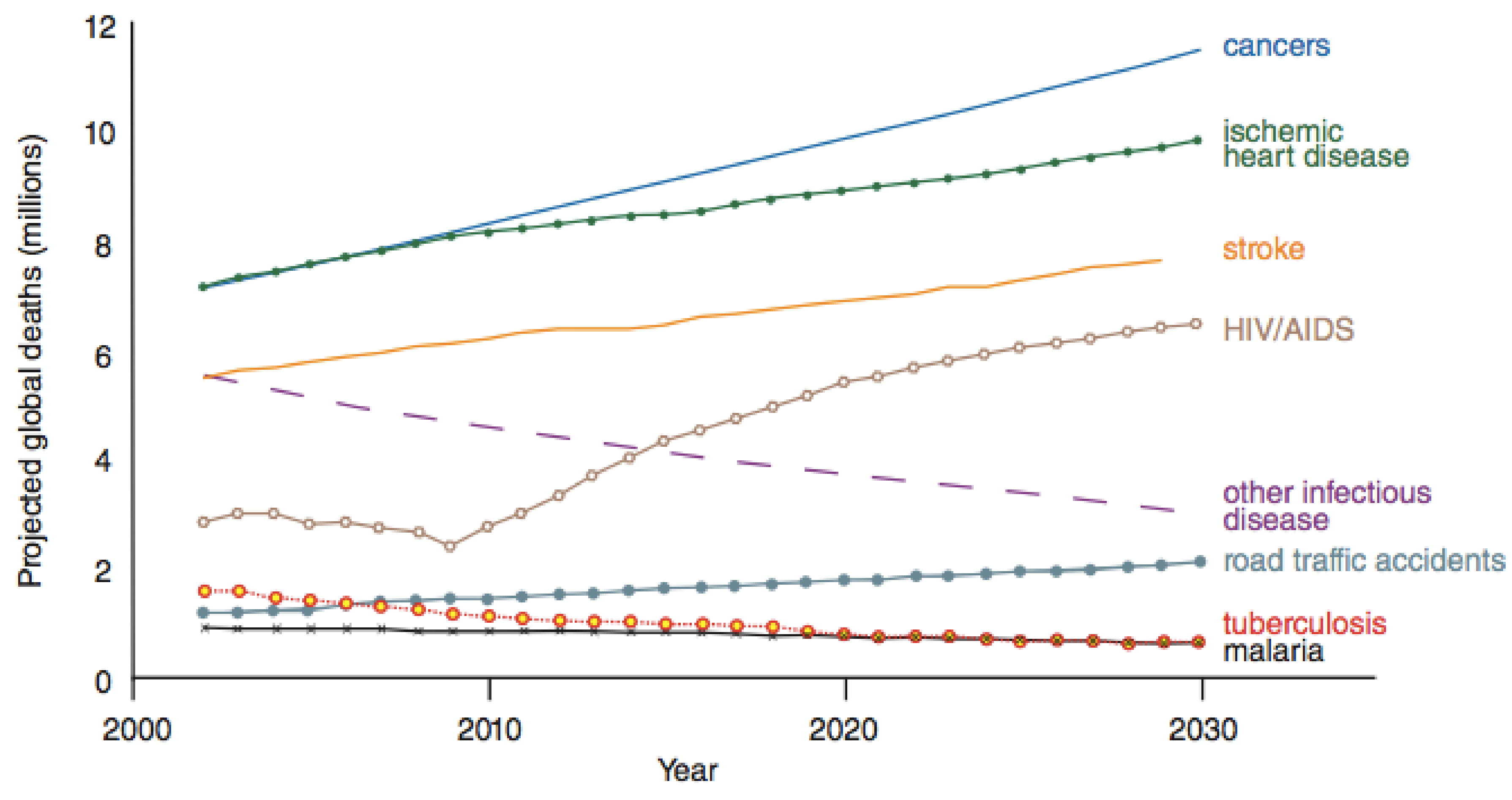


FIGURE 21.2 Projected global deaths for selected causes of death, 2002–2030. Redrawn from the World Health Organization (World Health Statistics 2007, ⊕ http://www.who.int/whosis/whostat2007.pdf). Reproduced with permission from World Health Organization.

Percentage of global disability-adjusted life years (DALY) for various causes

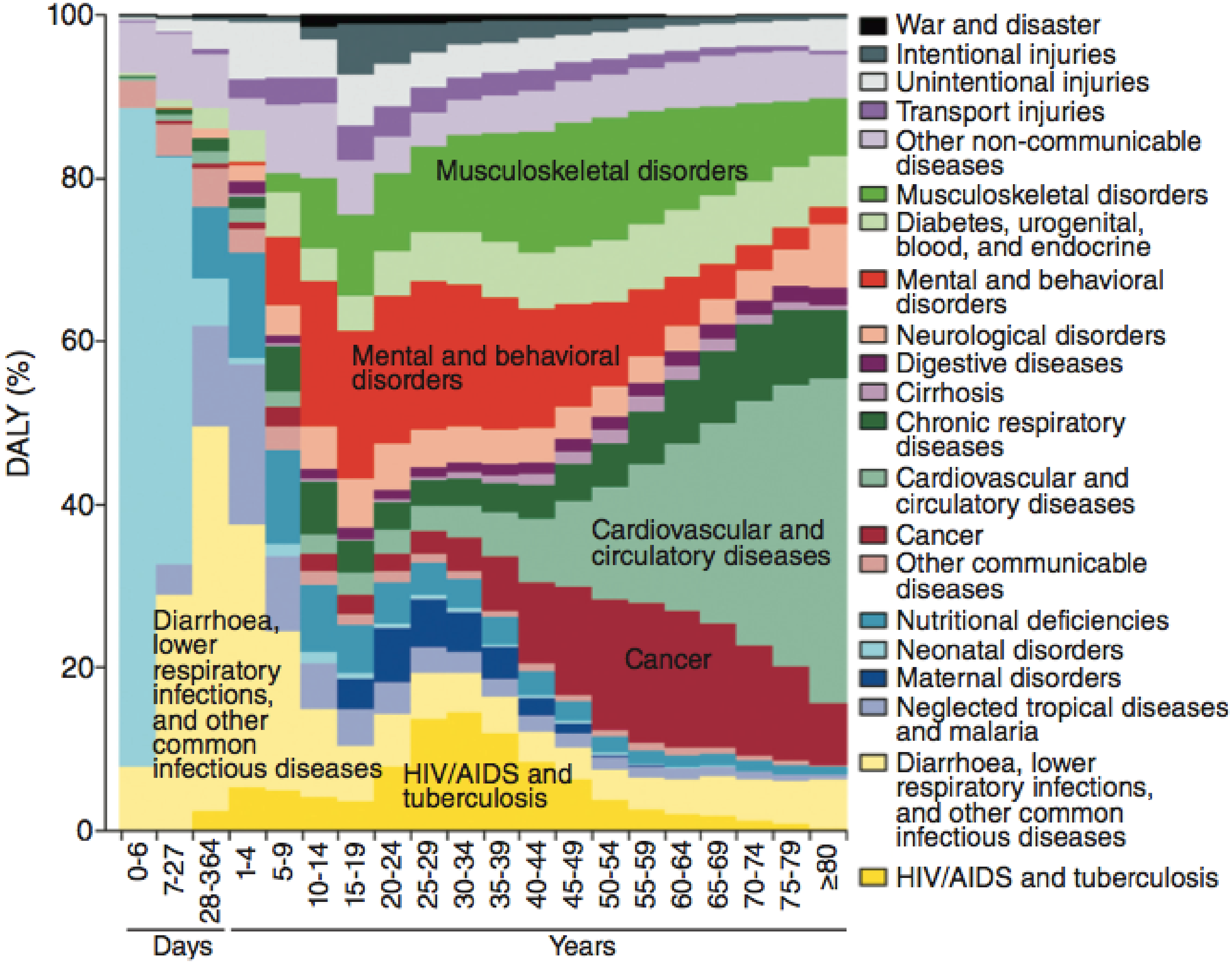


FIGURE 21.3 Percentage of global disability-adjusted life years (DALY) for various causes in 2010. Data are for females; results for males (not shown) are similar. Redrawn from Murray *et al.* (2012). Reproduced with permission from Elsevier.

Classification of disease

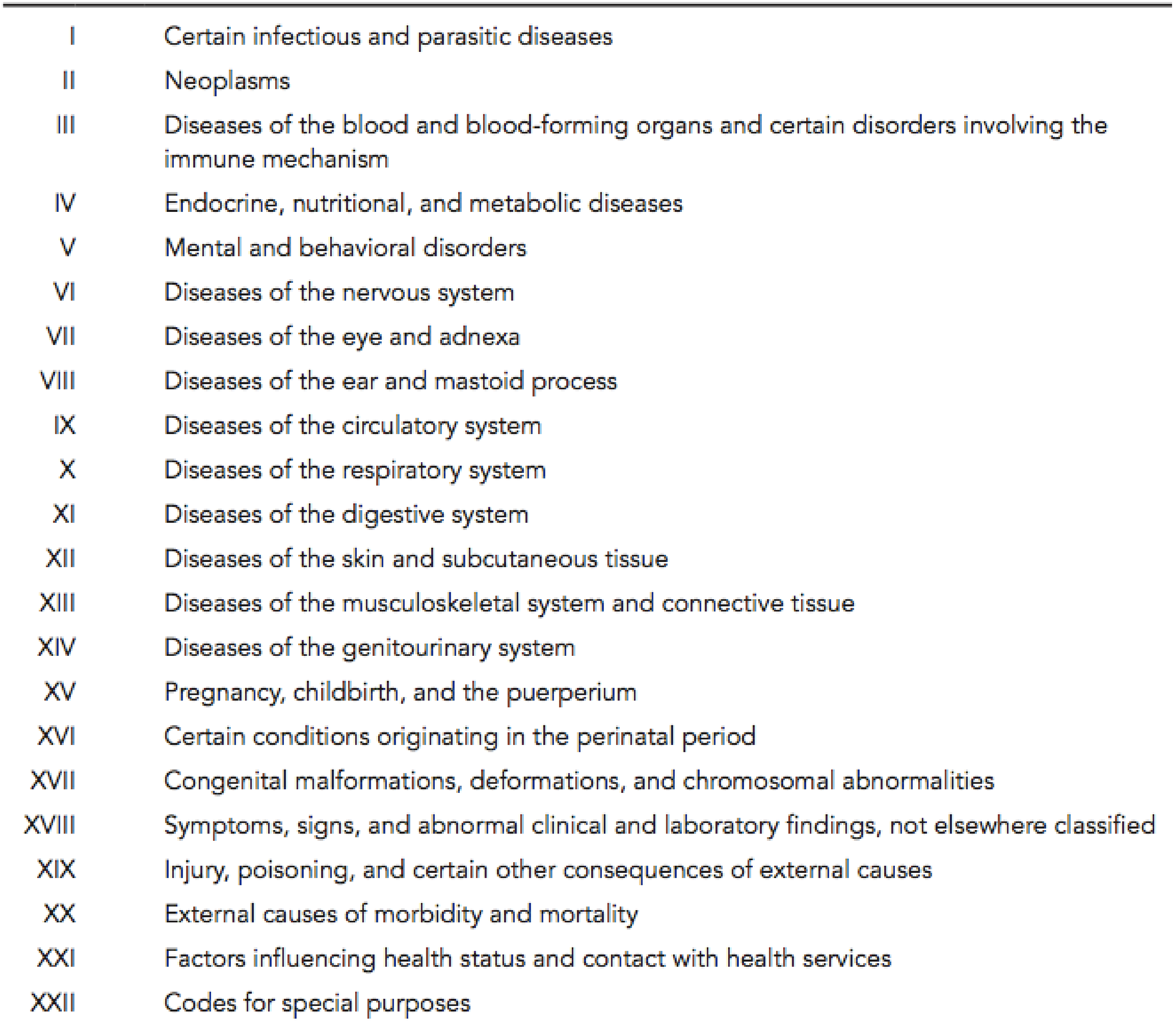
The International Statistical Classification of Diseases and Related Health Problems (ICD) is the main disease classification system used in health care. Examples of categories are:

1. Infectious and parasitic disease
2. Neoplasms (A new growth of abnormal tissue)
3. Endocrine, nutritional, and metabolic diseases…
4. Diseases of the blood and blood-forming organs
5. Mental disorders
6. Diseases of the nervous system and sense organs
7. Diseases of the circulatory system
8. Diseases of the respiratory system
9. Diseases of the digestive system

The **endocrine** system is made up of the pituitary gland, thyroid gland, parathyroid glands, adrenal glands, pancreas, ovaries (in females) and testicles (in males)

See http://www.who.int/whosis/icd10/

ICD classification system (ICD-10)



I	Certain infectious and parasitic diseases
II	Neoplasms
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	Endocrine, nutritional, and metabolic diseases
V	Mental and behavioral disorders
VI	Diseases of the nervous system
VII	Diseases of the eye and adnexa
VIII	Diseases of the ear and mastoid process
IX	Diseases of the circulatory system
X	Diseases of the respiratory system
XI	Diseases of the digestive system
XII	Diseases of the skin and subcutaneous tissue
XIII	Diseases of the musculoskeletal system and connective tissue
XIV	Diseases of the genitourinary system
XV	Pregnancy, childbirth, and the puerperium
XVI	Certain conditions originating in the perinatal period
XVII	Congenital malformations, deformations, and chromosomal abnormalities
XVIII	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified
XIX	Injury, poisoning, and certain other consequences of external causes
XX	External causes of morbidity and mortality
XXI	Factors influencing health status and contact with health services
XXII	Codes for special purposes

Outline

Human genetic disease: a consequence of DNA variation

Categories of disease

Disease databases

Approaches to identifying disease-associated genes and loci

Human disease genes in model organisms

Functional classification of disease genes

Perspective

Categories of disease

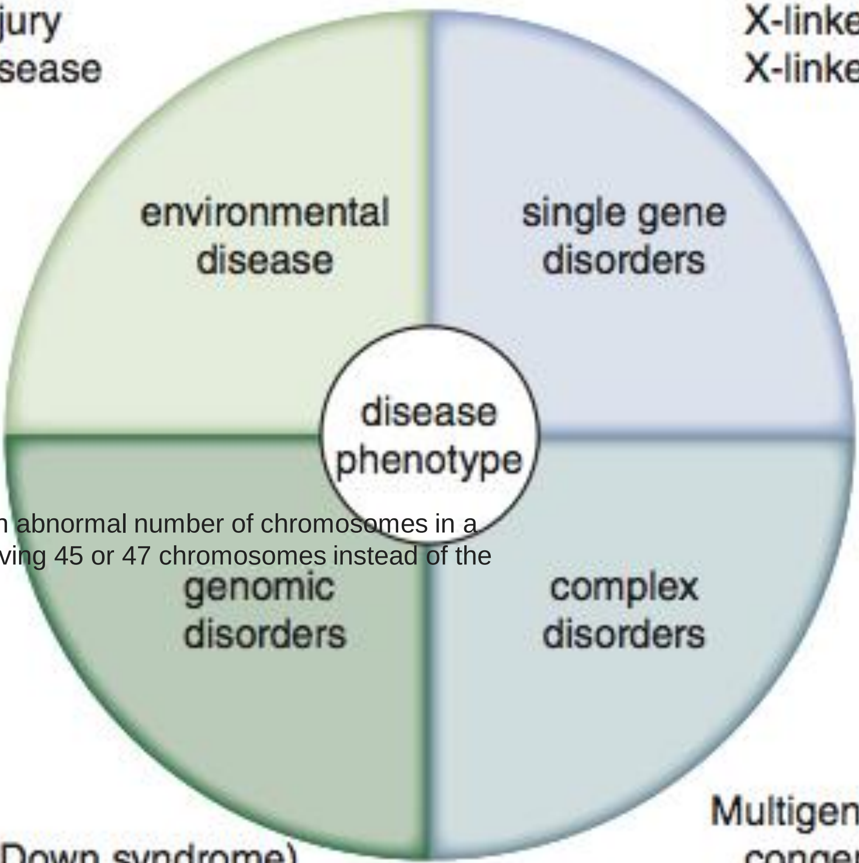
We can consider four main categories of human disease.

- single-gene (monogenic) disease
- complex disease
- genomic disease
- environmental disease

Categorization of disease based on cause

Examples:
- Malnutrition
- Lead poisoning
- Traumatic injury
- Infectious disease

Mendelian disorders	11/1000
autosomal dominant	6/1000
autosomal recessive	3/1000
X-linked recessive	1/1000
X-linked mental retardation	1/1000



Aneuploidy is the presence of an abnormal number of chromosomes in a cell, for example a human cell having 45 or 47 chromosomes instead of the usual 46.

Examples:
- Trisomy 21 (Down syndrome)
- Monosomy
- Segmental aneuploidy
- Microdeletion syndromes
- Microduplication syndromes

Multigenic disorders	~630/1000
congentical anomoalies	30/1000
CNS disorders	100/1000
cardiovascular	500/1000

Central nervous system (CNS)

Categories of disease

Single gene disorders	rare	multigenic
autosomal dominant		
autosomal recessive		
X-linked recessive		
Complex disorders	common	multigenic
congenital anomalies		
CNS		
cardiovascular		
Chromosomal disorders	common	multigenic
Infectious disease	common	multigenic
Environmental disease	common	multigenic

(1) Monogenic (single gene) disorders

Previously, a large distinction was made between monogenic (single gene) and polygenic (complex) disorders. They are now seen to be more on a continuum.

We may define a single-gene disorder as a disorder that is caused primarily by mutation(s) in a single gene. However, as we will see below, all monogenic disorders involve many genes.

A1000 Genomes paper (2010) suggests that "on average, each person is found to carry approximately 250-300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders."

PMID 20981092

(1) Monogenic (single gene) disorders

Autosomal dominant

BRCA1, BRCA2	1:1000
Huntington chorea	1:2,500
Tuberous sclerosis	1:15,000

An autosome is any of the numbered chromosomes, as opposed to the sex chromosomes. Humans have 22 pairs of **autosomes** and one pair of sex chromosomes

Autosomal recessive

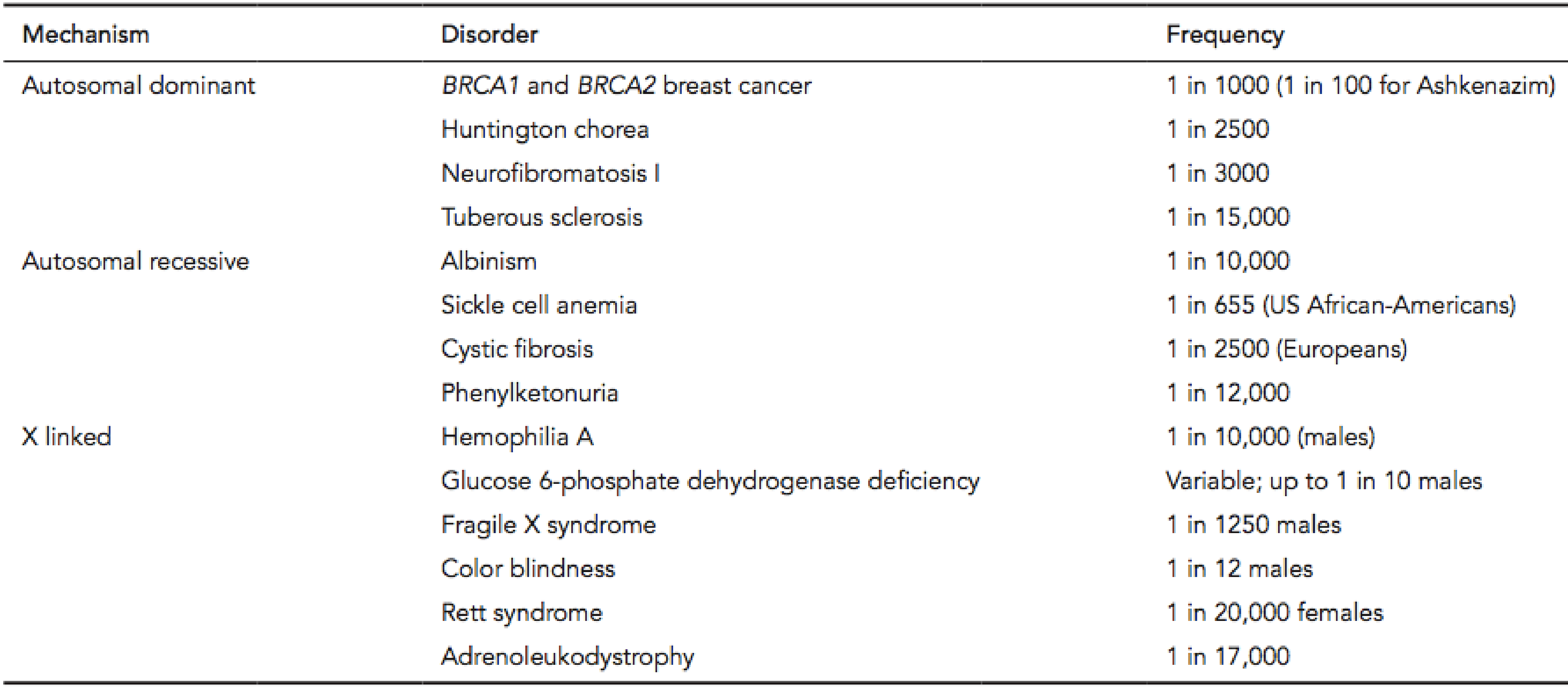
Albinism	1:10,000
Sickle cell anemia	1:655 (U.S. Afr.Am)
Cystic fibrosis	1:2,500 (Europeans)
Phenylketonuria	1:12,000

Albinism is a congenital disorder characterized in humans by the complete or partial absence of pigment in the skin, hair and eyes. **Phenylketonuria** (PKU) is an inborn error of metabolism that results in decreased metabolism of the amino acid phenylalanine. Untreated, PKU can lead to intellectual disability, seizures, behavioral problems, and mental disorders.

X-linked

Hemophilia A	1:10,000 (males)
Rett Syndrome	1:10,000 (females)
Fragile X Syndrome	1:1,250 (males)

Monogenic disorders: examples



Mechanism	Disorder	Frequency
Autosomal dominant	*BRCA1* and *BRCA2* breast cancer	1 in 1000 (1 in 100 for Ashkenazim)
	Huntington chorea	1 in 2500
	Neurofibromatosis I	1 in 3000
	Tuberous sclerosis	1 in 15,000
Autosomal recessive	Albinism	1 in 10,000
	Sickle cell anemia	1 in 655 (US African-Americans)
	Cystic fibrosis	1 in 2500 (Europeans)
	Phenylketonuria	1 in 12,000
X linked	Hemophilia A	1 in 10,000 (males)
	Glucose 6-phosphate dehydrogenase deficiency	Variable; up to 1 in 10 males
	Fragile X syndrome	1 in 1250 males
	Color blindness	1 in 12 males
	Rett syndrome	1 in 20,000 females
	Adrenoleukodystrophy	1 in 17,000

Monogenic (single gene) disorders

Sickle cell anemia is an example of a single gene disorder.

It is caused by mutations in beta globin (HBB). We saw that the E6V mutation is very common.
This mutation causes hemoglobin molecules ($\alpha_2\beta_2$) to aggregate, giving red blood cells a sickled appearance.
(A sickle, bagging hook or reaping-hook.)

This single gene disorder is unusually prevalent because the heterozygous state confers protection to those exposed to the malaria parasite.

You can read Linus Pauling's 1949 article describing the abnormal electrophoretic mobility of HBB on-line at http://profiles.nlm.nih.gov/MM/B/B/R/L/

A monogenic disorder: Rett Syndrome

Rett syndrome (RTT) is another example of a single gene disorder. We will discuss the following aspects:

Clinical presentation
Neurobiology
Gene defect: *MECP2*, a transcriptional repressor (Xq28)
OMIM entry
Locus-specific database entry
Single nucleotide polymorphisms (SNPs)

Rett Syndrome: Clinical Presentation

Normal pre- and perinatal development

Neurocognitive regression

- Deceleration of head and brain growth
- Loss of speech and social skills (autistic)
- Loss of purposeful hand movements
- Truncal ataxia
- Repetitive hand movements
- Seizures

Rett Syndrome: Neurobiology

- Decreased Total Brain Volume
- Reduced Cortical Thickness
- Nigrostriatal Pathology (Brain)
- (https://en.wikipedia.org/wiki/Nigrostriatal_pathway
- Basal Forebrain Cholinergic System

The **cholinergic system** is composed of organized nerve cells that use the neurotransmitter acetylcholine in the transduction of action potentials.

- Glutamatergic Abnormalities

(A **glutamatergic** agent (or drug) is a chemical that directly modulates the excitatory amino acid (**glutamate**/aspartate) system in the body or brain.

- Disruption of Neuronal Markers in olfactory epithelium

Rett Syndrome: Genetics

- Affects only females (~1/10,000)
- X-linked male-lethal? No: mutations arise in father
- "Genetic Lethality"
- >99% of cases are sporadic

•Sporadic: occurring at irregular intervals or only in a few places; scattered or isolated.

- Twins: MZ - 7/8 DZ - 2/13

•**MZ twins** develop when one egg is fertilised by a single sperm and during the first two weeks after conception, the developing embryo splits into two. As a result, two, genetically identical babies develop. **DZ twins** occur when two eggs are released at a single ovulation and are fertilised by two different sperm.

- Rare mother - daughter affected pairs documented
- X exclusion mapping: Xq28

•Exclusion mapping is a technique used to map the location of a gene by successively eliminating regions of the chromosome that cannot contain the gene.

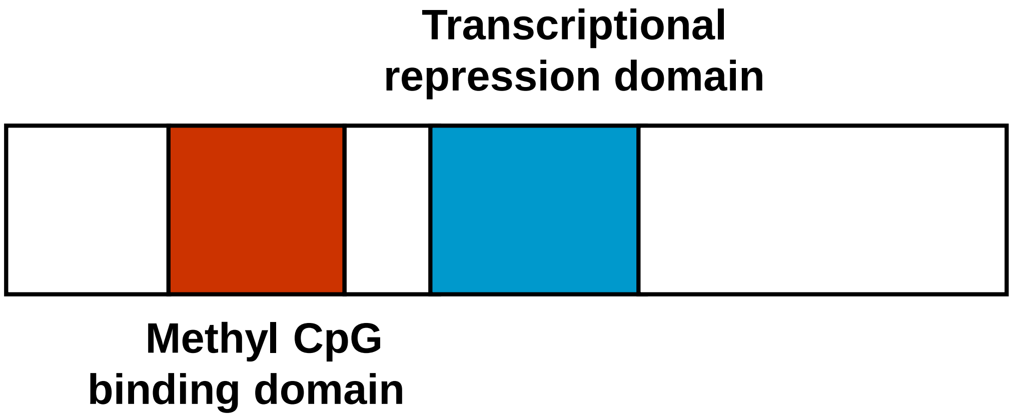
- Linkage analysis: Xq28

•Genetic **linkage analysis** is a powerful tool to detect the chromosomal location of disease genes

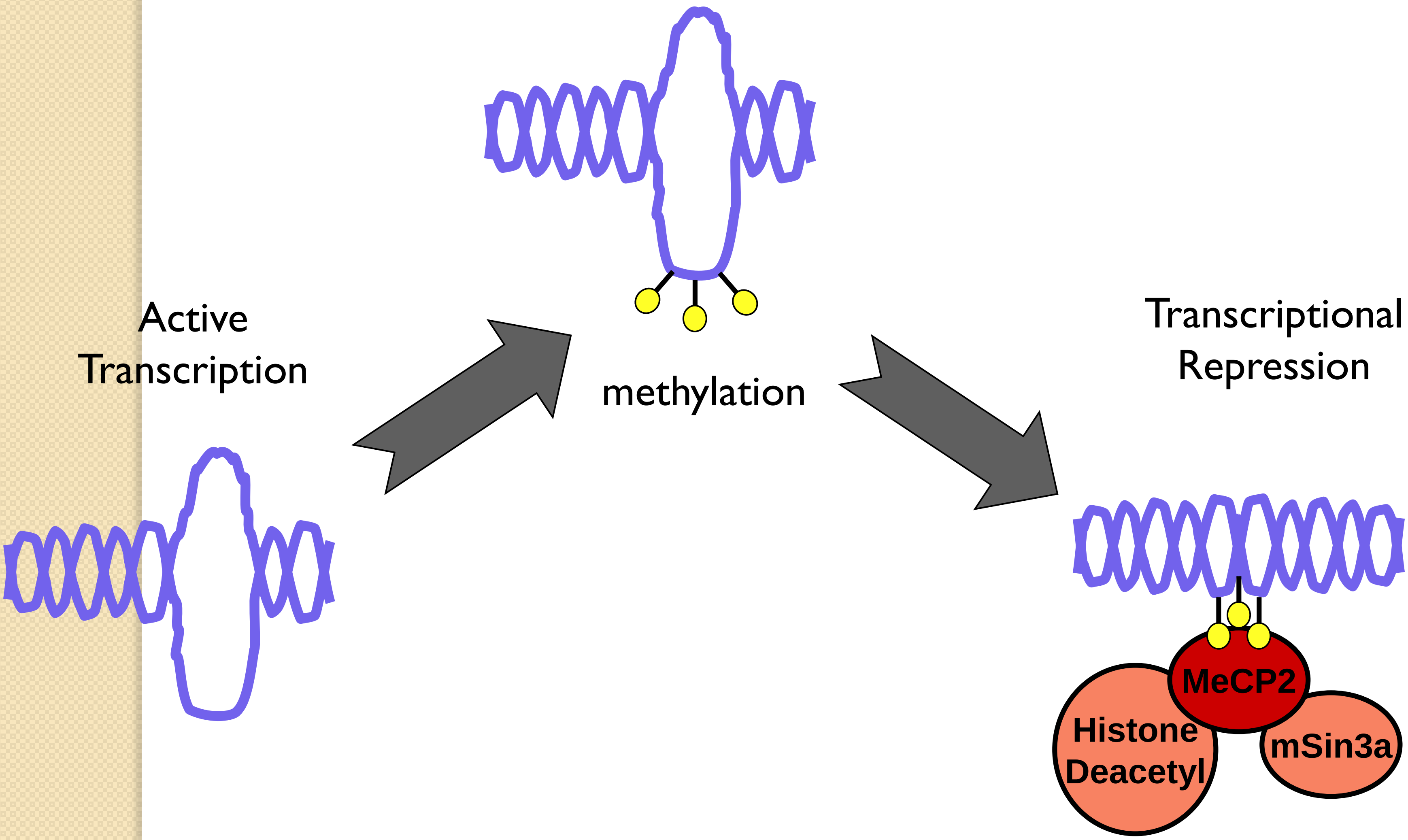
Mutations in *MECP2* cause Rett Syndrome

Rett Syndrome is Caused by Mutations in X-linked *MECP2*, Encoding Methyl-CpG-Binding Protein

R.E. Amir et al. (*Nature Genetics* 1999)



Overview of MeCP2 Function


Active Transcription

methylation

Transcriptional Repression

Disease principles highlighted by RTT

- Phenotype in males (severe neonatal encephalopathy, often fatal) does not resemble that of females

Neonatal encephalopathy (NE), is defined by signs and symptoms of abnormal neurological function in the first few days of life in an infant born at term.

- Females may be spared a more severe phenotype because of random X chromosome inactivation. In all females, each cell chooses to express either the maternal or paternal X chromosome, early in life. Thus RTT females are a mosaic of cells expressing normal and mutated copies of MECP2.

- X-inactivation patterns in females are normally about 50-50. However they may be skewed 99-1, allowing a female to be a carrier. Several females, spared by skewing, have given birth to affected daughters.

(2) Complex disorders

Multiple genes are involved. The combination of mutations in multiple genes define the disease.

Complex diseases are non-Mendelian: they show familial aggregation, but not segregation. This means that they are heritable, but it is not easy to identify the responsible genes in pedigrees (e.g. by linkage analysis).

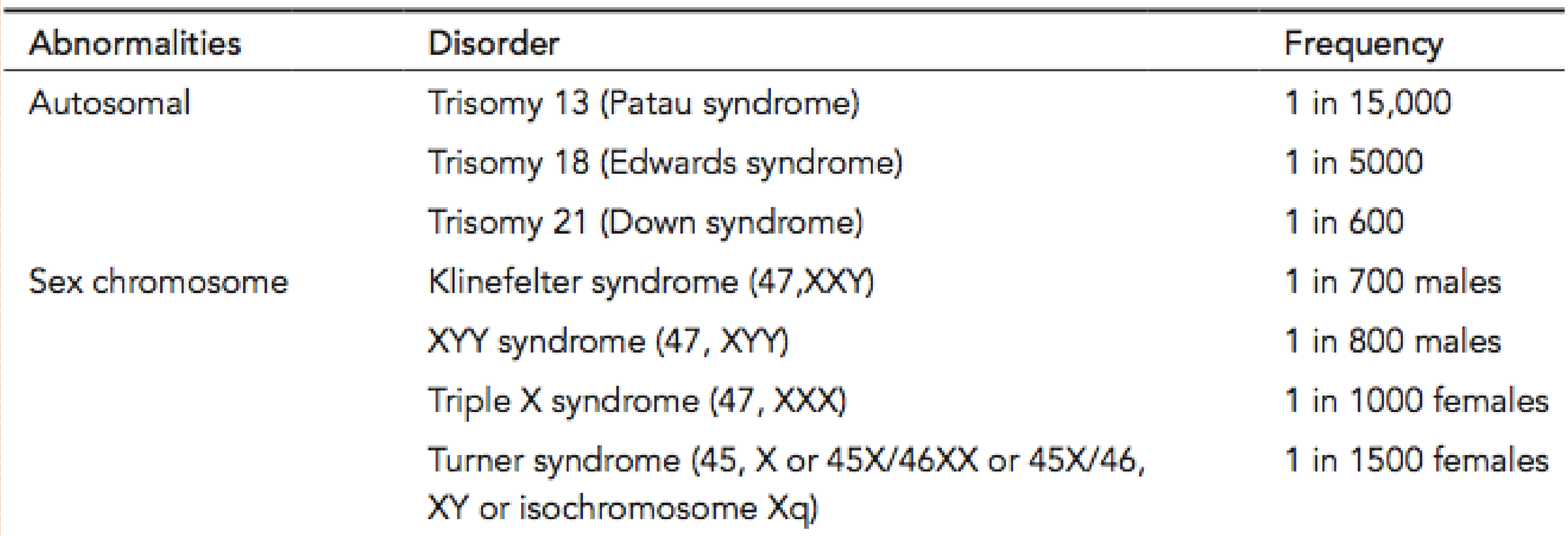
Susceptibility alleles have a high population frequency. Examples are asthma, autism, high blood pressure, obesity, osteoporosis.

Osteoporosis, which literally means porous bone, is a disease in which the density and quality of bone are reduced.

(3) Genomic (chromosomal) disorders

Many diseases are caused by deletions, duplications, or rearrangements of chromosomal DNA. In addition, aneuploidy can occur (having an abnormal number of chromosomes).

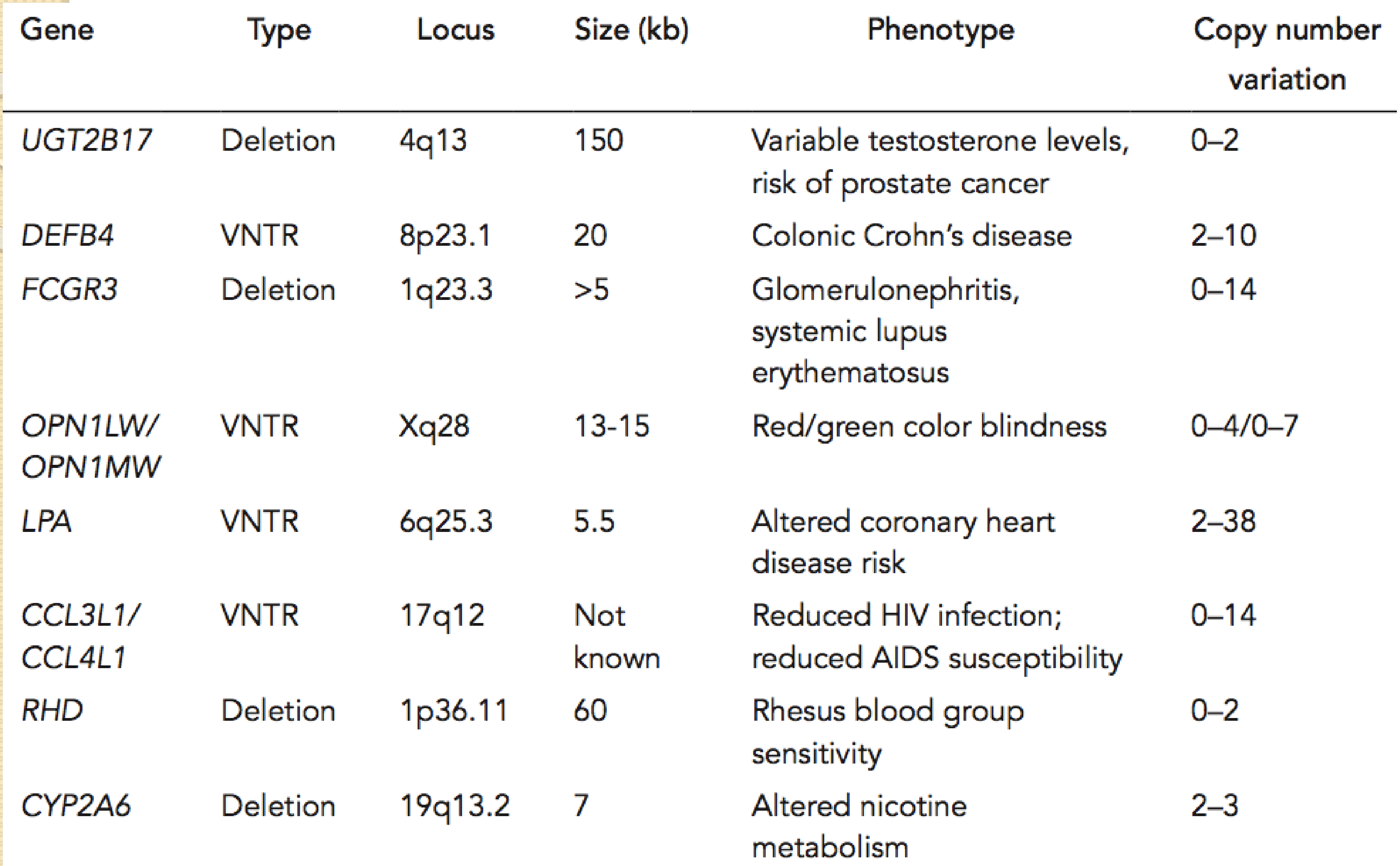
Frequency of chromosomal aneuploidies among liveborn infants



Abnormalities	Disorder	Frequency
Autosomal	Trisomy 13 (Patau syndrome)	1 in 15,000
	Trisomy 18 (Edwards syndrome)	1 in 5000
	Trisomy 21 (Down syndrome)	1 in 600
Sex chromosome	Klinefelter syndrome (47,XXY)	1 in 700 males
	XYY syndrome (47, XYY)	1 in 800 males
	Triple X syndrome (47, XXX)	1 in 1000 females
	Turner syndrome (45, X or 45X/46XX or 45X/46, XY or isochromosome Xq)	1 in 1500 females

Source: Beaudet et al. (2001) with permission from McGraw Hill.

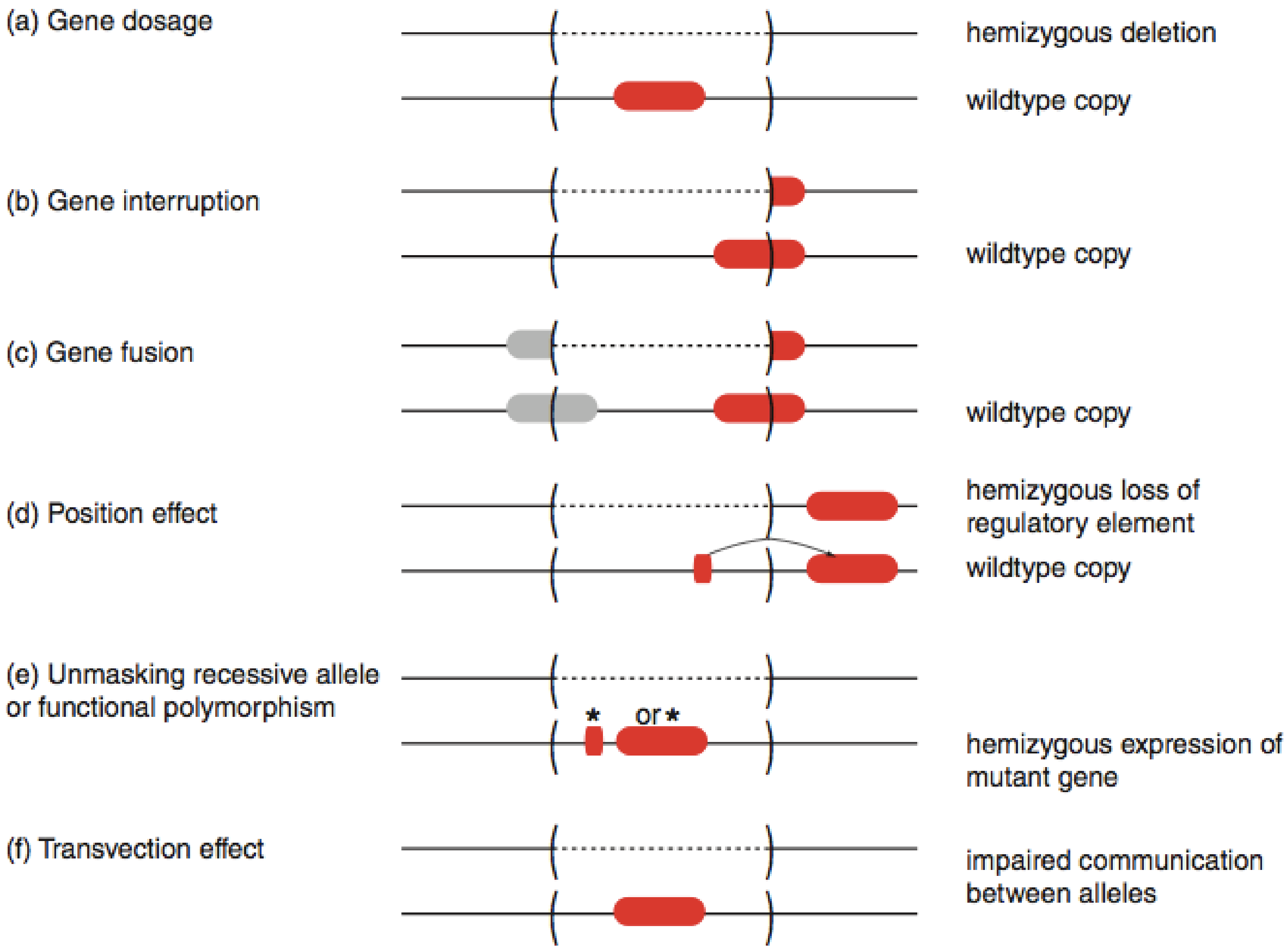
Common structural polymorphisms and disease



Gene	Type	Locus	Size (kb)	Phenotype	Copy number variation
UGT2B17	Deletion	4q13	150	Variable testosterone levels, risk of prostate cancer	0–2
DEFB4	VNTR	8p23.1	20	Colonic Crohn's disease	2–10
FCGR3	Deletion	1q23.3	>5	Glomerulonephritis, systemic lupus erythematosus	0–14
OPN1LW/ OPN1MW	VNTR	Xq28	13-15	Red/green color blindness	0–4/0–7
LPA	VNTR	6q25.3	5.5	Altered coronary heart disease risk	2–38
CCL3L1/ CCL4L1	VNTR	17q12	Not known	Reduced HIV infection; reduced AIDS susceptibility	0–14
RHD	Deletion	1p36.11	60	Rhesus blood group sensitivity	0–2
CYP2A6	Deletion	19q13.2	7	Altered nicotine metabolism	2–3

Source: Human Genome Structural Variation Working Group (2007). Reproduced with permission from Macmillan Publishers.

Models for the molecular mechanisms of genomic disorders



Transvection is an epigenetic phenomenon that results from an interaction between an allele on one chromosome and the corresponding allele on the homologous chromosome.

Categories of disease: (4) environmental

Example:

Lead poisoning is an environmental disease. It is common (about 9% of children have high blood levels).

But two children exposed to the same dose of lead may have entirely different phenotypes.

This susceptibility has a genetic basis.

Conclusion: genes affect susceptibility to environmental insults, and infectious disease. Even single-gene disorders involve many genes in their phenotypic expression.

Disease and genetic background

Individuals from particular geographic origins may have increased risk for disease:

Tay–Sachs disease, becomes apparent around three to six months of age with the baby losing the ability to turn over, sit, or crawl.

- Tay-Sachs disease is prevalent among Ashkenazi Jews.
- About 8% of the African-American population are carriers of a mutant *HBB* gene.
- Males rather than females are susceptible to Alport disease, male pattern baldness, and prostate cancer.
- Cystic fibrosis affects ~30,000 people in the United States with ~12 million carriers, and is the most common fatal genetic disease in that country. While it affects all groups, Caucasians of northern European ancestry are particularly susceptible.

Other categories of disease: Organellar

Diseases can be classified based on the affected organelle (or cell type or organ).

Mitochondria

Over 100 disease-causing mutations identified. The next slide shows a morbid map of the mitochondrial genome.

Peroxisomes

Mutations affect either perixosome function or peroxisome biogenesis; yeast provide a model

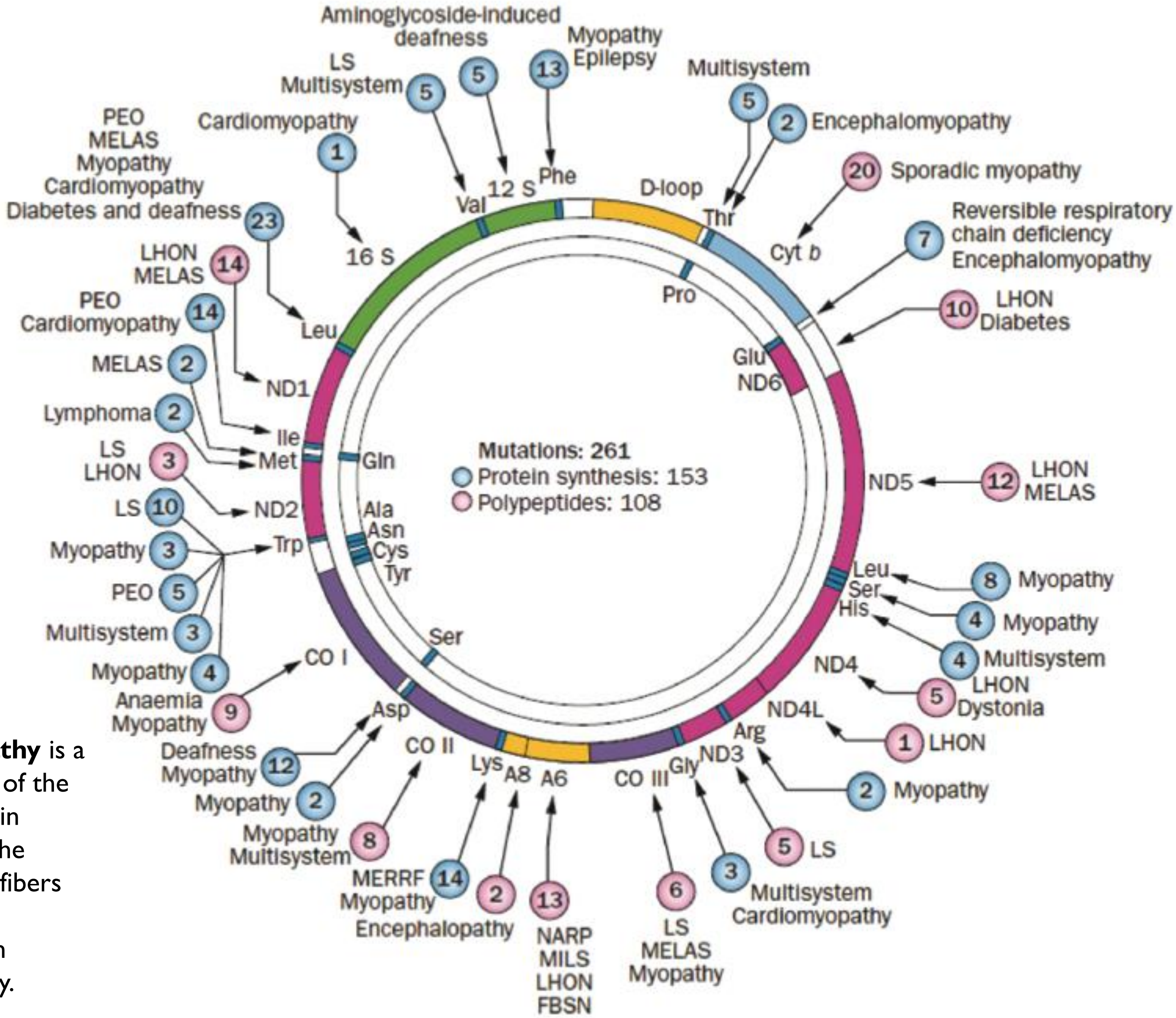
Peroxisomes are small, membrane-enclosed organelles that contain enzymes involved in a variety of metabolic reactions, including several aspects of energy metabolism.

Lysosomes

Many lysosomal storage diseases

A **lysosome** is a membrane-bound cell organelle that contains digestive enzymes. **Lysosomes** are involved with various cell processes. They break down excess or worn-out cell parts. They may be used to destroy invading viruses and bacteria.

Morbidity map of the human mitochondrial genome

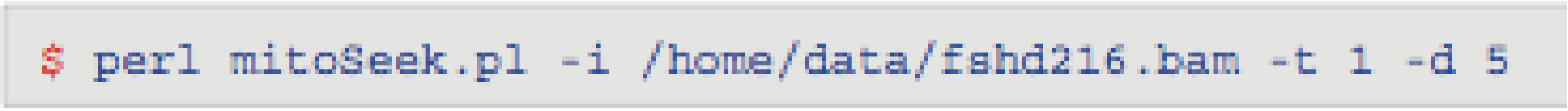


Myopathy is a disease of the muscle in which the muscle fibers do not function properly.

Colored sections represent protein-coding genes. *Source:* DiMauro *et al*. (2013).

Mitoseek: assess mitochondrial variation from whole exome (or genome) sequence data

The MitoSeek program allows you to input a BAM file. Mitochondrial DNA is so abundant that it is often incidentally sequenced with high coverage. It assembles the ~16.5 kb mitochondrial genome and reports variation such **as heteroplasmy**.



```
$ perl mitoSeek.pl -i /home/data/fshd216.bam -t 1 -d 5
```

Heteroplasmy is the presence of more than one type of organellar genome (mitochondrial DNA or plastid DNA) within a cell or individual. It is an important factor in considering the severity of mitochondrial diseases.

Somatic mosaic disease

Mosaicism is the occurrence of genetically distinct populations of cells within an organism (derived from a single zygote, to distinguish it from chimerism).

Mosaicism: the property or state of being composed of cells of two genetically different types.

Chimerism is a condition whereby a person has not one but two complete genomes (sets of DNA) in their body

Genetic changes may involve somatic cells such as skin or liver (somatic mosaicism), or they may involve germline cells (germline mosaicism, also called gonadal mosaicism).

Postzygotic, somatic, mosaic mutations have indeed been identified for diseases including the McCune- Albright syndrome (*GNAS* mutations), the Proteus syndrome (*AKT* mutations), and Sturge-Weber syndrome (mutations in *GNAQ*).

Sturge-Weber syndrome (SWS) is a neurological **disorder** marked by a distinctive port-wine stain on the forehead, scalp, or around the eye.

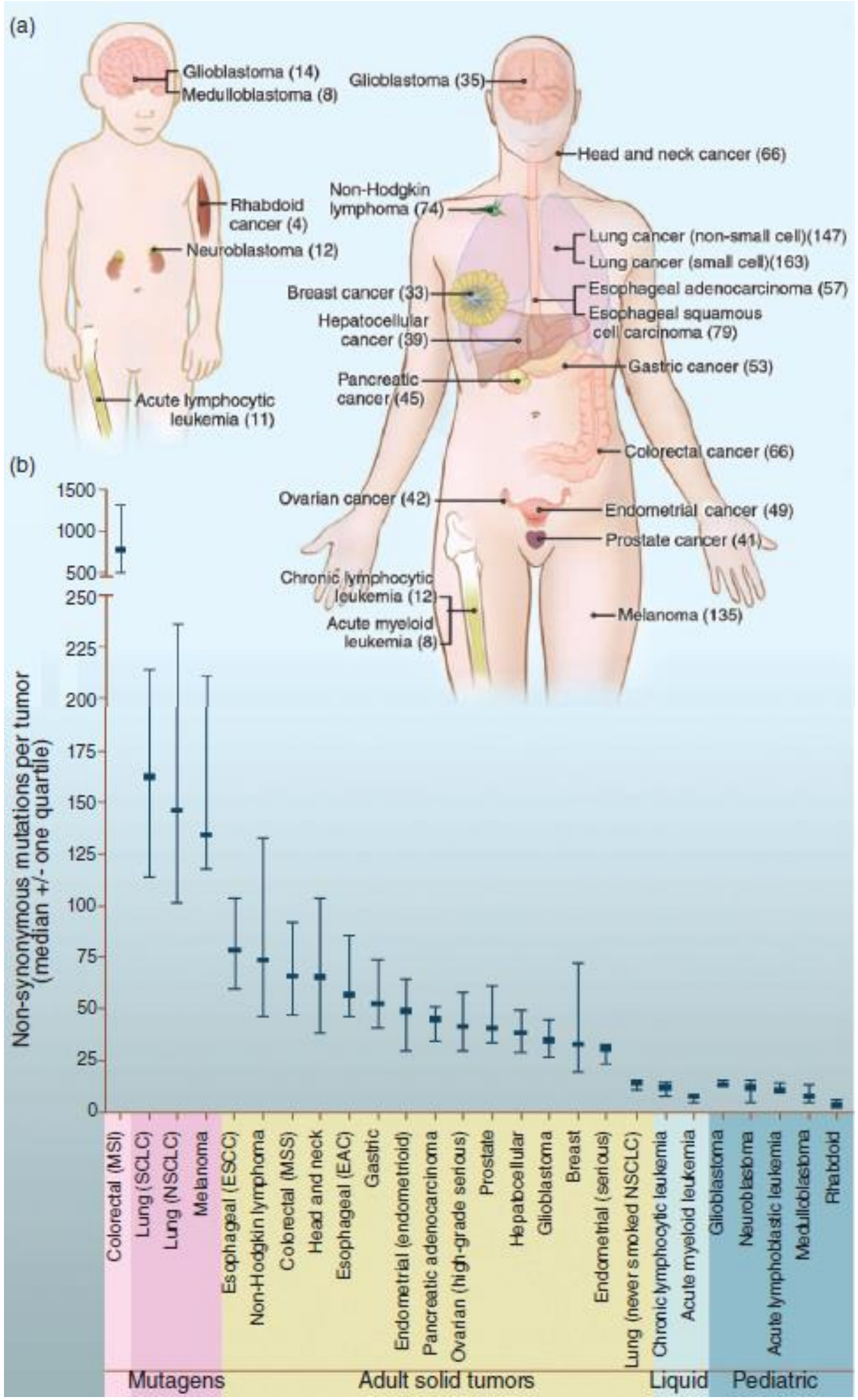
Cancer: a somatic mosaic disease

- Cancer is a somatic mosaic disease, arising from a clone having somatic mutations and leading to malignant transformation.
- Cancer occurs when **DNA mutations confer selective advantage to cells that proliferate.**
- Knudson (1971) introduced a two-hit hypothesis of cancer, suggesting that for dominantly inherited retinoblastoma one mutation is inherited through the germ cells while a second somatic mutation occurs; for a nonhereditary form of cancer two somatic mutations occur.
- There are six hallmarks of cancer, described by Hanahan and Weinberg (2011): proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, induction of angiogenesis, and inactivating invasion and metastasis.
- **Angiogenesis** is the formation of new blood vessels

Cancer: a somatic mosaic disease

- COSMIC (catalogue of somatic mutations in cancer) includes information on ~1 million cancer samples, >1.6 million mutations, and various types of mutations (fusions, genomic rearrangements, and copy number variants).
- >200 types of cancer and many disease mechanisms
- The landscape of cancer includes two types of mutations.
- “Driver” mutations confer a selective growth advantage to cells, are implicated as causing the neoplastic process, and are positively selected for during tumorogenesis.
- “Passenger” mutations are retained by chance but confer no selective advantage and do not contribute to oncogenesis.

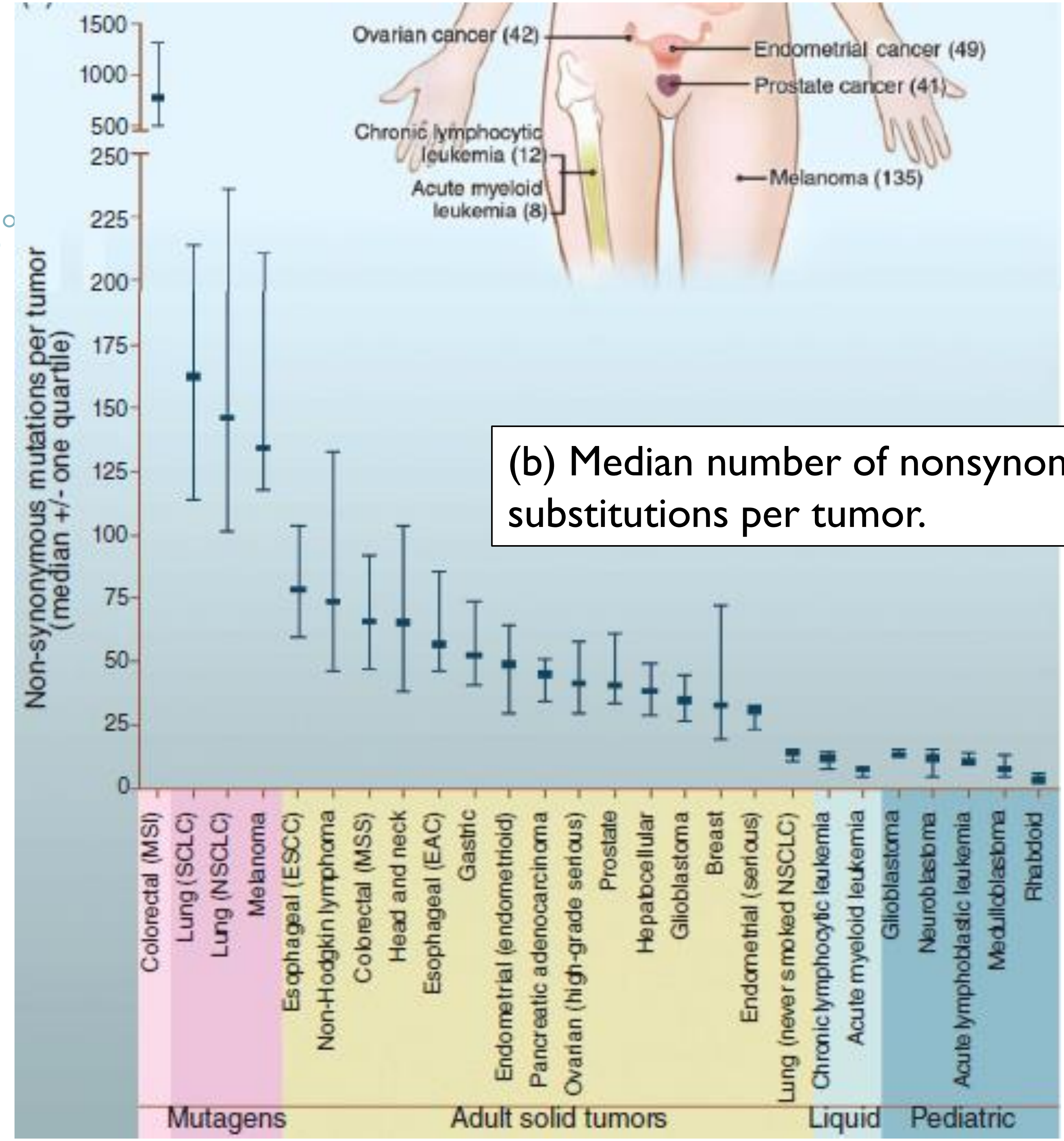
Somatic mutations in representative human cancers, based on genome-wide sequencing studies



(a) The genomes of adult (right) and pediatric (left) cancers. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al*. (2013).

(b) Median number of nonsynonymous substitutions per tumor.

A nonsynonymous substitution is a nucleotide mutation that alters the amino acid sequence of a protein.



(b) Median number of nonsynonymous substitutions per tumor.

Outline

Human genetic disease: a consequence of DNA variation

Categories of disease

Disease databases

Approaches to identifying disease-associated genes and loci

Human disease genes in model organisms

Functional classification of disease genes

Perspective

The principal disease database: OMIM

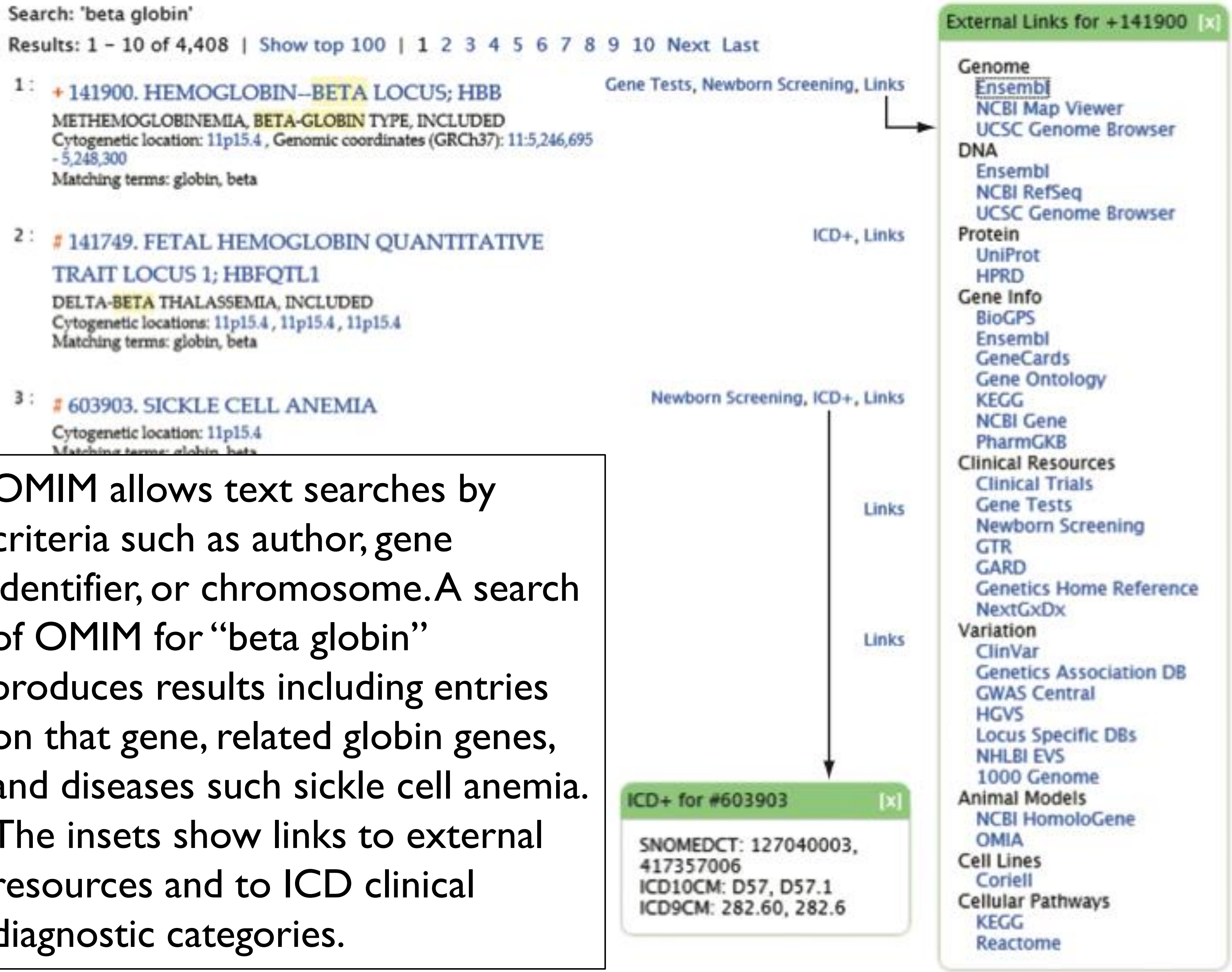
Online Mendelian Inheritance in Man (OMIM) is a comprehensive database for human genes and genetic disorders, with a focus on monogenic disorders.

It was started as MIM by Victor McKusick at JHU (1966). Ada Hamosh currently directs OMIM.

OMIM went online at NCBI in 1995. It is integrated with Entrez, NCBI Gene, Map View, and PubMed.

OMIM has a focus on Mendelian disorders.
There are few entries on chromosomal diseases.

Online Mendelian Inheritance in Man (OMIM)



Search: 'beta globin'

Results: 1 – 10 of 4,408 | Show top 100 | 1 2 3 4 5 6 7 8 9 10 Next Last

1 : + 141900. HEMOGLOBIN--BETA LOCUS; HBB — Gene Tests, Newborn Screening, Links

METHEMOGLOBINEMIA, BETA-GLOBIN TYPE, INCLUDED
Cytogenetic location: 11p15.4 , Genomic coordinates (GRCh37): 11:5,246,695 - 5,248,300
Matching terms: globin, beta

2 : # 141749. FETAL HEMOGLOBIN QUANTITATIVE TRAIT LOCUS 1; HBFQTL1 — ICD+, Links

DELTA-BETA THALASSEMIA, INCLUDED
Cytogenetic locations: 11p15.4 , 11p15.4 , 11p15.4
Matching terms: globin, beta

3 : # 603903. SICKLE CELL ANEMIA — Newborn Screening, ICD+, Links

Cytogenetic location: 11p15.4
Matching terms: globin, beta

Links

Links

ICD+ for #603903 [x]

SNOMEDCT: 127040003, 417357006
ICD10CM: D57, D57.1
ICD9CM: 282.60, 282.6

External Links for +141900 [x]

Genome
- Ensembl
- NCBI Map Viewer
- UCSC Genome Browser

DNA
- Ensembl
- NCBI RefSeq
- UCSC Genome Browser

Protein
- UniProt
- HPRD

Gene Info
- BioGPS
- Ensembl
- GeneCards
- Gene Ontology
- KEGG
- NCBI Gene
- PharmGKB

Clinical Resources
- Clinical Trials
- Gene Tests
- Newborn Screening
- GTR
- GARD
- Genetics Home Reference
- NextGxDx

Variation
- ClinVar
- Genetics Association DB
- GWAS Central
- HGVS
- Locus Specific DBs
- NHLBI EVS
- 1000 Genome

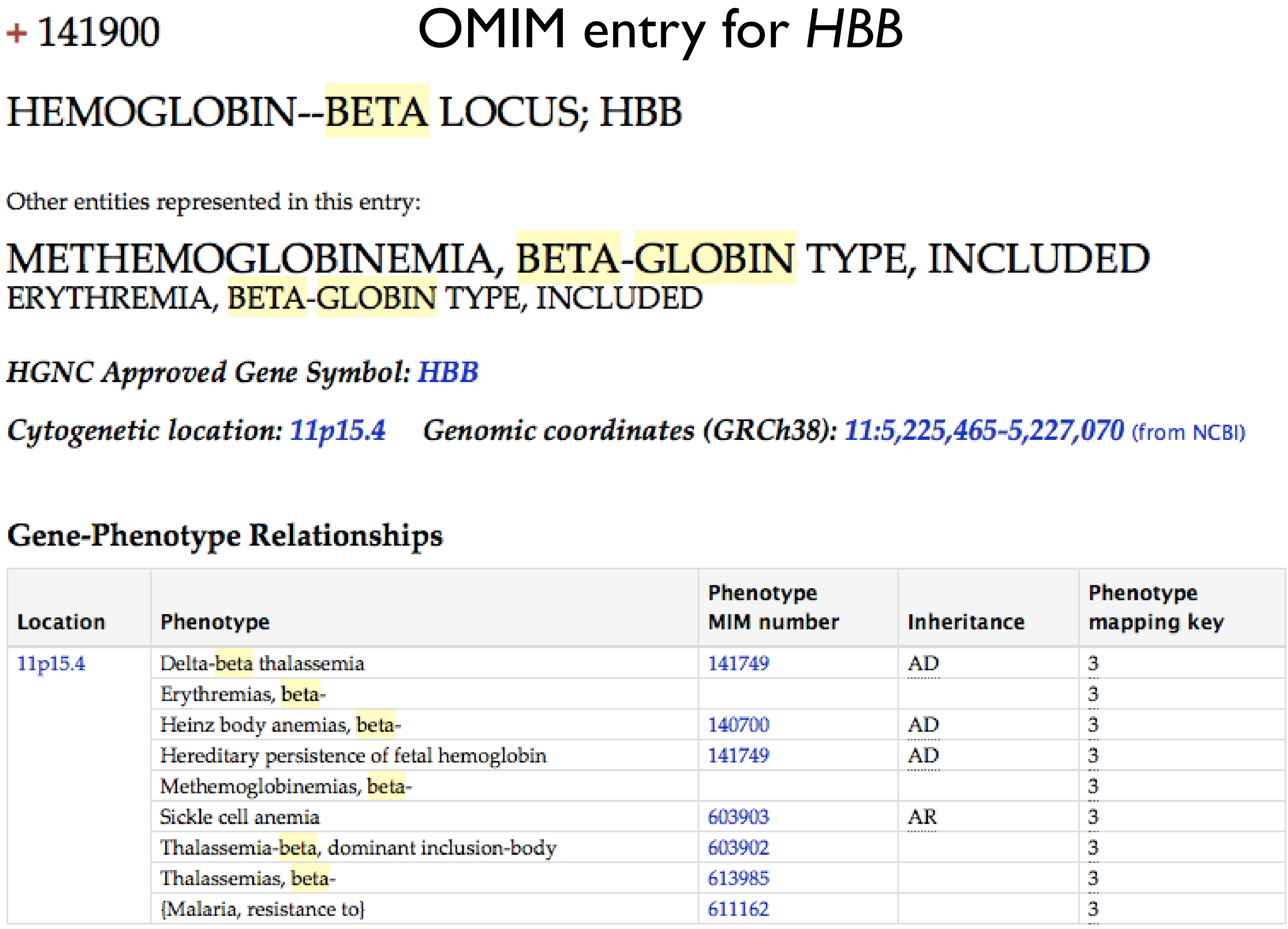
Animal Models
- NCBI HomoloGene
- OMIA

Cell Lines
- Coriell

Cellular Pathways
- KEGG
- Reactome

OMIM allows text searches by criteria such as author, gene identifier, or chromosome. A search of OMIM for "beta globin" produces results including entries on that gene, related globin genes, and diseases such sickle cell anemia. The insets show links to external resources and to ICD clinical diagnostic categories.

OMIM entry for *HBB*



+ 141900

HEMOGLOBIN--BETA LOCUS; HBB

Other entities represented in this entry:

METHEMOGLOBINEMIA, BETA-GLOBIN TYPE, INCLUDED
ERYTHREMIA, BETA-GLOBIN TYPE, INCLUDED

HGNC Approved Gene Symbol: *HBB*

Cytogenetic location: *11p15.4* *Genomic coordinates (GRCh38):* *11:5,225,465-5,227,070* (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
11p15.4	Delta-beta thalassemia	141749	AD	3
	Erythremias, beta-			3
	Heinz body anemias, beta-	140700	AD	3
	Hereditary persistence of fetal hemoglobin	141749	AD	3
	Methemoglobinemias, beta-			3
	Sickle cell anemia	603903	AR	3
	Thalassemia-beta, dominant inclusion-body	603902		3
	Thalassemias, beta-	613985		3
	{Malaria, resistance to}	611162		3

The OMIM entry for beta globin includes the OMIM identifier (+141900) and a variety of information such as clinical features, a description of animal models, and allelic variants.

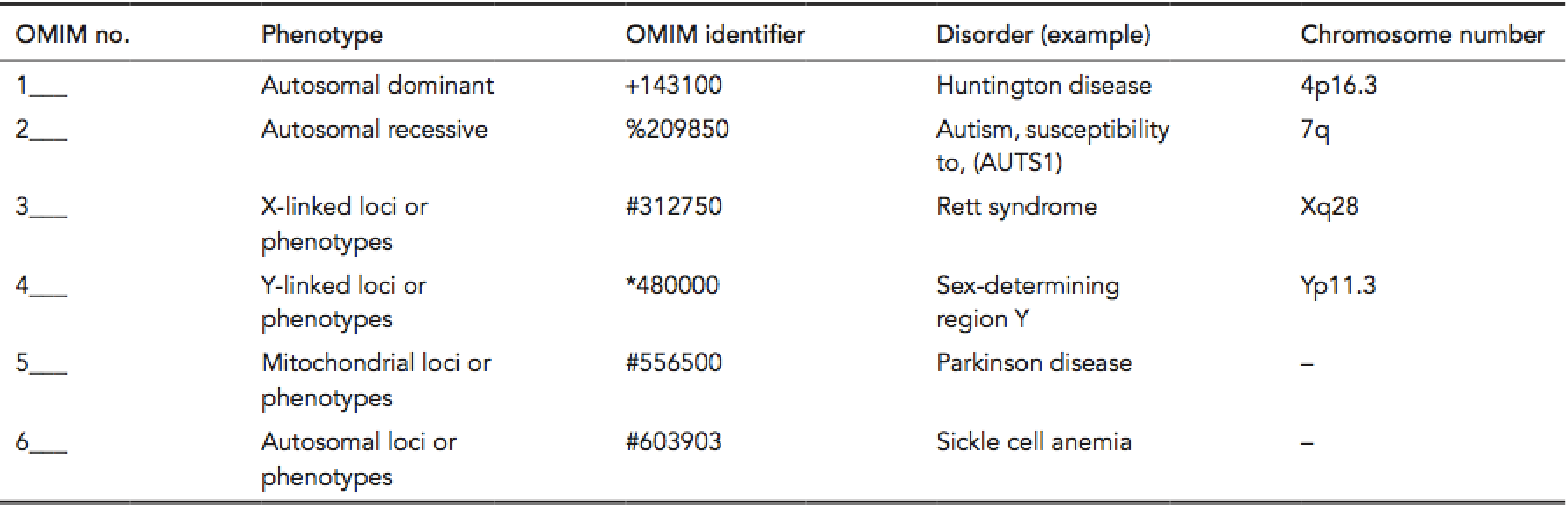
OMIM allelic variants

Most OMIM allelic variants represent disease-causing mutations.

They are selected for OMIM based on criteria such as historical importance, high population frequency, or involving an unusual pathogenetic mechanism.

Some allelic variants simply represent polymorphisms.

OMIM numbering system



OMIM no.	Phenotype	OMIM identifier	Disorder (example)	Chromosome number
1___	Autosomal dominant	+143100	Huntington disease	4p16.3
2___	Autosomal recessive	%209850	Autism, susceptibility to, (AUTS1)	7q
3___	X-linked loci or phenotypes	#312750	Rett syndrome	Xq28
4___	Y-linked loci or phenotypes	*480000	Sex-determining region Y	Yp11.3
5___	Mitochondrial loci or phenotypes	#556500	Parkinson disease	–
6___	Autosomal loci or phenotypes	#603903	Sickle cell anemia	–

OMIM number beginning 1 or 2 implies it entered the database before May 1994; OMIM number beginning 6 implies it was created after May 1994; + indicates a gene of known sequence and a phenotype; % indicates a confirmed Mendelian phenotype (or phenotypic locus) for which the underlying molecular basis is not known; # indicates a descriptive entry (usually of phenotype); * preceding entry indicates a gene of known sequence.

Other central mutation databases

GeneCards (Weizmann)

--collects and integrates information from several dozen independent databases such as OMIM, GenBank, UniGene, Ensembl, MIPS.

--visit http://www.genecards.org/

Human Gene Mutation Database (HGMD)

--major mutation database

--commercial access (fee-based)

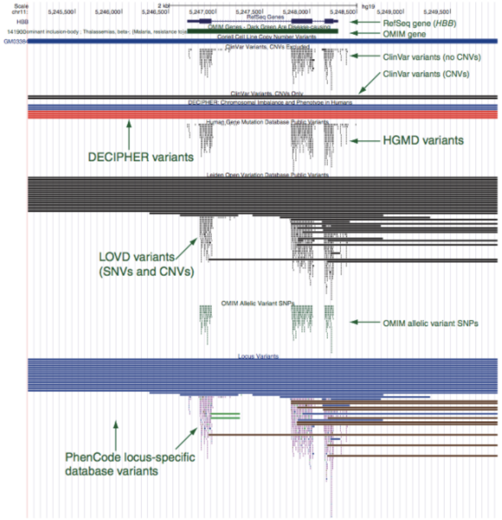
--visit http://www.hgmd.cf.ac.uk/

ClinVar (NCBI)

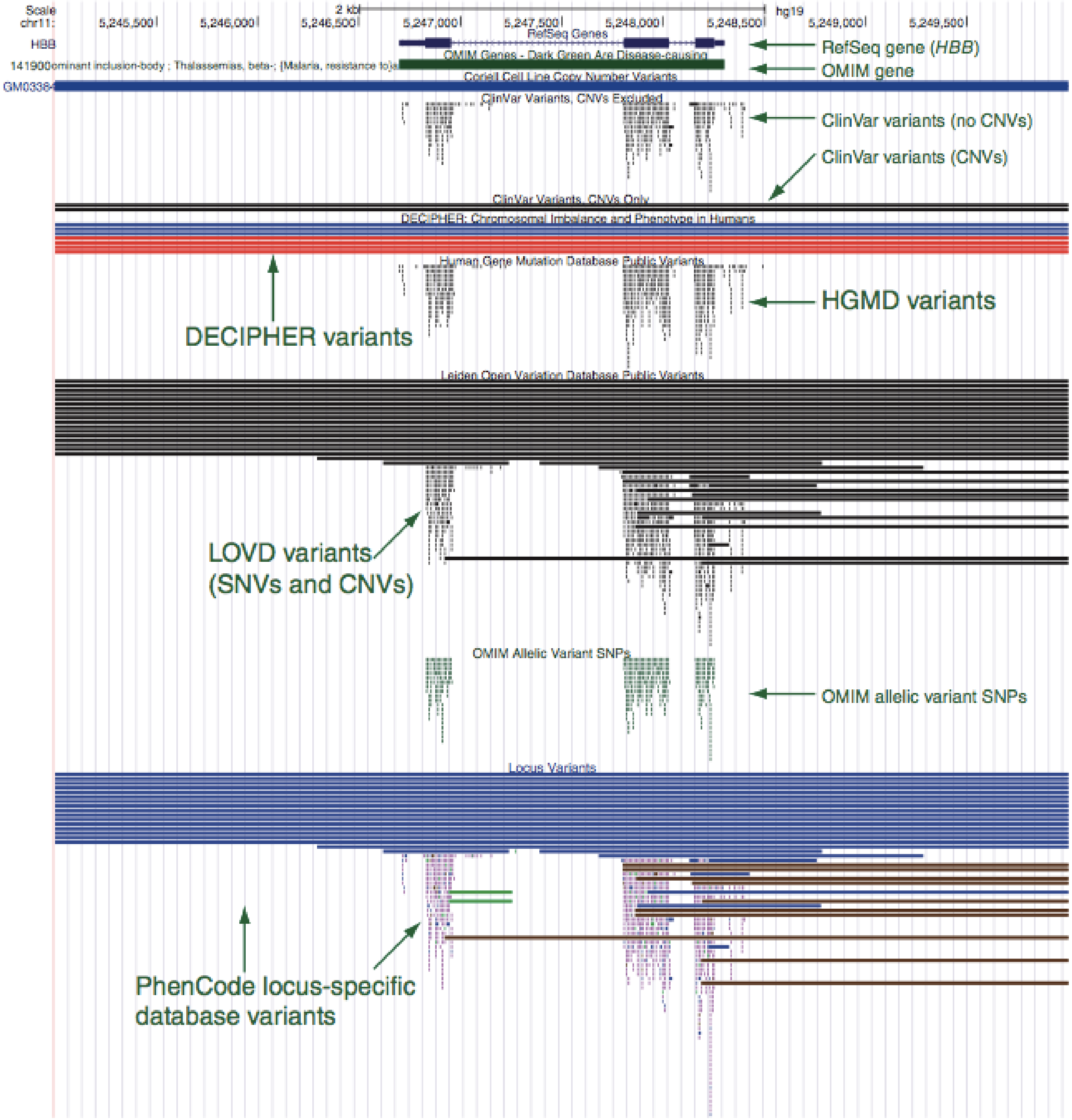
- The ClinVar database provides data on human variants and their relationship to disease.
- It links to the NIH Genetic Testing Registry (GTR), MedGen, Gene, OMIM, and PubMed.
- GTR centralizes genetic test information.
- MedGen organizes human medical genetics information, for example providing several hundred entries on medical conditions relevant to a query for hemoglobin.

Human disease resources:

UCSC Genome Browser includes tracks to display data from disease databases


Scale
chr11:
2 kb
hg19
5,245,500
5,246,000
5,246,500
5,247,000
5,247,500
5,248,000
5,248,500
5,249,000
5,249,500
RefSeq Genes
HBB
RefSeq gene (*HBB*)
OMIM Genes - Dark Green Are Disease-causing
141900 ominant inclusion-body ; Thalassemias, beta-; {Malaria, resistance to}
OMIM gene
Coriell Cell Line Copy Number Variants
GM03384
ClinVar Variants, CNVs Excluded
ClinVar variants (no CNVs)
ClinVar variants (CNVs)
ClinVar Variants, CNVs Only
DECIPHER: Chromosomal Imbalance and Phenotype in Humans
DECIPHER variants
Human Gene Mutation Database Public Variants
HGMD variants
Leiden Open Variation Database Public Variants
LOVD variants
(SNVs and CNVs)
OMIM Allelic Variant SNPs
OMIM allelic variant SNPs
Locus Variants
PhenCode locus-specific
database variants

A 5000 base pair region is shown (chr11:5,245,001–5,250,000) including *HBB* as shown by the RefSeq Genes track. The OMIM entry is shaded dark green, indicating it has disease-causing variants. HGMD, ClinVar, OMIM, and PhenCode entries are displayed at squish density, with similar profiles and with the majority of variants overlapping the exons (thick blue rectangles of the RefSeq track). Copy number variants (CNVs) are displayed in a separate ClinVar track, in the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) database, in the Coriell track displaying cell lines (and/or genomic DNA samples) available to the research community, and in the Leiden Open Variation Database (LOVD) which includes both single-nucleotide variants (SNVs) and CNVs.


Scale
chr11:
2 kb
hg19
5,245,500
5,246,000
5,246,500
5,247,000
5,247,500
5,248,000
5,248,500
5,249,000
5,249,500
RefSeq Genes
HBB
OMIM Genes - Dark Green Are Disease-causing
141900ominant inclusion-body ; Thalassemias, beta-; {Malaria, resistance to}a
Coriell Cell Line Copy Number Variants
GM03364
ClinVar Variants, CNVs Excluded
ClinVar Variants, CNVs Only
DECIPHER: Chromosomal Imbalance and Phenotype in Humans
Human Gene Mutation Database Public Variants
Leiden Open Variation Database Public Variants
OMIM Allelic Variant SNPs
Locus Variants
RefSeq gene (HBB)
OMIM gene
ClinVar variants (no CNVs)
ClinVar variants (CNVs)
HGMD variants
DECIPHER variants
LOVD variants
(SNVs and CNVs)
OMIM allelic variant SNPs
PhenCode locus-specific
database variants

https://www.lovd.nl/

Two kinds of disease mutation databases

[1] Central

- OMIM
- GeneCards
- Human Gene Mutation Database

[2] Locus-specific databases (mutation databases)

- Describe one gene in depth
- Complementary to central databases
- Offer specialist expertise
- There are hundreds of locus-specific databases

Locus-specific (mutation) databases: HGVS and LOVD

There are two main gateways to large numbers of locus-specific databases.

[1] The Human Genome Variation Society (HGVS) provides access to 1600 locus-specific mutation databases. It offers many additional database resources. See: http://www.hgvs.org/content/databases-tools

[2] The Leiden Open Variation Database (LOVD) has emerged as a platform supporting thousands of locus-specific databases. See: http://www.lovd.nl/3.0/home

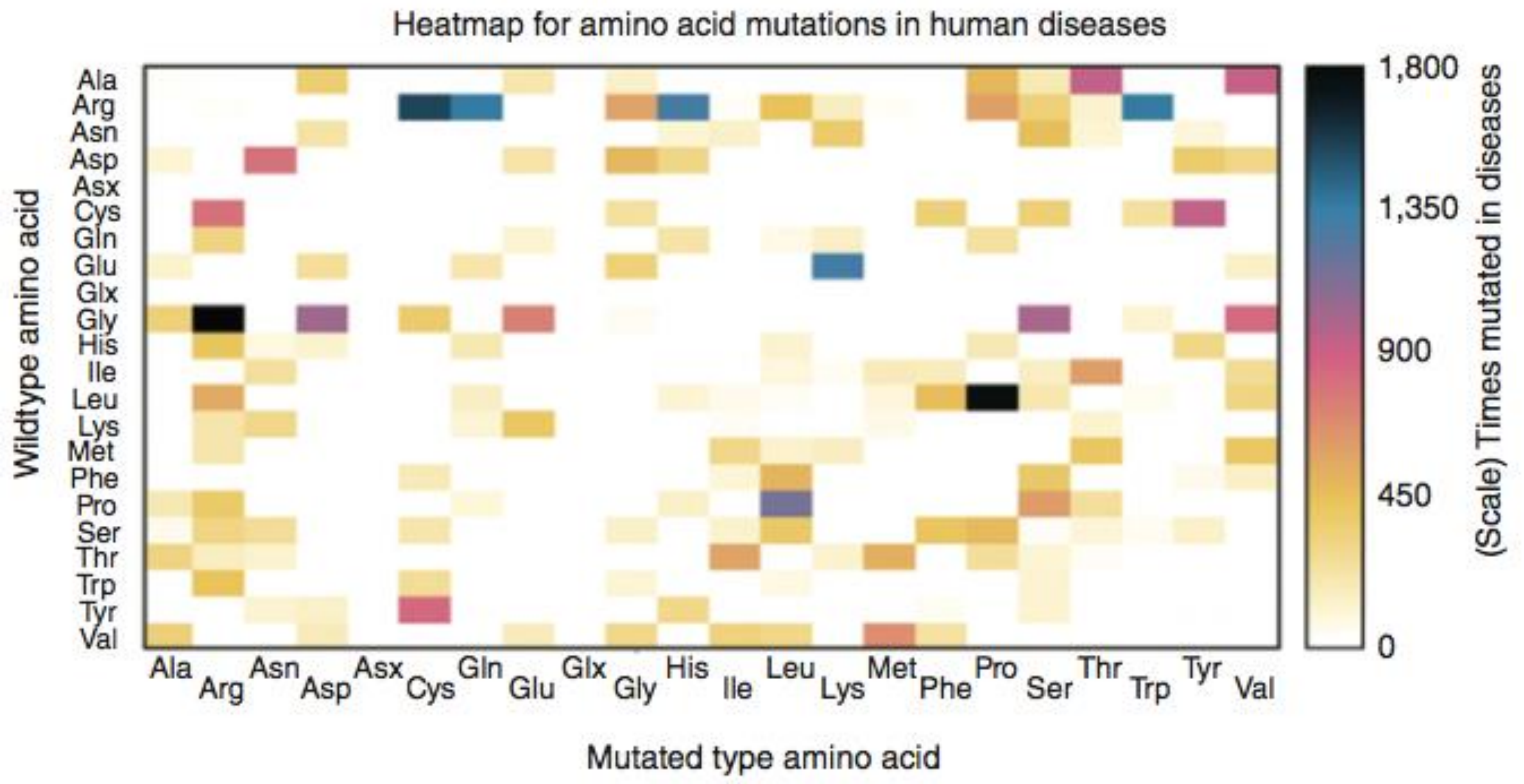
Disease and amino acid subsitutions

Information in disease databases allows us to explore the amino acid substitutions that occur in human disease. The three most common substitutions were:

leucine to proline
glycine to arginine
arginine to cysteine

From the BLOSUM62 matrix these have scores of −3, −2, and 0.

Heat map of amino acid variants in human diseases



The observed frequencies of wildtype transitions to mutated variants that are implicated in human disease are shown. The variants are from OMIM, HGMD,UniProt/Swiss-Prot, and ClinVar. Redrawn from Peterson *et al.* (2013).

Outline

Human genetic disease: a consequence of DNA variation

Categories of disease

Disease databases

Approaches to identifying disease-associated genes and loci

Human disease genes in model organisms

Functional classification of disease genes

Perspective

Four approaches to identifying disease genes

Linkage analysis

Genome-wide association studies (GWAS)

Identification of chromosomal abnormalities

Genomic DNA sequencing

Four approaches: [1] Linkage analysis

- A genetic linkage map displays genetic information in reference to linkage groups (chromosomes).
- The mapping units are centiMorgans, based on recombination frequency between polymorphic markers such as SNPs or microsatellites.
- One cM equals one recombination event in 100 meioses; for the human genome, the recombination rate is typically 1–2 cM/Mb.

- https://www.youtube.com/watch?v=ftrJh44ndkQ
- http://asia.ensembl.org/Homo_sapiens/Tools/LD

Four approaches: [1] Linkage analysis

- In linkage studies, genetic markers are used to search for coinheritance of chromosomal regions within families, that is, polymorphic markers that flank a disease locus segregate with the disease in families. Two genes that are in proximity on a chromosome will usually cosegregate during meiosis.
- By following the pattern of transmission of a large set of markers in a large pedigree, linkage analysis can be used to localize a disease gene based on its linkage to a genetic marker locus.
- Huntington's disease, a progressive degenerative disorder, was the first autosomal disorder for which linkage analysis was used to identify the disease locus.

Four approaches: [2] GWAS

- It is difficult to identify the genetic causes of common human diseases that involve multiple genes, each of which may make only a small contribution to the disease risk.
- Genome-wide association studies (GWAS) uses SNP markers to identify disease loci.
- In **family-based** designs, markers are measured in probands and unaffected individuals to identify differences in the frequency of variants.
- In **population-based** designs, a large number of unrelated cases and controls are studied (typically hundreds or thousands in each group). Larger sample sizes offer increased statistical power.

Single nucleotide polymorphisms (SNPs)

SNPs are the most common type of genetic variation in humans. They account for 90% of the variation between individuals.

Most are neutral polymorphisms. Some cause disease. The density of SNPs is about 1 every 100 to 300 bases.

SNPs may occur anywhere: in coding regions (cSNPs), in introns, in regulatory regions of genes, or in intergenic regions. In coding regions, changes may be synonymous or non-synonymous.

SNPs and disease

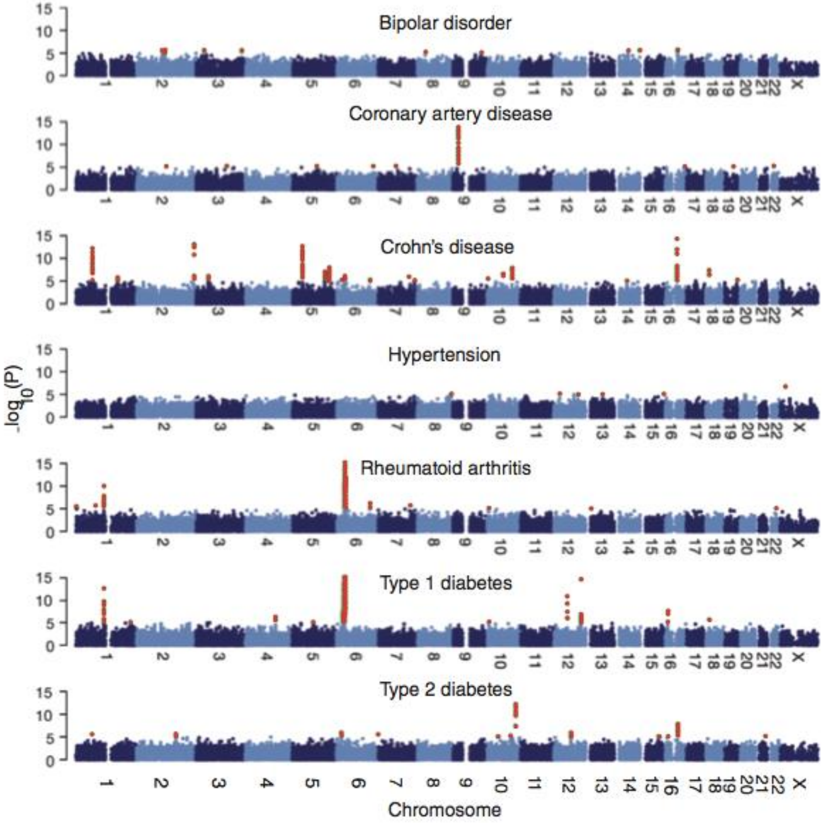
SNPs may be informative with respect to disease:

[1] Functional variation. A SNP associated with a nonsynonymous substitution in a coding region will change the amino acid sequence of a protein.

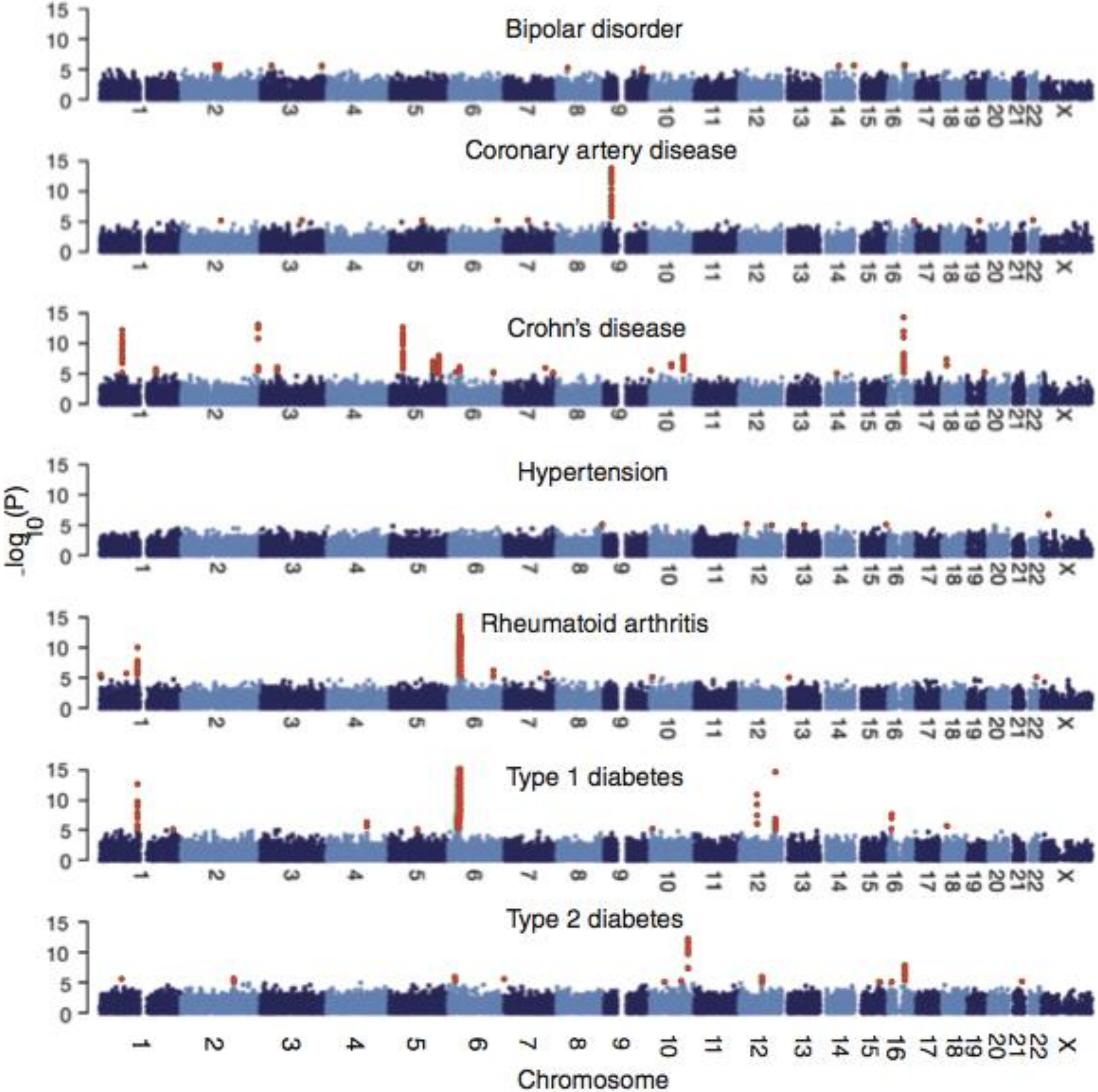
[2] Regulatory variation. A SNP in a noncoding region can influence gene expression.

[3] Association. SNPs can be used in whole-genome association studies. SNP frequency is compared between affected and control populations.

Results of a genome-wide association study using 16,179 individuals to search for genes contributing to seven common familial disorders



For each of seven diseases, the *y* axis shows the –log10 p value for SNPs that were positive for quality control criteria. The *x* axis shows the chromosomes. *p* values <1×10–5 are high-lighted in red. Panels are truncated at –log10(*p* value) = 15. Redrawn from Wellcome Trust Case Control Consortium (2007).


Bipolar disorder
Coronary artery disease
Crohn's disease
Hypertension
Rheumatoid arthritis
Type 1 diabetes
Type 2 diabetes
$-\log_{10}(P)$
0
5
10
15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosome

Four approaches: [3] Chromosomal abnormalities

- Common chromosomal aberrations in early development include the gain or loss of whole chromosomes. Such structural abnormalities may be detected by standard cytogenetic approaches such as karyotype analysis and **fluorescence *in situ* hybridization** (FISH).
- These techniques may also reveal large-scale duplications, deletions, or rearrangements.
- Spectral karyotyping/multiplex-FISH (SKY/M-FISH) permits each chromosome to be visualized, facilitating the identification of abnormal karyotypes.
- Array comparative genomic hybridization detects chromosomal abnormalities.
- Whole genome sequencing has emerged as a powerful tool to detect structural variation.

Four approaches: [4] Genome sequencing: monogenic

- Whole exome sequencing (WES) has been useful for identifying variants that cause monogenic disorders.
- Mendelian diseases are typically caused primarily by mutations affecting the coding region of a gene.
- The yield of whole-exome sequencing has therefore been high:
- Focus is on a small subset of the genome (~60 megabases), enriched for functionally relevant loci.
- Motivation to perform WES: is less than whole genome sequencing (WGS), and data analysis is relatively simpler.

Four approaches: [4] Genome sequencing: complex disorders

- Whole genome sequencing (WGS) detects 3-4 million single nucleotide variants (SNVs) per individual, substantially more than in a SNP array
- Trio-based WES or WGS often used to study complex diseases
- Interpretation of variants relevant to the phenotype is challenging

Outline

Human genetic disease: a consequence of DNA variation

Categories of disease

Disease databases

Approaches to identifying disease-associated genes and loci

Human disease genes in model organisms

Functional classification of disease genes

Perspective

Human disease genes in model organisms

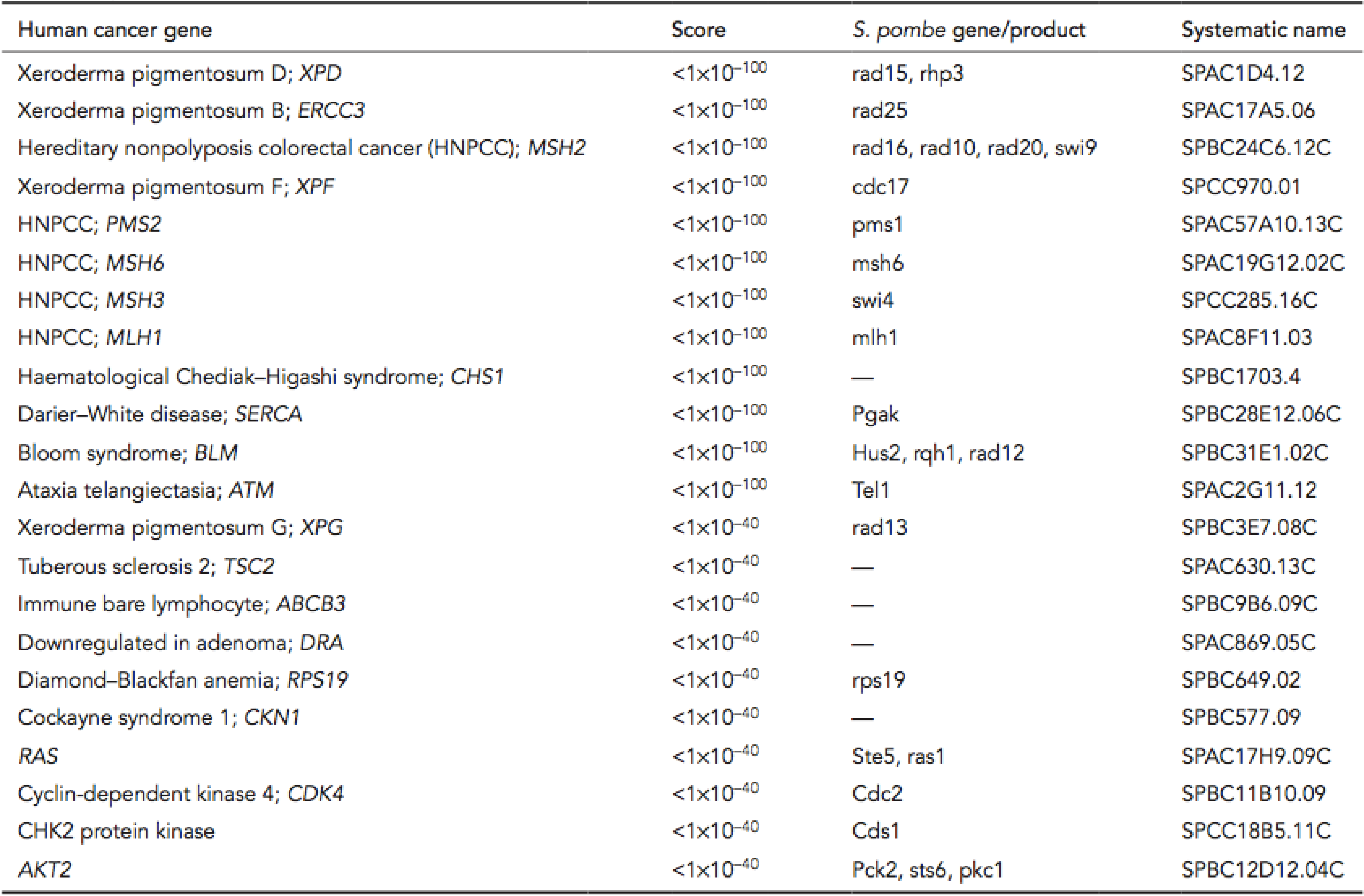
Once a human disease gene is identified in a model organism, its function can be studied (e.g. by gene knockouts). Most genome projects include an analysis of human disease gene orthologs.

As genomes have been sequenced we can

- identify orthologs of human disease genes. This facilitates comparative studies.
- identify instances in which a human gene has a disease-associated variant, and the model organism has that variant as its wildtype form. Such findings can help us understand the functional consequences and evolutionary history of mutations.

Schizosaccharomyces pombe genes related to human disease genes

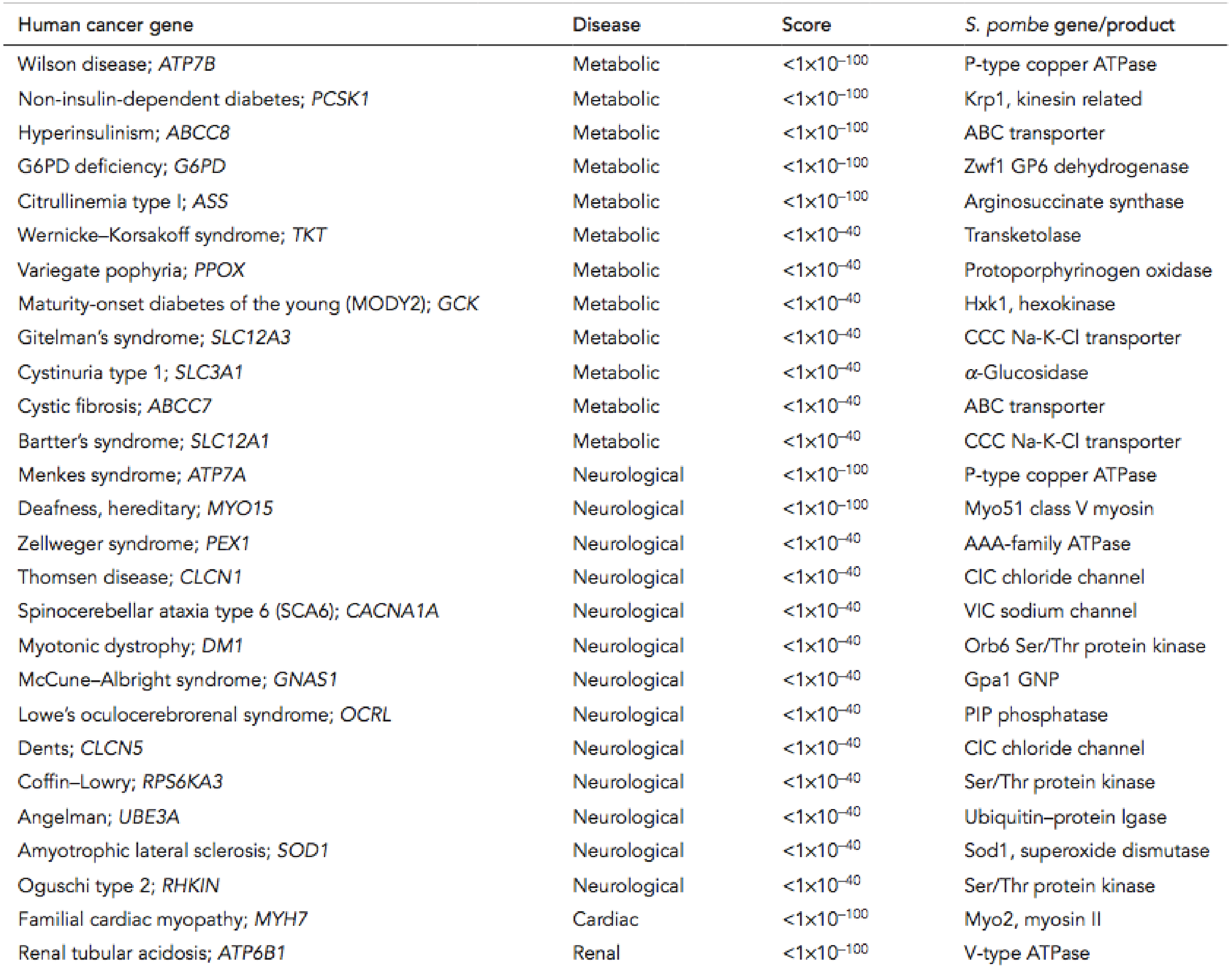
petite-negative yeast



Human cancer gene	Score	*S. pombe* gene/product	Systematic name
Xeroderma pigmentosum D; *XPD*	$<1\times10^{-100}$	rad15, rhp3	SPAC1D4.12
Xeroderma pigmentosum B; *ERCC3*	$<1\times10^{-100}$	rad25	SPAC17A5.06
Hereditary nonpolyposis colorectal cancer (HNPCC); *MSH2*	$<1\times10^{-100}$	rad16, rad10, rad20, swi9	SPBC24C6.12C
Xeroderma pigmentosum F; *XPF*	$<1\times10^{-100}$	cdc17	SPCC970.01
HNPCC; *PMS2*	$<1\times10^{-100}$	pms1	SPAC57A10.13C
HNPCC; *MSH6*	$<1\times10^{-100}$	msh6	SPAC19G12.02C
HNPCC; *MSH3*	$<1\times10^{-100}$	swi4	SPCC285.16C
HNPCC; *MLH1*	$<1\times10^{-100}$	mlh1	SPAC8F11.03
Haematological Chediak–Higashi syndrome; *CHS1*	$<1\times10^{-100}$	—	SPBC1703.4
Darier–White disease; *SERCA*	$<1\times10^{-100}$	Pgak	SPBC28E12.06C
Bloom syndrome; *BLM*	$<1\times10^{-100}$	Hus2, rqh1, rad12	SPBC31E1.02C
Ataxia telangiectasia; *ATM*	$<1\times10^{-100}$	Tel1	SPAC2G11.12
Xeroderma pigmentosum G; *XPG*	$<1\times10^{-40}$	rad13	SPBC3E7.08C
Tuberous sclerosis 2; *TSC2*	$<1\times10^{-40}$	—	SPAC630.13C
Immune bare lymphocyte; *ABCB3*	$<1\times10^{-40}$	—	SPBC9B6.09C
Downregulated in adenoma; *DRA*	$<1\times10^{-40}$	—	SPAC869.05C
Diamond–Blackfan anemia; *RPS19*	$<1\times10^{-40}$	rps19	SPBC649.02
Cockayne syndrome 1; *CKN1*	$<1\times10^{-40}$	—	SPBC577.09
RAS	$<1\times10^{-40}$	Ste5, ras1	SPAC17H9.09C
Cyclin-dependent kinase 4; *CDK4*	$<1\times10^{-40}$	Cdc2	SPBC11B10.09
CHK2 protein kinase	$<1\times10^{-40}$	Cds1	SPCC18B5.11C
AKT2	$<1\times10^{-40}$	Pck2, sts6, pkc1	SPBC12D12.04C

Score is the expect value from a BLAST search. Adapted from Wood et al. (2002).

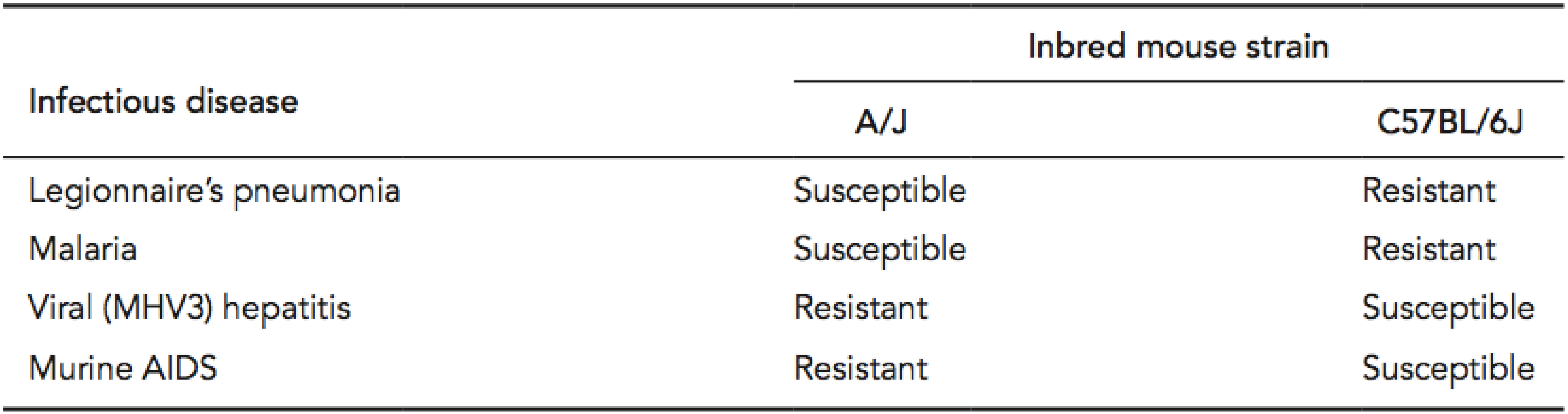
Schizosaccharomyces pombe genes related to human disease genes



Human cancer gene	Disease	Score	*S. pombe* gene/product
Wilson disease; *ATP7B*	Metabolic	$<1\times10^{-100}$	P-type copper ATPase
Non-insulin-dependent diabetes; *PCSK1*	Metabolic	$<1\times10^{-100}$	Krp1, kinesin related
Hyperinsulinism; *ABCC8*	Metabolic	$<1\times10^{-100}$	ABC transporter
G6PD deficiency; *G6PD*	Metabolic	$<1\times10^{-100}$	Zwf1 GP6 dehydrogenase
Citrullinemia type I; *ASS*	Metabolic	$<1\times10^{-100}$	Arginosuccinate synthase
Wernicke–Korsakoff syndrome; *TKT*	Metabolic	$<1\times10^{-40}$	Transketolase
Variegate pophyria; *PPOX*	Metabolic	$<1\times10^{-40}$	Protoporphyrinogen oxidase
Maturity-onset diabetes of the young (MODY2); *GCK*	Metabolic	$<1\times10^{-40}$	Hxk1, hexokinase
Gitelman's syndrome; *SLC12A3*	Metabolic	$<1\times10^{-40}$	CCC Na-K-Cl transporter
Cystinuria type 1; *SLC3A1*	Metabolic	$<1\times10^{-40}$	α-Glucosidase
Cystic fibrosis; *ABCC7*	Metabolic	$<1\times10^{-40}$	ABC transporter
Bartter's syndrome; *SLC12A1*	Metabolic	$<1\times10^{-40}$	CCC Na-K-Cl transporter
Menkes syndrome; *ATP7A*	Neurological	$<1\times10^{-100}$	P-type copper ATPase
Deafness, hereditary; *MYO15*	Neurological	$<1\times10^{-100}$	Myo51 class V myosin
Zellweger syndrome; *PEX1*	Neurological	$<1\times10^{-40}$	AAA-family ATPase
Thomsen disease; *CLCN1*	Neurological	$<1\times10^{-40}$	ClC chloride channel
Spinocerebellar ataxia type 6 (SCA6); *CACNA1A*	Neurological	$<1\times10^{-40}$	VIC sodium channel
Myotonic dystrophy; *DM1*	Neurological	$<1\times10^{-40}$	Orb6 Ser/Thr protein kinase
McCune–Albright syndrome; *GNAS1*	Neurological	$<1\times10^{-40}$	Gpa1 GNP
Lowe's oculocerebrorenal syndrome; *OCRL*	Neurological	$<1\times10^{-40}$	PIP phosphatase
Dents; *CLCN5*	Neurological	$<1\times10^{-40}$	ClC chloride channel
Coffin–Lowry; *RPS6KA3*	Neurological	$<1\times10^{-40}$	Ser/Thr protein kinase
Angelman; *UBE3A*	Neurological	$<1\times10^{-40}$	Ubiquitin–protein lgase
Amyotrophic lateral sclerosis; *SOD1*	Neurological	$<1\times10^{-40}$	Sod1, superoxide dismutase
Oguschi type 2; *RHKIN*	Neurological	$<1\times10^{-40}$	Ser/Thr protein kinase
Familial cardiac myopathy; *MYH7*	Cardiac	$<1\times10^{-100}$	Myo2, myosin II
Renal tubular acidosis; *ATP6B1*	Renal	$<1\times10^{-100}$	V-type ATPase

Score is the expect value from a BLAST search. GNP: guanine nucleotide binding. Adapted from Wood et al. (2002).

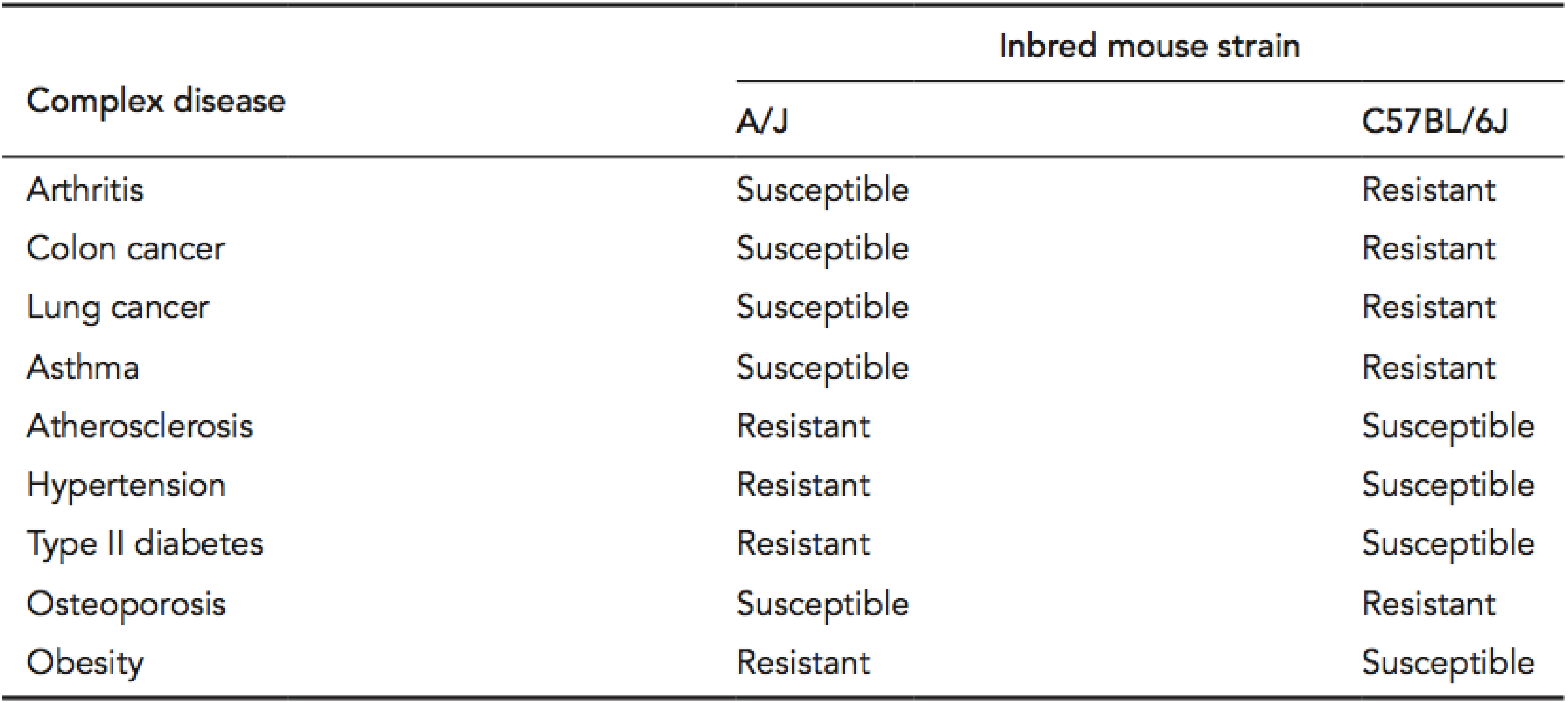
Infectious disease susceptibility of mouse strains



Infectious disease	Inbred mouse strain	
	A/J	C57BL/6J
Legionnaire's pneumonia	Susceptible	Resistant
Malaria	Susceptible	Resistant
Viral (MHV3) hepatitis	Resistant	Susceptible
Murine AIDS	Resistant	Susceptible

Understanding the genetic basis of disease susceptibility across mouse strains may help us to understand the disease process in humans.

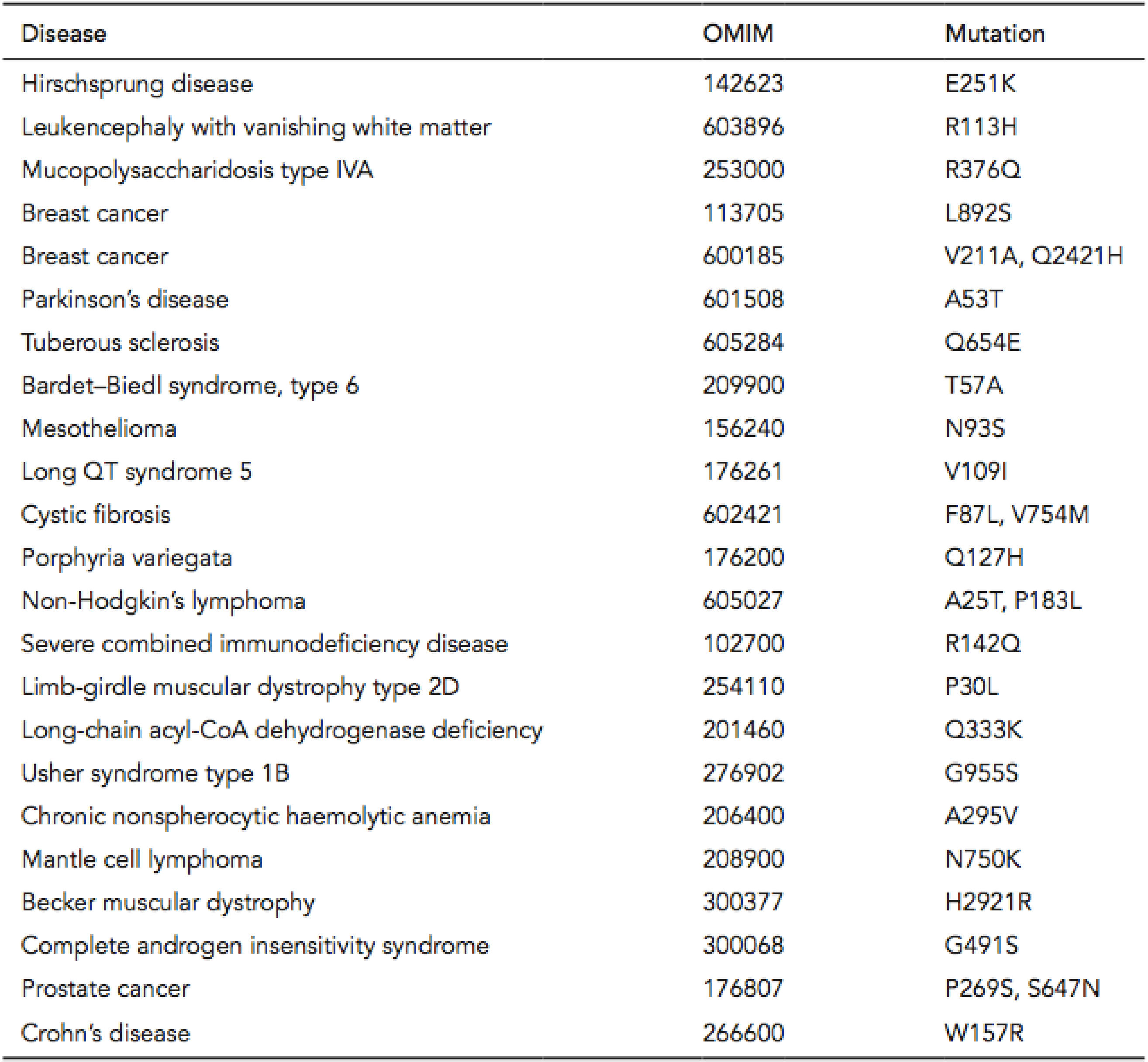
Common complex disease susceptibility of mouse strains



Complex disease	Inbred mouse strain	
	A/J	C57BL/6J
Arthritis	Susceptible	Resistant
Colon cancer	Susceptible	Resistant
Lung cancer	Susceptible	Resistant
Asthma	Susceptible	Resistant
Atherosclerosis	Resistant	Susceptible
Hypertension	Resistant	Susceptible
Type II diabetes	Resistant	Susceptible
Osteoporosis	Susceptible	Resistant
Obesity	Resistant	Susceptible

Understanding the genetic basis of disease susceptibility across mouse strains may help us to understand the disease process in humans.

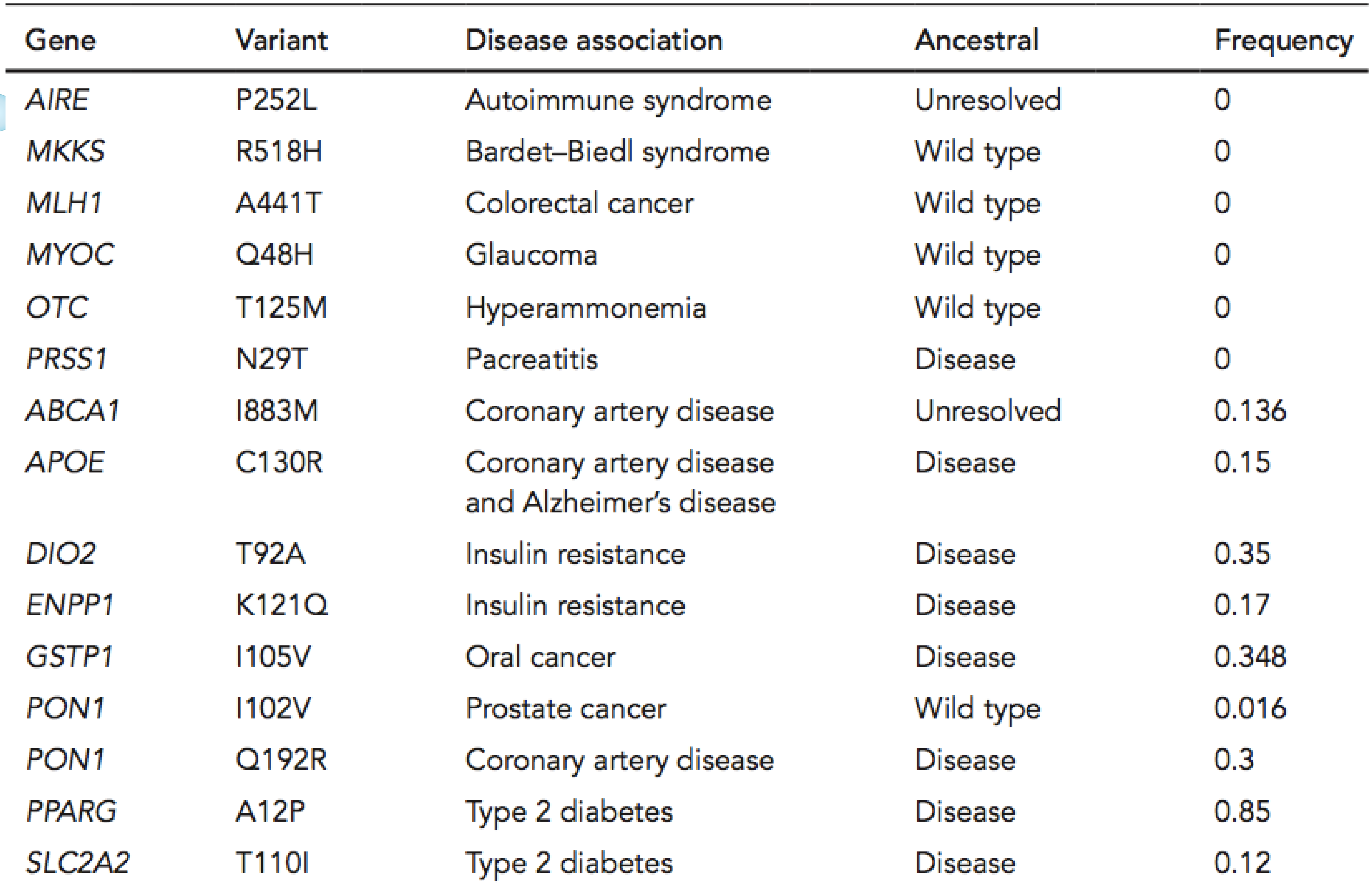
Human disease-associated sequence variants for which wildtype mouse sequence matches diseased human sequence



Disease	OMIM	Mutation
Hirschsprung disease	142623	E251K
Leukencephaly with vanishing white matter	603896	R113H
Mucopolysaccharidosis type IVA	253000	R376Q
Breast cancer	113705	L892S
Breast cancer	600185	V211A, Q2421H
Parkinson’s disease	601508	A53T
Tuberous sclerosis	605284	Q654E
Bardet–Biedl syndrome, type 6	209900	T57A
Mesothelioma	156240	N93S
Long QT syndrome 5	176261	V109I
Cystic fibrosis	602421	F87L, V754M
Porphyria variegata	176200	Q127H
Non-Hodgkin’s lymphoma	605027	A25T, P183L
Severe combined immunodeficiency disease	102700	R142Q
Limb-girdle muscular dystrophy type 2D	254110	P30L
Long-chain acyl-CoA dehydrogenase deficiency	201460	Q333K
Usher syndrome type 1B	276902	G955S
Chronic nonspherocytic haemolytic anemia	206400	A295V
Mantle cell lymphoma	208900	N750K
Becker muscular dystrophy	300377	H2921R
Complete androgen insensitivity syndrome	300068	G491S
Prostate cancer	176807	P269S, S647N
Crohn’s disease	266600	W157R

Adapted from Mouse Genome Sequencing Consortium et al. (2002)

Human disease variants matching the wildtype chimpanzee allele



Gene	Variant	Disease association	Ancestral	Frequency
AIRE	P252L	Autoimmune syndrome	Unresolved	0
MKKS	R518H	Bardet–Biedl syndrome	Wild type	0
MLH1	A441T	Colorectal cancer	Wild type	0
MYOC	Q48H	Glaucoma	Wild type	0
OTC	T125M	Hyperammonemia	Wild type	0
PRSS1	N29T	Pacreatitis	Disease	0
ABCA1	I883M	Coronary artery disease	Unresolved	0.136
APOE	C130R	Coronary artery disease and Alzheimer's disease	Disease	0.15
DIO2	T92A	Insulin resistance	Disease	0.35
ENPP1	K121Q	Insulin resistance	Disease	0.17
GSTP1	I105V	Oral cancer	Disease	0.348
PON1	I102V	Prostate cancer	Wild type	0.016
PON1	Q192R	Coronary artery disease	Disease	0.3
PPARG	A12P	Type 2 diabetes	Disease	0.85
SLC2A2	T110I	Type 2 diabetes	Disease	0.12

B&FG 3e
Tab. 21.17
Page 1060

Variants are listed as benign variant, codon number, disease/chimpanzee variant. ancestral variants are inferred using primate outgroups. Frequency is of the disease allele in humans. *Source:* Chimpanzee Sequencing and Analysis Consortium (2005).

Outline

Human genetic disease: a consequence of DNA variation

Categories of disease

Disease databases

Approaches to identifying disease-associated genes and loci

Human disease genes in model organisms

Functional classification of disease genes

Perspective

Perspective

There are several kinds of bioinformatics approaches to human disease:

- Human disease is a consequence of variation in DNA sequence. These variations are catalogued in databases of molecular sequences (such as GenBank, SRA, and ENA).
- Human disease databases have a major role in organizing information about disease genes. There are centralized databases, notably OMIM, ClinVar, and HGMD, as well as locus-specific mutation databases.
- Functional genomics screens provide insight into the mechanisms of disease genes and disease processes.