GUIDELINES FOR HUMAN GENE NOMENCLATURE

An International System for Human Gene Nomenclature (ISGN, 1987)

T.B. Shows[1], P.J. McAlpine[2], C. Boucheix[3], F.S. Collins[4], P.M. Conneally[5], J. Frezal[6], H. Gershowitz[4], P.N. Goodfellow[7], J.G. Hall[8], P. Issitt[9], C.A. Jones[10], B.B. Knowles[11], M. Lewis[12], V.A. McKusick[13], M. Meisler[4], N.E. Morton[14], P. Rubinstein[15], M.S. Schanfield[16], R.D. Schmickel[17], M.H. Skolnick[18], M.A. Spence[19], G.R. Sutherland[20], M. Traver[21], N. Van Cong[6], H.F. Willard[22].

[1]Roswell Park Memorial Institute, Buffalo; [2]University of Manitoba, Winnipeg; [3]Hopital Paul Brousse, Villejuif; [4]Univeristy of Michigan Medical School, Ann Arbor; [5]Indiana University Medical Center, Indianapolis; [6]Hopital des Enfants Malades, Paris; [7]Imperial Cancer Research Fund, London; [8]University of British Columbia, Vancouver; [9]Blood Services, S. Florida Region, Miami; [10]Roosevelt Institute of Cancer Research, Denver; [11]Wistar Institute, Philadelphia; [12]Rh Laboratory, Winnipeg; [13]Johns Hopkins Hospital, Baltimore; [14]Memorial Sloan-Kettering Cancer Center, New York City; [15]New York Blood Center, New York City; [16]Allo-Type Genetic Testing, Atlanta; [17]University of Pennsylvania Medical School, Philadelphia; [18]University of Utah School of Medicine, Salt Lake City; [19]University of California, Los Angeles; [20]Adelaide Children's Hospital, Adelaide; [21]University of Wisconsin Hospital, Madison; [22]University of Toronto, Toronto.

CONTENT PAGES

PROLOGUE

Extensive guidelines for a uniform gene nomenclature were first published in the Human Gene Mapping 5 Workshop (Edinburgh, 1979) in response to a plenary session directive by the Human Gene Mapping 4 Workshop (Winnipeg, 1977). These guidelines (Shows *et al.*, 1979) were adopted as the official system for human gene nomenclature by Human Gene Mapping 6 (Oslo, 1981) and revised at subsequent Human Gene Mapping (HGM) Workshops (Shows *et al.*, 1982; Shows *et al.*, 1984; McAlpine *et al.*, 1985). In response to HGM8 (Helsinki, 1985) directives, a nomenclature workshop was conducted to update guidelines with particular attention to blood group markers, other cell surface markers, inherited clinical syndromes, and DNA segment terminology.

A nomenclature workshop was held in September, 1986, at the University of Michigan, Ann Arbor, with funding from the Howard Hughes Medical Institute. The intent of this meeting was to update the 1979 guidelines where necessary and to propose new guidelines and directions for gene systems undergoing an evolution in understanding based on new technology. These guidelines were approved at HGM9 (Paris, 1987) and will be in effect at the time of publication.

The philosophy of the Human Nomenclature Committee is that gene nomenclature should evolve with new technology rather than be restrictive as sometimes occurs when historical and single gene nomenclature systems are applied.

ONE HUMAN GENOME - ONE GENE LANGUAGE

Certainly there exists a genetic and molecular basis for a single human gene language without dialects. All human nuclear genes as we know them follow the same genetic, molecular, and evolutionary principles. Each is located at a specific chromosomal site and arranged with other genes in a unique linear order on chromosomes. All genes exhibit sequence variability, even at multiple sites. The types of genetic and molecular changes detected for one gene are usually common to other genes. A gene may be represented once in the human genome or as closely related members of a gene family or complex. Human genes are not unique but homologous with genes in other mammalian species. Human genes follow simple Mendelian and somatic cell genetic laws of inheritance and function together to direct the biology of the organism. Thus it is reasonable and logical to develop a standard and consolidated human gene nomenclature system rather than have a human gene language based on different gene systems and consequently composed of different dialects.

GENE AND ALLELE TERMINOLOGY

Gene Names

The name of a gene should be specific and brief and, where possible, should convey information about the character or function of the gene. The name may indicate, for example, a morphological or disease character when the gene function is not known, or a biochemical, molecular, or antigenic property. Ultimately, a gene's name should evolve to indicate the exact gene function. It is conceivable that in the future the chromosomal location at the molecular level will be incorporated into the gene name and symbol. When the information known about a gene is extensive, it will not be possible to convey the characteristics about a gene through its name. In these circumstances, additional gene characteristics may be listed in a compendium cataloging human genes.

Gene Symbols

1. Genes are designated by upper-case Latin letters or by a combination of upper-case letters and Arabic numbers. Since symbols should be short to be useful and should not attempt to indicate all known information about a gene, a total of three characters to designate gene names is optimal; it is recommended that no more than five characters be used. Based on classical genetic guidelines, gene symbols always are either underlined or italicized. Gene symbols need not be italicized in catalogs of known genes. However, when fragments or synthesized segments of genes are referred to, symbols need not be italicized. New symbols must not duplicate existing gene symbols. Examples: *HPRT* (hypoxanthine phosphoribosyl transferase); *G6PD* (glucose-6-phosphate dehydrogenase); *ALB* (albumin); *RP* (retinitis pigmentosum); *CBP* (color blindness, protan); *HBB* (hemoglobin, beta polypeptide).

2. The first letter should be the same as that of the name of the gene to facilitate alphabetical listing and grouping.

3. The initial character should always be a letter. Subsequent characters of the symbol may be other letters or, if necessary, Arabic numerals.

4. All characters in a gene symbol should be written on the same line; thus, no superscripts or subscripts may be used.

5. No Roman numerals may be used. Roman numbers in previously used symbols should be changed to their Arabic equivalents.

6. Greek letters are not permitted in a gene symbol. All Greek symbols should be changed to letters in the Latin alphabet (Table I).

7. A Greek letter prefixing a gene name must be changed to its Latin alphabet equivalent and placed at the end of the gene symbol. This permits alphabetical ordering of the gene in listings with similar properties, such as substrate specificities. Examples *GLA* (galactosidase, alpha); *GLB* (galactosidase, beta).

8. Where gene products of similar function are encoded by different genes, the corresponding loci are designated by Arabic numerals placed immediately after the gene symbol, without any space between the letters

Table I. Greek-to-Latin alphabet conversion

Greek	Lower case	Latin upper case conversion
α	alpha	A
β	beta	B
γ	gamma	G
δ	delta	D
ε	epsilon	E
ζ	zeta	Z
η	eta	H
θ	theta	Q
ι	iota	I
κ	kappa	K
λ	lambda	L
μ	mu	M
ν	nu	N
ξ	xi	X
ο	omicron	O
π	pi	P
ρ	rho	R
σ	sigma	S
τ	tau	T
υ	upsilon	Y
φ	phi	F
χ	chi	C
ψ	psi	U
ω	omega	W

Table II. Single-letter amino acid symbols

Amino acid	Three-letter symbol	One-letter symbol
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Asn + Asp	Asx	B
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Gln + Glu	Glx	Z
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

and numbers used. Examples: *PGM1*, *PGM2*, *PGM3* (three loci for phosphoglucomutase activity); *ADH1*, *ADH2*, *ADH3* (three alcohol dehydrogenase loci); *HBA1*, *HBA2* (duplicated forms of the α-hemoglobin gene). However, single-letter suffixes may be used to designate these different loci only if they exist historically. Examples: *LDHA*, *LDHB*, *LDHC* (three lactate dehydrogenase loci).

9. A final character in the gene symbol may be used to specify a characteristic of the gene. While letters to specify tissue distribution have been used historically, Arabic numbers are now preferred as experience has shown that tissue specificity may not be as restricted as described initially.

10. If the name of a gene contains a character or property for which there is a recognized abbreviation, the abbreviation should be used; for example, the single-letter abbreviation for amino acids (Table II) used in aminoacyl residues or approved biochemical abbreviations such as GLC for glucose and GSH for glutathione.

Allele Symbols

1. The allele symbol should be limited to four characters, with an optimum of three characters. Only capital letters or Arabic numerals in any order should be used.

2. Allele designations are written on the same line as gene symbols.

3. The allele characters are separated from the locus characters by a new symbol--the asterisk--which serves to combine gene and allele symbols. There should be no spaces between gene, asterisk, and allele, and the entire symbol should be underlined or italicized.

In order to keep the gene and allele designations separated but together, a new character, the asterisk, has been introduced. Advantages of the asterisk are many. The asterisk is convenient, universal, and does not convey past genetic meaning such as the dash, space, or comma. The asterisk preceding a symbol indicates that it is an allele of a gene. Likewise, an asterisk following a symbol indicates that it is a gene. After the gene and allele symbols have been identified, the allele symbol preceded by an asterisk can be used separately in text.

For example: *PGM1*1*, *PGM1*2* (for alleles of *PGM1*, phosphoglucomutase-1); *ADA*1*, *ADA*2* (for alleles of *ADA*, adenosine deaminase); *ABO*A1*, *ABO*B* (for alleles at the *ABO* locus); *HBB*6V* (identifies the sickle cell mutation resulting from the substitution in the sixth codon of beta

hemoglobin, [β6 GLY→VAL] with the single-letter amino acid symbol); *HBA*68K* (identifies substitution [α-68 ASN→LYS] in HbG-Philadelphia); *BF*F1* (allele at the properdin factor B locus); *C6*A* (allele of complement C6); *PLG*A* (allele at plasminogen locus); *HP*1, HP*2* (for the alleles of the haptoglobin locus); *DM*D, DM*N* (alleles at the myotonic dystrophy locus, *D* = dominant and *N* = normal "wild" type); *CF*N, CF*R* (alleles at cystic fibrosis locus, *N* = normal and *R* = recessive); *TG*A1, TG*A2* (alleles at the thyroglobulin locus detected by probing TaqI restricted genomic DNA with the pCHT16/8 probe).

4. The allele symbol may convey additional information. The first allele in a series may be designated A or 1. The symbol may convey a morphological characteristic, biochemical property, cellular location, control property, or, ultimately, the amino acid nucleotide substitution (i.e., *HBB*6V*). No normal plus (+) symbol or variant minus (-) symbol, Roman numeral, or Greek symbol should be used. If the name of a geographic location is used in designating an allele, it should be limited to no more than a four-character symbol. If an allele exhibits no activity, this is indicated by an O (capital letter O) (see Enzyme and Protein Allele section). For optimal usage, allele symbols should be brief and need not summarize all information known about their genetic specificity.

5. If the information regarding the genetic specificity is too complex to be conveyed conveniently in a symbol (e.g., kinetic properties, amino acid substitutions, or subcellular localization), alleles may be designated by letter or number and the information conveyed in tables.

Printing Gene and Allele Symbols

Gene and allele symbols are underlined in manuscript and italicized in print. Italics need not be used in catalogs. It may be convenient in manuscripts, computer printouts and in printed text to designate a gene symbol by following it with an asterisk (e.g., ***PGM1****). When only allele symbols are displayed, they can be preceded by an asterisk. For example, for *PGM1*1*, the allele is printed as **1*. *(Phenotype symbols are discussed below.)*

GENOTYPE TERMINOLOGY

Genotypes are written in accordance with the above suggestions and follow historical precedents. A horizontal line or slash separates alleles and indicates chromosome homologs. Examples are given in three formats: using the horizontal (chromosome) line, complete description on a single line, and shortened designations.

1. Heterozygote for the common alleles at the *ADA* and *AMY1* loci:

$$\frac{ADA^*1}{ADA^*2}$$ or *ADA*1/ADA*2* or *ADA*1/*2*

$$\frac{AMY1^*A}{AMY1^*B}$$ or *AMY1*A/AMY1*B* or *AMY1*A/*B*

2. Genotype of an individual heterozygous at the *ADA* locus, homozygous at the *ADH1* locus, and heterozygous at the *AMY1* locus (all unlinked loci and separated by a semicolon):

$$\frac{ADA^*1}{ADA^*2}\ ;\ \frac{ADH1^*1}{ADH1^*1}\ ;\ \frac{AMY1^*A}{AMY1^*B}$$

or

*ADA*1/ADA*2; ADH1*1/ADH1*1; AMY1*A/AMY1*B*

or

*ADA*1/*2; ADH1*1/*1; AMY*A/*B*

3. Genotypes for sex-linked traits distinguish between males and females. At the glucose-6-phosphate dehydrogenase locus (*G6PD*), genotypes for heterozygous female and hemizygous males follow a similar pattern.

Females: $\frac{G6PD^*A}{G6PD^*B}$ or *G6PD*A/G6PD*B* or *G6PD*A/*B*

Males: *G6PD*A/Y* and *G6PD*B/Y*

(The Y identifies the male and maintains the diploid nature.)

Linkage and Phase

Horizontal lines or slashes separate alleles and indicate chromosome location.

1. Loci not located on the same chromosome are separated by a semicolon:

$$\frac{ENO1^*1}{ENO1^*2}\ ;\ \frac{PGM2^*1}{PGM2^*2}$$ or *ENO1*1/ENO1*2 ; PGM2*1/PGM2*2* or *ENO1*1/*2 ; PGM2*1/*2*

2. Loci on the same chromosome (linked or syntenic), where the phase is known, are joined by a horizontal line but separated by a space and listed in alphabetical order when gene order is not known:

$$\frac{AMY1^*A\ \ PGM1^*2}{AMY1^*B\ \ PGM1^*1}$$

For text, the loci can be printed on a single line, with a space separating genes in phase and a slash indicating different homologs:

*AMY1*A PGM1*2/AMY1*B PGM1*1.*

3. Loci on the same chromosome but phase not known are separated by a comma:

*PGM1*1, AMY1*A*

*PGM1*2, AMY1*B*

or printed on a single line with a separating comma: *PGM1*1/PGM1*2*, AMY1*A/AMY1*B. The linear order of genes is pter-*PGM1-AMY1*-cen, where a hyphen separates the linked genes.

4. If the linear order and phase of the genes on the same chromosome are known, they are listed in order from the end of the short arm to the end of the long arm of the chromosome and separated by a space:

*ENO1*1 PGM1*2 AMY1*B*

*ENO1*2 PGM1*1 AMY1*A*

or *ENO1*1 PGM1*2 AMY1*B/ENO1*2 PGM1*1 AMY1*A*. The linear order on chromosome 1 is pter-*ENO1-PGM1-AMY1*-cen.

5. If the gene order on the same chromosome is not known, then the loci are listed on the linear map alphabetically, separated by a comma, and enclosed by parentheses:

pter-*PGD-AK2-(ACTA, APOA2, REN)*-qter.

PHENOTYPE TERMINOLOGY

Phenotypic symbols should have the same characters as the genotypic and allelic symbols. They should not be underlined or italicized, and all characters should be written on the same line. The asterisk separating gene and allele characters in genotypes is omitted in phenotype symbols. Rather, a space separates gene and allele characters in phenotypes; a comma separates alleles.

Examples of genotypes and corresponding phenotypes:

Genotype	Phenotype
*ADA*1/ADA*1*	ADA 1
*ADA*1/ADA*2*	ADA 1,2
*ADA*2/ADA*2*	ADA 2
*PGM1*1/PGM1*2*	PGM1 1,2
*C3*F/C3*S*	C3 F,S
*C2*C/C2*QO*[1]	C2 C,QO
*HBB*A/HBB*6V*	HBB A,S
*ABO*A1/ABO*O*	ABO A1
*CF*N/CF*R*	CF N
*HD*N/HD*D*	HD D
*G6PD*A/G6PD*B*	G6PD A,B
*G6PD*A/Y*	G6PD A
*APOA1*A1/APOA1*A2*	APOA1 A1,A2
*INS*A1/INS*A2*	INS A1,A2

[1]See next section for quantitative variants.

ENZYMES AND PROTEINS

Gene Names

Names of genes coding for enzymes are based on those recommended by the Nomenclature Committee of the International Union of Biochemistry. Names of plasma proteins, hemoglobins, and specialized proteins are based on standard names and those recommended by their respective committees (Enzyme Nomenclature Recommendations, 1985; Putnam, 1975; Alper and Parkman, 1974). Gene names should be brief and specific and should convey the character or function of the gene. Cloned genes or segments of genes should assume the name of the gene that was cloned. Nomenclature of DNA segments recognized by cloned probes will be discussed below. For examples, see McAlpine *et al.*, 1985; Table I.

Gene Symbols

1. Gene symbols should follow the above guidelines. The symbol should be abbreviated as much as possible, with from one to four characters recommended, underlined in manuscript and italicized in print. Only upper-case Latin letters are to be used, with the first letter of the symbol the same as that of the gene name. Arabic numerals should not be used in gene symbols unless necessary and should not initiate a gene symbol. Symbols do not have superscripts, subscripts, Roman numerals, or Greek letters. Greek letters that are prefixes must be changed to their Latin equivalents (Table I) and placed at the end of the gene symbol. Examples: βGAL-1 becomes *GLB1*; HBδ becomes *HBD*; βGUS becomes *GUSB*; 6PGD becomes *PGD*.

2. When different genes code for proteins with similar functions, the same characters for the symbol should be used followed by an Arabic numeral to distinguish the loci. There should be no spaces in the symbol. Letter suffixes to designate different loci, including those appearing to have tissue-specific expression, should be avoided. Examples: *CA1*, *CA2*, *CA3* (three carbonic anhydrase loci); *HBA1*, *HBA2* (two hemoglobin α polypeptide loci); *AK1*, *AK2*, *AK3* (three adenylate kinase loci).

3. Whenever the name of any aminoacyl residue occurs in the description of a genetic character, it should be represented by its single-letter code in the gene symbol. Appropriate single-letter designations for amino acids may be found in Lehninger (1975) (Table II). Examples of aminoacyl-tRNA synthetases: *LARS* (leucyl-tRNA synthetase); *WARS* (tryptophanyl-tRNA synthetase) (see Table II). Other examples of enzyme gene loci are provided in Table III and follow the above guidelines.

4. For serum proteins, genetic locus symbols should conform to the general rules for designating loci given above. See Table IV.

5. Hemoglobin gene symbols are developed by converting the names of the polypeptides using the above guidelines. The letters in the symbol are capitalized; the Greek letter designations are converted to their Latin

equivalents and occur after the root of the hemoglobin gene symbol (*HB*).

Examples: *HBA* (Hbα chain locus)
HBG1 ($Hb^{A}\gamma$ polypeptide locus)
HBG2 ($Hb^{G}\gamma$ polypeptide locus)
HBZ (Hbζ polypeptide locus).

Table V provides other examples of the proposed terminology for hemoglobin loci.

6. Molecular technology has identified non-transcribed sequences that bear striking homologies to structural gene sequences. These sequences are termed pseudogenes. In order to show the relatedness of pseudogenes to functional genes, pseudogenes will be identified with the gene symbol of the structural gene followed by a P for pseudogene. In order to reserve P for pseudogenes, avoid using the letter P in the last position of a structural gene symbol where possible. Examples: *HBBP1* (hemoglobin, beta pseudogene 1); *ACTBP1* (actin, beta pseudogene 1); *ACTBP2* (actin, beta pseudogene 2), etc. Pseudogenes may be on different chromosomes or closely linked to the functional gene and occur in varying numbers.

Allele Symbols

1. Allele symbols follow the guidelines detailed above.

Examples:

*G6PD*A, G6PD*B; ADA*1, ADA*2; ENO1*1, ENO1*2*

*PGM2*1, PGM2*2; AMY1*A, AMY1*B; HP*1, HP*2.*

2. The allele symbol should be limited to a four-character symbol. If the name of a geographic location is used, it should be limited to a four-character symbol. Examples $G6PD^{Seattle}$ becomes *G6PD*SEAT*; $G6PD^{Andhra\ Pradesh}$ becomes *G6PD*AP*. Eventually these geographical allelic designations should be replaced by amino acid substitution terminology.

3. Where numbers or single letters of the alphabet are used to represent alleles or specificities, the order of their presentation in genotypes or phenotypes should always be in the normal sequence.

4. Only capitalized letters or Arabic numerals, or a combination of both, should be used; plus (+) and minus (-) signs should be avoided. Alleles representing quantitative (Q) variants of a given locus may be designated as follows: QE for those associated with elevated (E) quantities; QL for those associated with amounts that are low (L); QO (capital letter O) for those associated with no detectable quantity of the product, e.g., *PGM1*QO*. It is permissible to designate alleles controlling quantitative variation by the geographic location, number, or letter.

5. If allele symbols include a decimal, then the decimal must be dropped. For example: *C3 SO.4* becomes *C3*SO4*.

6. A proposed terminology for hemoglobin alleles is found in Table V. The ultimate goal of an evolving gene nomenclature is to identify alleles at a single locus by their amino acid substitutions or DNA structural differences. The amino acid substitutions that identify hemoglobin variants allow this goal to be realized. Usually, a four-character space allowance permits listing the location of the substitution with the amino acid number and the single-letter international code for aminoacyl residues that are substituted. If the allele information extends beyond four characters, as it could for double substitutions, fusion of hemoglobin genes, or extended chains, a shorthand identification is recommended (Table V). Additional information could be tabulated in catalogs with descriptions of human genes and alleles.

Phenotypes of Enzymes and Proteins

These symbols should be the same as gene and allele symbols but not underlined or italicized. The asterisk is omitted and the gene and allele characters are separated by a space.

Examples of genotypes and corresponding phenotypes:

Genotype	Phenotype
*G6PD*A/G6PD*B*	G6PD A,B
*PGK*1/PGK*2*	PGK 1,2
*ADA*1/ADA*QO* (silent allele)	ADA 1
*AK1*1/AK1*2*	AK1 1,2
*PGM1*1/PGM1*2*	PGM1 1,2
*HBB*A/HBB*6V*	HBB A,S
*HBA*A/HBA*16E*	HBA A,I
*HBB*6V/HBB*6K*	HBB S,C
*TF*C/TF*B1*	TF C,B1
*CP*A/CP*BNH*	CP A,BNH
*PI*M1/PI*S*	PI M1,S
*CHE1*A/CHE1*QO* (silent allele)	CHE1 A
*HP*1F/HP*1S*	HP 1F,1S[1]
*ALB*U/ALB*MEX*	ALB U,MEX

[1]Since 1F and 1S only apply to the "1" alleles, the "1" can be omitted, e.g., HP F,S.

Table III. Examples of gene symbols for enzymes[1] (See McAlpine et al., 1985; Table I).

Enzyme name	E.C. No.	Locus symbol
Aconitase (soluble)	4.2.1.3	*ACO1*
Aconitase (mitochondrial)	4.2.1.3	*ACO2*
Adenosine deaminase	3.5.4.4	*ADA*
α-Amylase (salivary)	3.2.1.1	*AMY1*
α-Amylase (pancreatic)	3.2.1.1	*AMY2*
Carbonic anhydrase-I	4.2.1.1	*CA1*
Carbonic anhydrase-II	4.2.1.1	*CA2*
Carbonic anhydrase-III	4.2.1.1	*CA3*
α-Galactosidase	3.2.1.22	*GLA*
β-Galactosidase-1	3.2.1.23	*GLB1*
β-Galactosidase-2	3.2.1.23	*GLB2*
Glutathione reductase	1.6.4.2	*GSR*
Hexosaminidase-α polypeptide	3.2.1.30	*HEXA*
Lactate dehydrogenase-A	1.1.1.27	*LDHA*
Pepsinogen	3.4.23.x[2]	*PGA*
6 Phosphogluconate dehydrogenase	1.1.1.44	*PGD*
Tryptophanyl-tRNA synthetase	6.1.1.2	*WARS*
UDP glucose pyrophosphorylase-1	2.7.7.9	*UGP1*
UDP glucose pyrophosphorylase-2	2.7.7.9	*UGP2*

[1]Enzyme loci have been tabulated with recommended and previous gene symbols (see Shows *et al*, 1979; Shows and McAlpine, 1979).
[2]Exact enzyme classification not known.

Table IV. Examples of gene symbols for proteins of known function

Protein	Symbol
Albumin	*ALB*
Antithrombin III	*AT3*
α_1-Antitrypsin (α_1-protease inhibitor)	*PI*
Ceruloplasmin	*CP*
C1 inhibitor	*C1NH*
complement component 2	*C2*
Group-specific component (vitamin D binding protein)	*GC*
Immunoglobulin α1 polypeptide	*IGHA1*
Immunoglobulin δ polypeptide	*IGHD*
Immunoglobulin ε polypeptide	*IGHE*
Immunoglobulin γ1 polypeptide	*IGHG1*
Immunoglobulin μ polypeptide	*IGHM*
Immunoglobulin κ polypeptide	*IGK*
Immunoglobulin λ polypeptide	*IGL*
Orosomucoid	*ORM*
Plasminogen activator, tissue	*PLAT*
Plasminogen activator, urokinase	*PLAU*

Table V. Suggested hemoglobin terminology

Hemoglobin loci
HBA1 and *HBA2* for the two Hb α chain loci
HBB for Hb β chain locus
HBD for Hb δ chain locus
HBE for the Hb ε chain locus
HBG1 for the Hb $^{A}\gamma$ chain locus
HBG2 for the Hb $^{G}\gamma$ chain locus
HBZ for the Hb ζ chain locus
Alleles (examples)
Single substitutions
*HBB*6V* for Hb S
*HBA*16E* for Hb I-Philadelphia
*HBD*69R* for Hb A - Indonesia
*HBG1*80Y* for Hb F-Victoria Jubilee (an $^{A}\gamma$ mutant)
*HBG2*117R* for Hb F-Malta-I (a $^{G}\gamma$ mutant)
Double substitutions
*HBB*6V142V* for Hb S-Travis
*HBB*6V73N* for Hb C-Harlem
Fusion of HB genes
*HBDB*LH* for Hb Lepore Hollandia
*HBBD*M1* for Hb Miyada (antiLepore)
*HBG1B*KE* for Hb Kenya (an $^{A}\gamma\beta$) hybrid
Extended chains
*HBA*CS* for Hb Constant Spring
*HBB*TAK* for Hb TAK
*HBA*GRA* for Hb Grady
Alleles with deletions
*HBB*LE* for Hb Leiden
*HBB*LY* for Hb Lyon
Thalassemia
*HBA1*THO* for αThal (α1 gene not active)
*HBA2*THO* for αThal (α2 gene not active)
*HBA2*THO HBA1*THO* for αThal (both α1 and α2 not active)
*HBB*THO* for β^{o}Thal (no β chain production)
*HBB*TH* for β^{+}Thal (β chain production decreased)
*HBD*THO* for δ^{o}Thal (no δ chain production)
*HBG1*THO* for $^{G}\gamma$Thal ($^{A}\gamma$ deleted)
*HBG2*THO* for $^{A}\gamma$Thal ($^{G}\gamma$ deleted)
*HBBD*THO* for $\beta^{o}\delta^{o}$Thal (no β and δ chain production)

INHERITED CLINICAL DISORDERS

The evolution of genetic knowledge and gene nomenclature in regard to a clinical syndrome is illustrated by dominant elliptocytosis (Table VI). In stage 1, there is recognition of a phenotype apparently inherited as a monogenic Mendelian trait eligible for a McKusick number (McKusick, 1984). Often at this time an acronym for the disorder is introduced, like SLO for Smith-Lemli-Opitz disease. Since this acronym is a good

initial choice for a gene symbol, it should not duplicate an accepted symbol. There may or may not be evidence at this stage of chromosomal location. In stage 2, two or more loci are identified by linkage heterogeneity, complementation, or biochemical evidence, but the gene symbols continue to contain information about the clinical phenotype. However, for some disorders (e.g., Huntington Disease) such heterogeneity may not be found. In stage 3, the loci are defined by identification of the gene products or functions and their clinical phenotypes must be encoded in alleles. In stage 4, the molecular basis of the mutation is identified. In stage 5, polymorphism in the normal allele is recognized. Elliptocytosis is currently at stage 3.

Table VI. Evolution of gene nomenclature

Stage	Clinical phenotype	Locus symbol	Allele
1. Mendelian	EL	*EL*	**D*
2. Locus heterogeneity	EL	*EL1*	**D*
		EL2	**D*
3. Molecular locus	EL	*PB41*	**EL*
		SPTA	**EL*
4. Molecular mutation	EL	*PB41*	**EL48G* (hypothetical)
5. Polymorphism	N	*PB41*	**N1,*N2* (hypothetical)

Symbols: *EL,* elliptocytosis; **EL,* elliptocytosis allele after the gene function is known; **D,* dominant allele; *PB41,* protein band 4.1; *SPTA,* spectrin, alpha; *EL48G,* hypothetical amino acid substitution defining the elliptocytosis mutation and allele; N - normal; **N1, *N2* - normal alleles 1 and 2.

Gene Names

The nomenclature of inherited genetic diseases should follow the generalized terminology for gene symbols. The symbol should be brief and yet convey the name and characteristics of the syndrome. When the biochemical abnormality is determined, the symbol describing the clinical characteristics of the syndrome should be changed to a gene symbol representing the biochemical property. When conveying a description of the disease, it is recommended that the McKusick number be given for further information concerning the genetic defect.

More and more syndromes, sporadic or inherited, or both, are being attributed to small chromosomal deletions or duplications. In *Drosophila,* there is a tradition of assigning gene symbols to deletions or duplications of apparently single loci, like *Bar* and *Notch.* On the other hand, syndromes clearly associated with multiple loci (like Down syndrome) should not be given gene symbols. We suggest that syndromes associated with regional deletion or duplication be assigned letters CR (for chromosome region) in place of S for syndrome (e.g., *PWCR* [for Prader-Willi syndrome], *BWCR* [for Beckwith Wiedemann syndrome], *LGCR* [for Langer-Giedion syndrome]) and that the cytological identification of the critical region be incorporated in the corresponding allele (e.g., *PWCR*15Q11;* Prader-Willi chromosome region is located on chromosome 15 in the q11 region).

Where an apparently homologous locus has been previously identified and named in another mammal, that gene symbol should be reserved and used in other species including humans: e.g., *HYP* for X-linked hypophosphatemia with vitamin D resistant rickets (see later section on homologous genes).

When a clinical phenotype is suspected to have an autosomal and X-linked form, the gene symbol for the latter may terminate with an X to indicate mode of inheritance (e.g., *RPX).* Genetic evidence may subsequently require designation of different loci as *RPX1, RPX2,* etc.

Multiple loci should not be anticipated (and locus symbols assigned) on merely clinical grounds, when the clinical phenotypes may well be due to allelic or modifying genes. Frequent changes in locus symbols should be avoided, so that symbols associated with clinical syndromes should be as familiar as possible yet consistent with the rule that molecular definition takes precedence over clinical definition.

Allele Symbols

1. The above guidelines for notation should be followed with respect to designating alleles, except that in the case of a rare genetic disease, where little may be known concerning the gene or the normal and mutant alleles, historical symbols may be necessary. Examples: *N* (normal "wild" type); *D* (dominant); *R* (recessive).

2. When describing sex-linked clinical traits in hemizygous males, the genotype designation is limited to the normal or mutant X allele; e.g., in *CBP*R/Y* (for protanopic male), the *R* designates a recessive allele, and the *Y* identifies the male and the diploid nature of the organism.

3. Phenotypic information which cannot be justified on genetic grounds should be assigned to allele symbols and not the locus symbol. Thus, type 1 and type 2 Smith-Lemli-Opitz syndrome should provisionally be attributed to alleles *SLO*1* and *SLO*2,* respectively, not to undemonstrated *SLO1* and *SLO2* loci.

4. Continuity with tradition can be maintained by assigning the first several characters (up to three) of the name of the allele to the clinical phenotype, followed by any molecular designations which may be known (Table VII).

Table VII. Examples of clinical gene and allele symbols

Disease or trait	Locus symbol	Allele symbol
hemoglobin S, structure proven	*HBB*	**S6V*
hemoglobin S, structure not proven	*HBB*	**S*
hemoglobin M (Osaka)	*HBB*	*M58Y
hemoglobin C (Harlem)	*HBB*	**C6V73N*
phenylketonuria (classical)	*PAH*	**PKU*
osteogenesis imperfecta type IV due to COL1A2 defect, type 4	*COL1A2*	**OI4*
Tay-Sachs, Ashkenazi form	*HEXA*	**TSA*
Sandhoff disease	*HEXB*	**SF*
Lesch-Nyhan syndrome	*HPRT*	**LN*
gout due to HPRT defect	*HPRT*	**G*

Phenotypes

Phenotypes characteristic of particular heterozygotes (genetic compounds) are designated in the usual way; for example, HBB S,C for sickle cell-hemoglobin C disease.

GENETIC NOMENCLATURE OF BLOOD GROUPS

Background: At several Human Gene Mapping Workshops there have been numerous requests for the Nomenclature Committee to propose a simplified, uniform, and more genetic nomenclature for blood group notations to be be used for presenting genetic information about an individual and in family studies. Recognizing the long history of blood group notations and blood banking symbols, we propose for consideration a system which requires few changes to historical precedents yet still conforms to the gene nomenclature guidelines (above).

The modifications proposed here are concerned only with symbols for genotypes and phenotypes in the genetic context and are not intended to replace blood bank usages. As geneticists, our concerns deal with the problem of genetic notations and not with the notation of blood group *specificities*. The latter issue is a technical one to be dealt with by experts in the blood banking field, for it is properly their concern. Over the years, as each system has been defined, certain conventions have arisen so that the notational symbols used in speaking or writing about each blood group system may be curiously distinct from one another. However, the red cell antigen systems constitute a similar set of genes, similar in that the same basic properties are characteristic of *all* red cell antigen systems and that the same laboratory techniques are applied to all of them. Therefore, there is no *a priori* reason why a uniform gene nomenclature can not be devised for the genes to which the red cell antigens are attributed.

In accordance with the general guidelines presented in this document, the modifications proposed for the blood group systems are minor, simplify past usage, and begin to utilize gene names and symbols incorporating gene-function information where available.

Recommendations

1. Blood group symbols for reporting genetic information should follow the human gene nomenclature guidelines as established by the "1979" report (Shows, *et al.*, 1979). Genetic examples for blood groups are given in Table VIII.

2. It is recommended that the ISBT (International Society of Blood Transfusion) proposal (Allen *et al.*, 1983; Allen, 1983; Lewis *et al.*, 1985) to assign upper-case letters and numbers to the blood group specificities be accepted and applied, except that for historical reasons, notation for the genes of the ABO system continue unchanged. In the RH system, allelic designations derived from the normal short form representations (such as *R1*, *R2*, etc.) may be used, as in Table VIII.

3. In systems which are characterized by the association of several specificities apparently encoded by linked genes, either a short form representation or the ISBT numbers may be used.

When the ISBT numbers are used to itemize each of the specificities, only those specificities which have been positively identified shall be noted. A listing of the antisera used (by specificity) should always be presented in the Materials and Methods section of publications.

4. The absence of all known specificities should be assigned the phenotype of "O".

5. It is recommended that as the gene products of specific loci are identified, the blood group loci be renamed in light of the new information, retaining, however, the conventional system name.

a) MN system - It is recommended that we continue with the traditional nomenclature since the complete molecular descriptions of the M and S loci are not available. Consideration should be given to renaming these loci when their functions become known.

b) We recommend that the locus for the Bombay phenotype be renamed fucosyl transferase 1 (*FUT1*) and the two alleles at that locus be designated **H* and **O* (following guidelines detailed in the previous section). When the secretor locus is more fully defined, it probably will be renamed fucosyl transferase 2 (*FUT2*).

6. In the GM system, individual specification may be used to record positive assignments only, or short forms may be used to note particular haplotypic contributions. A list of recommended short forms is given in Table VIII.

7. The biochemical nomenclature for Pl and P+ should be referred to the ISBT for consideration by a panel of experts.

Table VIII. Examples of blood group locus, allele, and phenotype symbols

System	Locus	Alleles	Genotype	Phenotype
ABO	*ABO*	*ABO*A1* *ABO*B*	*ABO*A1/*B*	ABO A1,B
Kidd	*JK*	*JK*A* *JK*B*	*JK*A/B*	JK A,B
H	*FUT1*	*FUT1*H* *FUT1*O*	*FUT1*H/*O*	FUT1 H
Colton	*CO*	*CO*0* *CO*1* *CO*2*	*CO*1/*0*	CO 1
RH	*RH*	*RH*1,2,5* *RH*1,3,4*	*RH*1,2,5/*1,3,4* *or*	RH 1,2,3,4,5
		*RH*R1* *RH*R2*	*RH*R1/*R2*	RH R1,R2
MNS	*MNSM*	*MNSM*1* *MNSM*2*	*MNSM*1 MNSS*3/MNSM*2 MNSS*4*	M 1,2 S 3,4
	MNSS	*MNSS*3* *MNSS*4*	Phase known: *MNSM*1 S*3/M*2 S*4* Phase not known: *MNSM*1/*2, MNSS*3/*4*	
IGH (GM)	*IGHG1*	*IGHG1*A* *IGHG1*F*	*IGHG1*A/IGHG3*G/IGHG1*F IGHG3*B* Phase known:	G1 A,F G3 B,G or
	IGHG3	*IGHG3*B* *IGHG3*G*	*IGHG1*A G3*G/G1*F G3*B* Phase not known: *IGHG1*A/*F, IGHG3*G/*B*	GM A,B,F,G

CELL SURFACE ANTIGENS (other than blood groups)

Genes encoding cell surface molecules of unknown function have been assigned to human chromosomes under various designations based on different methodologies and different levels of understanding these molecules. Since cell surface antigens have been primarily detected by polyclonal or monoclonal antibodies, each may not be a unique molecule encoded by a discrete gene. Often, cell surface molecules identified by one methodology prove to be identical to molecules identified by a different method. We have previously proposed gene symbol guidelines for markers whose functions are not initially known (Shows and McAlpine, 1982). These guidelines can be applied to cell surface antigens, external membrane proteins, proteins identified by 2-D electrophoresis, antigens recognized by monoclonal antibodies, and anonymous undefined DNA segments. These guidelines follow a three-stage hierarchical procedure for naming a gene (see Table IX).

In Phase 1, an initial local provisional symbol is given. In Phase 2, further characteristics of the marker become known, such as molecular weight or polymeric structure, and distinction from previously described markers is determined. In the final stage, Phase 3, the biological function of the gene is determined. This order of local, provisional, and then functional gene symbol provides a mechanism for naming markers when their functions are not initially known; for example, the cell surface antigens.

Phase 1: The local symbol is determined by the describing laboratory. If an antigen is recognized by a monoclonal antibody in a laboratory at the Imperial Cancer Research Fund and is the third such antibody described, the local gene symbol is *MIC3* (M: monoclonal; IC: Imperial Cancer Research Fund; 3: third antigen). If a polyclonal antibody or other procedure identifies a cell surface antigen, the procedure is the same except that the symbol is preceded by an "S" (surface antigen) (see Shows

and McAlpine, 1982). Each cell surface antigen will be identified by a laboratory designation until it is considered to be encoded by a unique gene and is given a formal designation (see below) and the assignment verified at a Human Gene Mapping Workshop.

Phase 2: During this phase, the cell surface molecule is further characterized and given a provisional gene symbol based on physicochemical-immunological-clinical criteria. If identity is determined for two or more previously described antigens, a new symbol is established and the former local symbols are retired. For example, if the antibodies which define the S15 and MER50 antigens encoded by genes on chromosome 12 appear to recognize the same molecule and, after comparison to products of all other genes previously assigned to chromosome 12, this molecule appears unique, the gene encoding this molecule will receive a new designation: M12S1, where M = membrane; 12 = chromosome number; S = surface; and 1 is the sequential number of such markers on chromosome 12 (see Table XI). These gene symbols will be determined by a cell surface antigen nomenclature committee and will be approved at the next Human Gene Mapping Workshop.

After careful assessment at workshops sponsored by the International Union of Immunological Societies and the World Health Organization (IUIS-WHO), human leukocyte differentiation antigens defined by monoclonal antibodies have been grouped into clusters of differentiation (CD). Genes encoding various of these CDs have been assigned to human chromosomes. It is our recommendation that if any of the previously assigned genes for cell surface antigens appear identical to those chromosomally assigned CD groups, that the CD designation be used (Table IX). In the case of cell surface antigens composed of multiple polypeptides, the gene symbol will be followed by "P" (for polypeptide) and an upper-case letter to designate the chain of the molecule assigned to the particular chromosome (see Table IX). Once genes for cell surface molecules or their component polypeptides have been chromosomally assigned on the basis of cloned DNA probes, designations will be used as appropriate until Phase 3 is reached. Genes encoding cell surface molecules, whether initially identified as antigens, external membrane proteins, or cloned DNA probes, will be designated by the appropriate M (membrane) or CD nomenclature.

Phase 3: When the function of the gene becomes known, the gene is named according to the guidelines for such function (see Table IX).

Following these guidelines, Table X lists examples of cell surface antigens and their gene nomenclature.

Table IX. Stages of nomenclature for genes encoding cell surface antigens

Stage 1 Local Laboratory Symbol - initial description

M (monoclonal), laboratory initials, number (sequential), * (allele)

Example: Gene defined by monoclonal antibody:
MSK11
*MSK11*1* (hypothetical)

S (surface), laboratory initials, number (sequential), *(allele)

Example: Gene defined by polyclonal antibody:
SRP1 (hypothetical)

Stage 2 Provisional Gene Symbol - approved gene symbols replace laboratory designations

A. M (membrane), chromosome number, S (surface), sequential number, * (allele)

Laboratory gene symbol becomes	Approved gene symbol
MSK11	*M5S1*
MSK30	*M14S1*
MIC2	*MXYS1*
MIC5	*MXS1*
S7 same as *S6*	*M7S1* (hypothetical)
MRP3 same as *MW15* (hypothetical)	*M13S2* (hypothetical)

B. CD (cluster differentiation [IUIS-WHO]), sequence number, upper-case letter to designate a subcluster, P (polypeptide), and upper-case letter to designate the chain of a multimeric molecule.

Current symbol becomes	CD symbol
MIC3	*CD9*
LEU1	*CD5*
T3D	*CD3PD*
T3E	*CD3E*
M12S1 (hypothetical)	*CD50* (hypothetical)

Stage 3 Functional Gene Symbol - M (membrane) and CD symbols are retired when a functional gene symbol is assigned.

Provisional symbol becomes	Functional gene symbol
CD35	*CR1*
M7S1 (formerly *S6* and *S7)*	*EGFR*

Table X. Cell surface antigens: Examples of antigens and their gene nomenclature

HGM Symbol	Markers	Mode	Status	Chromosomal Assignment
CD2	leukocyte antigen 5 (CD2p50, identified by monoclonal antibodies 9.6, T1 Leu5 and 35.1)	S	P	4
CD3PD	TiT3 complex, delta polypeptide (CD3p19-29)	S, RE	P	11p23-pter
HLADRB1	major histocompatibility complex, class II antigens, beta gene 1	S, RE	C	6p21.3
MDF1	antigen identified by monoclonal antibody A-3 A4	S	P	4
MDU1	antigen identified by monoclonal antibodies 4F2, TRA1.10, and TROP4	S	P	11q
MER1	antigen identified by monoclonal antibodies W6/34, 5C1, etc. (formerly lethal antigen 1)	D, S	C	11pter-p13
MFI3	antigen (glycoprotein, MW 86K) identified by monoclonal antibody 2B2	S	P	21
MIC10	antigen identified by monoclonal antibody TRA-2-10	S	P	1
MSK1	antigen identified by monoclonal antibody AJ9	S	P	1p22
THY1	thy-1 cell surface antigen	A	C	11q22.3

Testing Recommendation

The status of currently assigned genes coding for cell surface antigens, as well as those proposed for inclusion at a following Gene Mapping Workshop, should be considered only after an attempt has been made to determine their uniqueness and/or relatedness. Such definition of genes encoding cell surface antigens will be possible only if antibodies recognizing antigen specificities which appear to map to the same genetic region are exchanged among relevant laboratories for assessment of the reactivity of these reagents. Data obtained from these exchanges will be evaluated and used to devise formal recommendations to the Human Gene Mapping Nomenclature Committee. Examples are *S1 = MER1 = MIC1*; *S2 = MIC8 = MF16 = MDU1*; and *MSK5 = MSK6 = MIC2*.

(Further information regarding this testing process may be obtained from Claude Boucheix, Hopital Paul Brousse 16, 94800 Villejuif, France, or Barbara Knowles, Wistar Institute, Philadelphia, PA.)

GENETIC NOMENCLATURE OF DNA SEGMENTS

Naming Arbitrary DNA Fragments and Loci

The nomenclature for arbitrary DNA segments of unknown function was described at the HGM6 by Skolnick and Francke (1982) and Shows and McAlpine (1982). Modifications to this nomenclature were adapted by Willard *et al.* (1985) and include the following guidelines for determining each part of the symbol:

Part I: "D" for DNA

Part II: "1,2,.....22,X,Y,XY,N" for the chromosomal assignment, where XY is for segments homologous on the X and Y chromosomes, and "N" is for segments detected on multiple chromosomes other than X and Y.

Part III: A symbol indicating the complexity of the DNA segment detected by the probe, with "S" for a unique DNA segment, and "Z" for repetitive DNA segments found at a single chromosome site or "F" for small undefined families of homologous sequences found on multiple chromosomes.

Part IV: "1,2,3.....", a sequential number to give uniqueness to the above concatenated symbols.

Part V: When referring to a particular defined member of a small, multi-chromosome family, the site can be designated with the suffix "S1, S2, S3,...". Thus, the sites on chromosome 1 and 3 homologous to the DNF15 family are referred to as DNF15S1 and DNF15S2.

Proposed modifications and additions to these basic guidelines:

A. It is proposed that the DNF terminology for designation of families of related sequences be altered by replacing the N with the chromosome on which the specific segment is located, e.g., D11F10S2, which is a member of the same family as D3F10S3, would now include the indication of their chromosomal localization. The previous designation for these segments would have been DNF10S2 and DNF10S3.

B. Regarding the use of the symbols "P" for pseudogenes and "L" for "like" or cross-hybridizing sequences, the previous recommendations (Shows *et al.*, 1984; McAlpine *et al.*, 1985) still seem appropriate. We suggest that the "P" should only be used if there is strong evidence that the locus is actually a pseudogene. Such evidence might be provided by sequence data or hybridization results (i.e., for a processed pseudogene). In the absence of such data, the less specific designation "L" should be used.

C. Policy for naming expanded DNA segments.

A hierarchical nomenclature for designation of expanded DNA segments is proposed. This new nomenclature will be required, for example, when two previous DNA segments are connected by chromosome walking. There are four levels in the hierarchy, increasing in length:

1. Segments (see above part III).
2. Extended segments, which are produced by "chromosome walking" from an original, mapped segment to generate a contiguous stretch of cloned DNA spanning at least 100 kb.
3. Merged segments, which result from the molecular joining of two independently derived segments or extended segments. Merged segments will usually be several hundred kb in length.
4. Regional segments, which result from the molecular joining of one merged segment with any other independently derived segment. Regional segments will usually be approximately one million basepairs in length.
5. Joined segments will result from the molecular connection of two regions.

The nomenclature for expanding DNA segments will consist of replacing the S designation with E, M, R, or J (above). The number of the higher order members of the hierarchy will be derived from one of the original segment numbers in order to preserve the information regarding relationship to the original segment. Specifically, the following rules apply to the assignment of ascension numbers:

1. When unequal order segments are involved, the number of the highest order segment is conserved, i.e., J R M E S.
2. When equal order segments are involved, select the lower number.
3. When one of the original segments is designated by a locus name, that designation takes precedence.
4. When both segments have locus designations, the name for the new, larger segment will be determined by the nomenclature committee.

Example 1. To illustrate the system, consider two segments from chromosome 11, designated D11S83 and D11S16. If chromosome walking from D11S83 generated a contiguous stretch of DNA spanning more than 100 kb, then the entire stretch would acquire the new designation D11E83. (Note that the E number is the same as the S number rather than a new sequential number.)

If a further expansion of D11E83 resulted in overlap with the independent segment D11S16, then the resulting, contiguous merged segment would be given the designation D11M83. (Note the acquisition of number of the higher order precursor, the E.) Had the merger involved two walked segments, D11E83 and D11E16, then the resulting merged segment would have been designated D11M16, reflecting acquisition of the lower ascension number in the case of merging segments of the same hierarchical order.

Any further merging of D11M16 with additional segments, extended segments, or merged segments, in either direction, results in the generation of a larger unit designated a regional segment. The same rules for acquisition of ascension numbers apply to regional segments: if two like elements are merged, the region acquires the lower number of the two; for merging of two elements of unequal order, the number of the higher order segment is retained. This new region, for example, might be designated D11R16.

Example 2. If one of the segments includes a named locus, the resulting higher order segments retain the locus name as the last component of its nomenclature. For example, if D11M16 were to merge with the catalase locus (CAT), the resulting region would be designated: D11RCAT. If two identified loci are involved, refer the issue of how to designate the region to the Nomenclature Committee.

Additional examples of designations for expanded DNA segments

First segment	Second segment	Expanded segment
Example 1		
D11S83	"chromosome walk"	D11E83
D11S16	D11E83	D11M83
D11M83	D11S12	D11R83
D11R83	D11S40	D11R83
D11R83	D11R21	D11J21

Example 2

D1S5	D1S63	D1M5
AMY1	"chromosome walk"	D1EAMY1
D1M54	D1EAMY1	D1MAMY1
D1S32	D1MAMY1	D1RAMY1
D1RAMY1	D1R17	D1J21

Nomenclature for Molecularly Defined Alleles

With regard to DNA segments, each allele is defined by a specific probe, a specific restriction enzyme, and particular fragment length(s). Alleles are designated by a letter followed by an ascension number, as described in HGM8.

Allele designations

locus	alleles	probe	enzyme	fragments
D4S20	**A1*	CG6	*Sau*96	0.85
	**A2*			0.54, 0.31
	**B1*	CG6	*Bgl*II	6.5
	**B2*			*5.5*

Alleles of cloned genes whose molecular characteristics have been determined may be assigned symbols indicative of their structure followed by a sequential ascension number. Commonly used names may be retained in the allele designation followed by a molecular designation such as the following changes in codon number and amino acid or codon nucleotides:

DEL, deletion
SP, splicing alteration
FS, frameshift
TERM, termination codon
5PR, alteration in the 5' region upstream of the coding portion of the gene
3PR, alteration in the 3' region downstream of the gene
IV, alteration in the intervening sequence

Allele designations for molecular variation

locus	alleles	description
HBB	**S*	sickle cell allele
HBB	**S6V*	alternate designation
HBB	**S6GTG*	alternate designation
HBB	**DEL1*	first designated deletion
HBB	**DEL2*	second designated deletion

Note: A dictionary of alleles will be required for complete molecular information for each allele, such as noncoding base substitutions and the location of deletions.

PROTOONCOGENES

Acutely transforming mammalian and avian retroviruses possess specific genes (viral oncogenes) responsible for malignant conversion of cells. Mammalian cells, and cells from other organisms, possess normal cellular genes (protooncogenes) that are homologous to these acutely transforming retroviral oncogenes. Currently, there are many normal cellular protooncogenes that are known. Initially, these, and as yet unidentified protooncogenes, will be identified by the symbol of the retrovirus bearing the viral oncogene. For example, the oncogene obtained from the Abelson murine leukemia virus is designated v-*abl*, and the human cellular protooncogene will be designated *ABL* following human gene nomenclature guidelines. Additionally, the Harvey ras sarcoma virus oncogene is designated v-Ha-*ras*-1; the human protooncogene will be designated *HRAS1*. Gene symbols for other protooncogenes are *MOS*, *MYC*, *SRC*, *KRAS2*, *FES*, and *FMS* (see McAlpine *et al.*, 1985; Table I).

When the function of the protooncogenes becomes known, then the gene symbol acquires the functional symbol. For example:

SIS becomes *PDGFB* (platelet derived growth factor beta polypeptide chain)

ERBB becomes *EGFR* (epidermal growth factor receptor)

VIRUS - ASSOCIATED MARKERS

A hierarchical gene nomenclature for virus-associated markers and viral insertion sites has been described in detail (Shows *et al.*, 1979). Four categories of virus interaction with the human genome can be identified for the purpose of gene nomenclature and mapping: (1) integration, (2) sensitivity, (3) regulation, and (4) chromosome modification.

Integration is used to define sites at which a virus is integrated into human chromosomes of cells following an infection. The interaction does not describe an endogenous viral gene nor is it restricted to specific sites in the human genome. A viral integration site is given a gene symbol only when (1) a unique flanking DNA segment has been isolated and mapped, (2) a preferred site exists for integration following exogenous infection, or (3) there are endogenous viral sequences in the human genome. An example is hepatitis B virus integration site 1 (*HVBS1*) on human chromosome 11p.

Sensitivity describes the susceptibility of a cell to virus infection and defines the expression of a human gene; for example, human coronavirus sensitivity (*HCVS*) on chromosome 15q.

Regulation involves genes that alter in any way the expression of specific viruses.

Modification identifies site-specific morphological changes in human chromosomes resulting from virus

infection; for example, the adenovirus type 12 modification of chromosome 1q (*A12M1*).

Gene symbols for a viral associated gene or DNA segment should begin with the initial letter of the virus name. The letter "V" for virus should be included in the symbol and could be the first letter if necessary. A letter should be incorporated into the symbol to designate one of the four categories above: "I" (integration); "S" (sensitivity); "R" (regulation); or "M" (modification). The gene symbol should include the specific designation for a virus which follows standard viral terminology and which also conforms to gene symbol recommendations described above.

Examples:

A12M1 - Adenovirus type 12 modification site 1
BEVI - Baboon endogenous virus, M7, integration
E11S - ECHO 11 virus sensitivity
HCVS - Human coronavirus sensitivity
HVBS1 - Hepatitis B virus integration site 1
HV1S - Herpes simplex virus type 1 sensitivity
PVS - Polio virus sensitivity

MARKERS WHOSE FUNCTIONS ARE NOT INITIALLY KNOWN

It is expected that in addition to cell surface antigens and undefined DNA segments (above), there will be described numerous gene markers whose function is not initially understood but which will be identified. External membrane proteins and polypeptides identified by gel electrophoresis are examples of such markers.

Gene symbols for these markers should follow the step-wise procedure described previously (Shows and McAlpine, 1984, and above) for cell surface antigens and arbitrary DNA segments.

Examples: *M130* external membrane protein of 130KD
PSF3 protein spot in 2-D gels (MW 37KD) from a particular laboratory with its sequential number

FRAGILE SITES

Fragile sites are points on chromosomes which are seen as breaks or gaps on metaphase preparation (see Cytogenetic Committee Report, HGM9). Fragile sites may be present in some individuals and segregate in a Mendelian manner (the rare fragile sites) or be present on all chromosomes (the common fragile sites). Fragile sites are of use in gene mapping studies: they can be used in conjunction with *in situ* hybridization studies, and the rare ones can be used for linkage studies with other genetic markers.

Fragile sites have an accepted cytogenetic nomenclature (ISCN, 1985) and a locus designation (McAlpine *et al.*, 1985). Their gene nomenclature uses the letters FRA followed by a sequential letter of the alphabet. Each new fragile site described on a chromosome is then given the next letter. For example, the distamycin A sensitive fragile site on the long arm of chromosome 16 is designated *FRA16B*. Once a gene symbol has been attached to a fragile site, it should not change.

To further specify fragile sites, an allele system can be used. This incorporates the letters R for rare or C for common fragile site (arbitrarily, rare fragile sites are designated as being on less than 50% of chromosomes in the population) and the current chromosomal location. The non-fragile allele is designated N (normal). Thus, to specify a chromosome with the rare distamycin A type on the long arm of chromosome 16, the form *FRA16B*RQ221* would be used (see Table XI).

There are problems in using alternative systems of nomenclature which include the chromosomal location of the fragile site. Some chromosome bands contain more than one fragile site, and with changes in location, it could appear that new fragile sites are being described.

Table XI. Fragile site genetic nomenclature

Fragile site type	Gene Symbol	Alleles
common, aphidicolin type at 16q22.1	*FRA16C*	**C16Q221* **N*
rare, distamycin A type at 16q22.1	*FRA16B*	**R16Q221* **N*
folate type,at Xq27.3 associated with mental retardation	*FRAXA*	*RQ273 **N*

GUIDELINES FOR MITOCHONDRIAL GENE NOMENCLATURE

Each mitochondrion contains circular chromosomes composed of heavy strands (H) and light strands (L). The mitochondrial chromosome has been sequenced and contains 16,569 basepairs. The several genes that have been mapped, therefore, can be precisely located. The location is identified by the starting and ending nucleotide.

Since some of the mitochondrial genes perform similar functions as nuclear genes and would have the same gene symbols, it is necessary to distinguish mitochondrial genes from nuclear genes. All mitochondrial genes begin with "MT", and the remaining part of the symbol follows the gene nomenclature guidelines already described. For example, the tRNA's are identified by the single-letter designation for the corresponding amino acid (Table II). Table XII tabulates the human mitochondrial genes following the guidelines proposed here.

Table XII. Human gene nomenclature for the mitochondrial genome

Gene symbol	Marker name	Strand H, heavy L, light	Nucleotide location
MTATP6	ATPase subunit 6	H	8527-9207
MTATP8	ATPase subunit 8	H	8366-8572
*MTCOX*1*	cytochrome c oxidase subunit I	H	5904-7444
*MTCOX*2*	cytochrome c oxidase subunit II	H	7586-8262
*MTCOX*3*	cytochrome c oxidase subunit III	H	9207-9990
MTCYB	cytochrome b	H	14747-15887
MTHSP	H-strand promoter	H	545-567
MTLSP	L-strand promoter	L	392-435
MTND1	NADH dehydrogenase subunit 1	H	3307-4262
MTND2	NADH dehydrogenase subunit 2	H	4470-5511
MTND3	NADH dehydrogenase subunit 3	H	10059-10404
MTND4	NADH dehydrogenase subunit 4	H	10760-12137
MTND4L	NADH dehydrogenase subunit 4L	H	10470-10766
MTND5	NADH dehydrogenase subunit 5	H	12337-14148
MTND6	NADH dehydrogenase subunit 6	L	14149-14673
MTOHR	Origin of H-strand replication	H	about 200
MTOLR	Origin of L-strand replication	L	5729-5805
MTRNR1	12S rRNA	H	648-1601
MTRNR2	16S rRNA	H	1671-3229
MTTA	tRNA, alanine	L	5587-5655
MTTC	tRNA, cysteine	L	5761-5826
MTTD	tRNA, aspartic acid	H	7518-7585
MTTE	tRNA, glutamic acid	L	14674-14742
MTTF	tRNA, phenylalanine	H	577-647
MTTG	tRNA, glycine	H	9991-10058
MTTH	tRNA, histidine	H	12138-12206
MTTI	tRNA, isoleucine	H	4263-4331
MTTK	tRNA, lysine	H	8295-8364
MTTL1	tRNA, leucine-1	H	3230-3304
MTTL2	tRNA, leucine-2	H	12266-12336
MTTM	tRNA, methionine	H	4402-4469
MTTN	tRNA, asparagine	L	5657-5729
MTTP	tRNA, proline	L	15955-16023
MTTQ	tRNA, glutamine	L	4329-4400
MTTR	tRNA, arginine	H	10405-10469
MTTS1	tRNA, serine-1	L	7445-7516
MTTS2	tRNA, serine-2	H	12207-12265
MTTT	tRNA, threonine	H	15888-15953
MTTV	tRNA, valine	H	1602-1670
MTTW	tRNA, tryptophan	H	5512-5576
MTTY	tRNA, tyrosine	L	5826-5891

HOMOLOGOUS GENES IN DIFFERENT SPECIES

It is recommended that mammalian geneticists make every possible attempt to standardize terminology for homologous genes among different species. In particular, before choosing new gene symbols, it would be desirable to identify and use whenever possible preexisting homologous gene symbols in other species. Furthermore, we recommend as an editorial policy for genetic journals that authors must state when naming a new gene that they have consulted the standardized nomenclature of a species and have attempted to conform to homologous gene symbols of other species.

Sources for finding gene symbols of other mammalian species include *Genetic Variants and Strains of the Laboratory Mouse* (1981), M. Green; *Mouse News Letter II* (Feb. 1987); *Rat News Letter*; *Peromyscus Newsletter*;

Mendelian Inheritance in Man (1986), V. McKusick; *Human Gene Mapping Workshops*; *Cytogenetics and Cell Genetics*; and *Genetic Maps* (1987), S. J. O'Brien.

Determination of Homologous Gene Products

Three types of information may be available for the ascertainment of homology between gene products identified in different species: (1) molecular structure, as indicated by similarities in amino acid sequence, molecular weight, immunological cross-reaction, subunit structure, and formation of functional heteropolymeric molecules in interspecific somatic cell hybrids; (2) biological or biochemical function, such as tissue distribution, developmental time of appearance, similar pleiotropic effects, subcellular location, substrate specificity of enzymes, and response to inhibitors; (3) comparative gene linkage, that is, close linkage of homologous genes in different species as presumptive evidence for homology. This is particularly evident for X-linked loci.

Standardization of Nomenclature of Homologous Genes

1. Homologous genes should have the same gene and allele nomenclature.

2. When a locus or series of genes has been defined in one species, and it is reasonable to expect that in the future a homologous gene will be identified in man, we recommend that the designated symbol be "reserved" for the human loci. This should be done, also, in other species for genes first identified in man.

3. When necessary to distinguish the species of origin for homologous genes with the same gene symbol, the three-letter code for different species, comprising the initial letter of the specific name already established by the Committee on Standardization in Human Cytogenetics, is recommended. The code would be for use in publications only and not incorporated as part of the gene symbol. The species designation would be added as a prefix to the gene symbol. For example, HSA signifies *Homo sapiens* and MMU stands for *Mus musculus*. Examples of using the species designation with the gene symbol; human loci: *(HSA)G6PD*; *(HSA)HBB*; *(HSA)ALB*; homologous mouse loci: *(MMU)G6pd*; *(MMU)Hbb*; *(MMU)Alb*.

4. The agreement between human and mouse gene nomenclature for many homologous gene loci should be continued and extended.

MAINTAINING A STANDARDIZED HUMAN GENE NOMENCLATURE

The human gene nomenclature committee (T. B. Shows and P. J. McAlpine, Co-chairs) is a standing committee that functions through the Human Gene Mapping Workshops. This committee has the responsibility of (1) establishing guidelines for naming genes; (2) approving all gene symbols (not DNA segments or fragile sites); (3) bringing individual genetic systems into compliance with the standardized nomenclature; (4) maintaining a list of gene symbols for mapped genes; (5) maintaining a list of retired gene symbols; (6) maintaining a list of reserved gene symbols; (7) coordinating guidelines for markers generated by new methodology (i.e., DNA segments and antigens identified by monoclonal antibodies); (8) resolving gene nomenclature issues among laboratories and for different species; (9) creating subcommittees to establish nomenclature for sets of new and problem markers; (10) meeting on a regular basis for updating and maintaining human gene symbols; and (11) promoting and publishing human gene nomenclature guidelines. In view of the increasing interest and activity in human genetics at both the research and applied levels, an international standing committee which functions on a continuous basis is now a necessity.

ACKNOWLEDGEMENTS

We wish to thank the Howard Hughes Medical Institute for funding the Human Gene Nomenclature Meeting necessary to create this document.

REFERENCES

Allen, F.H. Jr.; Anstee, D.J.; Bird, G.W.G. *et al.*:ISBT working party on terminology for red cell surface antigens. Preliminary report. Vox Sang. 42:164-165 (1983).

Allen, F.H.Jr.: Report of the ISBT working party on terminology for red cell surface antigens. ISBT Newsletter. 18:(1983).

Alper, C.A. and Parkman, R.: Serum proteins and other genetic markers of the blood. *In* D. G. Nathan and F. S. Oski, eds.: *Hematology of Infancy and Childhood*, pp. 818-849 (W. B. Saunders Co., Philadelphia 1974).

Enzyme Nomenclature. Recommendations (1984) of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzyme Catalysed Reactions. (Academic Press, Inc., Orlando, San Diego, New York, London, Toronto, Montreal, Sydney, Tokyo 1984).

Green, M.C., editor: *General Variants and Strains of the Laboratory Mouse*. (Gustav Fischer Verlag Publisher, 1981).

Instructions to Authors. J. Biol. Chem. 251:1-10 (1976).

ISCN 1985: An International System for Human Cytogenetic Nomenclature (1985). Cytogenetic and Cell Genetics, S. Karger, publisher.

Lehninger, A. L.: Biochemistry, 2nd Ed., p. 72 (Worth Publishers, Inc., New York, 1975).

Lewis, M.; Allen, F.H. Jr.; Anstee, D.J.; Bird, G.W.G.; Brodheim, E.; Contreras, M.; Crookston, M.; Dahr, W.; Engelfriet, C.P.; Giles, C.M.; Issitt, P.D.; Jorgensen, F.; Kornstad, L.; Leikola, J.; Lubenko, A.; Marsh, W.L.; Moore, B.P.L.; Morel, P.; Moulds, J.J.; Nevanlinna, H.; Nordhagen, R.; Rosenfield, R.E.; Sabo, B.; Salmon, C.; Seidl, S.; Tippett, P.A.; Walker, R.H. and Yasuda, J.:

ISBT working party on terminology for red cell surface antigens: Munich report. Vox Sang 49:171-175 (1985).

McAlpine, P.J., Shows, T.B., Miller, R.L., and Pakstis, A.J. The 1985 catalog of mapped genes and report of the nomenclature committee. Cytogenet. Cell Genet. 40:8-66 (1985).

McKusick, V. A.: Mendelian inheritance in man, 7th Ed. (The Johns Hopkins University Press, Baltimore, 1986).

Mouse News Letter 77: Feb., 1987. (Oxford University Press, Oxford, UK)

O'Brien, S.J., ed. Genetic Maps 4:1-755 (Cold Spring Harbor Laboratory, 1987).

Putnam, F.: Alpha, beta, gamma, omega-the roster of the plasma proteins. *In* F. W. Putnam, ed.: *The Plasma Proteins*, Vol. 1, pp. 58-125 (Academic Press, New York, 1975).

Recommendations for nomenclature of hemoglobins. Hemoglobin 3:1-11, 1979.

Shows, T.B.; Alper, C.A.; Bootsma, D.; Dorf, M.; Douglas,T.; Huisman, T.; Kit, S.; Klinger, H.P.; Kozak, C.; Lalley, P.A.; Lindsley, D.; McAlpine, P.J.; McDougall, J.K.; Meera Khan, P.; Meisler, M.; Morton, N.E.; Opitz, J.M.; Partridge, C.W.; Payne, R.; Roderick, T.H.; Rubinstein, P.; Ruddle, F.H.; Shaw, M.; Spranger, J.W. and Weiss, K.: International system for human gene nomenclature (ISGN, HGM5) Cytogenet. Cell Genet. 25:96-116 (1979).

Shows, T.B. and McAlpine, P.J. The 1981 catalogue of assigned human genetic markers and report of the nomenclature committee. Cytogenet. Cell Genet. 32:221-245 (1982).

Shows, T.B.; McAlpine, P.J., and Miller, R.L.: The 1983 catalog of mapped human genetic markers and report of the nomenclature committee. (HGM7) Cytogenet. Cell Genet. 37:340-393 (1984).

Skolnick, M.H. and Francke, U. Report of the committee on human gene mapping by recombinant DNA techniques. Cytogenet. Cell Genet. 32:194-204 (1982).