Associations between human disease genes and overlapping gene groups and multiple amino acid runs

Samuel Karlin*†, Chingfer Chen*, Andrew J. Gentles*, and Michael Cleary‡

Departments of *Mathematics and ‡Pathology, Stanford University, Stanford, CA 94305

Contributed by Samuel Karlin, October 30, 2002

Overlapping gene groups (OGGs) arise when exons of one gene are contained within the introns of another. Typically, the two overlapping genes are encoded on opposite DNA strands. OGGs are often associated with specific disease phenotypes. In this report, we identify genes with OGG architecture and genes encoding multiple long amino acid runs and examine their relations to diseases. OGGs appear to be susceptible to genomic rearrangements as happens commonly with the loci of the DiGeorge syndrome on human chromosome 22. We also examine the degree of conservation of OGGs between human and mouse. Our analyses suggest that (*i*) a high proportion of genes in OGG regions are disease-associated, (*ii*) genomic rearrangements are likely to occur within OGGs, possibly as a consequence of anomalous sequence features prevalent in these regions, and (*iii*) multiple amino acid runs are also frequently associated with pathologies.

The study of the association between human diseases and their underlying molecular causes is of considerable medical importance. Some disease-associated genes represent essential genes whose functional impairment is deleterious. However, nonessential genes may also induce disease phenotypes by means of dominant-negative effects or gain of toxic function. Diseases caused by deletions in noncoding regions may relate to gene regulation. Several genetic disease mechanisms can be distinguished, including (*i*) haplo-insufficiency (1), wherein loss of one gene copy results in insufficient gene product for normal function. In general, haploinsufficiency indicates that both alleles are necessary for proper biological function. In diseases such as Down's syndrome and Charcot-Marie-Tooth disease, gene overexpression can result from trisomy. In these cases it is the number of functional gene copies which is critical. (*ii*) Altered chromosome structure (e.g., segmental duplications and deletions, break point clusters, inversions, and translocations) can be related to disease. (*iii*) Toxic gain or loss of function can arise through alteration of protein binding sites, protein misfolding, or inappropriate aggregation as occurs in the polyglutamine trinucleotide repeat diseases. Our studies emphasize diseases associated with chromosomal sequence anomalies, the occurrence of overlapping gene groups (OGGs), and genes encoding multiple long amino acid runs.

Overlapping Gene Groups

There appears to be a strong correlation between genes associated with human diseases and overlapping groups of genes and/or genes that encode multiple amino acid runs (see examples below). OGGs are distributed in the current human genome Ensembl (www.ensembl.org) annotation as shown in Table 1. Here we review examples in chromosomes (Chr) 21 and 22. There are at least 10 OGGs in Chr 21, according to the Riken annotation (ref. 2; Table 2), and at least 34 OGGs in Chr 22 (Sanger data release 3.1, ref. 3; Table 3). Tables 2 and 3 also indicate associations with known diseases. More OGGs may emerge as the genome annotation is refined.

OGG loci may be susceptible to genomic rearrangements, as occurs with the loci of the DiGeorge syndrome (DGS) region of Chr 22. Such rearrangements may be mediated by recombination events based on region-specific low copy repeats. The DGS region of 22q11.2 is particularly rich with segmental duplications, which can induce deletions, translocations, and genomic instability (4). There are several anomalous sequence features associated with OGGs, including Alu sequences intersecting exons, pseudogenes occupying introns, and single-exon (intronless) genes that often result from a processed multiexon gene.

At least 28 genes in Chr 21 are related to diseases, as characterized in the GeneCards database (5), as are 64 genes in Chr 22. Specific disorders that have been mapped to genes on Chr 21 and that involve OGG structures include: amyotrophic lateral sclerosis (ALS, Lou Gehrig's disease), linked to the GRIK1 ionotrophic kainate 1 glutamate receptor gene at 21q22 (6, 7); homocystinuria, a metabolic disorder linked to the cystathionine beta-synthase (CBS) gene (8); genes of the Down's Syndrome Critical Region (DSCR) (9–11); and the gene for amyloid beta (A4) precursor protein (APP) at location 21q21, associated with Alzheimer's disease (12).

We illustrate examples of OGGs in Fig. 1. The overlapping structure of the GRIK1 gene is shown in the first example: ORF41 (two exons) and ORF9 (two exons) overlap GRIK1 (17 exons) in the intron between exons 8 and 9, and in the intron between exons 1 and 2, respectively. There is some overlap between the first exon of ORF41 and exon 8 of GRIK1. The GRIK1 (GLUR5) locus at 21q 22.1 (13) coincides with the localization of the mutant gene causing ALS. The structure and function of the glutamate receptor subunits GLUR2, GLUR5, and GLUR6 are altered by RNA editing, converting the codon CAG (coding for glutamine) to the codon CGG (arginine), which may be important in controlling the rate of calcium flux in different states of the brain (13). Another prominent example in Chr 21 is the overlap between the genes CBS and PKNOX1. All 14 exons of the 30-kb-long CBS gene are located in the last intron of the gene for the homeobox protein PKNOX1 (11 exons). The gene U2AF1 (eight exons) is situated 5′ to the CBS gene in the same long intron of PKNOX1. The metabolic disorder homocystinuria is due to cystathionine beta-synthase deficiency and manifests as disorders of the eyes, central nervous system, skeletal systems, and vascular systems. The exons of overlapping genes tend to lie within large introns, usually the boundary (first or last) introns of another gene structure.

In Chr 22, two OGGs are associated with genes of the DGS region: CLTCL1/DVL1L1 (clathrin heavy polypeptide-like 1/human homolog to the 3′ end of *Drosophila dishevelled* segment-polarity gene) and TR/COMT (thioredoxin reductase beta/catechol-*O*-methyltransferase) (14, 15). DGS is related to one or more large deletions from Chr 22 apparently generated by recombination at meiosis. The 22q11 region of Chr 22 is susceptible to rearrangements associated with several genetic

Abbreviations: Chr, chromosome; ALS, amyotrophic lateral sclerosis; CBS, cystathionine beta-synthase; DGS, DiGeorge syndrome; OGG, overlapping gene group; Ψg, pseudogene.

†To whom correspondence should be addressed. E-mail: karlin@math.stanford.edu.

www.pnas.org/cgi/doi/10.1073/pnas.262658799

Table 1. Numbers of overlapping gene groups in the Ensembl annotation of human chromosomes

Chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
OGGs	144	75	80	47	58	69	89	47	54	58	95	78	30	46	43	67	83	23	97	36	10*	34*	21	4

*Numbers for Chrs 21 and 22 are from the Riken and Sanger annotations (see text).

disorders and malignant tumors. These include the cat eye syndrome (CES), part of the velocardiofacial syndrome (VCFS), DGS, and the der(22) chromosomal translocation (16). Many VCFS/DGS patients have a similar 3-Mb deletion and some have sporadically dispersed short deletions or translocations. DGS apparently results from haplo-insufficiency effects, and, in particular, the transcription factor Tbox-1 gene (TBX1) has been documented as one major contributing factor in congenital heart defects (4). How the observed OGG contributes to any DGS-related phenotype is unknown. In general, as in most gene deletion syndromes, a large majority of patients with DGS and Smith-Magenis syndrome have a common deletion interval, which may reflect meiotic unequal crossing-over mediated by flanking low copy number repeats. However, although patients with these conditions have almost identical deletions, there is substantial clinical variability. Galili *et al.* (14) verified synteny between a 150-kb region on mouse Chr 16 and the portion of 22q11 most commonly deleted in DGS.

Another major OGG of Chr 22 connects TIMP3 (tissue inhibitor of metalloproteinase; refs. 17 and 18) with SYN3 (synapsin-III, a membrane protein possibly involved in regulating neurotransmitter release) and is shown in Fig. 1 *Lower.* TIMP3 is associated with Sorsby fundus dystrophy and is a zinc-binding endopeptidase localized to the extracellular matrix that is expressed in many tissues, but is especially abundant in the placenta. Further OGG examples include: TR and COMT (putatively connected with schizophrenia); SERPIND1 (heparin cofactor II associated with thrombophilia) and PIK4CA (phosphatidylinositol 4-kinase α-subunit); and RTDR1 (rhabdoid tumor deletion region protein 1) and GNAZ (guanine nucleotide binding protein α-z).

There are several OGG-like structures involving sequences related to BCR (breakpoint cluster region fusion gene), which connects the distal part of Chr 22 to the q-arm of Chr 9 in the Philadelphia translocation, causing chronic myeloid leukemia. BCR itself overlaps with the F-box protein pseudogene FBXW3. In addition, there are seven BCR-like pseudogenes on Chr 22, of which two appear in OGGs, as shown in Table 4. There are five non-OGG BCR-like pseudogenes: AP000550.6, AP000552.3, BCRL4, AP000354.4, and BCRL6. It is striking that the eight BCR-related sequences of Chr 22 cluster within a 6-Mb stretch.

OGGs and Anomalous Sequence Features

There is a conspicuous association of disease genes with OGGs involving intrusions of Alu, and/or pseudogene, and/or single-

Table 2. Overlapping gene groups in human Chr 21 (Riken annotation)

Gene locus	No. of exons, strand	Description	Relations
(1) GRIK1	17, −	Glutamate receptor	GRIK1 is linked to ALS; two Alus overlap the same internal exon of GRIK1; two claudin intronless genes, CLDN17 and CLDN8, are immediately 5′ to GRIK1
(2) ORF41	2, +	Spliced EST	
(3) ORF9	2, +	Spliced EST	
(1) TIAM1	29, −	T-lymphoma invasion and metastasis-inducing TIAM1 protein	The Ψg (UBE3AP2) is immediately 3′ to TIAM1; the Ψg (BTRC2P) and then the disease gene SOD1 (causing ALS) is immediately 5′ to TIAM1
(2) PRED31	3, +	Predicted gene	
(1) ITSN	12, +	Intersectin-1 SH3 domain protein	An Alu sequence overlaps with the 3′ boundary exon of ITSN, which is overexpressed in the brain in Down's syndrome, suggesting a gene dosage contribution
(2) ATP50	7, −	ATP synthase OSCP subunit, oligomycin sensitivity conferring protein	
(1) DSCR1	4, −	Down's syndrome candidate region protein, proline-rich protein	DSCR1 may play a role in central nervous system development; the intronless gene KCNE1 (Lange-Nielsen syndrome) is immediately 3′ to DSCR1
(2) PRED39	5, +	Predicted gene	
(1) BACE2 (ASP2)	9, +	β-site APP-cleaving enzyme 2	Decreased expression of ASP2 reduces amyloid β-peptide production, a precursor in amyloid plaque formation
(2) PRED43	4, −	Predicted gene	
(1) PKNOX1	11, +	Homeobox-containing protein	CBS is associated with homocystinuria; an Alu overlaps an internal exon; the disease gene crystallin α-A and the progressive myoclonus epilepsy (EPM1) critical region of 21q22.3 are immediately 3′ to PKNOX1
(2) CBS	14, −	CBS	
(3) U2AF1	8, −	U2 snRNP auxiliary factor small subunit	
(1) HSF2BP	9, −	Heat shock transcription factor 2 binding protein	The Ψg (RPL31P) and H2BFS are within the same boundary intron of HSF2BP; H2BFS is a single-exon gene
(2) H2BFS	1, +	H2B histone family S member	
(1) ORF30	2, +	ORF	An Alu sequence overlaps with the 3′ exon of ORF30; the Ψg (IMMTP) is immediately 3′ to ORF30
(2) ORF29	4, −	Spliced mRNA	
(3) ORF31	6, +	Spliced EST	
(4) PRED53	7, −	Predicted gene	
(1) ADARB1	13, +	dsRNA adenosine deaminase	An Alu sequence overlaps with an internal exon of ADARB1; at least four isoforms of ADARB1 have been identified
(2) PRED57	3, −	Predicted gene	
(3) PRED58	4, −	Predicted gene	
(1) PCBP3	11, +	Poly (rC)-binding protein 3	The disease gene COL6A1 (Bethlem myopathy) is immediately 3′ to PCBP3
(2) PRED62	4, −	Putative gene containing transmembrane domain	

Ψg is the notation for pseudogene. There are four single-exon genes (CLDN17, CLDN8, KCNE1, and H2BFS) contained in, or proximal to, overlapping gene groups. There are eight recognized disease genes (GRIK1, SOD1, ITSN, DSCR1, KCNE1, CBS, CRYAA, and COL6A1) related to the overlapping gene groups.

Table 3. Overlapping gene groups in human Chr 22 (Sanger annotation)

Gene locus	No. of exons, strand	Description	Relations
(1) AP000546.2	7,−	Similar to Wp:CE19906 and C2 genomic clone Em:AC002038	Ψg AP000546.1 and gene (2) are in the same boundary intron of gene (1); the Ψg AP000545.1 is immediately 3′ to gene (1)
(2) AP000547.1*	3,+	Similar to Tr:O96017 protein kinase	
(1) AC008101.3	7,+	Human cDNA for KAIA2502 protein	Ψg AC008101.2 is in an internal intron of (2)
(2) AC008101.5	3,−	Matches ESTs	
(1) CLTCL1	33,−	Clathrin-heavy polypeptide-like 1	CLTCL1 may play a role in hypertonia in VCFS; DVL1L1 is deleted in DGS and is partly responsible for catch-22 syndrome; two Ψgs (AC000081.1; AC000094.2) are in two internal introns of (1); gene (2) is intronless
(2) DVL1L1*	1,−	Human homologue sequences to the 3′ end of *D. dishevelled* segment-polarity gene are deleted in the DGS	
(1) TR	18,−	Thioredoxin reductase beta	Gene (2) is involved in 22q11 deletion syndrome (inc. VCFS/DGS); Ψg AC000078.2 is in an intron of TR
(2) COMT	6,+	Catechol-*O*-methyltransferase	
(1) AC006547.4	14,+	Matches ENCORE sequence	The first and last exons of the genes overlap by 3 bp
(2) AC006547.2	12,−	Similar to Tr:P70222 mouse HTF9C	
(1) AC007731.1*	5,+	Matches ESTs—novel LCR gene	Gene (1) is in an intron of USP18
(2) USP18*	9,−	Ubiquitin-specific protease 18	
(1) SERPIND1	4,+	Heparin cofactor II (HCF2)	SERPIND1/HCF2 is related to thrombophilia
(2) PIK4CA	55,−	Phosphatidylinositol 4-kinase α-subunit	
(1) GNAZ	3,+	G protein α-subunit	GNAZ is in an intron of RTDR1
(2) RTDR1	7,−	Rhabdoid tumor deletion region protein 1	RTDR1 and the gene RAB36 are often deleted in pediatric rhabdoid tumors
(1) AP000344.6*	9,+	Matches Incyte ESTs (imperfect match)	
(2) AP000344.2	6,−	Matching EST cluster	
(1) AP000346.5*	1,+	Similar to Tr:O70122 mouse sodium-glucose cotransporter	Gene (1) is in intron of gene (2)
(2) AP000346.6	7,−	*Homo sapiens* mRNA	
(1) AP000348.3	6,+	Matches EST cluster	
(2) AP000348.4	4,−	Similar to Sw:Q03667 and Sw:Q09254	
(1) SMARCB1	9,+	SWI/SNF related matrix-associated actin-dependent regulator of chromatin subfamily b, member 1	The end two exons of the two genes overlap SMARCB1 is a tumor suppressor gene that is inactivated in certain malignant rhabdoid tumors
(2) AP000350.1	7,−	Similar to *H. sapiens* CGI-101 protein mRNA (AF151859).	
(1) bK221G9.4	14,+	Matches EST cluster	Ψg bK221G9.1 and (2) are in and intron of (1)
(2) bK243E7.3*	3,−	Matches EST sequences	
(1) DJ268D13.2*	1,+	Similar to Sw:P25112 *H. sapiens* 40S ribosomal protein S28	Gene (1) is in an intron of gene (2) SEZ6L is a membrane protein located in a region that is often deleted in small cell lung cancers and in advanced non-small cell cancers
(2) SEZ6L	15,+	Seizure related gene 6 (mouse)-like	
(1) bK1048E9.5	7,+	Matches EST cluster	An Alu sequence overlaps the 3′ terminal exon of gene (2)
(2) bK445C9.6	15,−	Similar to mouse tuftelin-interacting protein 10 mRNA AF097181	
(1) HMG1L10*	1,+	High mobility group protein 1-like 10	Gene (1) is intronless
(2) TPST2	7,−	Tyrosylprotein sulfotransferase 2	
(1) dJ353E16.2	4,+	Matches ESTs	
(2) cB42E1.1	23,−	Novel protein	
(1) CHEK2	15,−	Protein kinase Chk2	Checkpoint protein Chk2 is involved in Li-Fraumeni syndrome (familial cancer and diverse tumor types) and somatic osteosarcoma
(2) dJ366L4.2	6,+	Matches ESTs	
(1) RFPL1S	1,−	RET finger protein-like 1 antisense	Gene (1) is intronless
(2) RFPL1	2,+	ret finger protein-like 1	
(1) LIF	3,−	Leukemia inhibitory factor	LIF has the capacity to induce terminal differentiation in leukemic cells
(2) AC004264.3	1,+	*H. sapiens* clone IMAGE:3355596	Gene (2) is intronless
(1) AC004997.11	1,−	Human mRNA for KIAA1656 protein	Gene (1) is intronless
(2) AC004997.9	11,+	Matches ESTs	
(3) SF3A1	16,−	Pre-mRNA splicing factor SF3a subunit	
(1) AC004542.4*	2,+	Matches ESTs	ZNF278 (MAZR) is a transcriptional repressor that undergoes fusion with the Ewing sarcoma gene EWS in small round cell tumors
(2) dJ430N8.1	27,−	KIAA0852	
(3) AC005003.5*	1,+	*H. sapiens* clone MGC:15705	Gene (3) is intronless
(4) ZNF278	4,−	Zinc finger protein 278	
(1) dJ858B16.1	32,+	Human mRNA for KIAA0542 protein	Three Ψgs (bA247I13.5; bA247I13.6; bA247I13.3) are within three different internal introns of gene (1)
(2) PISD	8,−	Phosphatidylserine decarboxylase	
(1) cN44A4.2*	3,−	*H. sapiens* novel gene	The Ψg bK440B3,1 is in an intron of gene (2)
(2) YWHAH	2,+	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein ε	YWHAH is the ε isoform of 14-3-3 signal transduction protein and could be associated with neuropsychiatric disorders

Table 3. (continued)

Gene locus	No. of exons, strand	Description	Relations
(1) RFPL3	2,+	ret finger protein-like 3	Two Ψgs (dJ90G24.5 and dJ149A16.5) are immediately 5′ to RFPL3
(2) RFPL3S	4,−	ret finger protein-like 3 antisense	
(1) SYN3	14,−	Synapsin-III	TIMP3 is related to Sorsby Fundus dystrophy; a Ψg (dJ309122.3) overlaps with the 3′ terminal exon of TIMP3; two Ψgs (bK415G2.2, dJ302D9.1) are immediately 5′ to SYN3
(2) TIMP3	5,+	Tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy pseudoinflammatory)	
(1) UQCRFSL1*	1,+	Ubiquinol-cytochrome *c* reductase Rieske iron-sulfur polypeptide-like 1	Gene (1) is intronless
(2) dJ370M22.3*	4,−	Similar to human epsin 2a mRNA	
(1) MKL1	15,−	Megacaryocytic acute leukemia protein	Gene (2) and the Ψg bK229A8.1 are in the same intron of (1); Ψg dJ1042K10.6 is in another intron
(2) dJ591N18.1*	2,+	Cytochrome *c* oxidase subunit VIb	MLK1 is associated with acute leukemia by translocation *t*(1;22)(p13,q13) with RBM15
(1) dJ408N23.2	1,−	Similar to mouse Mrj, encodes a DnaJ-related cochaperone that is essential for murine placental development	An Alu sequence overlaps the 3′ terminal exon of gene (2); gene (1) is intronless
(2) dJ1057D18.1	10,+	Similar to yeast hypothetical peptidase	
(1) dJ756G23.3	17,+	Similar to Tr:Q24191 *Drosophila* transcriptional repressor protein	
(2) dJ756G23.1*	6,−	Similar to mouse chondroadherin	
(1) ACO2	18,+	Aconitase 2, mitochondrial	Aconitase 2 is a tricarboxylic acid (TCA) cycle gene
(2) dJ347H13.5	7,−	Novel protein similar to yeast DNA-directed RNA pol III 25-kDa subunit	
(1) dJ345P10.4	32,−	Human mRNA for KIAA1672 protein	Gene (2) and two Ψgs (dJ345P10.1 and dJ388M5.1) are in introns of gene (1)
(2) HMG17L1*	2,+	High-mobility group (nonhistone chromosomal) protein 17 like	
(1) U51561.2	1,−	*H. sapiens* cDNA FLJ32756 fis	Gene (1) is intronless
(2) C22orf4	13,+	Similar to Tr:Q92680	
(1) CHKL	11,−	Choline kinase-like	
(2) U62317.15	2,+	Matches ESTs	

Genes associated with diseases in OGGs of Chr 22 are CLTCL1, DVL1L1, COMT, SERPIND1, RTDR1, SMARCB1, SEZ6L, CHEK2, LIF, ZNF278 (MAZR), YWHAH, TIMP3, and MKL1.

*Partial gene.

exon (intronless) sequences (see Tables 2 and 3). Also, two single-exon genes, CLDN17 (claudin) and CLDN8, that contribute to tight junction formations are immediately 5′ to GRIK1. SOD1 (superoxide dismutase) follows the OGG of TIAM1 and a pseudogene (BTRC2P) in Chr 21, and is within 2 Mb of GRIK1. Reduction in SOD1 activity might be expected to lead to an accumulation of toxic superoxide radicals, which can cause familial ALS (19). Allelic variants of GRIK1 further contribute to the pathogenesis of juvenile absence epilepsy (13). In the CBS/PKNOX1 OGG, an Alu sequence overlaps with an exon of the PKNOX1 gene. These may predispose the gene to detrimental rearrangements. It is further documented that the CBS gene can undergo alternative splicing in its 5′ UTR (8). Another gene, CRYAA (crystalline), which can produce a cataract phenotype, is directly 3′ to PKNOX1.

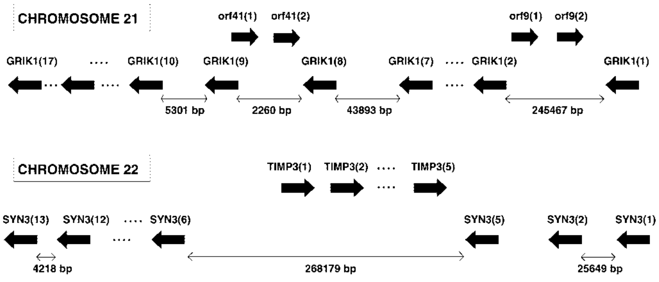

Fig. 1. Examples of OGGs in Chrs 21 and 22.

In Chr 21, only 34 (of 12,168) Alu elements overlap exons. Twenty Alu elements are either totally within or envelop a complete exon, four of which are internal exons. Four Alus overlap internal exons, whereas the other 10 overlap boundary exons mostly in UTRs. In Chr 22 (23,675 Alus), there are only 165 instances, involving 87 genes, of an Alu overlapping an exon. Of these, 5 involve an internal coding exon, 98 are with a noncoding exon, and 62 are with boundary exons that contain a translation initiation or termination codon. In 3 cases, the Alu completely envelops an exon, in 141 it is contained within an exon, and in 21 cases the Alu and exon overlap. However, there are 12,367 instances of an Alu being contained in an intron of a gene on Chr 22, involving 408 of the 546 coding genes. These results are broadly consistent with other studies of the occurrence of transposable elements in coding genes (20). Although Alu insertions and other transposition events have been shown

Table 4. Three OGGs associated with BCR-like genes/pseudogenes in Chr 22

Genes in OGG	Description
AC008103.3	cDNA DKFZp434K191
AC008103.2	BCR-related sequence
AC007050.4	Similar to human cDNA DKFZp434p211
BCRL5	Breakpoint cluster region-like 5
BCR	Active BCR-related gene (Philadelphia translocation)
FBXW3	F-box protein FBX3

Table 5. OGGs of TIMP and synapsin in human and mouse

Human		Mouse	
Genes in OGG	Chr	Genes in OGG	Chr
TIMP3	22	TIMP3	10
SYN3		SYN3	
TIMP1	X	TIMP1	X
SYN1		SYN1	
TIMP4	3	TIMP4	6
SYN2		SYN2	
TIMP2	17	TIMP2	11
Similar to mouse testis-specific protein		Testis-specific protein	

The four pairs of OGGs in mouse chromosomes are the homologues of the corresponding human OGG pairs.

to generate null alleles through insertional transposition, these appear to be uncommon mechanisms for human diseases (21), except possibly in the context of OGG structures. In Chr 21, there are no pseudogenes overlapping exons. In Chr 22, there are 11 pseudogenes that show some overlap with exons of coding genes. In eight of these cases, the overlap is between an exon and an intron of a multiexon pseudogene. There is one intronless pseudogene that partially overlaps with a gene exon sequence, and two intronless pseudogenes contained within exons; however, all of the exons of the genes involved are untranslated. No pseudogenes overlap with coding exons. Pseudogene sequences are biased toward highly expressed genes, emphasizing ribosomal protein genes (22, 23).

Concurrence of OGGs, Pseudogenes, and Disease Genes

It appears that genes containing pseudogenes in introns or Alu elements overlapping with exons have a strong disposition for disease on Chr 22. There are 546 genes annotated (Sanger data), with 64 disease genes [GeneCards (5)], and 34 OGGs involving 71 genes, of which 13 are disease-associated. Thus, the fraction of genes with an associated disease from among the OGG collection is $13/71 = 0.18$. There are $64 - 13 = 51$ remaining known disease genes of a total of $546 - 71 = 475$ genes not associated with OGGs. The fraction of disease genes in non-OGG surroundings is thus $(64 - 13)/(546 - 71) = 0.10$. An analogous calculation indicates that disease genes seem to have a higher chance of overlapping with pseudogenes: 49 genes overlap with pseudogenes in Chr 22, including 12 of the 64 known disease genes. Thus $12/64 = 0.19$ of disease genes in Chr 22 are associated with pseudogenes, compared with $(49 - 12)/(546 - 64) = 0.08$ of genes with no known disease association.

The OGG of the gene combination TIMP3 and SYN3 is conserved in mouse and in *Drosophila* (17). A number of OGG structures based on the Ensembl data collection connect additional TIMP subunits and synapsin subunits in the human and mouse genomes. These are displayed in Table 5. Another OGG present in both human Chr 22 and mouse is the gene pair TR and COMT (see Table 3).

Multiple Amino Acid Runs

There are 192 human protein sequences [of 10,651 $\geq$ 200 aa long, extracted from RefSeq (www.ncbi.nlm.nih.gov/LocusLink/refseq.html)] that have multiple amino acid runs (24). More than 40% of these proteins are associated with diseases, as identified in OMIM (25). All established human CAG triplet repeat (polyglutamine) diseases (26), together with some potential new ones, qualify as having multiple runs, not just of glutamine. In addition, many proteins related to leukemia and other cancers have multiple runs: 14 cancer-related proteins (e.g., adenomatous polyposis coli, breast carcinoma-associated antigen, and matrix metalloproteinase 24); 10 leukemia-related proteins often resulting from chromosomal translocations (listed in Table 6); 14 channel proteins, mainly voltage-gated Ca^{2+} and K^+ channel proteins; 6 proteases, including acrosin, calpain 4, and some metalloproteinases; and a variety of disease syndrome-related proteins (e.g., Wiskott-Aldrich syndrome and cat eye syndrome). A key aspect of 82 of the 192 human protein sequences is their role in transcription, translation, and developmental regulation. Strikingly, many of these proteins are homeotic homologs of *Drosophila* developmental sequences and transcription factors, including *timeless*, *trithorax*, *frizzled*, *dead ringer* (*retained*), and *diaphanous 3*.

In marked contrast, no metabolic enzymes (e.g., glycolysis, tricarboxylic acid cycle, and pentose phosphate pathway), structural proteins (e.g., actin, myosin, and troponin 1), or housekeeping proteins contain multiple runs. However, several structural-regulatory proteins do have multiple runs, including ankyrin 3, nucleolin, SMARCA2 (actin-dependent regulator of chromatin), and synapsin II, which may function in the regulation of neurotransmitter release.

Prokaryote protein analogs/homologs in the human genome do not have multiple amino acid runs. On this basis, multiple

Table 6. Leukaemia-related proteins containing multiple amino acid runs

Gene, description	Chr	Amino acid runs	Disease association
ALK, anaplastic lymphoma kinase Ki-1	2	G8 G5 G6 G6	Important role in the development of the brain; anaplastic large cell lymphomas, caused by 2;5 translocation
BRD2, bromodomain-containing protein 2	6	E6 E5 E5 S5 S11	Mitogen-activated kinase, possibly part of a signal transduction pathway involved in growth control upregulated in certain leukemias
CREBBP, CREB-binding protein	16	Q5 Q18	Augments the activity of phosphorylated CREB to activate transcription of cAMP-responsive genes Rubinstein-Taybi syndrome, leukemias
D6S51E, HLA-B-associated transcript-2	6	P5 P5 G6 G6 P5	Limited to cell lines of leukemic origin
KIAA0304 (MLL2) gene product	19	G5 P6 P5 P5 P5 Q5 P7	Called trithorax homolog 2 (MLL2) in SWISS-PROT, but has little similarity to the other MLL2 gene in this list
LAF4, lymphoid nuclear protein 4	2	S7 S5 S8	Tissue-restricted nuclear transcriptional activator lymphoid tissue; acute lymphoblastic leukemia, Burkitt's lymphoma
MLL2, myeloid/lymphoid or mixed-lineage leukemia 2	12	E5 A5 Q5 Q6 Q9 Q14 Q6 Q11 Q6 Q14 Q6 Q8 Q5 Q5 Q7 Q5 Q8 Q7 Q6 Q10 Q8 Q7 Q5 Q5	Leukemia
MLLT6, trithorax (*Drosophila*) homolog	17	G6 G5 S11 A8	Acute leukemias (by chromosomal translocations)
MN1, meningioma 1	22	Q5 Q5 P5 Q28 P5 G5 G7 G5	May play role in tumor suppression; highest expression in skeletal muscle; acute myeloid leukemia by a chromosomal translocation
ZNF220, zinc finger protein 220	8	E6 E5 E5 E5 S6 P6 P7	May represent a chromatin-associated acetyltransferase; acute myeloid leukemia (by translocation)

CREB, cAMP response element-binding protein.

runs in human proteins may be a recent evolutionary outcome, concomitant with complex brain or heart development. Multiple runs are, however, substantially conserved between human and mouse proteins. Of 56 SWISS-PROT mouse proteins that have multiple runs, 52 have a human homolog. In 43 cases (83%), the human homolog also has multiple runs; in 10%, the human homolog has more than one run but does not meet the criterion for multiple runs; and in the remaining 7% [DDX9 (ATP-dependent RNA helicase A), DUS8 (neuronal tyrosine threonine phosphatase 1), HOXD9 (homeobox protein), and UBF1 (nucleolar transcription factor 1)], the human protein has one or no runs. Examples of human/mouse proteins that share multiple runs are CREB-binding protein, *diaphanous* and *even-skipped* homologs, anaplastic lymphoma kinase, *myc*-associated zinc finger (MAZ), and two zinc finger proteins of the cerebellum (ZIC2 and ZIC3). The disease genes meningioma 1 (MN1), Ran GTPase activating protein 1 (RANGAP1), and the cat eye syndrome region (CECR) of Chr 22 encode proteins with an abundance of multiple long homopeptides, multiple charge clusters, and a large count of multiplets (amino acid doublets, triplets, etc.). These sequence properties could induce neurological phenotypes (24).

We conclude that the majority of OGGs and genes encoding significantly many amino acid long runs are potentially associated with disease. We also hypothesize that OGGs increase the potential for genomic rearrangements and/or disruption of transcription regulation, and may predispose these gene groups to contain disease-related genes at substantially higher frequency than non-OGG genes. The presence of an OGG may cause difficulties in transcription, in fostering complex gene rearrangements, and redundancies in propensity to mutations and mutational hot spots (partly dependent on the presence of Alus), and in generating gene dosage imbalances. Alu (and other transposable elements) are innately mobile and, like pseudogenes, are heavily prone to mutation (21). Also, many single-exon genes, like pseudogenes, derive often from the processing of multiexon genes (see ref. 22). Thus, human disease genes tend to be associated with disrupting Alu sequences, and/or pseudogenes, and/or proximal single-exon genes. Extant OGGs and consequent rearrangements appear as a novel configuration of many disease genes. Experimental studies are required to confirm these observations and elucidate the underlying mechanisms.

This work was supported in part by National Institutes of Health Grants 5R01GM10452-38 and 5R01HG00335-15 (to S.K.).

1. Seidman, J. G. & Seidman, C. (2002) *J. Clin. Invest.* **109,** 451–455.
2. Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D. K., *et al.* (2000) *Nature* **405,** 311–319.
3. Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., *et al.* (1999) *Nature* **402,** 489–495.
4. Edelmann, L., Pandita, R. K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R. S., Magenis, E., Shprintzen, R. J. & Morrow, B. E. (1999) *Hum. Mol. Genet.* **8,** 1157–1167.
5. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. (1998) *Bioinformatics* **14,** 656–664.
6. Eubanks, J. H., Puranam, R. S., Kleckner, N. W., Bettler, B., Heinemann, S. F. & McNamara, J. O. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 178–182.
7. Paschen, W. & Djuricic, B. (1994) *Cell. Mol. Neurobiol.* **14,** 259–270.
8. Avramopoulos, D., Cox, T., Kraus, J. P., Chakravarti, A. & Antonarakis, S. E. (1993) *Hum. Genet.* **90,** 566–568.
9. Fuentes, J.-J., Pritchard, M. A., Planas, A. M., Bosch, A., Ferrer, I. & Estivill, X. (1995) *Hum. Mol. Genet.* **4,** 1935–1944.
10. Nakamura, A., Hattori, M. & Sakaki, Y. (1997) *J. Biochem.* **122,** 872–877.
11. Korenberg, J. R., Chen, X. N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, C., Dignan, P., Disteche, C., *et al.* (1994) *Proc. Natl. Acad. Sci. USA* **91,** 4997–5001.
12. Yanker, B. A. (1996) *Neuron* **16,** 921–932.
13. Sander, T., Hildmann, T., Kretz, R., Furst, R., Sailer, U., Bauer, G., Schmitz, B., Beck-Mannagetta, G., Wienker, T. F. & Janz, D. (1997) *Am. J. Med. Genet.* **74,** 416–421.
14. Galili, N., Baldwin, H. S., Lund, J., Reeves, R., Gong, W., Wang, Z., Roe, B. A., Emanuel, B. S., Nayak, S., Mickanin, C., *et al.* (1997) *Genome Res.* **7,** 17–26.
15. Schinke, M. & Izumo, S. (2001) *Nat. Genet.* **27,** 238–240.
16. Tapia-Páez, I., Kost-Alimova, M., Hu, P., Roe, B. A., Blennow, E., Fedorova, L., Imreh, S. & Dumanski, J. P. (2001) *Hum. Genet.* **109,** 167–177.
17. Weber, B. H., Vogt, G., Pruett, R. C., Stohr, H. & Felbor, U. (1994) *Nat. Genet.* **8,** 352–356.
18. Stohr, H., Roomp, K., Felbor, U. & Weber, B. H. (1995) *Genome Res.* **5,** 483–487.
19. Estevez, A. G., Crow, J. P., Sampson, J. B., Reiter, C., Zhuang, Y., Richardson, G. J., Tarpey, M. M., Barbeito, L. & Beckman, J. S. (1999) *Science* **286,** 2498–2500.
20. Nekrutenko, A. & Li, W.-H. (2001) *Trends Genet.* **17,** 619–621.
21. Deininger, P. L. & Batzer, M. A. (1999) *Mol. Genet. Metab.* **67,** 183–193.
22. Chen, C., Gentles, A. J., Jurka, J. & Karlin, S. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2930–2935.
23. Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone, P., Echols, N., Johnson, T. & Gerstein, M. (2002) *Genome Res.* **12,** 272–280.
24. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. & Gentles, A. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 333–338.
25. Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D. & McKusick, V. A. (2002) *Nucleic Acids Res.* **30,** 52–55.
26. Zoghbi, H. Y. & Orr, H. T. (2000) *Annu. Rev. Neurosci.* **23,** 217–247.