



## Pathological Consequences of Sequence Duplications in the Human Genome

Richard Mazzarella and David Schlessinger

*Genome Res.* 1998 8: 1007-1021

Access the most recent version at doi:[10.1101/gr.8.10.1007](https://doi.org/10.1101/gr.8.10.1007)

---

**References** This article cites 113 articles, 25 of which can be accessed free at:  
<http://genome.cshlp.org/content/8/10/1007.full.html#ref-list-1>

### License

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Cold Spring Harbor Laboratory Press

## REVIEW

# Pathological Consequences of Sequence Duplications in the Human Genome

Richard Mazzarella<sup>1</sup> and David Schlessinger<sup>2,3</sup>

<sup>1</sup>Institute for Biomedical Computing and Center for Genetics in Medicine, Washington University School of Medicine, St. Louis, Missouri 63110 USA; <sup>2</sup>Laboratory of Genetics, Gerontology Research Center, National Institute on Aging, Baltimore, Maryland 21224 USA

As large-scale sequencing accumulates momentum, an increasing number of instances are being revealed in which genes or other relatively rare sequences are duplicated, either in tandem or at nearby locations. Such duplications are a source of considerable polymorphism in populations, and also increase the evolutionary possibilities for the coregulation of juxtaposed sequences. As a further consequence, they promote inversions and deletions that are responsible for significant inherited pathology. Here we review known examples of genomic duplications present on the human X chromosome and autosomes.

Gene duplication is an important mechanism in the evolutionary process. As analyzed by Ohno in his classic monograph (1970), these duplication events liberate copies of the gene to diverge and take up new functional roles in the organism, while the master gene is constrained to preserve its original role. Families of genes, often including those in clustered repeats, have been encountered since the beginning of molecular biological analysis. The notable early examples included the HOX genes (Krumlauf 1994); members of the immunoglobulin superfamily, such as the immunoglobulins, the T-cell receptors, and the major histocompatibility complex (MHC) genes (Hood et al. 1985); the globin genes (Orkin and Kazazian 1984); and the small and large rRNAs (Srivastava and Schlessinger 1991). Evolution has played with the regulatory possibilities, resulting in mechanisms as varied as the globin switch, immunoglobulin diversity, and the succession of expression of tandem HOX genes during development. Recently we reviewed some features of the resulting distribution of repetitive elements and genes in the human genome (Mazzarella and Schlessinger 1997). In this discussion we expand on some of the consequences of sequence duplications as they relate to human physiology and disease.

The ability to detect and analyze duplication in genomes has been expanded enormously by the explosive progress of long-range sequencing analyses. On a statistical and comparative level, the inference of repetitive elements and motifs has shown that a variety of sequences of unknown function, as well

as functional segments of genes, are spread through the genome (for an interesting recent discussion, see Babbitt and Gerlt 1997). Duplication and divergence are central to the generation of diversity and new genes. Duplication can involve rare noncoding or coding sequences, and can occur with or without associated clustering. Thus, members of the actin and tubulin families are scattered in the genome, but globin genes show a more complex pattern involving clustering. As part of the pattern, for example, restriction mapping of the short arm of chromosome 16 has revealed that three different alleles of the  $\alpha$ -globin gene lie, respectively, 170, 350, and 430 kb from the telomere (Wilkie et al. 1991). Polymorphic length variation at this locus is postulated to have arisen by nonhomologous exchanges between the subtelomeric repeats on different chromosomes. Furthermore, heterozygosity for the telomere polymorphism may have an effect on meiotic segregation. Because most nonhomologous pairing resulting in nondisjunction occurs at telomeres, trisomy of chromosome 16 may be more frequent in heterozygotes for the subtelomeric region (Speed 1988). Interestingly, trisomy of chromosome 16 is the most common trisomy seen in early natural abortuses (Bond and Chandley 1983).

A variety of specialized selective pressures could promote the development of sequence clusters. They include

1. An increase in expression. This can occur by simple expansion of tandem repeats, like the rRNA and 5S genes, or by the duplication of sequences at nearby but noncontiguous positions, as exemplified in the discussion below.

<sup>3</sup>Corresponding author.  
E-MAIL [schlessingerd@grc.nia.nih.gov](mailto:schlessingerd@grc.nia.nih.gov); FAX (401) 558-8331.

## MAZZARELLA AND SCHLESSINGER

2. Preservation of function of genes on the Y chromosome, which must retain activity in the face of accumulating mutations with no available cognate chromosome to rescue defects by recombination (Muller 1964; Rice 1987); recently a number of genes on the Y chromosome have been shown to be duplicated many times in tandem (Lahn and Page 1997).

## X-linked Clusters and Pathology

Repeated sequences can mediate local deletions, duplications, and inversions, with a number of consequences for genome diversity and genetic pathology; the range and seriousness of such events is increased when repeats occur near one another but

not directly juxtaposed. Perhaps because it is one of the first regions of the genome to be analyzed in detail, the telomeric cytogenetic band q28 of the X chromosome shows a number of clusters. It also affords corresponding instances in which significant human pathology results from the interaction of the duplicated sequences (Table 1). Figure 1 provides a sketch of current information about some clustered sequences on the X chromosome. Figure 2 illustrates recombinational events that have been detected in Xq28. More than 10% of 2.5 Mb of sequence determined to date for Xq28 is duplicated at least once. The duplications are usually nearby in the genome, but also include a 26.5-kb segment found on both chromosome 16p11.1 and Xq28 (Eichler et al. 1996).

Table 1. Sequence Duplications and Consequences

Duplication	Location	Consequences
Glucocerebrosidase gene 3.3-kb repeats	1q21 4q35	Gaucher disease 1. facioscapulohumeral muscular dystrophy? 2. polymorphic copy number
<i>SMA</i> gene 21-hydroxylase and complement 4 genes	5q11.2–q13.3 6p21.3	spinal muscular atrophy adrenal hyperplasia III (21-hydroxylase deficiency)
<i>GTF2I</i> gene <i>CYP11B2</i> and <i>CYP11B1</i> genes	7q11.23 8q21	William's syndrome glucocorticoid-remediable aldosteronism
$\beta$ and $\beta$ -like genes Homologous sequences?	11p15.5 15q11–q13	$\beta$ -thalassemia Prader-Willi and Angelman syndromes
Subtelomeric repeats $\alpha$ gene <i>PKD1</i> gene	16pter 16p13.3 16p13.1	trisomy of chromosome 16? $\alpha$ -thalassemia complicates the analysis of polycystic kidney disease 1 patients
17-kb repeats 200 kb repeats (gene cluster) Low copy repeats?	17p11.2–p12 17p11.2 22q11	Charcot-Marie-tooth type 1A Smith-Magenis syndrome DiGeorge syndrome Velocardiofacial syndrome
<i>CYP2D6</i> gene Homologous sequences? Homologous sequences Homologous sequences? Homologous sequences	22q13.1 Xp21.3 Xp22.32 Xq22 Xq25	debrisoquine metabolism deficiency dosage-sensitive sex reversal X-linked ichthyosis Pelizaeus-Merzbacher disease lymphoproliferative disorder?
Iduronate-2-sulphatase gene Red and Green pigment genes	Xq28 Xq28	Hunter's syndrome 1. color blindness 2. polymorphic copy number
11.3-kb inverted repeats	Xq28	1. Emery-Dreifuss syndrome 2. polymorphic inversion of 38 kb
9.5-kb repeats 13-bp repeats	Xq28 mitochondria	hemophilia A Kearns-Sayre syndrome

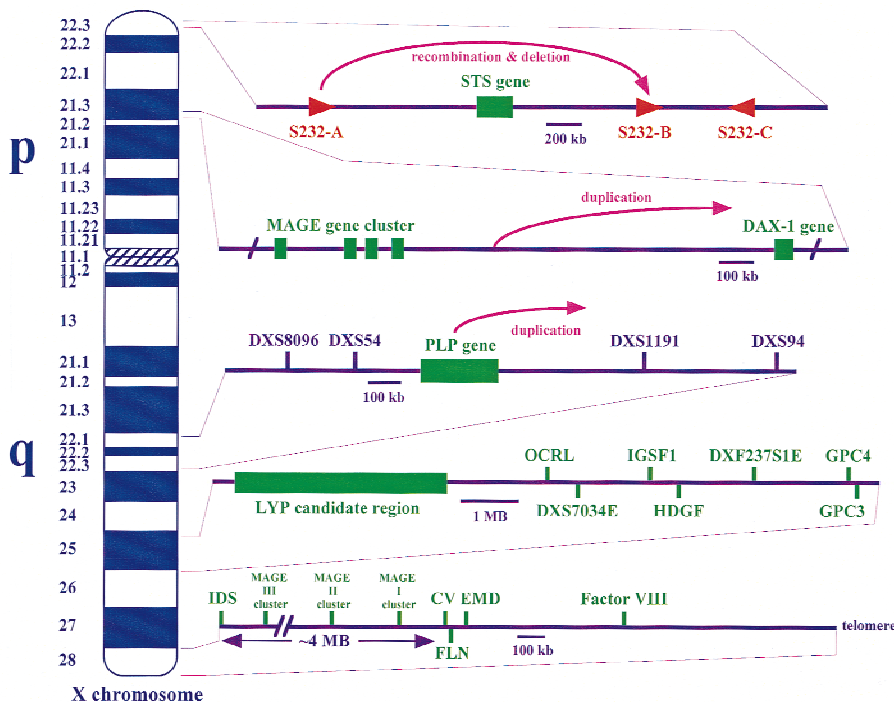


Figure 1 X chromosome regions with duplicated sequences involved in pathology. The regions are (from *top to bottom*) the X-linked ichthyosis region in Xp22.3, the dosage-sensitive sex reversal candidate region in Xp21.3, the Pelizaeus-Merzbacher region in Xq22, the lymphoproliferative syndrome in Xq25, and regions in Xq28 (further diagrammed in Fig. 2). The genes and probes indicated are all described in the Genome Database and include (STS) steroid sulfatase; (S232-A,B,D) three repetitive sequences that cross-hybridize with genomic clone CRI-S232; (*MAGE*) melanoma antigen gene family; (*DAX-1*), critical region for adrenal hypoplasia congenita; (*PLP*) proteolipid protein; (*DXS8096*, *DXS54*, *DXS1191*, and *DXS94*) polymorphic markers; (*LYP*) lymphoproliferative syndrome locus; (*OCRL*) oculocerbrorenal (Lowe) syndrome gene; (*IGSF1*) immunoglobulin superfamily gene 1; (*HDGF*) hepatoma-derived growth factor gene; (*GPC3*) glypican 3; (*DXS7034E* and *DXF237S1E*) expressed sequence tags; (*GPC4*) glypican 4; (*IDS*) iduronate-2-sulfatase gene; (*CV*) color vision genes; (*FLN*) filamin gene; (*EMD*) Emery-Dreifuss muscular dystrophy (emerin) gene; and *factor VIII* gene.

The classic locus for tandem repeats is color vision. Extensive polymorphism is associated with color blindness in ~1 in 12 white men and 1 in 200 women [McKusick 1994; OMIM no. 303800 (green pigment) and OMIM no. 303900 (red pigment)]; The Online Mendelian Inheritance in Man (OMIM), edited by Dr. Victor A. McKusick and colleagues, is found at URL <http://www.ncbi.nlm.nih.gov/omim/>. The vast majority of color blindness results from variations in a tandem set of one to four red and one to seven green pigment genes (Nathans et al. 1992; Neitz and Neitz 1995), and phenotypic differences in color vision between individuals are the direct result of the ratio of expressed red genes to green genes. The six exons of the red and green pigment genes and their intragenic regions are 98%

identical, suggesting that they arose recently in evolution by duplication events (Vollrath et al. 1988). The spectral differences between the two pigments are attributable to base changes in exon 5. In addition, the two genes are distinguished by a major polymorphism, in which the red gene has a 1.8-kb insertion in intron 1, with a resultant increase from 13.4 to 15.2 kb in genomic span compared to the green gene. Green pigment genes are proposed to be duplicated by homologous recombination in intergenic crossover events, whereas red genes are duplicated in intragenic events (Neitz and Neitz 1995).

Emery-Dreifuss muscular dystrophy (*EMD*; Emery 1989; OMIM no. 310300) results from lesions in the emerin gene, which is comprised of 6 exons spanning ~2 kb in a GC-rich region of Xq28 (Bione et al. 1994). This 220-kb region of the genome is gene-dense, containing at least 14 known genes (Chen et al. 1996). Located adjacent to *EMD* and transcribed in the opposite direction is the filamin gene (*FLN*). This gene encodes an actin binding protein, with 48 exons spanning ~26 kb. Flanking a 38-kb segment containing the two genes are two 11.3-kb inverted repeats that are 99.2% identical (Chen et al. 1996). It has been shown that these repeats can lead to a complete deletion of the emerin gene as well as a partial duplication of the adjacent *FLN* gene, apparently resulting from mispairing of the inverted repeats followed by a double recombination event (Small et al. 1997).

Recombination between the inverted repeats also apparently contributes to the persistence of both the homogeneity of the 11.3-kb sequences and the inversion of the intervening 38-kb region containing the *FLN* and *EMD* genes. The inversion is frequent enough to make 33% of females heterozygous for the region (i.e., having one X chromosome

## MAZZARELLA AND SCHLESSINGER

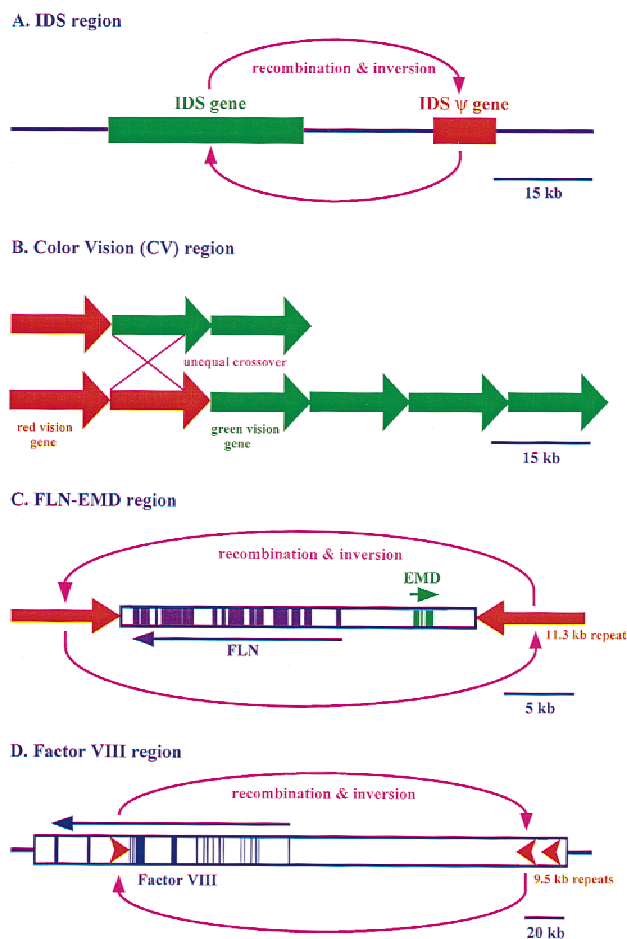


Figure 2 Proposed recombination events between homologous sequences affecting genes in Xq28. Recombinations lead to, respectively (A) Hunter syndrome; (B) color blindness, (C) polymorphism (inversion) and occasional further crossover/deletion to give EMD muscular dystrophy, and (D) hemophilia A.

with the region in one orientation and the other X with the opposite orientation; Small et al. 1997). The data also suggest an explanation for reported discrepancies between genetic and physical map distances in this region of Xq28 (Small et al. 1997).

The Xq28 region also contains at least two other regions in which nearby but nontandem duplications are involved in inherited disease. Hunter syndrome (mucopolysaccharidosis type II) is an X-linked lysosomal storage disorder caused by a deficiency in the activity of the enzyme iduronate-2-sulfatase (IDS) (Young and Harper 1982; OMIM no. 309900). About 3 kb of the *IDS* gene is duplicated 20 kb distal to the active gene (Bondeson et al. 1995a,b; Timms et al. 1995, 1997), and a significant fraction of Hunter syndrome cases (15%) are caused by recombination between the gene and its pseudogene, with the consequent deletion of the intervening

material (Bondeson et al. 1995b). In addition to localizing the duplicated segment, genomic sequencing found several nearby genes that are affected by more extensive deletions in severe Hunter syndrome cases with additional phenotypes.

Hemophilia A (coagulant factor VIII deficiency; OMIM no. 306700) is another of the paradigmatic X-linked recessive disorders. The 26 exons of the factor VIII gene are scattered in 180 kb of genomic DNA and are transcribed in the telomeric to centromeric direction. A CpG island is located ~10 kb downstream from exon 22, in the largest 32-kb intron of the gene. The CpG island appears to function as a bidirectional promoter encoding two different transcripts, referred to as factor VIII-associated genes *A* and *B* (Levinson et al. 1990, 1992). The part of intron 22 that contains the CpG island is repeated in extragenic copies situated ~300 kb and 400 kb telomeric to the 5' end of the factor VIII gene (Naylor et al. 1995). DNA sequencing and chemical mismatch analysis have demonstrated that these three repeat units are 9.5 kb long and 99.9% identical. About 45% of the cases of the severe form of hemophilia A arise by recombinational inversion occurring between the intragenic copy and one of the extragenic copies of the sequence (Lakich et al. 1993; Naylor et al. 1993; Tuddenham et al. 1994). This results from homologous mispairing and a single crossover event. As a consequence, the factor VIII gene becomes disrupted, with exons 1–22 dissociated from and flipped to an orientation opposite that of exons 23–26.

The melanoma antigen gene (*MAGE*) family is comprised of 12 genes found in three clusters of four genes, all in Xq28 (Rogner et al. 1995), and an additional cluster of 4 genes located in Xp21.3 (Lurquin et al. 1997). The coding region of each *MAGE* gene is a single exon. Those in Xq28 are 69–98% identical, and those in Xp21.3 are 66–81% identical; there is 45–63% identity between the genes at the two locations. These genes encode a melanoma antigen, with products detected from six of the genes in Xq28 (*MAGE A1, A2, A3, A4, A6, and A12*) in lung cancers, sarcomas, leukemias, colon cancers, and breast carcinomas (van der Bruggen et al. 1991; De Plaen et al. 1994; De Smet et al. 1994). Similarly, two of the genes from the Xp21.3 region (*MAGE B1* and *B2*) are expressed in a significant fraction of tumors from different histological origins (Lurquin et al. 1997). The identification of the tumor-specific antigen genes within these clusters is significant, because they might be candidates for immunotherapeutic intervention. In addition, the *MAGE* genes could be involved in hereditary disease as they

could again provoke gene dosage changes in Xq28, or in Xp21.3, where the cluster maps within the critical region for the dosage-sensitive sex (DSS) reversal [locus duplication of that region results in a male-to-female sex reversal phenotype (Bardoni et al. 1994; OMIM no. 300018)].

X-linked ichthyosis (OMIM no. 308100) provides a well-characterized example of pathology caused by duplications at some distance from one another. This disease was mapped to Xp22.32, and ~90% of ichthyosis patients were found to be deleted for the entire steroid sulfatase gene (Ballabio et al. 1989; Shapiro et al. 1989). Molecular analysis of the region revealed four homologous sequence elements, one distal and three proximal to steroid sulfatase (STS), distributed over 2.5 Mb. Subsequent studies showed that the majority of deletion patient breakpoints occurred within these homologous sequences, indicating recombination between these noncoding duplicated elements (Yen et al. 1990).

Pelizaeus-Merzbacher disease (PMD) is located in Xq22, and in many individuals is caused by a duplication of the proteolipid protein (*PLP*) gene (Woodward et al. 1998; OMIM no. 312080). Analysis of patient DNAs has shown that the duplication can vary from 500 kb to 1.65 Mb in length, although the patients all share the same distal end, differing in the proximal end of the duplicated region. Because affected males are homozygous for a variety of polymorphic markers in the region, it appears that the duplicated alleles are derived from the same chromosome. Therefore, the duplication may arise by intrachromosomal rearrangement.

Another disease that may be caused by a recombination between homologous but relatively distant repeats is the X-linked lymphoproliferative disorder (OMIM no. 308240). Several patients have similarly sized deletions in Xq25, and mapping of the region in the vicinity of the breakpoints with PCR-based markers shows that some sequences are repeated in the areas that border the deletions (Porta et al. 1997). In addition, a gene of the immunoglobulin superfamily occurs outside of the deletion borders but near the disease locus, and exhibits striking homology to natural killer (NK) receptors (Mazzarella et al. 1998). This observation may be a coincidence, but lymphoproliferative patients are deficient in NK cell activity (Sullivan et al. 1980), and there may be functional clustering of genes in the region.

#### Autosomal Clusters and Pathophysiology

Comparable events documenting the relationship between sequence duplication and disease have

been observed on autosomes (Table 1). Figure 3 characterizes 12 instances, categorized roughly as arising from unequal crossover between clusters of related gene sequences (A–F), changes involving an intrachromosomal recombination step (G–K), and an example of a putative unequal crossover followed by gene conversion (L). Here we outline these cases further.

The hemoglobinopathies (Fig. 3A,B) are classic autosomal examples of sequence duplication leading to human pathology (Maniatis et al. 1980; Collins and Weissman 1984; Orkin and Kazazian 1984; Antonarakis et al. 1985; Higgs et al. 1989). The hemoglobin tetramer is composed of two  $\alpha$  (or  $\alpha$ -like) subunits and two  $\beta$  (or  $\beta$ -like) subunits. The  $\alpha$  gene cluster is found in chromosome 16p13.3 and is comprised of two active  $\alpha$  genes ( $\alpha_1$  and  $\alpha_2$ ), two pseudogenes ( $\psi\alpha_1$  and  $\psi\alpha_2$ ), and the embryonically expressed  $\alpha$ -like  $\zeta$  gene ( $\zeta_2$ ) and its pseudogene ( $\psi\zeta_1$ ). The majority of lesions of the  $\alpha$ -globin gene cluster are the products of deletion, resulting in  $\alpha$ -thalassemia (OMIM no. 141850; Fig. 3A). The  $\alpha_1$  and  $\alpha_2$  genes are nearly identical at the nucleotide level and encode identical proteins. The level of homology between the two genes extends ~1 kb into the 5' flanking region, and overall the two genes are highly homologous over 4 kb. Homologous exchanges appear to promote unequal crossover events, resulting in one chromosome with added globin genes and the other chromosome with less or no globin genes. As a result of such unequal genetic exchanges, individuals may have as many as 6  $\alpha$  genes, although the excess production of  $\alpha$ -globin appears to have no negative consequences.

The  $\beta$  gene cluster is located at 11p15.5 and is composed of the embryonic  $\epsilon$  gene, two fetal  $\gamma$  genes ( $G\gamma$  and  $A\gamma$ ), the  $\delta$  gene, and the  $\beta$  gene and its pseudogene ( $\psi\beta_1$ ). Deletions of the  $\beta$  cluster resulting in  $\beta$ -thalassemia (OMIM #141900; Fig. 3B) are rare as the homologous regions between the cluster members are limited to portions of the exons. One type of cataloged defect is of particular interest here: a deletion of ~7 kb observed in patients with Hb Lepore. Analysis of these individuals reveals a hybrid gene produced by fusion of the 5' portion of the  $\delta$  *globin* gene with the 3' portion of the  $\beta$  *globin* gene. This observation suggests that an unequal crossover has occurred between the adjacent genes. A similar recombination mechanism has been postulated for the fusion of  $A\gamma$  and  $\beta$  in patients with Hb Kenya.

Pathology based on numbers of repeats certainly can extend to dispersed gene families. The P-450 superfamily (involved in Fig. 3C,D,L) consists

## MAZZARELLA AND SCHLESSINGER

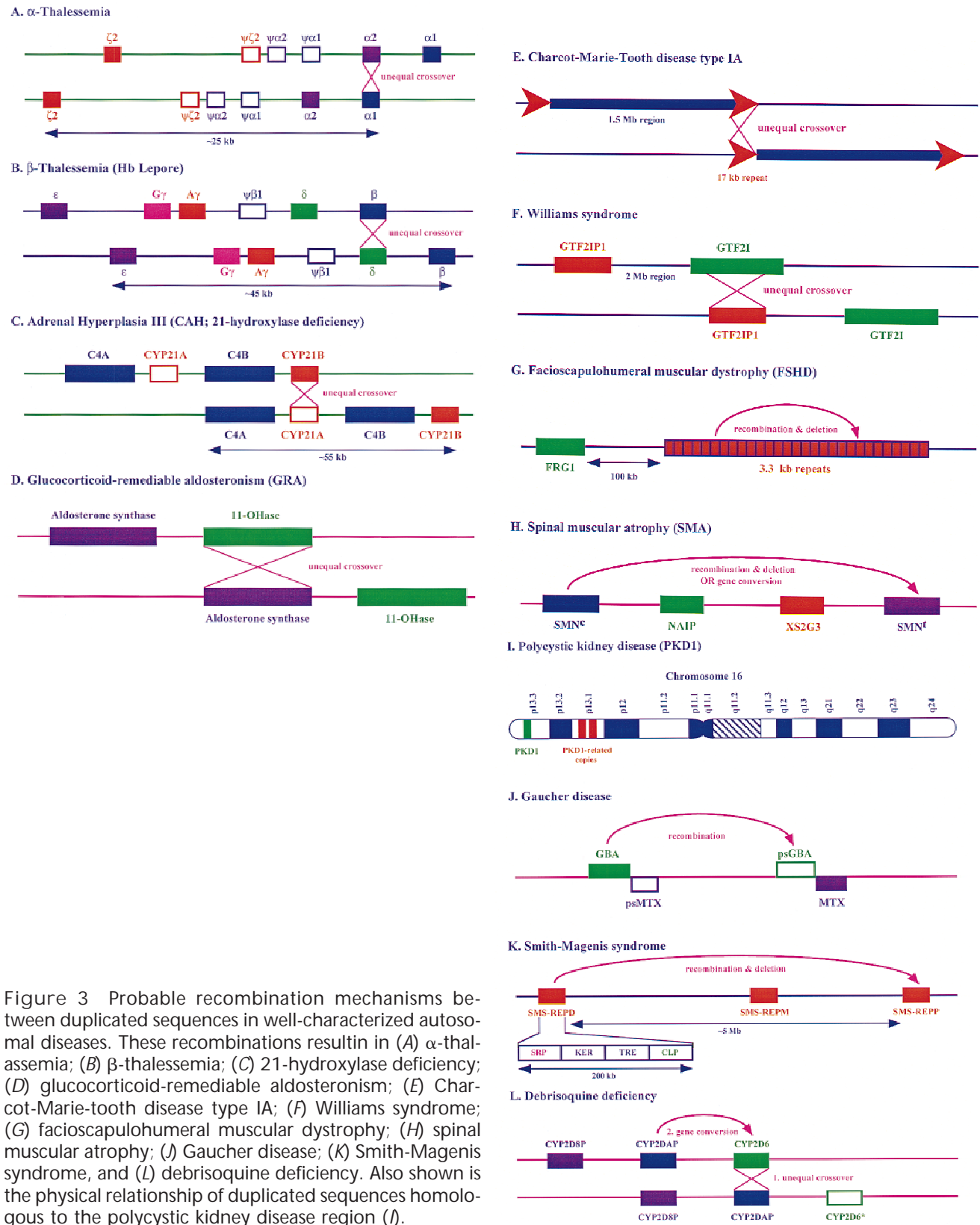


Figure 3 Probable recombination mechanisms between duplicated sequences in well-characterized autosomal diseases. These recombinations result in (A)  $\alpha$ -thalassemia; (B)  $\beta$ -thalassemia; (C) 21-hydroxylase deficiency; (D) glucocorticoid-remediable aldosteronism; (E) Charcot-Marie-tooth disease type IA; (F) Williams syndrome; (G) facioscapulohumeral muscular dystrophy; (H) spinal muscular atrophy; (J) Gaucher disease; (K) Smith-Magenis syndrome, and (L) debrisoquine deficiency. Also shown is the physical relationship of duplicated sequences homologous to the polycystic kidney disease region (I).

of >10 gene families and 100 genes that are localized to a least 6 different chromosomes, including 6, 7, 10, 15, 19, and 22 (Nebert et al. 1991). Four of these cytochrome P-450 enzyme families are responsible for the metabolism of numerous substrates including steroids and drugs (Nebert and Gonzalez 1987).

About 95% of the cases of congenital adrenal hyperplasia (CAH; adrenal hyperplasia III) are caused by deficiency of the enzyme 21-hydroxylase (21-OHase; OMIM no. 201910; Fig. 3C), which is one member of the cytochrome P-450 superfamily. This autosomal recessive disorder is based on events occurring within the MHC locus in 6p21.3 (Werkmeister et al. 1986). Molecular analysis of the region in normal individuals reveals two 21-OHase genes alternating with two complement four genes (*C4A* and *C4B*; Donohoue et al. 1986; Werkmeister et al. 1986). The 21-OHase genes include one inactive pseudogene (*CYP21A*) and the other (*CYP21B*) encoding the active gene product. Homology between the gene and the pseudogene is 98% in the coding regions and 96% in the intronic regions (White et al. 1985). The disease state arises from several different mechanisms including point mutation, deletion, duplication, and gene conversion. The latter three lesions probably result from recombinations at meiosis between the pseudogene and the active gene. One study has demonstrated that an unequal crossover between *CYP21A* and *CYP21B* genes results in deletion of the active gene, and that such crossovers occur at specific regions with the homologous genes (Donohoue et al. 1989).

Glucocorticoid-remediable aldosteronism (GRA) is a rare autosomal dominant disorder in 8q21 (also known as glucocorticoid-suppressible hyperaldosteronism; GSH; OMIM no. 103900; Fig. 3D; Lifton et al. 1992). Present in this region of the genome are another two members of the P-450 family, aldosterone synthase (*CYP11B2*) and steroid 11- $\beta$ -hydroxylase (*CYP11B1*), which are 95% identical and arranged in a head-to-tail configuration. The GRA disorder arises from an unequal crossover between the two genes, which produces a chimeric gene containing the 5' regulatory region of the 11- $\beta$ -hydroxylase gene fused to the coding sequence of aldosterone synthase. Aberrant expression of aldosterone synthase activity in the adrenal fasciculata results because its transcription is now controlled by adrenocorticotrophic hormone (ACTH) because of the 11- $\beta$ -hydroxylase regulatory sequences.

Pathology based on gene dosage and unequal crossing-over is seen in Charcot-Marie-Tooth disease type IA (CMT1A), a dominant peripheral neuropathy mapped to chromosome 17p11.2-p12

(OMIM no. 118220; Fig. 3E). A large 17-kb repeat is involved in the etiology of this disease. Two copies flank a 1.5-Mb region in unaffected individuals; but in patients, physical mapping of the region detected a tandem duplication of a 1.5-Mb segment (Pentao et al. 1992). Further analysis of chromosomes from patients showed that they contain three copies of the 17-kb repeat rather than the usual two. Duplication was suggested to have arisen by misalignment of the 17-kb repeat sequences followed by unequal crossover during meiosis. Thus, the duplication in most patients has been termed a kind of segmental trisomy (Matise et al. 1994; Schiavon et al. 1994). In striking support of this notion, Huxley et al. (1996) created a mouse model sharing many of the features of CMT1A by pronuclear injection of a yeast artificial chromosome (YAC) containing the locus.

Deletion rather than additional copies is involved in Williams syndrome, an autosomal dominant syndrome based on 7q11.23 (OMIM no. 194050; Fig. 3F). A similar deletion of ~2 Mb, apparently arising independently many times, has been characterized in many affected individuals (Nickerson et al. 1995; Osborne et al. 1996). About 90% of Williams patients are hemizygous for the elastin gene, having deleted the copy from one chromosome (Nickerson et al. 1995). The mechanism of deletion is unknown, but is again likely to involve the pairing of homologous sequences and a crossover that loses the intervening DNA. Recent evidence indicates that the homologous sequences may involve the *GTF2I* gene and its pseudogene. The *GTF2I* gene encodes the transcription initiator binding protein TFII-I, a phosphorylation substrate for the Bruton's tyrosine kinase, and maps near the telomeric breakpoint of the 2-Mb deletion; its pseudogene *GTF2IP* maps close to the centromeric breakpoint (Perez-Jurado et al. 1998).

Complex recombination and deletion events are thought to underlie facioscapulohumeral muscular dystrophy (FSHD), an autosomal dominant myopathy in 4q35 (Wijmenga et al. 1990; Lunt and Harper 1991; OMIM no. 158900; Fig. 3G). Analysis of a polymorphic *EcoRI* fragment tightly linked to the *FSHD* disease region revealed rearrangements in FSHD patients (Upadhyaya et al. 1991). Further analysis of this polymorphic marker showed that it can vary in size from 10 kb in affected individuals to 300 kb in normal individuals (Lee et al. 1995). The disease state has been correlated with the deletion of an integral number of 3.3-kb tandemly repeated units contained within the *EcoRI* fragment (van Deutekom et al. 1993). These repeats contain two

## MAZZARELLA AND SCHLESSINGER

known repetitive elements and two homeodomain motifs, although no corresponding transcripts have been detected (Hewitt et al. 1994). FISH experiments suggest that the repeated units are members of a 3.3-kb repeat family found in the heterochromatic regions of the genome (Lyle et al. 1995). This suggested that the deletion of an integral number of the repeats may lead to position effect variegation, repressing transcription of a nearby gene and thus leading to FSHD. A candidate gene (*FRG1*) has been identified ~100 kb centromeric to the repeat. It appears to belong to a multigene family with related sequences on multiple chromosomes, although there is thus far no evidence for the postulated repression of its transcription in patients (van Deutekom et al. 1996).

Recombination/deletion or gene conversion can be invoked in spinal muscular atrophy (SMA), an autosomal recessive disorder that is classified into three forms [Pearn 1980; Melki et al. 1990, 1994; OMIM no. 253300 (type I); OMIM no. 253550 (type II); OMIM no. 253400 (type III); Fig. 3H]. All three types of SMA map to 5q11.2–q13.3, a region of the genome containing multiple copies of different markers and genes (Thompson et al. 1995; Wirth et al. 1995). Three cDNAs found in the region have been used as probes to detect deletions in various SMA patients. Both copies of the neuronal apoptosis inhibitory protein (NAIP) gene and the *XS2G3* gene are deleted in ~50% of the patients with the most severe form of the disease (type I) and may contribute to the severity of the disease (Lefebvre et al. 1995; Roy et al. 1995). The third gene, the survival motor neuron (SMN) gene, is present in two nearly identical copies, referred to as the centromeric SMN gene (*SMN<sup>c</sup>*) and the telomeric SMN gene (*SMN<sup>t</sup>*) (van der Steege et al. 1995; McAndrew et al. 1997). The *SMN<sup>t</sup>* gene is absent in 95% of SMA patients, as a result either of sequence conversion (*SMN<sup>t</sup>* conversion to *SMN<sup>c</sup>*, giving rise to two *SMN<sup>c</sup>* copies) or *SMN<sup>t</sup>* gene deletion (DiDonato et al. 1997a). Sequence conversion is in fact known to be a common event in the milder forms of the disease (types II and III) (DiDonato et al. 1997b).

Comparably complex interactions of multiple copies of a long (50 kb) region are involved in autosomal dominant polycystic kidney disease (ADPKD), one of the most common genetic diseases, with a reported incidence of 1 in 1000 individuals (Gabow 1991; Fig. 3I). There is considerable variability in the age of onset and severity of the disease. Some of the variability can be explained by linkage to different genetic loci, with polycystic disease 1 (PKD1) occurring on 16p13.3 (OMIM no. 601313),

PKD2 on chromosome 4 (OMIM no. 173910), and PKD3 (OMIM no. 600666) as yet unmapped (Brook-Carter et al. 1994; Bogdanova et al. 1995; Daoust et al. 1995). In general, PKD1 appears to be a less severe form. The analysis of PKD1 has been complicated by the occurrence of at least three additional copies of a 50-kb region, containing the entire *PKD1* gene with the exception of 3.5 kb at the 3' end, on 16p13.1 (European Polycystic Kidney Disease Consortium 1994). The duplicated genomic regions are >95% identical. Interestingly, all of these copies produce polyadenylated transcripts but it is not known whether they encode proteins.

Gaucher disease (OMIM no. 230800; Fig. 3J) presents a relatively simple case of recombination-based deletion between repeated segments. The disease results from glucocerebrosidase deficiency and is the most common inherited lysosomal enzyme disorder. The glucocerebrosidase (*GBA*) gene is encoded by 11 exons (Choudary et al. 1985; Horowitz et al. 1989) and is located on chromosome 1q21 (Ginns et al. 1985). A pseudogene (*psGBA*) significantly contributes to the disease condition (Tsuji et al. 1987) and is located ~16 kb telomeric to *GBA* (Winfield et al. 1997). A number of mutations occurring in the pseudogene are detected in the encoded products of patients affected by the disease, apparently resulting from recombination between the two homologous sequences (Eyal et al. 1990; Latham et al. 1990; Zimran et al. 1990). In this case, the extent and evolutionary history of the duplication can be discerned partially. The duplication includes a second gene sequence for metaxin (*MTX*). One copy is adjacent and on the DNA strand opposite the *psGBA* gene; a corresponding pseudogene (*psMTX*) is nearby on the same strand. Analysis of sequence from the region indicates that the overall duplication extends ~14 kb, and occurred at an evolutionary time before the insertion of a 6.1-kb segment and several Alu sequences (Winfield et al. 1997).

A “common deletion” spanning 5 Mb is also seen in more than 90% of the patients affected with Smith-Magenis syndrome (SMS) in chromosome 17p11.2, ~500 kb proximal to the *CMT1A* disease region (Chen et al. 1997; OMIM no. 182290; Fig. 3K). Analysis of the region revealed three 200-kb low-copy repeats, two flanking the deletion and one in the middle of the deleted region. Further characterization of the repeats has shown that each repeat represents a gene cluster containing significant homologies to four different genes: coactosin-like protein (*CLP*), signal recognition particle (*SRP*), type-I keratin (*KER*), and the *TRE* oncogene (*TRE*). It is

unclear whether these genes are functional copies or pseudogenes. Examination of patient DNA showed that recombination almost always occurred between the proximal and distal repeats, presumably by intrachromosomal rearrangement, although other mechanisms are possible.

Still another instance involving the P-450 superfamily (Fig. 3, cf. L with C and D) involves a gene cluster of four members (the *CYP2D* subfamily cluster) localized at 22q13.1. It contains the functional *CYP2D6* gene and two highly homologous pseudogenes, and is important in the metabolism of ~20% of commonly prescribed drugs (Gonzalez et al. 1988; Kimura et al. 1989; OMIM no. 124030; Fig. 3L). Five to 10% of white populations are poor metabolizers of the antihypertensive debrisoquine and other plant alkaloids because of a genetic deficiency at the P-450 *CYP2D6* locus (Meyer et al. 1990). Several haplotypes of this gene cluster have been identified by restriction fragment length polymorphisms (Skoda et al. 1988). One of the haplotypes was found to contain four *CYP2D*-related genes, instead of the three found in most individuals (Heim and Meyer 1992). Comparison of the genes suggests that an early point mutation was followed by a crossover and gene conversion event. This would result in a net yield of three pseudogenes and a mutant *CYP2D6* gene, resulting in the deficient metabolism of debrisoquine and other drugs (Heim and Meyer 1992).

Homologous recombination between repeated sequences has also been implicated as the mechanism by which a "common deletion" is produced in Prader-Willi syndrome (PWS; OMIM no. 176270) and Angelman syndrome (AS; OMIM no. 105830) patients in chromosome 15q11–q13 (Christian et al. 1995; Huang et al. 1997). Seventy percent of the individuals afflicted with PWS and AS result from a ~4-Mb deletion in the parental and maternal genomes, respectively. In addition, this region is subject to duplications and supernumerary marker formation. The recent construction of a detailed YAC map encompassing the region should aid in the resolution of which elements are responsible for this chromosomal rearrangement (Christian et al. 1998).

Low-copy repeats that lower the dosage of critical genes may also be involved in the deletion events seen in DiGeorge syndrome (DGS; OMIM no. 188400) and Velocardiofacial syndrome (VCFS; OMIM no. 192430). These syndromes are caused by haplo insufficiency of genes in chromosome 22q11. DGS is the more severe of the two disorders, including the VCFS phenotype as well as additional abnormalities. About 80%–85% of the DGS/VCFS patients

have been shown to have deletions of more than 1 Mb. Using FISH it has been shown that several low-copy repeat families flank the *DGS/VCFS* locus (Halford et al. 1993). Recently, a novel transmembrane protein has been identified as deleted in >80% of VCFS patients (Sirotkin et al. 1997). Further molecular studies of the region should specify or discount the role of the low-copy repeats in the deletion mechanism.

Although we have concentrated on nuclear events, we note that comparable sequence duplications and comparable consequences are also observed in the human mitochondrial genome. For example, one-third of the patients with Kearns-Sayre syndrome (KSS) have a "common deletion" of their mitochondrial mtDNA, sometimes associated with a tandem duplication (Holt et al. 1988; Zeviani et al. 1988; Moraes et al. 1989; OMIM no. 530000). This common deletion was found to be mediated, presumably through homologous recombination, by a 13-bp repeated sequence present in normal mitochondrial DNA (Schon et al. 1989; Mita et al. 1990). Furthermore, it appears that duplications of the region are more prevalent in heart tissue, an observation possibly correlated with the extremely high numbers of mitochondria in cardiomyocytes (Fromenty et al. 1997).

The incidence and variety of such pathological changes in DNA would be still further sharply increased if deletions or additions that involve highly repetitive elements distributed throughout the genome were also included. These range from the expansion of microsatellite repeats to crossovers between copies of *Alu* or other repetitive elements, all of which have been excluded here to focus on locally duplicated longer sequence tracts.

#### Summary: Consequences of Clustering

Detailed structural analyses of the genome are increasingly revealing clusters of sequence that provide a snapshot of evolution generating new genetic possibilities. In simple cases, nearby repeated sequences can lead to deletions, inversions, and the production of considerable diversity. A second source of clustering arises when a sequence moves from one genomic site to another. This creates the possibility for coregulation of the juxtaposed sequences, even when they are quite dissimilar.

The significance of duplications is thus dependent on their frequency, dosage effects, and location, as well as the time at which they occurred in human evolution. For example, classic examples show that newly arising extra copies of a gene or a

## MAZZARELLA AND SCHLESSINGER

chromosome (as in the extreme case of trisomy 21) can be as detrimental as deletions. The repertoire of possible pathology then naturally increases for gene duplications that have had time to diverge. In an extreme example, two neighboring transporter genes on chromosome 7 that have diverged considerably cause two different diseases when mutated [Pendred syndrome (Everett et al. 1997) and congenital chloride diarrhea (Hoglund et al. 1996)].

The relative contributions of deletions/additions and point mutations to genetic pathology will depend on factors like the size of the gene, incidence and severity of the effects of lesions, and selective factors. In this context, nearby repeats increase qualitatively the frequency of dynamic changes in DNA composition; in instances such as Williams syndrome, color blindness, and hemophilia A, those changes are involved in a very large fraction of analyzed cases. The highest incidence can be quantitated in well-studied examples like hemophilia A, where the overall incidence is ~1 in 10,000 to 20,000, and ~40% of cases arise from repeat-catalyzed inversion events (Fig. 2; Table 1). This frequency is thus comparable to the combined incidence of all other deleterious changes in a gene that spans 180 kb of genomic DNA. In other cases, like Williams syndrome, rates of 1 in 100,000 are not uncommon.

To estimate the potential impact of the range of such effects, we can ask just how frequent are duplications that are evolutionarily deep-seated potential sources of pathology? Long-range sequencing on the X chromosome has progressed far enough to suggest that levels of 5–10% of the genome are duplicated at least once (as in Figs. 1 and 2). On autosomes, sequencing is only now beginning to accumulate rapidly (and regions with duplications are generally harder to sequence). Nevertheless, it is notable that in situ studies with cosmid probes on chromosome 7 found more than one site for the order of 10% of the cosmids (Green et al. 1994), and Korenberg et al. (quoted in Pennisi 1998) have found similar levels of cross-hybridizing loci for bacterial artificial chromosomes, particularly in gene-rich pericentromeric and subtelomeric regions.

Therefore, if we speculate that duplications comparable to those we discuss here are likely to be spread through the genome, then every individual will have an appreciable chance of having undergone such an event (that is, an inversion, addition/deletion, or deletion occurring at a rate of 1 in 10,000 to 100,000 per gene, and taking place in any of 10,000 susceptible genes). Thus, the limited number of examples shown here are a very small tip of a

very large iceberg. The precise determination of the extent of duplications is coming from the sequencing of the human genome that has already provided some of the examples, and will continue to provide the probe reagents to assess the range of incidence of variation in copy number, inversions, and deletions.

### What Are the Practical Consequences for Genetic Investigations?

First, workers investigating a genetic locus are well-advised to ask what are the neighboring genes. The number of instances in which the next gene is potentially relevant to functional analysis is going up very rapidly. For example, in two instances in which we have been recently involved in some of the studies, the X-linked anhidrotic ectodermal dysplasia (*EDA*) gene is juxtaposed with two other genes that show high levels of expression in skin (Kere et al. 1996) and the gene responsible for the Simpson-Golabi-Behmel syndrome, encoding glypican 3 (Pilia et al. 1996), turns out to be next to the glypican 4 (Watanabe et al. 1995) gene (work in progress; see GenBank accession no. AC00240).

Second, the processes of nearby duplications and interactions give rise to very appreciable diversity between individuals and populations.

Third, inversions, deletions, and other changes in DNA are favored by these clusters. They occur at frequencies on the order of  $10^{-4}$ – $10^{-6}$ , sufficient to result in very significant contributions to the comparable rates of incidence of genetic disease.

### ACKNOWLEDGMENTS

We thank our colleagues, including Lucio Luzzatto, Dan Longo, Reid Huber, and Giuseppe Pilia, for careful reading and suggestions.

### REFERENCES

- Antonarakis, S.E., H.H. Kazazian Jr., and S.H. Orkin. 1985. DNA polymorphism and molecular pathology of the human globin gene clusters. *Hum. Genet.* 69: 1–14.
- Babbitt, P.C. and J.A. Gerlt. 1997. Understanding enzyme superfamilies. *J. Biol. Chem.* 272: 30591–30594.
- Ballabio, A., R. Carozzo, G. Parerti, A. Gil, M. Zollo, M.G. Persico, E. Gillard, N. Affara, J. Yates, M.A. Ferguson-Smith, R.R. Frants, A.W. Eriksson, and G. Andria. 1989. Molecular heterogeneity of steroid sulfatase deficiency: A multicenter study on 57 unrelated patients, at DNA and protein levels. *Genomics* 4: 36–40.

## DUPLICATIONS AND GENETIC PATHOLOGY

- Bardoni, B., E. Zanaria, S. Guioli, G. Floridia, K.C. Worley, G. Tonini, E. Ferrante, G. Chiumello, E.R.B. McCabe, M.L. Fraccaro, O. Zuffardi, and G. Camerino. 1994. A dosage sensitive locus at chromosome Xp21 is involved in male to female sex reversal. *Nat. Genet.* 7: 497-501.
- Bione, S., E. Maestrini, S. Rivella, M. Mancini, S. Regis, G. Romeo, and D. Toniolo. 1994. Identification of a novel X-linked gene responsible for Emery-Dreifuss muscular dystrophy. *Nat. Genet.* 8: 323-327.
- Bogdanova, N., B. Dworniczak, D. Dragova, V. Todorov, D. Dimitrakov, K. Kalinov, J. Hallmayer, J. Horst, and L. Kalaydjieva. 1995. Genetic heterogeneity of polycystic kidney disease in Bulgaria. *Hum. Genet.* 95: 645-650.
- Bond, D.J. and A.C. Chandley. 1983. *Aneuploidy*. Oxford Monographs on Medical Genetics, No. 11. Oxford University Press, Oxford, UK.
- Bondeson, M.L., H. Malmgren, N. Dahl, B.M. Carlberg, and U. Pettersson. 1995a. Presence of an IDS-related locus (IDS2) in Xq28 complicates the mutational analysis of Hunter syndrome. *Eur. J. Hum. Genet.* 3: 219-227.
- Bondeson, M.L., N. Dahl, H. Malmgren, W.J. Kleijer, T. Tonnesen, B.M. Carlberg, and U. Pettersson. 1995b. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum. Mol. Genet.* 4: 615-621.
- Brook-Carter, P.T., B. Peral, C.J. Ward, P. Thompson, J. Hughes, M.M. Maheshwar, M. Nellist, V. Gamble, P.C. Harris, and J.R. Sampson. 1994. Deletion of the TSC2 and PKD1 genes associated with severe infantile polycystic kidney disease—A contiguous gene syndrome. *Nat. Genet.* 8: 328-332.
- Chen, E.Y., M. Zollo, R. Mazzarella, A. Ciccodicola, C.-N. Chen, L. Zuo, C. Heiner, F. Burrough, M. Repetto, D. Schlessinger, and M. D'Urso. 1996. Long-range sequence analysis in Xq28: Thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. *Hum. Mol. Genet.* 5: 659-668.
- Chen, K.-S., P. Manian, T. Koeuth, L. Potocki, Q. Zhao, A.C. Chinault, C.C. Lee, and J.R. Lupski. 1997. Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat. Genet.* 17: 154-163.
- Choudary, P.V., E.I. Ginns, and J.A. Barranger. 1985. Molecular cloning and analysis of the human  $\beta$ -glucocerebrosidase gene. *DNA* 4: 74.
- Christian, S.L., W.P. Robinson, B. Huang, A. Mutirangura, M.R. Line, M. Nakao, U. Surti, A. Chakravarti, and D.H. Ledbetter. 1995. Molecular characterization of two proximal deletion breakpoint regions in both Prader-Willi and Angelman syndrome patients. *Am. J. Hum. Genet.* 57: 40-48.
- Christian, S.L., N.K. Bhatt, S.A. Martin, J.S. Sutcliffe, T. Kubota, B. Huang, A. Mutirangura, A.C. Chinault, A.L. Beaudet, and D. Ledbetter. 1998. Integrated YAC contig map of the Prader-Willi/Angelman region on chromosome 15q11-q13 with average STS spacing of 35 kb. *Genome Res.* 8: 146-157.
- Collins, F.S. and S.M. Weissman. 1984. The molecular genetics of human hemoglobin. *Prog. Nucleic Acid Res. Mol. Biol.* 31: 315-462.
- Daoust, M.C., D.M. Reynolds, D.G. Bichet, and S. Somlo. 1995. Evidence for a third genetic locus for autosomal dominant polycystic kidney disease. *Genomics* 25: 733-736.
- De Plaen, E., K. Arden, C. Traversari, J.J. Gaforio, J.-P. Szikora, C. De Smet, F. Brasseur, P. van der Bruggen, B. Lethe, C. Lurquin, et al. 1994. Structure, chromosomal localization and expression of twelve genes of the MAGE family. *Immunogenetics* 40: 360-369.
- De Smet, C., C. Lurquin, P. van der Bruggen, E. De Plaen, F. Brasseur, and T. Boon. 1994. Sequence and expression pattern of the human MAGE2 gene. *Immunogenetics* 42: 282-290.
- DiDonato, C.J., X.-N. Chen, D. Noya, J.R. Korenberg, J.H. Nadeau, and L.R. Simard. 1997a. Cloning, characterization, and copy number of the murine survival motor neurone gene: Homolog of the spinal muscular atrophy-determining gene. *Genome Res.* 7: 339-352.
- DiDonato, C.J., S.E. Ingraham, J.R. Mendell, T.W. Prior, S. Lenard, R.T. Moxley, J. Florence, and A.H. Burghes. 1997b. Deletions and conversion in spinal muscular atrophy patients: Is there a relationship to severity? *Ann. Neurol.* 41: 230-237.
- Donohoue, P.A., C. Van Dop, N. Jospe, and C.J. Migeon. 1986. Congenital adrenal hyperplasia: Molecular mechanisms resulting in 21-hydroxylase deficiency. *Acta Endocrinol. (Suppl.)* 279: 315-320.
- Donohoue, P., N. Jospe, C.J. Migeon, and C. Van Dop. 1989. Two distinct areas of unequal crossingover within the steroid 21-hydroxylase genes produce absence of CYP21B. *Genomics* 5: 397-406.
- Eichler, E.E., F. Lu, Y. Shen, R. Antonacci, V. Jurecic, N.A. Doggett, R.K. Moyzis, A. Baldini, R.A. Gibbs, and D.L. Nelson. 1996. Duplication of a gene-rich cluster between 16p11.1 and Xq28: A novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* 5: 899-912.
- Emery, A.E.H. 1989. Emery-Dreifuss syndrome. *J. Med. Genet.* 26: 637-641. European Polycystic Kidney Disease Consortium. 1994. The polycystic kidney disease I gene encodes a 14-kb transcript and lies within a duplicated region of chromosome 16. *Cell* 75: 1305-1315.
- Everett, L.A., B. Glaser, J.C. Beck, J.R. Idol, A. Buchs, M. Heyman, F. Adawi, E. Hazani, E. Nassir, A.D. Baxevanis, V.C. Sheffield, and E.D. Green. 1997. Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nat. Genet.* 17: 411-422.

## MAZZARELLA AND SCHLESSINGER

- Eyal, N., S. Wilder, and M. Horowitz. 1990. Prevalent and rare mutations among Gaucher patients. *Gene* 96: 277–283.
- Fromenty, B., R. Carrozzo, S. Shanske, and E.A. Schon. 1997. High proportions of mtDNA duplications in patients with Kearns-Sayre syndrome occur in the heart. *Am. J. Med. Genet.* 71: 443–452.
- Gabow, P.A. 1991. Autosomal dominant polycystic kidney disease—more than a renal disease. *Am. J. Kidney Dis.* 16: 403–413.
- Ginns, E.I., P.V. Choudary, S. Tsuji, B. Martin, B. Stubblefield, J. Sawyer, J. Hozier, and J.A. Barranger. 1985. Gene mapping and leader peptide sequence of human glucocerebrosidase: Implications for Gaucher disease. *Proc. Natl. Acad. Sci.* 82: 7101–7105.
- Gonzalez, F.J., F. Vilbois, J.P. Hardwick, O.W. McBride, D.W. Nebert, H.V. Gelboin, and U.A. Meyer. 1988. Human debrisoquine 4-hydroxylase (P450IID1): cDNA and deduced amino acid sequence and assignment of the CYP2D locus to chromosome 22. *Genomics* 2: 174–179.
- Green, E.D., J.R. Idol, R.M. Mohr-Tidwell, V.V. Braden, D.C. Peluso, R.S. Fulton, H.F. Massa, C.L. Magness, A.M. Wilson, J. Kimura et al. 1994. Integration of physical, genetic and cytogenetic maps of human chromosome 7: Isolation and analysis of yeast artificial chromosome clones for 117 mapped genetic markers. *Hum. Mol. Genet.* 3: 489–501.
- Halford, S., E. Lindsay, M. Nayudu, A.H. Carey, A. Baldini, and P.J. Scambler. 1993. Low-copy-number repeat sequences flank the DiGeorge/velo-cardio-facial syndrome loci at 22q11. *Hum. Mol. Genet.* 2: 191–196.
- Heim, M.H. and U.A. Meyer. 1992. Evolution of a highly polymorphic human cytochrome P450 gene cluster. *Genomics* 14: 49–58.
- Hewitt, J.E., R. Lyle, L.N. Clark, E.M. Valleley, T.J. Wright, C. Wijmenga, J.C.T. Van Deutekom, F. Francis, P.T. Sharpe, M. Hofker, R.R. Frants, and R. Williamson. 1994. Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. *Hum. Mol. Genet.* 3: 1287–1295.
- Higgs, D.R., M.A. Vickers, A.O.M. Wilkie, I.-M. Pretorius, A.P. Jarman, and D.J. Weatherall. 1989. A review of the molecular genetics of the human  $\alpha$ -globin gene cluster. *Blood* 73: 1081–1104.
- Hoglund, P., S. Haila, J. Socha, L. Tomaszewski, U. Saarialho-Kere, M.L. Karjalainen-Lindsberg, K. Airola, C. Holmberg, A. de la Chapelle, and J. Kere. 1996. Mutations of the Down-regulated in adenoma (DRA) gene cause congenital chloride diarrhoea. *Nat. Genet.* 14: 316–319.
- Holt, I.J., A.E. Harding, and J.A. Morgan-Hughes. 1988. Deletions of mitochondrial DNA in patients with mitochondrial myopathies. *Nature* 331: 717–719.
- Hood, L., M. Kronenberg, and T. Hunkapillar. 1985. T cell antigen receptors and the immunoglobulin supergene family. *Cell* 40: 225–229.
- Horowitz, M., S. Wilder, Z. Horowitz, O. Reiner, T. Gelbart, and E. Beutler. 1989. The human glucocerebrosidase gene and pseudogene: Structure and evolution. *Genomics* 4: 87–96.
- Huang, B., J.A. Crolla, S.L. Christian, M.E. Wolf-Ledbetter, M.E. Macha, P.N. Papenhausen, and D.H. Ledbetter. 1997. Refined molecular characterization of the breakpoints in small inv dup(15) chromosomes. *Hum. Genet.* 99: 11–17.
- Huxley, C., E. Passage, A. Manson, G. Putzu, D. Figarella-Branger, J.F. Pellissier, and M. Fontes. 1996. Construction of a mouse model of Charcot-Marie-Tooth disease type 1A by pronuclear injection of human YAC DNA. *Hum. Mol. Genet.* 5: 563–569.
- Kere, J., A.K. Srivastava, O. Montonen, J. Zonana, N. Thomas, B. Ferguson, F. Munoz, D. Morgan, A. Clarke, P. Baybayan et al. 1996. X-linked anhidrotic (hypohidrotic) ectodermal dysplasia is caused by mutation in a novel transmembrane protein. *Nat. Genet.* 13: 409–416.
- Kimura, S., M. Umeno, R.C. Skoda, U.A. Meyer, and F.J. Gonzalez. 1989. The human debrisoquine 4-hydroxylase (CYP2D6) locus: Sequence and identification of the polymorphic CYP2D6 gene, a related gene and a pseudogene. *Am. J. Hum. Genet.* 45: 889–904.
- Krumlauf, R. 1994. Hox genes in vertebrate development. *Cell* 78: 191–201.
- Lahn, B.T. and D.C. Page. 1997. Functional coherence of the human Y chromosome. *Science* 278: 675–680.
- Lakich, D., H.H. Kazazian, S.E. Antonarakis, and J. Gitschier. 1993. Inversions disrupting the factor VIII gene are a common cause of severe hemophilia A. *Nat. Genet.* 5: 236–241.
- Latham, T., G.A. Grabowski, B.D.M. Theophilus, and F.I. Smith. 1990. Complex alleles of the acid  $\beta$ -glucocerebrosidase gene in Gaucher disease. *Am. J. Hum. Genet.* 47: 79–86.
- Lee, J.H., K. Goto, K. Sahashi, I. Nonaka, C. Matsuda, and K. Arahata. 1995. Cloning and mapping of a very short (10-kb) EcoRI fragment associated with facioscapulohumeral muscular dystrophy (FSHD). *Muscle & Nerve (Suppl.)* 2: S27–S31.
- Lefebvre, S., L. Burglen, S. Reboullet, O. Clermont, P. Bulet, L. Viollet, B. Benichou, C. Cruand, P. Millasseau, M. Zeviani et al. 1995. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 80: 155–165.
- Levinson, B., S. Kenwrick, D. Lakich, G. Hammonds, Jr., and J. Gitschier. 1990. A transcribed gene in an intron of the human factor VIII gene. *Genomics* 7: 1–11.
- Levinson, B., S. Kenwrick, P. Gamel, K. Fisher, and J.

## DUPLICATIONS AND GENETIC PATHOLOGY

- Gitschier. 1992. Evidence for a third transcript from the human factor VIII gene. *Genomics* 14: 585–589.
- Lifton, R.P., R.G. Dluhy, M. Powers, G.M. Rich, S. Cook, S. Ulick, and J.-M. Lalouel. 1992. A chimaeric 11 $\beta$ -hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. *Nature* 355: 262–265.
- Lunt, P.W. and P.S. Harper. 1991. Genetic counseling in facioscapulohumeral muscular dystrophy. *J. Med. Genet.* 28: 655–664.
- Lurquin, C., C. De Smet, F. Brasseur, F. Muscatelli, V. Martelange, E. De Plaen, R. Brasseur, A.P. Monaco, and T. Boon. 1997. Two members of the human MAGEB gene family located in Xp21.3 are expressed in tumors of various histological origins. *Genomics* 46: 397–408.
- Lyle, R., T.J. Wright, L.N. Clark, and J.E. Hewitt. 1995. The FSHD-associated repeat, D4Z4, is a member of a dispersed family of homeobox-containing repeats, subsets of which are clustered on the short arms of the acrocentric chromosomes. *Genomics* 28: 389–397.
- Maniatis, T., E.F. Fritsch, J. Lauer, and R.M. Lawn. 1980. The molecular genetics of human hemoglobins. *Annu. Rev. Genet.* 14: 145–178.
- Matise, T.C., A. Chakravarti, P.I. Patel, J.R. Lupski, E. Nelis, V. Timmerman, C. Van Broeckhoven, and D.E. Weeks. 1994. Detection of tandem duplications and implications for linkage analysis. *Am. J. Hum. Genet.* 54: 1110–1121.
- Mazzarella, R. and D. Schlessinger. 1997. Duplication and distribution of repetitive elements and non-unique regions in the human genome. *Gene* 205: 29–38.
- Mazzarella, R., G. Pengue, J. Jones, C. Jones, and D. Schlessinger. 1998. Cloning and expression of an immunoglobulin superfamily gene (IGSF1) in Xq25. *Genomics* 48: 157–162.
- McAndrew, P.E., D.W. Parsons, L.R. Simard, C. Rochette, P.N. Ray, J.R. Mendell, T.W. Prior, and A.H.M. Burghes. 1997. Identification of proximal spinal muscular atrophy carriers and patients by analysis of SMN<sup>d</sup> and SMN<sup>c</sup> gene copy number. *Am. J. Hum. Genet.* 60: 1411–1422.
- McKusick, V.A. 1994. *Mendelian inheritance in man. Catalogs of human genes and genetic disorders*, 11th ed. Johns Hopkins University Press, Baltimore, MD.
- Melki, J., S. Abdelhak, P. Sheth, M.F. Bachelot, P. Burlet, A. Marcadet, J. Aicardi, A. Borois, J.P. Carriere, M. Fardeau et al. 1990. Gene for chronic proximal spinal muscular atrophies maps to chromosome 5q. *Nature* 344: 767–768.
- Melki, J., S. Lefebvre, L. Burglen, P. Burlet, O. Clemont, P. Millasseau, S. Reboullet, B. Benichou, M. Zeviani, D. Le Paslier et al. 1994. De novo and inherited deletions of the 5q13 region in spinal muscular atrophies. *Science* 264: 1474–1477.
- Meyer, U.A., U.M. Zanger, D. Grant, and M. Blum. 1990. Genetic polymorphism of drug metabolism. *Adv. Drug Res.* 19: 197–241.
- Mita, S., R. Rizzuto, C.T. Moraes, S. Shanske, E. Arnaudo, G.M. Fabrizi, Y. Koga, S. DiMauro, and E.A. Schon. 1990. Recombination via flanking direct repeats is a major cause of large-scale deletions of human mitochondrial DNA. *Nucleic Acids Res.* 18: 561–567.
- Moraes, C.T., S. DiMauro, M. Zeviani, A. Lombes, S. Shanske, A.F. Miranda, H. Nakase, E. Bonilla, L.C. Werneck, S. Servidet et al. 1989. Mitochondrial DNA deletions in progressive external ophthalmoplegia and Kearns-Sayre syndrome. *N. Engl. J. Med.* 320: 1293–1299.
- Muller, H.J. 1964. The relation of recombination to mutational advance. *Mutat. Res.* 1: 2–9.
- Nathans, J., S.L. Merbs, C.H. Sung, C.J. Weitz, and Y. Wang. 1992. Molecular genetics of human visual pigments. *Annu. Rev. Genet.* 26: 403–424.
- Naylor, J., A. Brinke, S. Hassock, P.M. Green, and F. Giannelli. 1993. Characteristic mRNA abnormality found in half the patients with severe hemophilia A is due to large DNA inversions. *Hum. Mol. Genet.* 2: 1773–1778.
- Naylor, J.A., D. Buck, P. Green, H. Williamson, D. Bentley, and F. Giannelli. 1995. Investigation of the factor VIII intron 22 repeated region (int22h) and the associated inversion junctions. *Hum. Mol. Genet.* 4: 1217–1224.
- Nebert, D.W. and F.J. Gonzalez. 1987. P450 genes: Structure, evolution and regulation. *Annu. Rev. Biochem.* 56: 945–993.
- Nebert, D.W., D.R. Nelson, M.J. Coon, R.W. Estabrook, R. Feyereisen, Y. Fukui-Kuriyama, F.J. Gonzalez, F.P. Guengerich, I.C. Gunsalus, E.F. Johnson et al. 1991. The P450 superfamily: Update on new sequences, gene mapping and recommended nomenclature. *DNA* 10: 1–14.
- Neitz, M. and J. Neitz. 1995. Numbers and ratios of visual pigment genes for normal red-green color vision. *Science* 267: 1013–1016.
- Nickerson, E., F. Greenberg, M.T. Keating, C. McCaskill, and L.G. Shaffer. 1995. Deletions of the elastin gene at 7q11.23 occur in approximately 90% of patients with Williams syndrome. *Am. J. Hum. Genet.* 56: 1156–1161.
- Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York, NY.
- OMIM, Online Mendelian Inheritance in Man. Center for Medical Genetics, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 1997. <http://www.ncbi.nlm.nih.gov/omim/>.
- Orkin, S.H. and H.H. Kazazian. 1984. The mutation and polymorphism of the human  $\beta$ -globin gene and its surrounding DNA. *Annu. Rev. Genet.* 18: 131–171.

## MAZZARELLA AND SCHLESSINGER

- Osborne, L.R., D. Duane Martindale, S.W. Scherer, X.-M. Shi, J. Huizenga, H.H.Q. Heng, T. Costa, B. Pober, L. Lew, J. Brinkman, J.M. Rommens, B. Koop, and L.-C. Tsui. 1996. Identification of genes from a 500 kb region at 7q11.23 that is commonly deleted in Williams syndrome. *Genomics* 36: 328–336.
- Pearn, J. 1980. Classification of spinal muscular atrophies. *Lancet* 1: 919–922.
- Pennisi, E. 1998. Genome research: Sifting through and making sense of genome sequences. *Science* 280: 1692–1693.
- Pentao, L., C.A. Wise, A.C. Chinault, P.I. Patel, and J.R. Lupski. 1992. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nat. Genet.* 2: 292–300.
- Perez-Jurado, L.A., Y.K. Wang, R. Peoples, A. Coloma, J. Cruces, and U.A. Francke. 1998. A duplicated gene in the breakpoint regions of the 7q11.23 Williams-Beuren syndrome deletion encodes the initiator binding protein TFII-I and BAP-135, a phosphorylation target of BTK. *Hum. Mol. Genet.* 7: 325–334.
- Pilia, G., R.M. Hughes-Benzie, A. MacKenzie, P. Baybayan, E.Y. Chen, R. Huber, G. Neri, A. Cao, A. Forabosco, and D. Schlessinger. 1996. Mutations in GPC3, a glypican gene, cause the Simpson-Golabi-Behmel overgrowth syndrome. *Nat. Genet.* 12: 241–247.
- Porta, G., S. MacMillan, R. Nagaraja, S. Mumm, I. Zucchi, G. Pilia, S. Maio, T. Featherstone, and D. Schlessinger. 1997. 4.5 Mb YAC/STS contig at 50 kb resolution, spanning Xq25 deletions in two patients with lymphoproliferative syndrome (LYP). *Genome Res.* 7: 27–36.
- Rice, W.R. 1987. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* 116: 161–167.
- Rogner, U.C., K. Wilke, E. Steck, B. Korn, and A. Poustka. 1995. The melanoma antigen gene (MAGE) family is clustered in the chromosomal band Xq28. *Genomics* 29: 725–731.
- Roy, N., M.S. Mahadevan, M. McLean, G. Shutler, Z. Yaraghi, R. Farahani, S. Baird, A. Besner-Johnston, C. Lefebvre, X. Kang et al. 1995. The gene for neuronal apoptosis inhibitory protein is partially deleted in individuals with spinal muscular atrophy. *Cell* 80: 167–178.
- Schiavon, F., M.L. Mostacciolo, F. Saad, L. Merlini, G. Siciliano, C. Angelini, and G.A. Danieli. 1994. Non-radioactive detection of 17p11.2 duplication in CMT1A: A study of 78 patients. *J. Med. Genet.* 31: 880–883.
- Schon, E.A., R. Rizzuto, C.T. Moraes, H. Nakase, M. Zeviani, and S. DiMauro. 1989. A direct repeat is a hotspot for large-scale deletions of human mitochondrial DNA. *Science* 244: 346–349.
- Shapiro, L.J., P. Yen, D. Pomerantz, E. Martin, L. Rolewic, and T. Mohandas. 1989. Molecular studies of deletions at the human steroid sulfatase locus. *Proc. Natl. Acad. Sci.* 86: 8477–8481.
- Sirotkin, H., B. Morrow, B. Saint-Jore, A. Puech, R. Das Gupta, S.R. Patanjali, A. Skoultchi, S.M. Weissman, and R. Kucherlapti. 1997. Identification, characterization, and precise mapping of a human gene encoding a novel membrane-spanning protein from the 22q11 region deleted in Velo-Cardio-Facial syndrome. *Genomics* 42: 245–251.
- Skoda, R.C., F.J. Gonzalez, A. Demierre, and U.A. Meyer. 1988. Two mutant alleles of the human cytochrome P450db1 gene (P450IID1) associated with genetically deficient metabolism of debrisoquine and other drugs. *Proc. Natl. Acad. Sci.* 85: 5240–5243.
- Small, K., J. Iber, and S. Warren. 1997. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat. Genet.* 16: 96–99.
- Speed, R.M. 1988. The possible role of meiotic pairing anomalies in the atresia of human fetal oocytes. *Hum. Genet.* 78: 260–266.
- Srivastava, A.K. and D. Schlessinger. 1991. Structure and organization of ribosomal DNA. *Biochimie* 73: 631–638.
- Sullivan, J.L., K.S. Byron, F.E. Brewster, and D.T. Purtilo. 1980. Deficient natural killer cell activity in X-linked lymphoproliferative syndrome. *Science* 210: 543–545.
- Thompson, T.G., C.J. DiDonato, L.R. Simard, S.E. Ingraham, A.H.M. Burghes, T.O. Crawford, C. Rochette, J.R. Mendell, and J.J. Wasmuth. 1995. A novel cDNA detects homozygous microdeletions in greater than 50% of type I spinal muscular atrophy patients. *Nat. Genet.* 9: 56–62.
- Timms, K.M., F. Lu, Y. Shen, C.A. Pierson, D.M. Muzny, Y. Gu, D.L. Nelson, and R.A. Gibbs. 1995. 130 kb of DNA sequence reveals two new genes and a regional duplication distal to the human iduronate-2-sulfate sulfatase locus. *Genome Res.* 5: 71–78.
- Timms, K.M., M.-L. Bondeson, M.A. Ansari-Lari, K. Lagerstedt, D.M. Muzny, S.P. Dugan-Rocha, D.L. Nelson, U. Pettersson, and R.A. Gibbs. 1997. Molecular and phenotypic variation in patients with severe Hunter syndrome. *Hum. Mol. Genet.* 6: 479–486.
- Tuddenham, E.G., R. Schwaab, J. Seehafer, D.S. Millar, J. Gitschier, M. Higuchi, S. Bidichandani, J.M. Connor, L.W. Hoyer, A. Yoshioka et al. 1994. Hemophilia A: Database of nucleotide substitutions, deletions, insertions and rearrangements of the factor VIII gene, second edition. *Nucleic Acids Res.* 22: 3511–3533 (corrected: *Nucleic Acids Res.* 22: 4851–4868).
- Tsui, S., P.V. Choudary, B.M. Martin, B.K. Stubblefield, J.A. Mayor, J.A. Barranger, and E.I. Ginns. 1987. A mutation in the human glucocerebrosidase gene in neuronopathic Gaucher's disease. *N. Engl. J. Med.* 316: 570–575.
- Upadhyaya, M., P.W. Lunt, M. Sarfarazi, W. Broadhead, J. Daniels, M. Owen, and P.S. Harper. 1991. A closely linked

## DUPLICATIONS AND GENETIC PATHOLOGY

- DNA marker for facioscapulohumeral disease on chromosome 4q. *J. Med. Genet.* 28: 665–671.
- van der Bruggen, P., C. Traversari, P. Chomez, C. Lurquin, E. De Plaen, B. Van den Eynde, A. Knuth, and T. Boon. 1991. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* 254: 1643–1647.
- van der Steege, G., P.M. Grootsholten, P. van der Vlies, T.G. Draaijers, J. Osinga, J.M. Cobben, H. Scheffer, and C.H.C.M. Buys. 1995. PCR-based DNA test to confirm clinical diagnosis of autosomal recessive spinal muscular atrophy. *Lancet* 345: 985–986.
- van Deutekom, J.C.T., C. Wijmenga, E.A.E. Van Tienhoven, A.M. Gruter, J.E. Hewitt, G.W. Padberg, G.-J. Van Ommen, M.H. Hofker, and R.R. Frants. 1993. FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum. Mol. Genet.* 2: 2037–2042.
- van Deutekom, J.C.T., R.J.L.F. Lemmers, P.K. Grewal, M. van Geel, S. Romberg, H.G. Dauwse, T.J. Wright, G.W. Padberg, M.H. Hofker, J.E. Hewitt, and R.R. Frants. 1996. Identification of the first gene (FRG1) from the FSHS region on human chromosome 4q35. *Hum. Mol. Genet.* 5: 581–590.
- Vollrath, D., J. Nathans, and R.W. Davis. 1988. Tandem array of human visual pigment genes at Xq28. *Science* 240: 1669–1672.
- Watanabe, K., H. Yamada, and Y. Yamaguchi. 1995. K-glypican: A novel GPI-anchored heparan sulfate proteoglycan that is highly expressed in developing brain and kidney. *J. Cell Biol.* 130: 1207–1218.
- Werkmeister, J.W., M.I. New, B. Dupont, and P.C. White. 1986. Frequent deletion and duplication of the steroid 21-hydroxylase genes. *Am. J. Hum. Genet.* 39: 461–469.
- White, P.C., D. Growwberger, B.J. Onufer, D.D. Chaplin, M.I. New, B. Dupont, and J.L. Strominger. 1985. Two genes encoding steroid 21-hydroxylase are located near the fourth component of complement. *Proc. Natl. Acad. Sci.* 81: 1089–1093.
- Wijmenga, C., R.R. Frants, O.F. Brouwer, P. Moerer, J.L. Weber, and G.W. Padberg. 1990. Location of facioscapulohumeral dystrophy and the scapulothoracic syndrome. *Lancet* 336: 651–653.
- Wilkie, A.O., D.R. Higgs, K.A. Rack, V.J. Buckle, N.K. Spurr, N. Fischel-Ghodsian, I. Ceccherini, W.R.A. Brown, and P.C. Harris. 1991. Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* 64: 595–606.
- Winfield, S.L., N. Tayebi, B.M. Martin, E.I. Ginns, and E. Sidransky. 1997. Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: Implications for Gaucher disease. *Genome Res.* 7: 1020–1026.
- Wirth, B., E. Hahnen, K. Morgan, C.J. DiDonato, A. Dadze, S. Rudnik-Schoneborn, L.R. Simard, K. Zerres, and A.H.M. Burghes. 1995. Allelic association and deletions in autosomal recessive proximal spinal muscular atrophy: Association of marker genotype with disease severity and candidate cDNAs. *Hum. Mol. Genet.* 4: 1273–1284.
- Woodward, K., E. Kendall, D. Vetrie, and S. Malcom. 1998. Pelizaeus-Merzbacher disease: Identification of Xq22 proteolipid-protein duplications and characterization of breakpoints by interphase FISH. *Am. J. Hum. Genet.* 63: 207–217.
- Yen, P.H., X.-M. Li, S.-P. Tsai, C. Johnson, T. Mohandas, and L.J. Shapiro. 1990. Frequent deletions of the human X chromosome distal short arm result from recombination between low copy repetitive elements. *Cell* 61: 603–610.
- Young, I.D. and P.S. Harper. 1982. Incidence of Hunter's syndrome. *Hum. Genet.* 60: 391–392.
- Zeviani, M., C.T. Moraes, S. DiMauro, H. Nakase, E. Bonilla, E.A. Schon, and L.P. Rowland. 1988. Deletions of mitochondrial DNA in Kearns-Sayre syndrome. *Neurology* 38: 1339–1346.
- Zimran, A., J. Sorge, E. Gross, M. Kubitz, C. West, and E. Beutler. 1990. A glucocerebrosidase fusion gene in Gaucher disease. Implications for the molecular anatomy, pathogenesis, and diagnosis of this disorder. *J. Clin. Invest.* 85: 219–222.