WILEY ENCYCLOPEDIA OF

Molecular Medicine

volume 5

WILEY ENCYCLOPEDIA OF

MOLECULAR MEDICINE

VOLUME 5

A Wiley-Interscience Publication
John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of molecular medicine.

p.; cm.

"Wiley-Interscience publication."

Incluces indexes.

ISBN 0-471-37494-6 (set : cloth : alk. paper)–ISBN 0-471-37497-0 (v. 1 : alk. paper) –ISBN 0-471-37498-9 (v. 2 : alk. paper)–ISBN 0-471-37496-2 (v. 3 : alk. paper)— ISBN 0-471-37495-4 (v. 4 : alk. paper)–ISBN 0-471-20306-8 (v. 5 : alk. paper)

1. Molecular biology–Encyclopedias. 2. Biotechnology–Encyclopedias. 3. Pathology, Molecular–Encyclopedias. I. John Wiley & Sons.

[DNLM: 1. Molecular Biology–Encyclopedias–English. 2. Biotechnology–Encyclopedias–English. 3. Clinical Medicine–Encyclopedias–English. 4. Genetics, Medical–Encyclopedias–English. QH 506 E568 2002]

QH506 .E536 2002

572.8′03—dc21 2001026922

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

S100 PROTEIN FAMILY

Claus W. Heizmann
B.W. Schäfer
University of Zurich
Zürich, Switzerland

One of the largest subfamilies of elongation factor (EF)-hand Ca^{2+}-binding proteins is constituted by the S100 proteins. They now comprise 18 different human members, which display a diverse pattern of cell- and tissue-specific distribution consistent with their pleiotropic intra- and extracellular functions. These functions are brought about through interaction with diverse target proteins, probably sometimes in a concerted manner with the ubiquitously expressed calmodulin.

A wide range of diseases, including cardiomyopathies, neurological diseases, chronic inflammation, and cancer were recently linked to deregulation of S100 gene expression. Hence, S100 proteins are currently considered for their potential use in clinical diagnostics as well as for therapeutic interventions.

PROTEIN STRUCTURE AND BIOCHEMICAL PROPERTIES

S100 proteins are characterized by two distinct EF-hand calcium-binding motifs with different affinities (Fig. 1). Usually, the carboxy-terminal EF-hand is referred to as the canonical Ca^{2+}-binding loop with a rather high affinity, whereas the amino-terminal loop is S100-specific and displays lower Ca^{2+}-binding affinity. These EF-hands are flanked by hydrophobic regions, responsible for interaction with target proteins, and separated by a central hinge region (1–3).

Generally S100 proteins can form homo- or heterodimers and bind four Ca^{2+} per dimer. Several S100 proteins also bind Zn^{2+}, however, to a different site. Zn^{2+}-binding is able to modify the Ca^{2+} affinity in a few cases. Furthermore, S100B and S100A5 also bind Cu^{2+}. Therefore there is considerable variation in the biochemical and metal-binding properties of S100 proteins, the physiological consequences of which remain to be investigated (2,3).

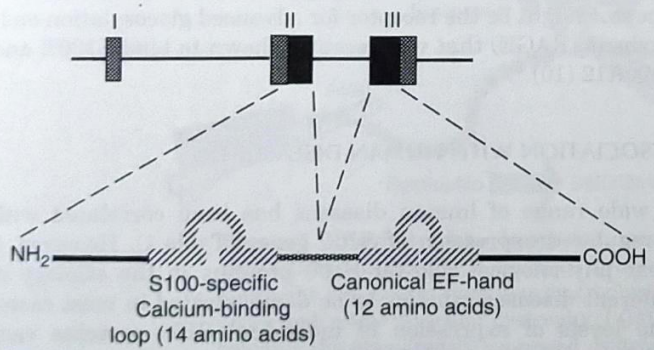

Figure 1. Generic *S100* gene and protein structure. A typical *S100* gene (upper panel) is composed of three exons, whereby exon 1 is not translated (grey box). Exons 2 and 3 are translated (black boxes) into protein (lower panel), which is composed of two Ca^{2+}-binding loops (grey), hydrophobic flanking regions (black), and a central hinge region (punctated).

GENOMIC ORGANIZATION

The structural organization of *S100* genes is highly conserved and a typical *S100* gene consists of three exons. Although the first exon carries exclusively 5′ untranslated sequences, the second exon contains the ATG and codes for the N-terminal EF-hand, and the third exon encodes the carboxy-terminal canonical EF-hand. A summary of all known *S100* gene structures is shown in Figure 2. Exceptions are *S100A4*, containing two alternatively spliced first exons and *S100A11* in which the first exon already contains coding sequences (1,4). Interestingly, two genes required during terminal differentiation of the epidermis contain an N-terminal S100 fusion (trichohyalin and profillagrin), which also confer to the generic *S100* gene structure (5).

So far, 13 *S100* genes and the two fusion genes are organized in a gene cluster on human chromosome 1q21 (Fig. 3), which led to the introduction of a new nomenclature (4). Recently, a similarly organized cluster of *S100* genes was found on mouse

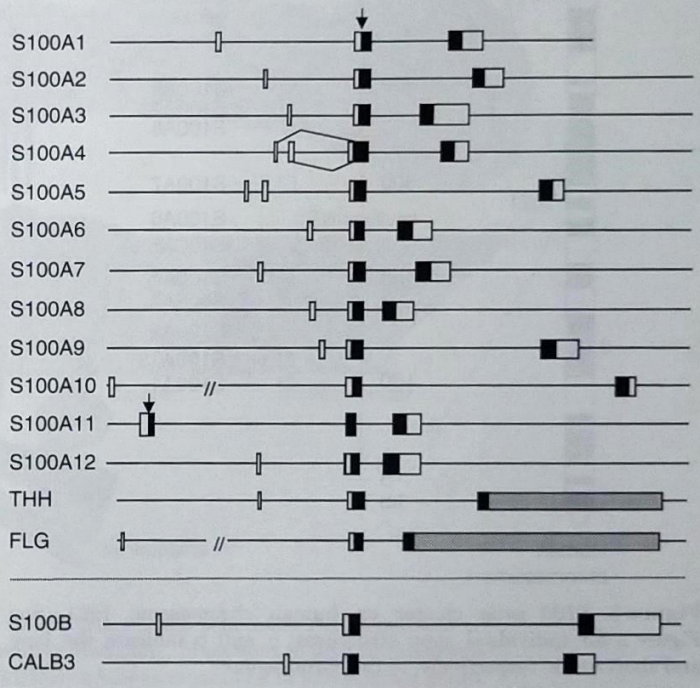

Figure 2. Known genomic structures of human *S100* genes. White boxes indicate untranslated exons, whereas black boxes represent coding sequences. The ATG start is indicated by an arrow. Genes above the separating line are located on human chromosome 1q21, whereas *S100B* and *CALB3* are located on chromosome 21 and X, respectively. Trichohyalin (*THH*) and profillagrin (*FLG*) are members of the epidermal differentiation genes and represent fusion genes between an *S100* gene and the structural part (grey box).

Table 1. Selected Functions of S100 Proteins

Protein	Postulated Functions	Possible Disease Association
S100A1	Muscle contraction	Cardiomyopathies
S100A2	Nuclear functions	Cancer
S100A3	Hair shaft formation	Cancer
S100A4	Cell motility	Cancer
S100A5	Unknown	Unknown
S100A6	Tumor progression, prolactin secretion	Cancer
S100A7	Keratinocyte differentiation	Psoriasis
S100A8/A9	Chemotactic activities	Inflammation, cystic fibrosis
S100A10	Neurotransmitter release	Inflammation
S100A11	Organization of early endosomes	Skin diseases, ocular melanoma
S100A12	Differentiation of epithelial cells	Mooren's ulcer
S100A13	Complexes with FGF-1	Unknown
S100B	Cell motility, phosphorylation, neurotrophic activity	Down's syndrome, melanoma, neurological disorders
S100P	Unknown	Unknown
Calbindin 9K	Ca^{2+} buffer and transport	Vitamin D deficiency

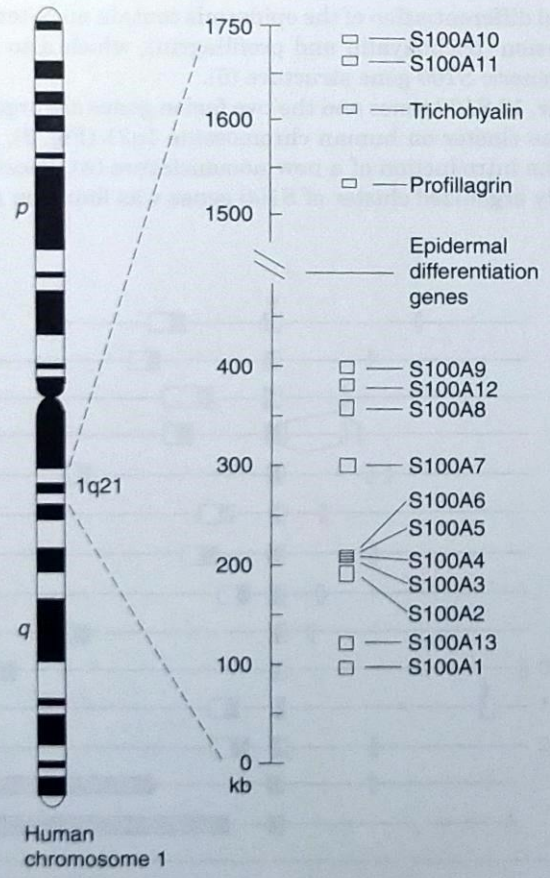

Figure 3. *S100* gene cluster on human chromosome 1q21. See Figure 2 for individual gene structures; q and p indicate the long and short arms, respectively, of the chromosome.

chromosome 3, indicating that the clustered organization is evolutionarily conserved (6).

On human chromosome 1q21, a number of abnormalities such as deletions, rearrangements, or translocations are associated with neoplasia, suggesting that the expression of *S100* genes might be altered in human cancer. Furthermore, the clustered organization poses the question whether each gene is regulated by its own elements or by possible superior locus control elements as suggested for the epidermal differentiation genes.

BIOLOGICAL FUNCTIONS

S100 proteins are involved in a large number of cellular activities such as signal transduction, cell differentiation, regulation of cell motility, transcription and cell cycle progression (Table 1), which are brought about through modulation of target proteins in a Ca^{2+}-, and possibly also in a Zn^{2+}- and Cu^{2+}-dependent manner.

In general one can distinguish intracellular functions of S100 proteins such as regulation of protein phosphorylation, cytoskeletal assembly, transcription, or enzyme activities from extracellular functions. These extracellular functions include chemotactic activity (S100A8 and S100A9) (7,8) as well as neurotrophic activities (S100B) (9). However, the mechanisms of secretion and the nature of high-affinity surface receptors are largely not known. One good candidate for such a receptor might be the receptor for advanced glycosylation end-products (RAGE) that was recently shown to bind S100B and S100A12 (10).

ASSOCIATION WITH HUMAN DISEASES

A wide range of human diseases has been correlated with deregulated expression of *S100* genes (Table 1). However, a clear physiological role for S100 proteins in the etiology of different diseases remains to be demonstrated in most cases. The levels of expression of individual S100 proteins vary most remarkably in different types of tumors, for example, enhanced expression of S100A4 can induce a higher metastatic phenotype in a transgenic mouse model (11) and appears to have a prognostic significance in lung cancer (12). On the other hand, S100A2 expression improves survival in squamous cell carcinomas (13). A second example for a clear disease

association is S100A1, which is downregulated in human cardiomyopathies and probably affects Ca^{2+} homeostasis in end-stage human heart failure (14). Indeed preliminary results from our laboratory indicate that the measurement of S100A1 in serum might be a putative indicator of myocardial infarction (15).

CONCLUSION

So far, S100 proteins have been extensively characterized on the biochemical level. In addition, several three-dimensional structures are now available, and a number of different target proteins have been identified. Finally, their gene structure as well as genomic organization is now known. This knowledge will provide the necessary foundation to probe the physiological functions of this fascinating protein family through the design of specific inhibitory drugs or genetic manipulation in animals. These experiments should reveal if S100 proteins will also contribute to novel therapeutic solutions.

BIBLIOGRAPHY

1. B.W. Schäfer and C.W. Heizmann, *Trends Biochem. Sci.* **21**, 134–140 (1996).
2. C.W. Heizmann and J.A. Cox, *BioMetals* **11**, 383–397 (1998).
3. L. Mäler, B.C.M. Potts, and W.J. Chazin, *J. Biomol. NMR* **13**, 233–247 (1999).
4. B.W. Schäfer et al., *Genomics* **25**, 638–643 (1995).
5. A.P. South et al., *J. Invest. Dermatol.* **112**, 910–918 (1999).
6. K. Ridinger et al., *Biochim. Biophys. Acta* **1448**, 254–263 (1998).
7. C. Geczy, *Biochim. Biophys. Acta* **1313**, 246–252 (1996).
8. C. Kerkhoff, M. Klempt, and C. Sorg, *Biochim. Biophys. Acta* **1448**, 200–211 (1998).
9. L.J. Van Eldik and W.S.T. Griffin, *Biochim. Biophys. Acta* **1223**, 398–403 (1994).
10. M.A. Hofmann et al., *Cell* **97**, 889–901 (1999).
11. R. Barraclough, *Biochim. Biophys. Acta* **1448**, 190–199 (1998).
12. K. Kimura et al., *Int. J. Oncol.* **16**, 1125–1131 (2000).
13. L. Lauriola et al., *Int. J. Cancer* **89**, 345–349 (2000).
14. A. Remppis et al., *Biochim. Biophys. Acta* **1313**, 253–257 (1996).
15. R. Kiewitz et al., *Biochem. Biophys. Res. Commun.* **274**, 865–871 (2000).

SARCOGLYCANS

ELIZABETH M. MCNALLY
University of Chicago
Chicago, Illinois

The dystrophin-glycoprotein complex (DGC) is present at the plasma membrane of skeletal and cardiac muscle. Composed of cytoskeletal elements and transmembrane proteins, the DGC is a multiprotein complex that links the cytoskeleton to the membrane and the extracellular matrix (Fig. 1). The DGC appears to be a multifunctional complex with both

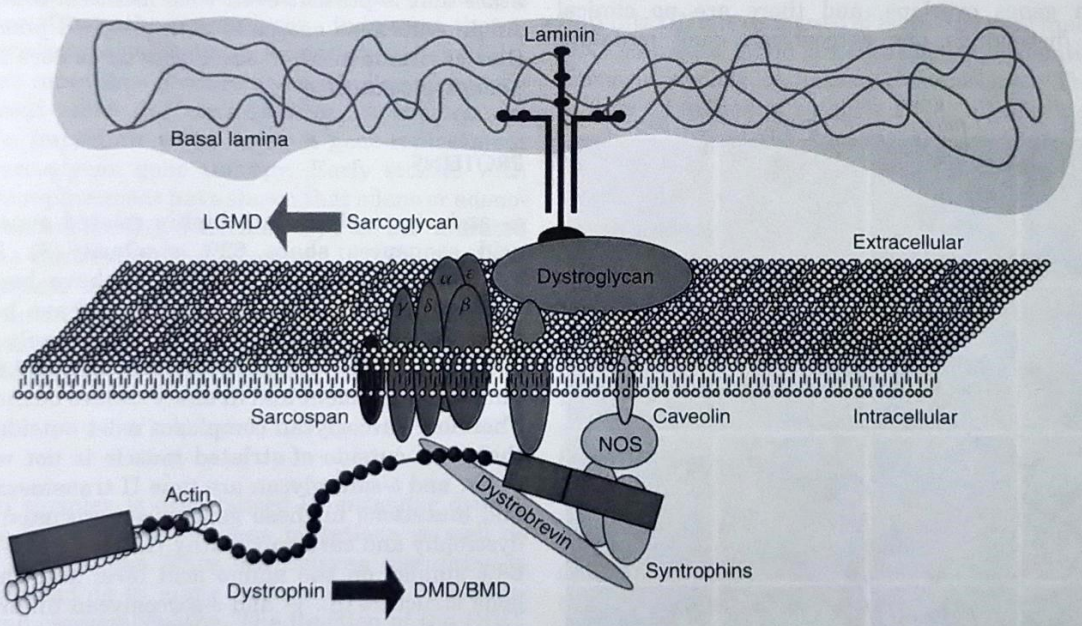

Figure 1. The dystrophin-glycoprotein complex (DGC). Shown is a schematic representation of the DGC and its role in the inherited forms of muscular dystrophy. Mutations in dystrophin cause Duchenne and Becker muscular dystrophy (DMD/BMD) whereas mutations in sarcoglycan genes cause limb girdle muscular dystrophy (LGMD). The DGC is isolated from muscle membranes and contains the ubiquitously expressed dystroglycan, as well as dystrobrevin and the syntrophins. Nitric oxide synthase (NOS) and caveolin also interact with DGC proteins. In skeletal and cardiac muscle, the DGC links cytoplasmic actin to the membrane and the extracellular matrix through dystroglycan's binding to laminin. Mutations in dystrophin or the sarcoglycan genes produce a secondary instability of sarcoglycan. Disruption of the sarcoglycan and the tetraspanin protein, sarcospan, is a common feature in many forms of muscular dystrophy and cardiomyopathy (see color figure).

mechanical and signaling functions. Skeletal and cardiac muscle are unique tissues in that they undergo contraction that is accompanied by considerable displacement of the tissue. The DGC is well positioned to be a "mechano-signaling" regulator of stretch, contraction, and the resultant force that is produced against the plasma and basement membranes.

Sarcoglycan is a subcomplex within the DGC, and mutations in the sarcoglycan genes are associated with autosomal recessive forms of muscular dystrophy (1,2). The sarcoglycan subunits can be separated from other DGC components in the presence of the detergent β-D-octyl-glucoside confirming that sarcoglycan is itself a complex (3). Early biochemical characterization suggested that sarcoglycan was three proteins of molecular mass 50 kD, 43 kD, and 35 kD. However, this approach did not fully account for the complexity of the sarcoglycan complex. The availability of protein and nucleotide databases revealed that the 50-kD component included both α- and ε-sarcoglycan and that the 35-kD protein was composed of the highly related γ-sarcoglycan and δ-sarcoglycan. β-sarcoglycan is a 43-kD protein that has only weak homology to γ- and δ-sarcoglycan. Each sarcoglycan subunit has a single transmembrane domain with substantial extracellular sequence and a short cytoplasmic tail. Mutations in sarcoglycan genes often produce a secondary instability of the remaining sarcoglycan proteins (Fig. 2), although exceptions to this generalization have been noted (4).

GENETICS

Generally, the clinical spectrum associated with mutations in sarcoglycan genes overlaps, and there are no clinical features that distinguish one sarcoglycan gene mutation from another (5). The range of phenotype found with sarcoglycan gene mutations is very similar to that seen with dystrophin mutations in the X-linked Duchenne/Becker muscular dystrophy (DMD/BMD). One notable exception is that dystrophin mutations produce cognitive impairment whereas sarcoglycan gene mutations do not. This is consistent with dystrophin's expression in brain, heart, and muscle and sarcoglycan's expression only in heart and muscle. In striated muscle, mutations in dystrophin also lead to disruption of the sarcoglycan complex. Thus sarcoglycan disintegration is a common element and is an important mediator of the dystrophic process because it affects both cardiac and skeletal muscle.

Mutations in sarcoglycan genes produce progressive, proximal muscle weakness that primarily affects the limbs and the trunk. These disorders have been collectively referred to as the limb girdle muscular dystrophies (LGMDs) (6). Autosomal recessive inheritance characterizes sarcoglycan gene mutations, although autosomal recessive LGMDs arise from mutations in non-sarcoglycan genes including mutations in the muscle-specific protease, calpain, and the novel membrane protein, dysferlin (7). The pathogenetic relationship, if any, between sarcoglycan, calpain, and dysferlin is not known. For sarcoglycan-associated LGMDs the age of onset is often in the first decade, and, in its most severe forms, confinement to wheelchair with concomitant loss of ambulation often occurs by the late second or third decade. Cardiomyopathy, typically of the dilated type, may also develop from sarcoglycan gene mutations. It should be noted that a wide variation in phenotype is present in these disorders, a variation that is present even with identical mutations (8). That single mutations associate with a varied phenotype suggests that environmental or other genetic factors may modify the phenotypic outcome.

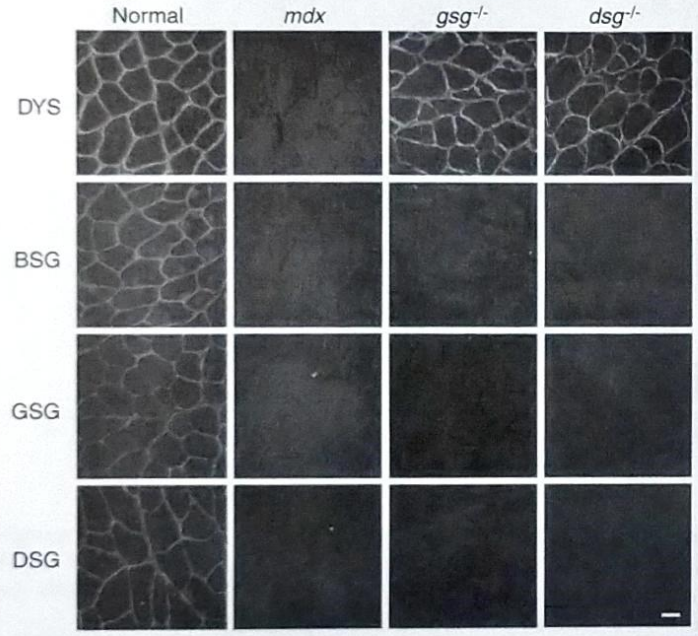

Figure 2. Disruption of sarcoglycan is common to many forms of muscular dystrophy. Immunostaining for dystrophin (DYS) and β-, γ-, and δ-sarcoglycan (BSG, GSG and DSG, respectively) in normal mice and mice deficient for dystrophin (*mdx*), γ-sarcoglycan ($gsg^{-/-}$) and δ-sarcoglycan ($dsg^{-/-}$) is shown. These genetically distinct forms of muscular dystrophy each have in common the loss of the sarcoglycan complex.

PROTEINS

α- and ε-sarcoglycan are highly related genes whose amino acid sequences share 62% similarity (8). Many different nonsense and missense mutations have been described in α-sarcoglycan, whereas no mutations have been found in ε-sarcoglycan. α-sarcoglycan is expressed only in cardiac and skeletal muscle, whereas ε-sarcoglycan is highly expressed during development and in many tissues outside of muscle (8). Therefore sarcoglycan complexes exist outside of muscle, but their role outside of striated muscle is not well understood. β-, γ-, and δ-sarcoglycan are type II transmembrane proteins, and mutations in these genes are associated with muscular dystrophy and cardiomyopathy (1,2). γ- and δ-sarcoglycan are 68% similar on the amino acid level and have an identical gene structure (8). γ- and δ-sarcoglycan differ with regard to smooth muscle expression. The smooth muscle sarcoglycan complex is composed of β-, δ-, and ε-sarcoglycan suggesting that a sarcoglycan trimer may be stable or that additional sarcoglycans may exist in smooth muscle (9).

CELL BIOLOGY

The function of sarcoglycan is not known. Its primary structure suggests a cell surface receptor, but ligands have not been identified. A muscle-specific form of the actin-binding protein

filamin, filamin C, was recently identified as a cytoplasmic-binding protein of γ- and δ-sarcoglycan (10). Redistribution of filamin occurs in response to mutations in γ- and δ-sarcoglycan and this may contribute to muscular dystrophy. α-sarcoglycan appears to have ecto-ATPase properties, whereas ε-sarcoglycan does not. These data suggest a regulatory role for sarcoglycan. Mutations in sarcoglycan do not affect dystrophin expression or localization (Fig. 2) but may lead to more subtle abnormalities within the DGC. In this regard the recent generation of mouse models of sarcoglycan deficiency has proved useful for the study of sarcoglycan function. Mice lacking γ-sarcoglycan develop cardiomyopathy and muscular dystrophy (11) as do mice lacking β- or δ-sarcoglycan (12–15). Mice lacking α-sarcoglycan appear to develop only muscular dystrophy and little or no cardiomyopathy, similar to patients with α-sarcoglycan gene mutations. Interestingly, studies of isolated muscle lacking individual sarcoglycans differ in the degree of muscle injury that results from muscle contraction. In response to eccentric contraction of isolated muscle, δ- or α-sarcoglycan-deficient muscle has a decrease in force production and an increase in membrane damage indicative of cellular injury (14). In contrast, eccentric contraction of γ-sarcoglycan-deficient muscle shows normal parameters; yet it shows a severe muscular dystrophy phenotype (16). This suggests that "nonmechanical" functions, potentially signaling in nature, may be sufficient to cause the phenotype of muscular dystrophy. Moreover, when γ-sarcoglycan is eliminated, α-, β-, and δ- sarcoglycan are still expressed, but at reduced levels (14). The residual α-, β-, and δ-sarcoglycan assemble and are glycosylated, but are not stable at the plasma membrane. In contrast, the loss of δ-sarcoglycan produces a complete absence of the other sarcoglycans indicating that δ-sarcoglycan is critical for sarcoglycan formation and stability. Thus sarcoglycan mutations produce a very similar phenotype but have different molecular consequences for the muscle, and this will be important when using a gene-replacement approach for sarcoglycan gene therapy. Early studies with sarcoglycan gene replacement have shown that adeno or adeno-associated viruses delivering a normal copy of the mutated sarcoglycan gene can restore the secondary instability of the sarcoglycan complex. It is likely that sarcoglycan mutations may be more amenable to gene therapy than dystrophin mutations because sarcoglycan proteins are relatively small and thus gene delivery is less problematic.

CONCLUSION

The sarcoglycan complex is destabilized in muscular dystrophies that arise from dystrophin gene mutations as well as sarcoglycan gene mutations. Thus, sarcoglycan disruption is a common factor leading to many different forms of muscular dystrophy and cardiomyopathy. The function of the DGC, including the subcomplex sarcoglycan, includes both mechanical and nonmechanical functions. The DGC stabilizes the muscle and heart plasma membranes in response to mechanical contraction, but, in addition, probably has functions that include both repair and maintenance of membrane integrity. Indeed the multifunctional, mechanosignaling nature of the DGC is recapitulated with the sarcoglycan complex itself because the sarcoglycan complex also has functions related to mechanical membrane stability as well as signaling and growth functions.

BIBLIOGRAPHY

1. C.G. Bonnemann et al., *Curr. Opin. Pediatr.* **8**, 569–582 (1996).
2. V. Straub et al., *Curr. Opin. Neurol.* **10**, 168–175 (1997).
3. E. Ozawa et al., *Muscle Nerve* **21**, 421–438 (1998).
4. M. Vainzof et al., *Hum. Mol. Genet.* **5**, 1963–1969 (1996).
5. D.J. Duggan et al., *N. Engl. J. Med.* **336**, 618–624 (1997).
6. K.M. Bushby et al., *Neuromuscul. Disord.* **5**, 337–343 (1995).
7. K.M. Bushby, *Brain* **122**, 1403–1420 (1999).
8. A.A. Hack et al., *Microsc. Res. Tech.* **48**, 167–180 (2000).
9. V. Straub et al., *J. Biol. Chem.* **274**, 27989–27996 (1999).
10. T.G. Thompson et al., *J. Cell Biol.* **148**, 115–126 (2000).
11. A.A. Hack et al., *J. Cell Biol.* **142**, 1279–1287 (1998).
12. R. Coral-Vazquez et al., *Cell* **98**, 465–474 (1999).
13. M. Durbeej et al., *Mol. Cell* **5**, 141–151 (2000).
14. A.A. Hack et al., *J. Cell Sci.* **113**, 2535–2544 (2000).
15. K. Araishi et al., *Hum. Mol. Genet.* **8**, 1589–1598 (1999).
16. A.A. Hack et al., *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10723–10728 (1999).

SATELLITE DNA

Brynn Levy
Teresa Dunn
Peter E. Warburton
Mount Sinai School of Medicine
New York, New York

Satellite DNA is composed of large arrays of tandemly repeated DNA elements, found largely as heterochromatic blocks in centromeric regions of human chromosomes. There are many different families of satellite DNA, all of which comprise as much as 10% of the human genome. Classical satellites 1, 2, and 3 are found primarily in large variable heterochromatic regions on human chromosomes. Alpha satellite DNA is found at the centromere of every human chromosome and may play a functional role in centromere formation. Other satellite DNA families include beta and gamma satellite DNA. Evolutionary processes acting on satellite DNA have played a role in shaping our chromosomes, leading to large arrays of highly homologous repeat units at particular chromosomal locations. Satellite DNA repeats can be fluorescently labeled and utilized as fluorescence in situ hybridization (FISH) probes for chromosome identification and/or enumeration. Clinical cytogenetic laboratories routinely use such probes for pre- and postnatal identification of cytogenetic abnormalities such as trisomies or supernumerary marker chromosomes. ICF (immunodeficiency, centromere instability, facial abnormalities) and Robert's syndromes are rare disorders that manifest as distinct chromosome abnormalities affecting the satellite DNA–containing heterochromatin.

SATELLITE DNA FAMILIES

Satellite DNA was historically defined as *human genomic DNA fractions* that had different buoyant densities on CsSO4 (cesium sulfate) gradients from the bulk of genomic DNA. Several satellite DNA fractions consisting of heterogeneous mixtures of repetitive DNA sequences were identified and

referred to as *classical satellites 1,2,* and 3. In situ hybridization of these different fractions to rodent–human somatic cell hybrids showed several large distinct blocks at specific chromosomal locations in the pericentromeric regions of human chromosome 1, 9, 16, and Y. The major locations of satellite I, II, and III DNA corresponds to large blocks of heterochromatin in human chromosomes, which can be readily visualized by their relatively intense staining with the fluorescent dyes DAPI and distamycin, as shown in Figure 1. Other major locations for satellite DNA include the short arms of acrocentric chromosomes 13, 14, 15, 21, and 22, both proximal and distal to the rDNA (ribosomal DNA) arrays.

Restriction enzyme analysis of these different satellite DNA fractions revealed several patterns consisting of ladders of repetitive units, some of which were specific to certain chromosomes. The heterogeneous nature of satellite fractions 1, 2, and 3 prompted their sequence analysis, resulting in identification of predominant repeats found in each density gradient fraction, called *satellites 1, 2,* and 3, respectively. Satellite 1 contains a 42-bp repeat arranged as alternating 17 bp-(ACATAAAATAT$^{C}/_{G}$AAAGT) and 25 bp-(ACCCAAAT$^{A}/_{G}$T$^{A}/_{G}$TATTATATACTGT) repeat units. Satellite 2 is identified as poorly conserved ATTCCATTCG repeats. Satellite 3 is defined as ATTCC repeats occasionally interspersed with A$^{T}/_{C}$TCGGGTTG.

In situ hybridizations with specific DNA clones have localized these satellites to several specific chromosomal regions. Satellite 1 was localized to the pericentromeric regions of chromosomes 3 and 4, and the short arms of the acrocentric chromosomes in regions both proximal and distal to the rDNA arrays. Satellite 2 was localized to the large heterochromatic regions of chromosomes 1 (1qh) and 16, with less prominent domains at the pericentromeric regions of chromosomes 2 and 10. Satellite 3 is localized to the variable heterochromatic regions of chromosome 1 and 9, the long arm of the Y chromosome, and the short arms of the acrocentric chromosomes proximal to the rDNA arrays. Molecular approaches to map these satellite DNA domains have confirmed these localizations, and uncovered additional small domains such as satellite III in the pericentromeric region of chromosome 10.

Alpha satellite DNA has been found at the centromere of every normal human and primate chromosome examined. It was first identified as a highly repetitive DNA fraction from the African Green Monkey and is the most extensively studied of all human satellite DNA families. Its hierarchical repeat unit organization serves as a conceptual framework for the organization of tandemly repeated satellite DNAs. The fundamental repeat units of alpha satellite DNA are approximately 171-bp monomers that display up to 40% divergence from each other. These monomers are tandemly organized into distinct linear arrangements called higher-order repeat units (HORs) ranging from 2 to more than 30 monomers. At the centromeres of each homologous chromosome pair, a particular HOR is in turn tandemly repeated up to several hundred times to form arrays as large as several million base pairs. For example, on chromosome 17, a 2.7-kb 16 monomer HOR is repeated approximately 1,000 times to form arrays of approximately 3,000 kb. HORs from a particular centromere are generally less than 5% diverged from each other, and thus can be used as specific FISH probes in order to identify individual chromosomes. Alpha satellite DNA also contains a 9-bp degenerate motif in a subset of its monomers, which serves as the binding site for a satellite DNA–binding protein called Centromere Protein-B (CENP-B), which is seen at most human and mammalian centromeres.

Several additional families of satellite DNA have been described and they are also generally found at the chromosomal locations typical of satellite DNA, for example, centromeric regions, the short arms of acrocentrics, and the Y chromosome. Beta satellite DNA is based on a 68-bp monomer, and individual subsets have been shown to be chromosome-specific by FISH. Gamma satellite DNA is based on a 220-bp monomer and has thus far been observed at the centromeres of chromosomes 8 and X. Additional families include a 48-bp satellite DNA, found on acrocentric chromosomes, and the Sn5 satellite family, which can be found in the pericentromeric regions of chromosomes 2 and 20 as well as on the acrocentric chromosomes.

Figure 1. DAPI/distamycin staining of human chromosomes show the prominent satellite DNA containing regions on chromosomes 1,9,15,16, and Y.

SATELLITE DNA AND CHROMOSOME EVOLUTION

The evolution of satellite DNAs is inextricably linked to the structure of mammalian chromosomes. The large arrays of satellite DNA found on human chromosomes are shaped by evolutionary processes such as unequal crossing-over and gene conversion. Such processes lead to homogenization of repeat units and expansion of arrays within a species or particular chromosomal location. Satellite DNAs are thought to accumulate at centromeres because these regions are transcriptionally inert (due to formation of a centromere/kinetochore). Crossing-over and recombination events in these regions will therefore have no adverse effect on the organism. A similar situation exists for the Y chromosome, which contains few genes and has also accumulated satellite DNAs. The alpha satellite DNA–binding protein CENP-B has been observed to share homology with the tigger family of ancient transposases, prompting theories that remnant 3′ nicking activity of CENP-B may accelerate the expansion and homogenization of alpha satellite DNA arrays that contain the CENP-B binding site.

The effect of satellite DNA evolution on chromosome structure will depend on the relative rates of exchange between sister chromatids and homologous chromosomes (intrachromosomal) and nonhomologous chromosomes (interchromosomal). In the case of human alpha satellite DNA the homogenization of repeat units at specific centromeres of each human chromosome suggest a relatively high frequency of intrachromosomal exchanges. A notable exception are the shared homologous alpha satellite DNA subsets found on certain pairs of human acrocentric chromosomes for example, 13/21 and 14/22. These shared alpha satellite DNA subsets, and indeed