

# Phenotypes, genotypes and disease susceptibility associated with gene copy number variations: complement C4 CNVs in European American healthy subjects and those with systemic lupus erythematosus

Y.L. Wu<sup>a,b</sup> Y. Yang<sup>a</sup> E.K. Chung<sup>a</sup> B. Zhou<sup>a</sup> K.J. Kitzmiller<sup>a,b</sup> S.L. Savelli<sup>a,c</sup>  
H.N. Nagaraja<sup>d</sup> D.J. Birmingham<sup>b,e</sup> B.P. Tsao<sup>f</sup> B.H. Rovin<sup>e</sup> L.A. Hebert<sup>e</sup>  
C.Y. Yu<sup>a-c</sup>

<sup>a</sup>Center for Molecular and Human Genetics, The Research Institute at Nationwide Children's Research Institute

<sup>b</sup>Integrated Biomedical Science Graduate Program, <sup>c</sup>Department of Pediatrics, <sup>d</sup>Department of Statistics,

<sup>e</sup>Department of Internal Medicine, The Ohio State University, Columbus, OH; <sup>f</sup>University of California, Los Angeles, CA (USA)

Accepted in revised form for publication by H. Kehrer-Sawatzki and D.N. Cooper, 13 June 2008.

**Abstract.** A new paradigm in human genetics is high frequencies of inter-individual variations in copy numbers of specific genomic DNA segments. Such common copy number variation (CNV) loci often contain genes engaged in host-environment interaction including those involved in immune effector functions. DNA sequences within a CNV locus often share a high degree of identity but beneficial or deleterious polymorphic variants are present among different individuals. Thus, common gene CNVs can contribute, both qualitatively and quantitatively, to a spectrum of phenotypic variants. In this review we describe the phenotypic and genotypic diversities of complement C4 created by copy number variations of RCCX modules (*RP-C4-CYP21-TNX*) and size dichotomy of C4 genes. A direct outcome of C4

CNV is the generation of two classes of polymorphic proteins, C4A and C4B, with differential chemical reactivities towards peptide or carbohydrate antigens, and a range of C4 plasma protein concentrations (from 15 to 70 mg/dl) among healthy subjects. Deliberate molecular genetic studies enabled development of definitive techniques to determine exact patterns of RCCX modular variations, copy numbers of long and short C4A and C4B genes by Southern blot analyses or by real-time quantitative PCR. It is found that in healthy European Americans, the total C4 gene copy number per diploid genome ranges from 2 to 6: 60.8% of people with four copies of C4 genes, 27.2% with less than four copies, and 12% with more than four copies. Such a distribution is skewed towards the low copy number side in patients with systemic lupus erythematosus (SLE), a prototypic autoimmune disease with complex etiology. In SLE, the frequency of individuals with less than four copies of C4 is significantly increased (42.2%), while the frequency of those with more than four copies is decreased (6%). This decrease in total C4 gene copy number in SLE is due to increases in homozygous and heterozygous deficiencies of C4A but not C4B. Therefore, it is concluded that lower copy number of C4 is a risk factor for and higher gene copy number of C4 is a protective factor against SLE disease susceptibility.

This work was supported by the National Institute of Arthritis, Musculoskeletal and Skin Diseases grants 1R01 AR050078 and 1R01 AR054459, a pilot grant from the Lupus Foundation of America, and the National Institute of Diabetes, Digestive and Kidney Diseases grant P01 DK55546.

Request reprints from C. Yung Yu, D.Phil.  
Center for Molecular and Human Genetics  
The Research Institute at Nationwide Children's Research Institute  
Columbus, OH 43205 (USA)  
telephone: +1 614 722 2821; fax: +1 614 722 2817; e-mail: cyu@chi.osu.edu

Copyright © 2009 S. Karger AG, Basel

Comparative genomic hybridization (CGH) and molecular genetic experiments in the past few years have revealed an important phenomenon that had escaped the attention of most human geneticists: many genes in our genome exhibit an inborn, inter-individual variation in copy number. In a recent report from the Database of Genomic Variants (<http://projects.tcag.ca/variation/>, build 36, hg18), a total of 14,724 loci show genomic alterations that involve segments of DNA >1 kb (Iafate et al., 2004; Sebat et al., 2004; Redon et al., 2006; Eichler et al., 2007; Lupski, 2007; Wong et al., 2007). Many of those are copy number variation (CNV) loci that include genes engaged in host-environment interactions, and those for immune response and sensory functions. This discovery provides a new and exciting opportunity to examine the genetic basis of quantitative traits and complex diseases.

CGH experiments using oligonucleotides or bacterial artificial chromosome (BAC) microarrays are informative because they provide global and high throughput data on possible locations of CNVs in human genomes. However, the data obtained generally are not definitive. Thus, they require further molecular genetic experiments for validation of the genetic composition and the precise boundaries for each variant locus.

DNA sequences within each CNV locus may share over 95% sequence identities. Nevertheless, considerable sequence polymorphisms are usually present. Gene CNVs and their associated polymorphisms can contribute to qualitative and quantitative diversities in their gene products. For immune effector genes, such diversities can lead to differences in intrinsic strength of the defense system, and result in varying susceptibilities to autoimmune diseases (Yang et al., 2003; Yu et al., 2003). To understand the roles of CNVs on the genetic risk of a complex disease such as systemic lupus erythematosus (SLE), it is essential to apply accurate and definitive techniques so that polymorphic variant(s) of the CNV loci correlated with disease predisposition can be determined unambiguously.

There are two types of CNVs. One is *de novo* that is detectable in some neurologic disease patients such as those with autism or schizophrenia (Sebat et al., 2007; Sutrala et al., 2007; Cantor and Geschwind, 2008; Marshall et al., 2008). The other is common CNV, that exists with relatively high frequency (>5%) in general human populations. Some common CNVs can also be observed at orthologous loci of non-human primates. In this review, we give a specific example of a common gene CNV locus and how it contributes to a complex pattern of phenotypes and genotypes and increases susceptibility to SLE.

### **Functional and phenotypic diversities of complement component C4**

Complement component C4 is the recognition (or non-enzymatic) subunit of the C3 convertase for the classical (or antibody-antigen) and the mannan binding lectin (MBL) activation pathways of the complement system (Reid and

Porter, 1981; Yu et al., 2003). Perhaps the most important structural feature of C4 (and its related protein C3) is the presence of a thioester bond (Law et al., 1980) between the sulfhydryl group of Cys-991 and the carbonyl group of Gln-994 that is hidden in the native protein but becomes exposed upon activation by proteolytic cleavage of C4, and release of the C4a peptide. The carbonyl group of Gln-994 in this activated subunit, known as C4b, will undergo nucleophilic attack to its target molecule, forming a covalent ester or amide linkage (Dodds et al., 1996). Downstream consequences of C4 activation include the generation of anaphylatoxins C3a and C5a, the activation of C3 and C5, the formation of the C5 convertase, and the assembly of the membrane attack complex on microbes (Walport, 2001). The deposition of activated C4 and C3 (i.e., C4b and C3b, respectively) via the covalent linkages on immune complexes also facilitates phagocytosis by macrophages through interactions with complement receptors (e.g. CR1 and CR3) and immunoglobulin Fc-gamma receptors, immunoclearance by binding to erythrocyte CR1 in the circulation, and degradation of the immune complexes through the endothelial reticulo-cyte system in the liver or the spleen (Cornacoff et al., 1983; Gatenby et al., 1990; Birmingham and Hebert, 2001). Therefore, C4 is an important effector protein for both innate and adaptive immune systems among vertebrate animals. Over the past thirty years, many immunologists and geneticists have been intrigued if not puzzled by the qualitative and quantitative diversities of the C4 phenotypes and genotypes in health and in diseases (Porter, 1983).

### **Qualitative variations or polymorphisms of C4 proteins**

#### *C4A and C4B isotypes and allotypes*

Through immunofixation of EDTA-plasma proteins resolved by gross difference in electric charge by non-reducing agarose gel electrophoresis, multiple electrophoretic and serologic variants of complement C4 were observed. To date, close to 40 polymorphic variants for C4 have been demonstrated (Mauff et al., 1998). These variants can be categorized into two isotypes: one migrates faster in the agarose gel because the proteins are acidic and therefore they are named C4A; the other migrates slower because the proteins are basic and therefore named C4B (Awdeh and Alper, 1980; Sim and Cross, 1986). The most common allotypes for C4A are A3, A2, A4, A6 and A1; the most common allotypes for C4B are B1, B2, B3 and B5. Molecular cloning and sequencing of the *C4A* and *C4B* genes allowed the elucidation of the isotypic protein sequences specific for C4A and C4B at positions 1101–1106, which are PCPVLD for C4A, and LSPVIH for C4B (Yu et al., 1986). Through a hemolytic overlay assay that employed rabbit hemolysin (antibodies against sheep red blood cells), serum complement from guinea pig that was deficient in C4, and sheep red blood cells as target, it was shown that human C4B allotypes resolved by agarose gel electrophoresis can lyse sheep RBC three to four times faster than C4A allotypes (Mauff et al., 1983). Functional binding assays of simple

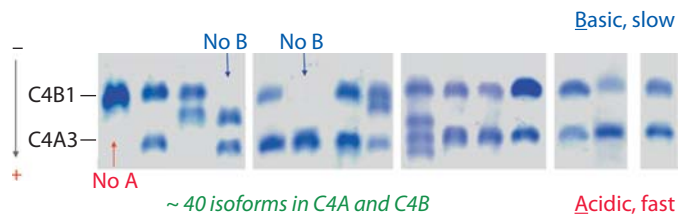
molecules such as glycine (with amino group) and glycerol (with hydroxyl group) using purified C4A and C4B proteins showed that activated C4A binds to glycine effectively and forms an amide bond, while activated C4B binds to glycerol very efficiently and forms an ester linkage with targets (Isenman and Young, 1984; Law et al., 1984). Site-directed mutagenesis of C4A or C4B for binding assays further showed that His-1106 plays a critical role in the catalysis and a short half-life (<1 s) for the transacylation activity by activated C4B thioester bond to its target. By contrast, Asp-1106 contributed to effective bindings of activated C4A to IgG immune aggregates (Carroll et al., 1990; Dodds et al., 1996). The C4 proteins in many vertebrates have a histidine residue at the orthologous position. Therefore, C4B is probably the more ancient protein and C4A is a likely recent addition or a gain of function (Dodds and Law, 1990; Martinez et al., 2001).

#### *Rodgers (Rg) and Chido (Ch) blood group antigens*

When a human subject undergoes a blood transfusion, the polymorphic C4 variants in the donor's blood can elicit an immune response in the recipient to generate allogenic antibodies. These alloantibodies can agglutinate the recipient red blood cells containing covalently-bound C4b or C4d fragments. Such observations led to the discovery of the Chido and the Rodgers blood groups (Middleton and Crookston, 1972; Longster and Giles, 1976; Giles et al., 1988). Anti-Chido antibodies were generated in recipients with C4B deficiency, and anti-Rodgers antibodies were generated in recipients with C4A deficiency. Through molecular cloning and sequencing of the genomic DNA corresponding to the polymorphic C4d region from human subjects with normal and reversed serologic antigenicities, it was found that the antigenic determinants for the Rodgers blood group reside on VDLL 1188–1191 (Rg1), N1157 (Rg3) and their combination (Rg2); while the Chido antigenic determinants reside on G1054 (Ch5), LSPVIH 1101–1106 (Ch4), S1157 (Ch6), ADLR 1188–1191, combinations of G1054 and LSPVIH 1101–1106 (Ch2), and of S1157 and ADLR 1188–1191 (Ch3). While most C4A are associated with Rg, and C4B with Ch, reverse associations of C4A with Ch (e.g. C4A1 with Ch1, 3, 5 and 6), and C4B and Rg (e.g., C4B5 with Rg1) were well documented (Yu et al., 1986, 1988).

#### **Quantitative variations of C4A and C4B proteins**

Figure 1 shows results of a C4 allotyping gel that demonstrates the qualitative and quantitative variations of C4A and C4B proteins from 15 human subjects. In lane 1 only the slow migrating C4B protein (C4B1) is present. In lane 6 only the fast migrating C4A protein (C4A3) is present. In lanes 2 and 12, there are more C4B proteins than C4A proteins. In lanes 5 and 14, there are more C4A proteins than C4B proteins. Polymorphic variants for C4A or C4B are discernible in lanes 3, 4, 8 and 9. How to explain the phenotypic variations of C4? Initially a monogenic model biallelic



**Fig. 1.** Polymorphism of complement C4A and C4B proteins. Human plasma proteins were digested with neuraminidase and carboxyl peptidase B to remove heterogeneities caused by glycosylations and incomplete processing of C-termini at the beta and alpha chains of C4 proteins. The proteins were resolved by high voltage agarose gel electrophoresis based on gross differences in electric charge, immunofixed by antisera against human C4, blotted to remove diffusible proteins and then stained. Plasma samples from 15 subjects are shown. The basic C4B proteins migrate slower than the acidic C4A proteins (Blanchong et al., 2000).

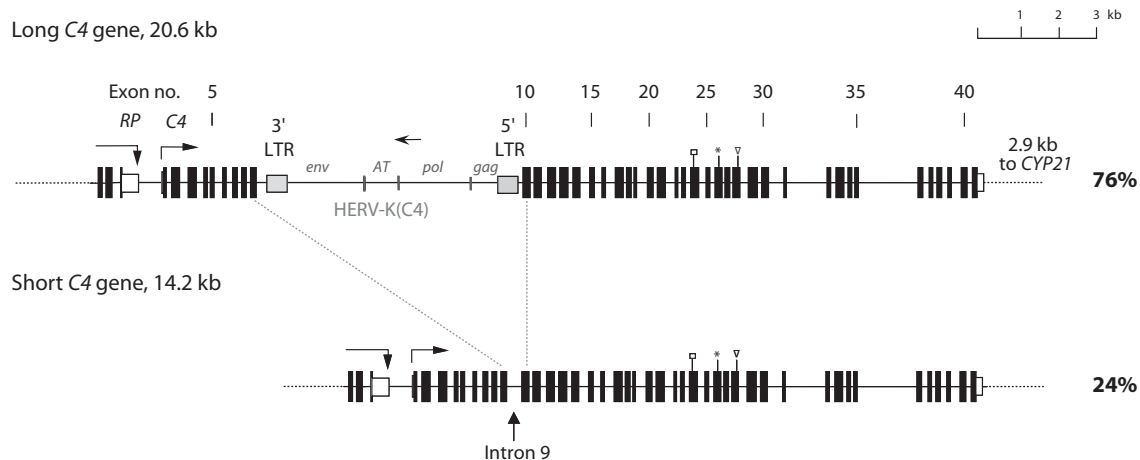
for C4A and C4B (or C4-fast and C4-slow) was proposed, but it did not explain the presence of three or four different proteins as observed in lanes 8 and 9, respectively. A two-locus model on each haplotype, one for C4A and one for C4B, was then proposed (O'Neill et al., 1978; Awdeh et al., 1979, 1983; Roos et al., 1982) and widely adopted. A 'partial deficiency for C4A' was proposed when the intensities of C4A protein appeared lower than those of C4B; a 'partial deficiency for C4B' was proposed when the intensities of C4B were lower than those of C4A. A 'null' allele was assigned in those cases with a partial C4A or C4B deficiency.

However, the C4A-C4B (two-locus) model can only explain about 50% of the C4 proteins in the human population. Family segregation and molecular genetic studies reveal unusually high frequencies of haplotypes with a single C4 gene coding for either C4A or C4B, haplotypes with two C4 loci coding for C4A only (C4A-C4A) or for C4B only (C4B-C4B), and three-locus haplotypes coding for both C4A and C4B. Such phenomenon is bewildering and cannot be easily explained by simple genetic models. Our concept of C4 genetics evolves gradually with deliberate molecular genetic and population studies in health and disease.

#### **Molecular basis of complement C4A and C4B genotypic variations**

##### *Dichotomous size variations of C4A and C4B genes*

The C4A and C4B proteins are each encoded by a 5.4-kb transcript assembled from 41 exons of C4 genes (Belt et al., 1984, 1985; Yu, 1991). The amino acid residues for the thioester bond are encoded by exon 24. The C4A/C4B isotypic amino acid residues at 1101–1106 are encoded by exon 26; the major Chido and Rodgers blood group antigens are encoded by exon 28 (Yu et al., 1986, 1988; Yu, 1991). Remarkably, there are two versions of C4 genes. The long form is 20.6 kb (L), the short version is 14.2 kb (S) (Fig. 2). The dif-



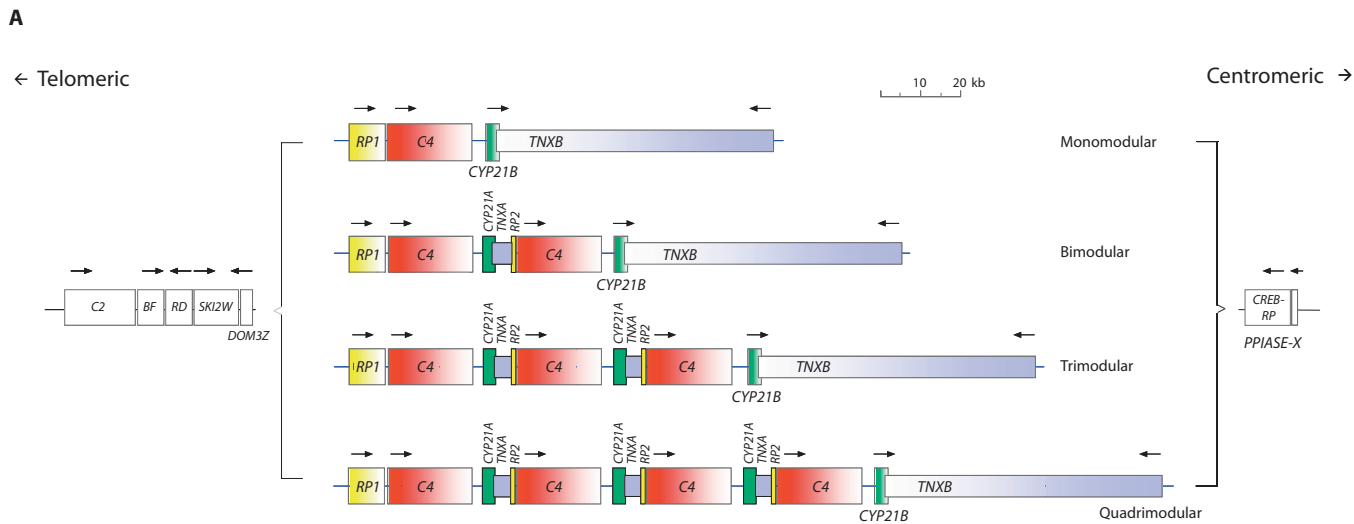
**Fig. 2.** Dichotomous size variation of human *C4* genes. Each human *C4* gene consists of 41 exons. The long *C4* gene contains the endogenous retrovirus HERV-K(C4) in intron 9. The thioester bond is encoded by exon 24, the C4A and C4B isotypic residues by exon 26, and the major Rodgers and Chido blood group antigenic determinants by exon 28. Among the European Americans, 76% of the *C4* genes belong to the long form and 24% belong to the short form (Yu, 1991).

ference is caused by the integration of an endogenous retrovirus HERV-K(C4) into intron 9 of the long genes. The configuration of the 6.36-kb HERV-K(C4) is opposite to the transcriptional orientation of the *C4* gene. There are two long terminal repeats of 548 bp and 550 bp flanking the DNA sequences similar to those for *gag-pol-env* in a typical retrovirus, but the reading frames are all closed due to multiple mutations (Dangel et al., 1994, 1995; Chu et al., 1995). In vitro reporter gene assays demonstrated that the regulatory sequence in the 3' LTR still maintains promoter activity that can drive transcription of an antisense RNA that is complementary to the 5' sequences of complement *C4* (Dangel et al., 1994; Mack et al., 2004). It was therefore hypothesized that HERV-K(C4) can modulate the expression of *C4* genes. Initially, it was thought that the *C4* genes coding for C4A belonged to the long version, while those coding for C4B could be a short gene. Now we know that a long *C4* gene can code for either a C4A or a C4B protein, as can a short gene, although most C4A genes are long, and a greater proportion of C4B genes are short.

*Modular or segmental duplications of RP-C4-CYP21-TNX (RCCX) in the human major histocompatibility complex (MHC)*

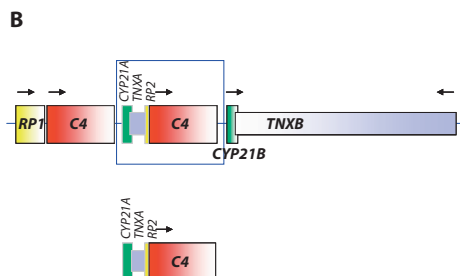
Determination and characterization of the genomic DNA sequences of *C4* and its neighboring genes from multiple subjects allowed a deliberate analysis of gene haplotypes in the class III region of the MHC. In about 17% of the MHC haplotypes from human subjects of Northern European ancestry, there is an intact *RPI* gene (also known as *STK19* for serine/threonine kinase 19), an intact gene complement *C4* that either codes for C4A or for C4B, an intact *CYP21B* (*CYP21A2*) gene (cytochrome P450 21-hydroxylase), and an intact *TNXB* gene that codes for extracellular matrix protein tenascin-X. In the remaining 83% of the

MHC haplotypes, there can be additional 1, 2 or 3 duplications of a genetic unit between *C4* and *CYP21B* (Fig. 3). The breakpoint of the segmental duplication was sequenced and identified to be at exon 7 of the *RPI* gene and intron 32 of the *TNXB* gene (Shen et al., 1994). Each duplication unit contains a *CYP21* gene that is mostly (but not always) a mutant gene (*CYP21A*, aka *CYP21A1P*), a 4.9-kb DNA fragment (*TNXA*) corresponding to intron 32 to exon 45 of *TNXB* that is fused to a 0.91-kb DNA fragment (*RP2*) corresponding to part of exons 7 to exon 9 of *RPI*, and an intact gene for complement *C4* that can be either *C4A* or *C4B*. Such a genetic duplication unit is termed RCCX module (Yang et al., 1999). Each RCCX module can be 32.7 kb or 26.3 kb in size, depending on the size of the *C4* gene present. The size variation of *C4* genes and the copy number variation of RCCX modules lead to a repertoire of physical length variants for the MHC class III region that probably play an important role in promoting gene conversion or non-allelic homologous recombinations between the length variants during meiosis (Blanchong et al., 2000, 2001; Chung et al., 2002a). There are two probable outcomes. One is the generation of qualitative and quantitative diversity of complement *C4*, which may confer an intrinsic difference in the strength of the immune effector system to respond to microbial infections (Yang et al., 2003). The other is a genetic burden for susceptibilities to autoimmune diseases such as SLE with low gene copy number of *C4*, genetic diseases such as congenital adrenal hyperplasia (CAH) by which deleterious mutations in *CYP21A* are incorporated in *CYP21B*, or in some patients with Ehlers-Danlos syndrome or CAH who acquired a 120-bp deletion spanning intron 35 to exon 36 into *TNXB* from *TNXA* via a recombination or a gene conversion-like event (Collier et al., 1993; Rupert et al., 1999; Yang et al., 1999; Blanchong et al., 2000; Schalkwijk et al., 2001).



Genes upstream and downstream of C4 undergo concurrent duplications:

- RP:** Ser/Thr nuclear protein kinase  
**C4:** Complement component C4  
**CYP21:** Steroid 21-hydroxylase  
**TNX:** Extracellular matrix protein tenascin-X



Duplication is discretely modular,  
 each duplicated RCCX is either 32.7 kb or 26.3 kb in size.  
 Each RCCX module adds:

- One *CYP21A* (pseudogene) or *CYP21B* (functional)
- Gene fragment for tenascin *TNXA* (4.9 kb)
- Gene fragment for Ser/Thr kinase *RP2* (0.9 kb)
- **One functional C4 gene: C4A or C4B; long or short**

**Fig. 3.** Modular variation of *RP-C4-CYP21-TNX* (RCCX) in the central region of the human major histocompatibility complex (MHC) on chromosome 6. **(A)** MHC haplotypes with 1, 2, 3 or 4 RCCX modules. **(B)** Details of a duplicated RCCX module (Yu and Whitacre, 2004).

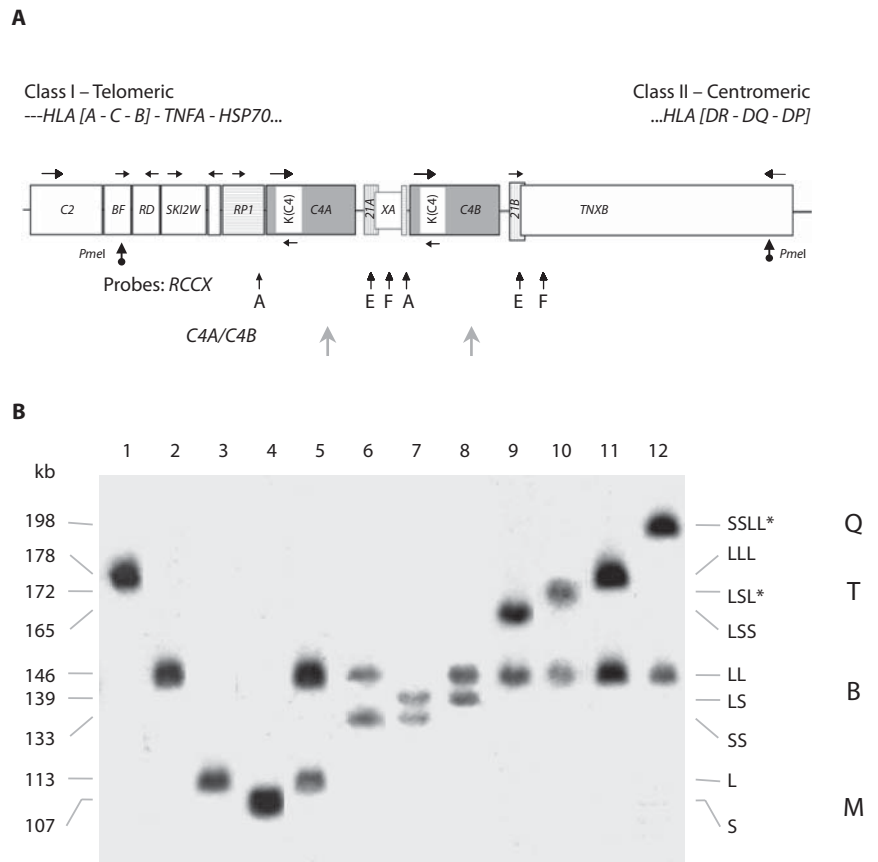
### Techniques to elucidate C4 gene copy number variations

Two levels of genetic diversities often accompany gene CNV. One is the continuous variation in the gene copy number, the second is the polymorphic variants present in the constituents of a CNV locus. To study the role of CNV

in a phenotype or a disease risk, an accurate and definitive answer to the number of genes and the specific variants present in the study subject is essential. Typically, independent strategies are required to validate and confirm the presence and the accuracy of a CNV call.

### Genomic Southern blot analyses

The path taken to unravel the complexity of RCCX modular variation, and the *C4* CNV and its associated polymorphisms, was crooked but filled with surprises. Early restriction fragment length polymorphism (RFLP) analyses using genomic DNA digested with *TaqI* enzyme, processed by Southern blot techniques and hybridized to a cDNA probe corresponding to the 5' region of the *C4* transcript revealed the presence of four restriction fragments of 7.0, 6.4, 6.0 and 5.4 kb. Based on limited structural and genetic information, it was suggested that the 7.0-kb fragment corresponded to a *C4A* gene, the 6.0- and 5.4-kb fragments represented polymorphic variants of *C4B*, while the 6.4-kb fragment marked a haplotype with a single *C4B* gene and a ~30-kb deletion of genomic sequence containing *C4A* (Carroll et al., 1985; Schneider et al., 1986). Subsequent deliberate sequence analyses revealed that the 7.0- and the 6.4-kb *TaqI* fragments actually represent an *RP1* gene linked to a long *C4* gene and to a short *C4* gene, respectively; while the 6.0- and the 5.4-kb fragments represent a *TNXA-RP2* hybrid segment linked to a long *C4* gene and to a short *C4* gene, respectively (Yu and Campbell, 1987; Shen et al., 1994; Yang et al., 1999). Both long *C4* gene and short *C4* gene can code for either *C4A* or *C4B*. The specific DNA sequences defining *C4A*



**Fig. 4.** Demonstration of RCCX modular variations by pulsed-field gel electrophoresis (PFGE) of *PmeI* digested genomic DNA. **(A)** Genetic map of the MHC complement gene cluster. Horizontal arrows show the transcriptional orientations of genes. Upward arrows show the locations of the *PmeI* restriction sites, or the locations of DNA probes for hybridization. **(B)** Demonstration of quadrimodular (Q), trimodular (T), bimodular (B) and monomodular (M) RCCX haplotypes by *PmeI* PFGE of genomic DNA from 12 human subjects. Lanes 1–4 are from subjects with homozygous trimodular LLL, bimodular LL, monomodular L, and monomodular S, respectively. Lanes 5–12 are from subjects who are heterozygous in RCCX with different combinations of haplotypes.

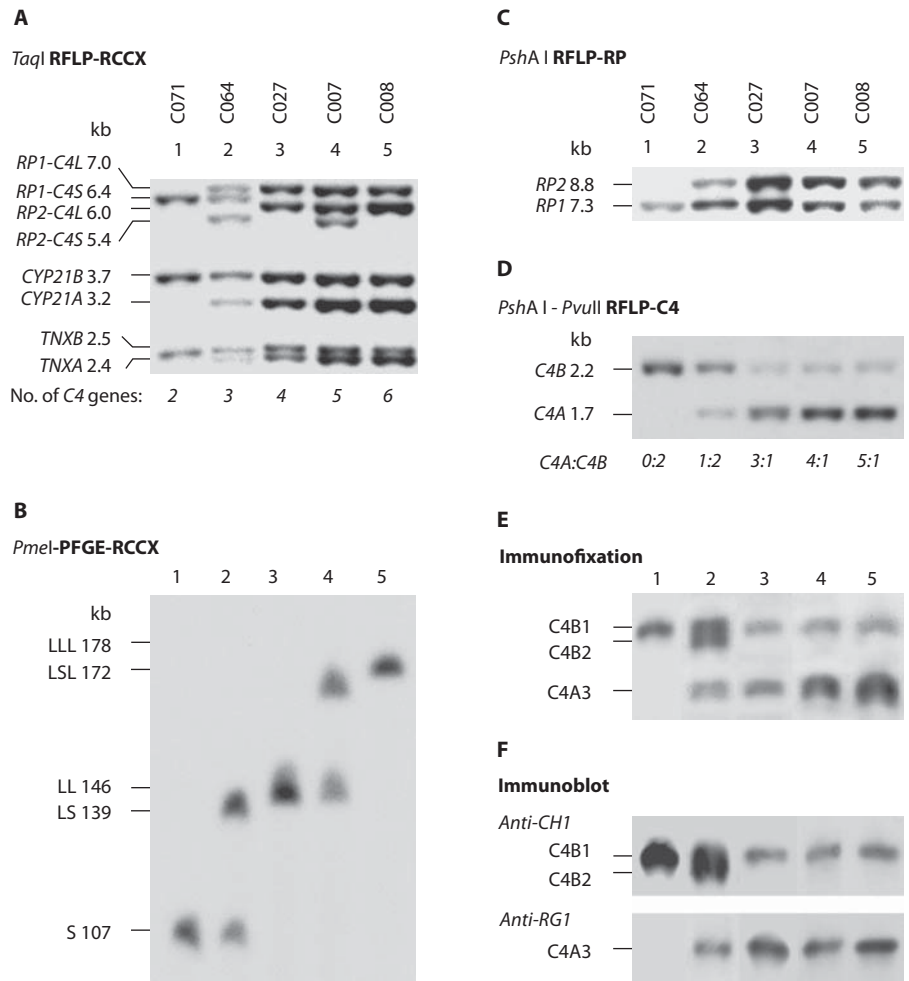
and *C4B* (coding for amino acid residues 1101–1106) can be distinguished by restriction enzymes *NlaIV* or *PshAI*, with the latter being a more robust enzyme and therefore preferred in Southern blot analysis (Yu and Campbell, 1987; Chung et al., 2002b).

To clearly demonstrate the variations in the number of RCCX modules, a long range mapping approach was applied. Through deliberate analyses of rare cutting restriction enzyme sites unaffected by methylations of CpG sequences at the MHC class III region, a single *PmeI* restriction site (recognition sequence: GTTTAAAC) was found in the complement factor B gene (*CFB*), and another *PmeI* site was found at the 5' region of the *TNXB* gene, which is beyond the breakpoint of RCCX duplication. Using intact genomic DNA trapped in low gelling temperature agarose for *PmeI* restriction digests and resolved by pulsed-field gel electrophoresis and Southern blot analysis, one can clearly demonstrate the presence of 1, 2, 3 or 4 RCCX modules with long or short *C4* genes present in a haplotype (Chung et al., 2002a, b). As shown in Fig. 4, subjects with monomodular RCCX haplotype can be represented by a 107-kb *PmeI* fragment if the *C4* gene is short or by a 113-kb fragment if the *C4* gene is long. An additional RCCX module will increase the *PmeI* fragment by 33 kb (with long *C4*) or by 26 kb (with short *C4*).

Employing three specific probes (probes A, E and F, Fig. 4A) for hybridization in a *TaqI* RFLP analysis to inter-

rogate (a) the combinations of *RP1/ RP2* genes with long or short *C4*, (b) a single nucleotide polymorphism frequently associated with *CYP21A* (*CYP21A1P*) and *CYP21B* (*CYP21A2*) genes, and (c) a 120-bp deletion frequently present in the *TNXA* but not in *TNXB*, enables an accurate determination of the copy numbers of RCCX modules with refined information on long and short *C4* genes, plus specifics of both *CYP21* and *TNX* (Fig. 5A). Independent Southern blot analysis of *PshAI/PvuII*-digested genomic DNA hybridized to a *C4d*-specific probe yielded relative gene copy numbers of *C4A* and *C4B* (Fig. 5D). Taken together with immunofixation experiments to elucidate the *C4A* and *C4B* allotypes, immunoblot experiments for the presence of the Ch1 and Rg1 blood group antigenic determinants, and radial immunodiffusion (or ELISA) for the plasma protein concentrations, we have the tools to precisely examine the genotypic and phenotypic variations of *C4* and RCCX modules (Chung et al., 2005). We have applied this strategy to study the genotypes and phenotypes in >1,000 human subjects with and without SLE (Yang et al., 2007). The strength of the Southern blot approach is that it elucidates the presence and absence, as well as the quantities of the actual genomic DNA segments being duplicated. Therefore, the results can be definite and conclusive. The limitation of genomic Southern blot is that the procedures require relatively large quantities of high quality DNA, i.e., at least 5  $\mu$ g per reaction per sample, and the procedure usually takes

**Fig. 5.** Genotypic and phenotypic variations of human C4 from five human subjects with 2, 3, 4, 5 or 6 copies of C4 genes. **(A)** *TaqI* RFLP to demonstrate the configurations of *RP1* linked to a long C4 or a short C4, the presence and relative quantities of *RP2* linked to a long C4 or a short C4, the presence and relative quantities of *CYP21B* and *CYP21A*, and the presence and relative quantities of *TNXB* and *TNXA*. **(B)** Demonstration of RCCX haplotypes by *PmeI* PFGE. Lane 1 is homozygous S/S; lane 2 is heterozygous LS/S; lane 3 is homozygous LL/LL; lane 4 is heterozygous LSL/LL; lane 5 is homozygous LLL/LLL. **(C)** *PshAI* RFLP demonstrating the relative quantities of *RP1* and *RP2*. *RP2* is not present in monomeric RCCX haplotypes. The number of *RP1* genes is constant (i.e., 2 copies per diploid genome) among all human subjects. The relative intensities of *RP2* to *RP1* restriction fragments give information about the number of duplicated RCCX modules present in an individual. **(D)** *PshAI-PvuII* RFLP to determine the presence and relative quantities of *C4A* and *C4B* genes. **(E)** Immunofixation of EDTA-plasma resolved by high voltage agarose gel electrophoresis to demonstrate *C4A* and *C4B* protein polymorphisms, using polyclonal serum against human C4. **(F)** Immunoblot experiments to show C4 proteins associated with Chido or with Rodgers blood group antigens. A monoclonal antibody against Ch1 was used in the upper panel; a monoclonal antibody against Rg1 was used in the lower panel (Chung et al., 2002b).

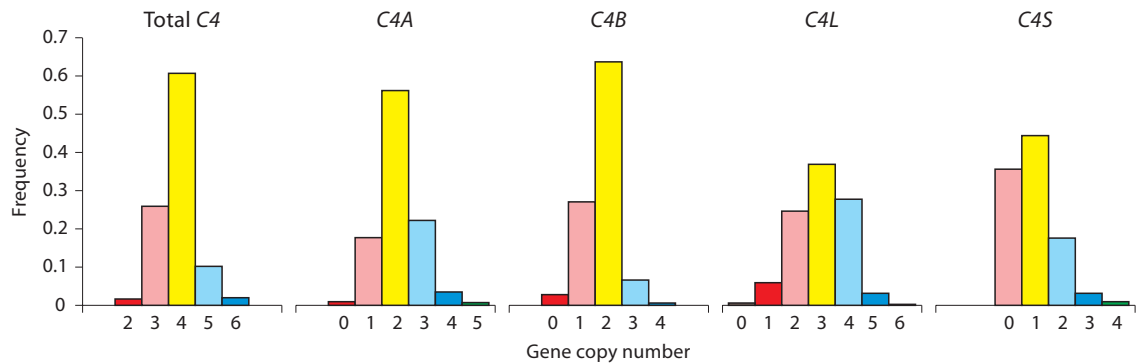


longer than three days to complete. For collaborative clinical studies of genetic risk factors involved in disease susceptibility, one would need to study hundreds to thousands of samples for multiple susceptibility genes. The demand for microgram quantity of genomic DNA per subject is usually prohibitive. Therefore, though the Southern blot was highly informative for characterizing the complexity of C4 CNV, it soon became clear that a sensitive, accurate and efficient alternative strategy was needed to interrogate C4 gene CNV in available cohorts of controls and patients.

#### PCR-based analyses of C4 CNVs

Initial efforts to determine the quantitative variation of genetic variants or SNPs by RFLP analyses of PCR-amplified genomic DNA were hampered by artifacts such as the formation of heteroduplexes during the PCR process. A creative strategy known as 'hot-stop PCR' circumvents pitfalls caused by heteroduplexes but the method is labor-intensive (Uejima et al., 2000; Chung et al., 2002b). More recently, a series of TaqMan-based real-time PCR methods have been developed and vigorously validated to decipher the C4 CNV and its associated polymorphisms (Wu et al., 2007). The fundamental concept for quantitative real-time PCR is that

the number of cycles taken to amplify an amplicon exponentially ( $C_T$  or threshold cycle) is dependent upon the initial quantity of DNA template present in a reaction. Thus gene copy number variation is one of the parameters that determine the  $C_T$ . With appropriately controlled template DNA samples, the higher the gene copy number (GCN) of an amplicon, the smaller is the  $C_T$  required to exponentially amplify a target. With application of an internal control that has a constant gene copy number among different samples, one can determine the difference of  $C_T$  between a target amplicon and an internal control amplicon ( $\Delta C_T$ ), and with a reference sample which has a known copy number of the same target gene, one can calculate the copy number of the target amplicon in an unknown sample by comparing the  $\Delta C_T$  between the reference sample and that of the unknown sample. This method is commonly known as comparative  $C_T$  method or  $\Delta\Delta C_T$  method. Several factors can significantly affect the amplification efficiencies between the target and control amplicons and create ambiguous or inaccurate results. First and foremost is the difference in amplification efficiencies between the target and control amplicons under a range of template concentrations. The second is the presence of impurities among different sam-



**Fig. 6.** Frequencies of gene copy number (GCN) groups for total *C4*, *C4A*, *C4B*, long *C4* (*C4L*) and short *C4* (*C4S*) in healthy subjects. The GCN frequencies were calculated from data based on >500 female and male healthy subjects of European ancestry from central Ohio (Yang et al., 2007).

ples differentially affecting either or both target and control amplicons. The third is the intrinsic difficulty to distinguish the  $\Delta C_T$  for high copy numbers of genes namely, 5 copies versus 6 copies. To overcome potential artifacts and ensure accuracies, we have developed a series of assays with multiple levels of controls to interrogate the copy number variations of human *C4*, which are briefly described as follows:

1. Instead of the comparative  $C_T$  method, the relative standard curve method is employed. Standard curves of  $C_T$  versus relative copy numbers of DNA template for the target and internal control amplicons are constructed by titrating a DNA sample that contains a known molar ratio of the target gene fragment to the internal gene fragments by six orders of magnitude (logs). In our assays, we use genomic cosmid DNA samples that contain an equal molar ratio of internal control amplicon (*RPI*) and target amplicon (*C4A*, *C4B*, *C4L*, *C4S* or *TNXA-RP2*) to generate such standard curves. The relative copy numbers of DNA template of target and internal controls in each unknown sample are independently calculated referring back to the two standard curves based on their observed  $C_T$ . A molar ratio between the target and the internal controls in an unknown sample is obtained by taking the quotient of the relative copy numbers between the target and the internal control. Since the latter has an invariant gene copy number across all samples, the gene copy number of the target can be calculated by multiplying the molar ratio and the gene copy number of the internal control. This method is more robust in withstanding differences in amplification efficiencies and unequal DNA concentrations across different samples and has yielded consistent results for continuous CNVs.

2. To ensure high accuracies for higher gene copy numbers, we also include DNA samples with different known gene copy numbers of the target gene in each experiment. A calibration curve is obtained by plotting the experimentally obtained gene copy numbers of these samples against their actual gene copy numbers. The gene copy numbers of the unknown samples are called after correcting the intrinsic

underestimation, especially for those with high copy numbers, by the calibration equation generated from this curve.

3. Two independent specific amplicons for *C4A* and *C4B* are designed using primers that specifically span the five defining SNPs in exon 26 for *C4A* and *C4B* genes.

4. Independent amplicon for *TNXA-RP2* is used to measure the total *C4* gene copy number. *TNXA-RP2* is only present in MHC haplotypes with duplicated RCCX modules. The number of total *C4* genes present = number of *TNXA-RP2* fragments + 2. This also increases the accuracy of results for high *C4* gene copy number by measuring the lower number target fragment (i.e., when total *C4* number equals 7, the number of *TNXA-RP2* is 5).

5. Independent amplicons for *C4* long and *C4* short are designed using specific primers for long and for short genes and a common TaqMan probe at intron 9 for both long and short genes.

6. Results of independent assays are validated by the equation:  $C4A + C4B = C4L + C4S = \text{number of } TNXA-RP2 + 2$ .

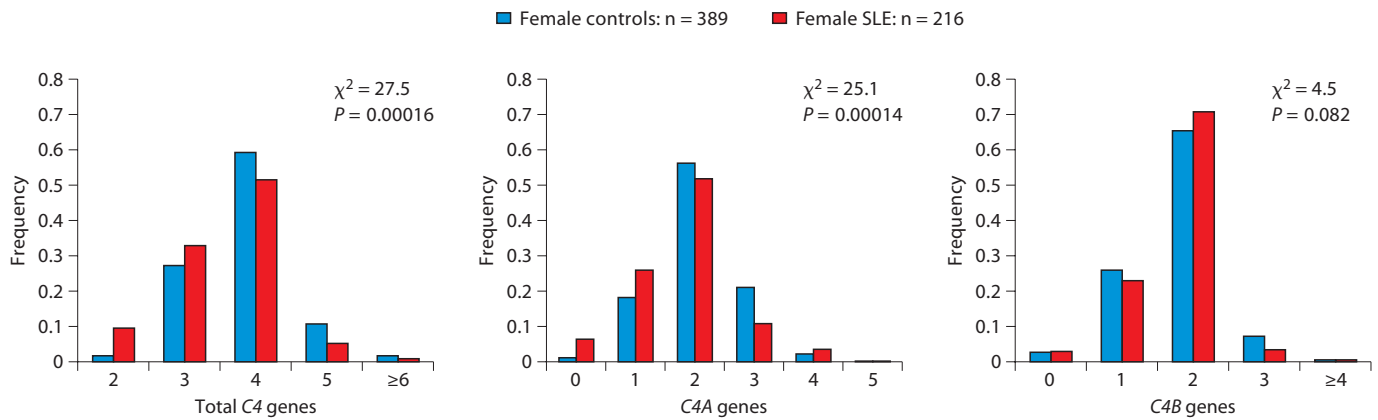
We have applied these strategies to over 2,000 human samples and the methods prove to be highly robust and reproducible. One important precaution to be taken is that primary genomic DNA is required for accurate results. It is found that whole genome amplified DNAs are not suitable for quantitative real-time PCR because of the widely unequal amplification among different genes during the process.

### C4 gene CNVs in healthy European Americans

We have determined *C4* gene copy numbers (GCN) in >500 healthy human subjects of Northern European ancestry from central Ohio. The distributions of the copy number groups for total *C4*, *C4A* and *C4B*, *C4L* and *C4S* are shown in Fig. 6.

In a diploid genome, the gene copy number of total *C4* varies from 2 to 6, although rare cases of seven or eight cop-





**Fig. 7.** Copy number variations of total *C4*, *C4A* and *C4B* in female SLE patients and female controls of European ancestry (Yang et al., 2007).

ies have been found (Blanchong et al., 2000; Chung et al., 2002a; Wu et al., 2007; Yang et al., 2007). While a good majority of healthy subjects have four copies of *C4* genes in a diploid genome (60.8%), about one-quarter have three copies (25.8%) and one-tenth have five copies (10.1%). Less than 4% of healthy subjects have either two copies (1.4%) or six copies (2.0%) of *C4* genes. While the most frequent number of total *C4* genes is four copies per diploid genome, it is intriguing to note that the distribution is skewed negatively towards the lower copy side. In the healthy population, 27.2% of human subjects have <4 copies of *C4* genes, but those with >4 copies only have a frequency of 12.0%.

The copy number of *C4A* genes in a diploid genome varies from 0 to 5. The most frequent GCN of *C4A* is two per diploid genome (56.0%). The *C4A* GCN distribution is relatively normal but slightly skewed towards the high copy number side (<2, 18.1%; >2, 25.9%). For *C4B*, the copy number varies from 0 to 4. The most frequent copy number of *C4B* is also two per diploid genome (63.6%) but the distribution is markedly skewed towards the low copy number side instead (<2, 29.9%; >2, 6.6%).

For long *C4* genes, the copy number varies from 0 to 6. The most frequent GCN is three (36.9%) per diploid genome and an even distribution pattern is observed (<3, 31.4%; >3, 31.8%). For short *C4* genes, the copy number varies from 0 to 4. The short gene distribution pattern is skewed significantly to the low copy number side: 35.1% of healthy subjects have no short *C4* genes, 44.4% have one copy and 17.4% have two copies per diploid genome. The frequency of subjects with more than two copies of short *C4* genes is only 3.0%.

### Lower total *C4* and *C4A* gene copy numbers in SLE patients

SLE is an autoimmune disease that predominantly affects women of childbearing age (Tsokos et al., 2007; Wallace and Hahn, 2007). The patient female to male ratio is about 9 to

1. We investigated the *C4* gene CNVs and associated polymorphisms in 216 female patients of European ancestry from central Ohio, and compared the results with those of 389 unrelated healthy controls who were gender-, race-, and geographically-matched. For the distribution of total *C4* and *C4A* GCN groups, there were significant increases in the low GCN groups, and decreases in the high GCN groups in SLE ( $P = 0.00016$  for total *C4*;  $P = 0.00014$  for *C4A*; Fig. 7). In essence, 42.2% of SLE patients had two or three copies of total *C4* genes, compared to 28.5% in female controls (Odds ratio: 1.823,  $P = 0.001$ ); 6% of SLE patients had five or six copies of *C4* genes, compared to 12% in female controls (Odds ratio: 0.466;  $P = 0.016$ ). The decrease in total *C4* GCN in SLE mainly contributes to a decrease in GCN in one of its two isotypes, *C4A*. There were significant increases in the frequencies of the *C4A* low copy number groups, and significant decreases in the frequencies of the *C4A* high copy number groups. Close to one-third (32.9%) of SLE patients had a homozygous or heterozygous deficiency (0 or 1 copy) of *C4A*, compared to 19.5% in matched controls (Odds ratio: 2.02,  $P = 0.0003$ ). On the other hand, 15.3% of SLE patients had 3, 4 or 5 copies of *C4A* genes, compared to 23.8% in healthy controls (Odds ratio: 0.574,  $P = 0.012$ ). The distributions of *C4B* GCN between European American SLE patients and controls were not significantly different. Similar results on lower copy numbers of total *C4* and *C4A* in SLE were obtained in a replication study using European American SLE patients from California.

Among the European Americans, there were good correlations between long *C4* genes and *C4A* ( $R = 0.695$ ), and between short *C4* genes and *C4B* ( $R = 0.437$ ) but those associations were not absolute. When the distributions of GCN groups for the long genes and the short genes were compared between SLE and controls, we observed an increase in the low copy number groups ( $\leq 2$ ), and a decrease in the medium and high copy number groups ( $\geq 3$ ) for the long genes ( $P = 0.036$ ). The GCN distribution patterns for the short genes were similar between SLE and controls (Yang et al., 2007).

## Concluding remarks

We are beginning a new era of studying genetic structural variations. An important element that drives genetic diversity is inter-individual gene CNV. This review illustrates a complex pattern of phenotypic and genetic variations that accompanies a specific CNV locus, namely, the RCCX in the central MHC on the short arm of chromosome 6. The focus is on the qualitative and quantitative diversities of immune functions created by gene CNV of complement C4. The C4A and C4B proteins probably evolved to have more efficient handling of antigens of different chemical nature, and the inter-individual quantitative variation further affords the flexibility to react to different microbes. The physiologic or pathogenic state of an individual can be determined by a delicate balance between a well-regulated immune response against infections, and an over-exuberant response that can lead to an autoimmune disease. It appears that among European Americans, low GCN of complement C4A is one of the genetic risk factors that can tip

the balance towards the autoimmune disease SLE. Further in-depth population studies are needed to determine the phenotypic and genotypic diversities created by C4 CNVs in different human races and ethnic groups, and the differences on disease susceptibilities to bacterial, viral, protozoan and fungal infections, and the disease risk and progression of autoimmune diseases. To assess the roles of CNVs in health and diseases, it is necessary to decipher the patterns of variation in a CNV locus, and to elucidate the associated polymorphisms, mutations, genetic recombinations that could have profound physiologic or functional impacts.

## Acknowledgements

We wish to thank past members of our group Drs. Carol Blanchong, Brad Baker, Drew Dangel, Kristi Rupert, Kapil Saxena, Liming Shen and Zhenyu Yang who contributed experimental data for this review. We would like to thank Dr. George Fust, Dr. Georges Hauptmann, Dr. Maisa Lokki, Dr. Karl Lotta and Dr. Joann Moulds for collaborations.

## References

- Awdeh ZL, Alper CA: Inherited structural polymorphism of the fourth component of human complement. *Proc Natl Acad Sci USA* 77:3576–3580 (1980).
- Awdeh ZL, Raum D, Alper CA: Genetic polymorphism of human complement C4 and detection of heterozygotes. *Nature* 282:205–208 (1979).
- Awdeh ZL, Raum D, Yunis EJ, Alper CA: Extended HLA/complement allele haplotypes: Evidence for T/t-like complex in man. *Proc Natl Acad Sci USA* 80:259–263 (1983).
- Belt KT, Carroll MC, Porter RR: The structural basis of the multiple forms of human complement component C4. *Cell* 36:907–914 (1984).
- Belt KT, Yu CY, Carroll MC, Porter RR: Polymorphism of human complement component C4. *Immunogenetics* 21:173–180 (1985).
- Birmingham DJ, Hebert LA: CR1 and CR1-like: the primate immune adherence receptors. *Immunol Rev* 180:100–111 (2001).
- Blanchong CA, Zhou B, Rupert KL, Chung EK, Jones KN, et al: Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in Caucasians: the load of RCCX genetic diversity on MHC-associated disease. *J Exp Med* 191:2183–2196 (2000).
- Blanchong CA, Chung EK, Rupert KL, Yang Y, Yang Z, et al: Genetic, structural and functional diversities of human complement components C4A and C4B and their mouse homologs, Slp and C4. *Int Immunopharmacol* 1:365–392 (2001).
- Cantor RM, Geschwind DH: Schizophrenia: genome, interrupted. *Neuron* 58:165–167 (2008).
- Carroll MC, Palsdottir A, Belt KT, Porter RR: Deletion of complement C4 and steroid 21-hydroxylase genes in the HLA class III region. *EMBO J* 4:2547–2552 (1985).
- Carroll MC, Fathallah DM, Bergamaschini L, Alicot EM, Isenman DE: Substitution of a single amino acid (aspartic acid for histidine) converts the functional activity of human complement C4B to C4A. *Proc Natl Acad Sci USA* 87:6868–6872 (1990).
- Chu X, Rittner C, Schneider PM: Length polymorphism of the human complement component C4 gene is due to an ancient retroviral integration. *Exp Clin Immunogenet* 12:74–81 (1995).
- Chung EK, Yang Y, Rennebohm RM, Lokki ML, Higgins GC, et al: Genetic sophistication of human complement C4A and C4B and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex (MHC). *Am J Hum Genet* 71:823–837 (2002a).
- Chung EK, Yang Y, Rupert KL, Jones KN, Rennebohm RM, et al: Determining the one, two, three or four long and short loci of human complement C4 in a major histocompatibility complex haplotype encoding for C4A or C4B proteins. *Am J Hum Genet* 71:810–822 (2002b).
- Chung EK, Wu YL, Yang Y, Zhou B, Yu CY: Human complement components C4A and C4B genetic diversities: complex genotypes and phenotypes, in Coligan JE, Bierer BE, Margulis DH, Shevach EM, Strober W (eds): *Current Protocols in Immunology*, pp 13.8.1–13.8.36 (John Wiley & Sons, Edison 2005).
- Collier S, Tassabehju M, Strachan T: A de novo pathological point mutation at the 21-hydroxylase locus: implications for gene conversion in the human genome. *Nat Genet* 3:260–265 (1993).
- Cornacoff JB, Hebert LA, Smead WL, VanAman ME, Birmingham DJ, Waxman FJ: Primate erythrocyte-immune complex-clearing mechanism. *J Clin Invest* 71:236–247 (1983).
- Dangel AW, Mendoza AR, Baker BJ, Daniel CM, Carroll MC, et al: The dichotomous size variation of human complement C4 gene is mediated by a novel family of endogenous retroviruses which also establishes species-specific genomic patterns among Old World primates. *Immunogenetics* 40:425–436 (1994).
- Dangel AW, Baker BJ, Mendoza AR, Yu CY: Complement component C4 gene intron 9 as a phylogenetic marker for primates: Long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics* 42:41–52 (1995).
- Dodds AW, Law SK: The complement component C4 of mammals. *Biochem J* 265:495–502 (1990).
- Dodds AW, Ren X-D, Willis AC, Law SKA: The reaction mechanism of the internal thioester in the human complement component C4. *Nature* 379:177–179 (1996).
- Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al: Completing the map of human genetic variation. *Nature* 447:161–165 (2007).
- Gatenby PA, Barbosa JE, Lachmann PJ: Differences between C4A and C4B in the handling of immune complexes: the enhancement of CR1 binding is more important than the inhibition of immunoprecipitation. *Clin Exp Immunol* 79:158–163 (1990).
- Giles CM, Uring-Lambert B, Goetz J, Hauptmann G, Fielder AHL, et al: Antigenic determinants expressed by human C4 allotypes: a study of 325 families provides evidence for the structural antigenic model. *Immunogenetics* 27:442–448 (1988).
- Iafraite AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al: Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951 (2004).
- Isenman DE, Young JR: The molecular basis for the differences in immune hemolysis activity of the Chido and Rodgers isotypes of human complement component C4. *J Immunol* 132:3019–3027 (1984).
- Law SK, Lichtenberg NA, Levine RP: Covalent binding and hemolytic activity of complement proteins. *Proc Natl Acad Sci USA* 77:7194–7198 (1980).
- Law SKA, Dodds AW, Porter RR: A comparison of the properties of two classes, C4A and C4B, of the human complement component C4. *EMBO J* 3:1819–1823 (1984).
- Longster G, Giles CM: A new antibody specificity, anti-Rga, reacting with a red cell, and serum antigen. *Vox Sang* 30:175–180 (1976).
- Lupski JR: Structural variation in the human genome. *N Engl J Med* 356:1169–1171 (2007).

- Mack M, Bender K, Schneider PM: Detection of retroviral antisense transcripts and promoter activity of the HERV-K(C4) insertion in the MHC class III region. *Immunogenetics* 56:321–332 (2004).
- Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al: Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 82:477–488 (2008).
- Martinez OP, Longman-Jacobsen N, Davies R, Chung EK, Gaudieri S, et al: Genetics of human complement C4 and evolution of the central MHC. *Frontiers Biosci* 6:d904–913 (2001).
- Mauff G, Alper CA, Awdeh Z, Batchelor JR, Bertrams T, et al: Statement on the nomenclature of human C4 allotypes. *Immunobiology* 164:184–191 (1983).
- Mauff G, Luther B, Schneider PM, Rittner C, Strandmann-Bellinghausen B, et al: Reference typing report for complement component C4. *Exp Clin Immunogenet* 15:249–260 (1998).
- Middleton J, Crookston M: Chido-substance in plasma. *Vox Sang* 23:256–261 (1972).
- O'Neill GJ, Yang SY, DuPont B: Two HLA-linked loci controlling the fourth component of human complement. *Proc Natl Acad Sci USA* 75:5165–5169 (1978).
- Porter RR: Complement polymorphism, the major histocompatibility complex and associated diseases: a speculation. *Mol Biol Med* 1:161–168 (1983).
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al: Global variation in copy number in the human genome. *Nature* 444:444–454 (2006).
- Reid KBM, Porter RR: The proteolytic activation systems of complement. *Ann Rev Biochem* 50:433–464 (1981).
- Roos MH, Mollenhauer E, Demant P, Rittner C: A molecular basis for the two locus model of human complement component C4. *Nature* 298:854–855 (1982).
- Rupert KL, Rennebohm RM, Yu CY: An unequal crossover between the RCCX modules of the human MHC leading to the presence of a *CYP21B* gene and a tenascin *TNXB/TNXA-RP2* recombinant between *C4A* and *C4B* genes in a patient with juvenile rheumatoid arthritis. *Exp Clin Immunogenet* 16:81–97 (1999).
- Schalkwijk J, Zweers M, Steijen PM, Dean WB, Taylor G, et al: A recessive form of the Ehlers-Danlos syndrome caused by tenascin-X deficiency. *New Engl J Med* 345:1167–1175 (2001).
- Schneider PM, Carroll MC, Alper CA, Rittner C, Whitehead AS, et al: Polymorphism of human complement C4 and steroid 21-hydroxylase genes. Restriction fragment length polymorphisms revealing structural deletions, homo-duplications, and size variants. *J Clin Invest* 78:650–657 (1986).
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al: Large-scale copy number polymorphism in the human genome. *Science* 305:525–528 (2004).
- Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al: Strong association of de novo copy number mutations with autism. *Science* 316:445–449 (2007).
- Shen L, Wu L-C, Sanlioglu S, Chen R, Mendoza AR, et al: Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region: molecular cloning, exon intron structure, composite retroposon, and breakpoint of gene duplication. *J Biol Chem* 269:8466–8476 (1994).
- Sim E, Cross S: Phenotyping of human complement component C4, a class III HLA antigen. *Biochem J* 239:763–767 (1986).
- Sutrala SR, Goossens D, Williams NM, Heyrman L, Adolffson R, et al: Gene copy number variation in schizophrenia. *Schizophr Res* 96:93–99 (2007).
- Tsokos GC, Gordon C, Smolen JS: *Systemic Lupus Erythematosus: a Companion to Rheumatology*, 1<sup>st</sup> ed (Mosby-Elsevier, Philadelphia 2007).
- Uejima H, Lee MP, Cui H, Feinberg AP: Hot-stop PCR: a simple and general assay for linear quantitation of allele ratios. *Nat Genet* 25:375–376 (2000).
- Wallace DJ, Hahn BH: *Dubios' Lupus Erythematosus*, 7<sup>th</sup> ed (Lippincott Williams & Wilkins, Philadelphia 2007).
- Walport MJ: Complement-part I. *New Engl J Med* 344:1058–1066 (2001).
- Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, et al: A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* 80:91–104 (2007).
- Wu YL, Savelli SL, Yang Y, Zhou B, Rovin BH, et al: Sensitive and specific real-time PCR assays to accurately determine copy-number variations (CNVs) of human complement *C4A*, *C4B*, *C4-Long*, *C4-Short* and *RCCX* modules: Elucidation of *C4* CNVs in 50 consanguineous subjects with defined HLA genotypes. *J Immunol* 179:3012–3025 (2007).
- Yang Y, Chung EK, Zhou B, Blanchong CA, Yu CY, et al: Diversity in intrinsic strengths of the human complement system: serum C4 protein concentrations correlate with *C4* gene size and polygenic variations, hemolytic activities and body mass index. *J Immunol* 171:2734–2745 (2003).
- Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, et al: Gene copy number variation and associated polymorphisms of complement component C4 in human systemic erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against European American SLE disease susceptibility. *Am J Hum Genet* 80:1037–1054 (2007).
- Yang Z, Mendoza AR, Welch TR, Zipf WB, Yu CY: Modular variations of HLA class III genes for serine/threonine kinase RP, complement C4, steroid 21-hydroxylase *CYP21* and tenascin *TNX* (*RCCX*): a mechanism for gene deletions and disease associations. *J Biol Chem* 274:12147–12156 (1999).
- Yu CY: The complete exon-intron structure of a human complement component C4A gene: DNA sequences, polymorphism, and linkage to the 21-hydroxylase gene. *J Immunol* 146:1057–1066 (1991).
- Yu CY, Campbell RD: Definitive RFLPs to distinguish between the human complement C4A/C4B isotypes and the major Rodgers/Chido determinants: application to the study of C4 null alleles. *Immunogenetics* 25:383–390 (1987).
- Yu CY, Whitacre CC: Sex, MHC and complement C4 in autoimmune diseases. *Trends Immunol* 25:694–699 (2004).
- Yu CY, Belt KT, Giles CM, Campbell RD, Porter RR: Structural basis of the polymorphism of human complement component C4A and C4B: gene size, reactivity and antigenicity. *EMBO J* 5:2873–2881 (1986).
- Yu CY, Campbell RD, Porter RR: A structural model for the location of the Rodgers and the Chido antigenic determinants and their correlation with the human complement C4A/C4B isotypes. *Immunogenetics* 27:399–405 (1988).
- Yu CY, Chung EK, Yang Y, Blanchong CA, Jacobsen N, et al: Dancing with complement C4 and the RP-C4-CYP21-TNX (*RCCX*) modules of the major histocompatibility complex. *Prog Nucl Acid Res Mol Biol* 75:217–292 (2003).