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


Center for Molecular and Human Genetics, The Research Institute at Nationwide Children’s Research Institute, Columbus, OH, USA

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 genes 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

Accepted in revised form for publication by H. Kehrer-Sawatzki and D.N. Cooper, 13 June 2008.
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 (Iafrate 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; Suturala et al., 2007; Cantor and Geschwind, 2008; Marshall et al., 2006). 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; Suturala et al., 2007; Cantor and Geschwind, 2008; Marshall et al., 2006). The other is common CNV, that exists with relative-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 mannann 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, immunoclearence by binding to erythocyte CR1 in the circulation, and degradation of the immune complexes through the endothelial reticuloocyte 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 allelotypes

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 (Awded and Alper, 1980; Sim and Cross, 1986). The most common allelotypes for C4A are A3, A2, A4, A6 and A1; the most common allelotypes 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 LSPV1H 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 allelotypes resolved by agarose gel electrophoresis can lyse sheep RBC three to four times faster than C4A allelotypes (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 GI054 (Ch5), LSPV1H 1101–1106 (Ch4), SI157 (Ch6), ADLR 1188–1191, combinations of GI054 and LSPV1H 1101–1106 (Ch2), and of SI157 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 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 24; 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).
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 human subjects of Northern European ancestry, there is an intact MHC haplotypes from human subjects of Northern European and 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).
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’/H11541 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 /H11011 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

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).
and C4B (coding for amino acid residues 1101–1106) can be
distinguished by restriction enzymes
\textit{Nla} IV or \textit{Psh} AI, with
the latter being a more robust enzyme and therefore pre-
ferred 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 ap-
pied. Through deliberate analyses of rare cutting restric-
tion enzyme sites unaffected by methylations of CpG se-
quences at the MHC class III region, a single
\textit{Pme} I restriction site (recognition sequence: GTTTAAAC) was found in
the complement factor B gene
(\textit{CFB}), and another
\textit{Pme} I site
was found at the 5′/H11541
region of the
\textit{TNXB} gene, which is be-
yond the breakpoint of RCCX duplication. Using intact ge-
nomic DNA trapped in low gelling temperature agarose for
\textit{Pme} I 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 \textit{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 \textit{Pme} I frag-
ment if the \textit{C4} gene is short or by a 113-kb fragment if the
\textit{C4} gene is long. An additional RCCX module will increase
the \textit{Pme} I fragment by 33 kb (with long \textit{C4}) or by 26 kb (with
short \textit{C4}).

Employing three specific probes (probes A, E and F,
Fig. 4A) for hybridization in a \textit{Taq} I RFLP analysis to inter-
rogate (a) the combinations of \textit{RP1}/\textit{RP2} genes with long or
short \textit{C4}, (b) a single nucleotide polymorphism frequently
associated with \textit{CYP21A} (\textit{CYP21A1P}) and \textit{CYP21B}
(\textit{CYP21A2}) genes, and (c) a 120-bp deletion frequently pres-
ent in the \textit{TNXA} but not in \textit{TNXB}, enables an accurate de-
termination of the copy numbers of RCCX modules with
refined information on long and short \textit{C4} genes, plus specif-
cics of both \textit{CYP21} and \textit{TNX} (Fig. 5A). Independent South-
ern blot analysis of \textit{Psh} AI/\textit{Pvu} II-digested genomic DNA hy-
bridized to a \textit{C4d}-specific probe yielded relative gene copy
numbers of \textit{C4A} and \textit{C4B} (Fig. 5D). Taken together with
immunofixation experiments to elucidate the
\textit{C4A} and \textit{C4B} allotypes, immunoblot experiments for the presence of the
\textit{Ch}1 and \textit{Rg}1 blood group antigenic determinants, and ra-
dial immunodiffusion (or ELISA) for the plasma protein
concentrations, we have the tools to precisely examine the genotypic and phenotypic variations of \textit{C4} and RCCX mod-
ules (Chung et al., 2005). We have applied this strategy to
study the genotypes and phenotypes in 1,000 human sub-
jects with and without SLE (Yang et al., 2007). The strength
of the Southern blot approach is that it elucidates the pres-
ence 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 rela-
tively large quantities of high quality DNA, i.e., at least 5 μg
per reaction per sample, and the procedure usually takes

![Fig. 4. Demonstration of RCCX modular varia-
tions by pulsed-field gel electrophoresis (PFGE) of
\textit{Pme} I digested genomic DNA. (A) Genetic map of
the MHC complement gene cluster. Horizontal ar-
rows show the transcriptional orientations of
genes. Upward arrows show the locations of the
\textit{Pme} I 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
\textit{Pme} I PFGE of genomic DNA from 12 human sub-
jects. Lanes 1–4 are from subjects with homozy-
gous trimodular LLL, bimodular LL, monomodu-
lar L, and monomodular S, respectively. Lanes 5–
12 are from subjects who are heterozygous in
RCCX with different combinations of haplotypes.
The fundamental concept for quantitative real-time PCR is that the number of cycles taken to amplify an amplicon exponentially (CT 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 CT. With appropriately controlled template quantity of DNA and an internal control amplicon, one can determine the CT method or comparative CT 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 and create ambiguous or inaccurate results. 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

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...
samples differentially affecting either or both target and control amplicons. The third is the intrinsic difficulty to distinguish the ΔC_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 (RP1) 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 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-
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 genes 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 (≤2), and a decrease in the medium and high copy number groups (≥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 Pust, Dr. Georges Hauptmann, Dr. Maisa Lokki, Dr. Karl Lotta and Dr. Joann Moulds for collaborations.

References


