Intraspecific Evolution of Human RCCX Copy Number Variation Traced by Haplotypes of the *CYP21A2* Gene

Zsófia Bánlaki[1], Julianna Anna Szabó[1], Ágnes Szilágyi[1], Attila Patócs[2], Zoltán Prohászka[1], George Füst[1,†], and Márton Doleschall[1,*]

[1]3rd Department of Internal Medicine, Semmelweis University, Budapest, Hungary

[2]Molecular Medicine Research Group, Hungarian Academy of Sciences and Semmelweis University, Budapest, Hungary

*Corresponding author: E-mail: doles@kut.sote.hu.

[†]Prof. George Fust departed this life in the summer of 2012.

Accepted: December 9, 2012

Data deposition: Nucleotide sequence data reported are available in the GenBank database under the accession numbers JN034382–JN034411 and JQ993310–JQ993314.

Abstract

The RCCX region is a complex, multiallelic, tandem copy number variation (CNV). Two complete genes, complement component 4 (*C4*) and steroid 21-hydroxylase (*CYP21A2,* formerly *CYP21B*), reside in its variable region. RCCX is prone to nonallelic homologous recombination (NAHR) such as unequal crossover, generating duplications and deletions of RCCX modules, and gene conversion. A series of allele-specific long-range polymerase chain reaction coupled to the whole-gene sequencing of *CYP21A2* was developed for molecular haplotyping. By means of the developed techniques, 35 different kinds of *CYP21A2* haplotype variant were experimentally determined from 112 unrelated European subjects. The number of the resolved *CYP21A2* haplotype variants was increased to 61 by bioinformatic haplotype reconstruction. The *CYP21A2* haplotype variants could be assigned to the haplotypic RCCX CNV structures (the copy number of RCCX modules) in most cases. The genealogy network constructed from the *CYP21A2* haplotype variants delineated the origin of RCCX structures. The different RCCX structures were located in tight groups. The minority of groups with identical RCCX structure occurred once in the network, implying monophyletic origin, but the majority of groups occurred several times and in different locations, indicating polyphyletic origin. The monophyletic groups were often created by single unequal crossover, whereas recurrent unequal crossover events generated some of the polyphyletic groups. As a result of recurrent NAHR events, more *CYP21A2* haplotype variants with different allele patterns belonged to the same RCCX structure. The intraspecific evolution of RCCX CNV described here has provided a reasonable expectation for that of complex, multiallelic, tandem CNVs in humans.

Key words: allele-specific long-range PCR, CNV, genealogy network, nonallelic homologous recombination.

Introduction

Copy number variations (CNVs) occupy a small proportion of the human genome but contribute significantly to genetic diversity (Redon et al. 2006; Conrad et al. 2010), greatly influence cellular phenotypes such as gene expression (Stranger et al. 2007), and are responsible for a wide spectrum of diseases and disease susceptibilities (Zhang et al. 2009). Multiallelic CNVs (greater than 2 possible haploid copy number [Conrad et al. 2010]) constitute a sizeable fraction of large CNVs, are highly enriched with gene content, and are closely associated with segmental duplications by virtue of their prevalent duplicated alleles (Redon et al. 2006; Conrad et al. 2010). Multiallelic CNVs may be considered as recent duplications during fixation phase and under the effect of neutral or positive evolutionary processes (Innan and Kondrashov 2010; Teshima and Innan 2012): Consequently, they play a significant role in gene and genome evolution (Hurles et al. 2008; Marques-Bonet et al. 2009). Underlying this rapid evolution, CNV alleles (copy number on a chromosome) with large, homologous, and tandem repeats are prone to rearrangements via nonallelic homologous recombination (NAHR) mechanisms (Hastings et al. 2009) such as unequal crossover and gene conversion. Unequal crossover facilitates

large structural rearrangements and copy number changes (Stankiewicz and Lupski 2002), whereas gene conversion mediates relatively short sequence transfers (Chen et al. 2007). By contrast, the relative contribution of rearrangement mechanisms to the emergence and maintenance of CNVs and their gene content is not well appreciated, in spite of the recent advances in our knowledge of the mutation mechanisms of genome-wide CNVs (Kidd et al. 2010). Furthermore, using genome-wide platforms, the multiallelic, tandem CNVs, as well as their duplicated gene contents, are very difficult to genotype directly and are poorly tagged by single-nucleotide polymorphisms (SNPs) (Conrad et al. 2010; Alkan et al. 2011; Campbell et al. 2011).

RCCX, often recognized by genome-wide CNV studies (Tuzun et al. 2005; Redon et al. 2006; Perry et al. 2008a; Conrad et al. 2010; Kato et al. 2010), is a complex, medium size (~30 kb per module), multiallelic, tandem CNV in the major histocompatibility complex (MHC) class III region (Horton et al. 2004), and it commonly consists of monomodular, bimodular, and trimodular CNV alleles with the prevalence of approximately 15%, 75%, and 10% in Europeans, respectively (Blanchong et al. 2000; Vatay et al. 2003). Four genes—serine/threonine kinase 19 (*STK19*), complement component 4 (*C4*), steroid 21-hydroxylase (*CYP21*), and tenascin-X (*TNX*)—reside close to each other in each module. Considering all modules, each of these genes usually materializes in the form of one active gene and zero, one or two pseudogenes determined by the module number, except for *C4*, which has only active copies. There is a functional difference among *C4* genes dividing them into *C4A* and *C4B* types, because five adjacent nucleotide substitutions cause four amino acid changes and immunological subfunctionalization (Szilagyi, Doleschall, et al. 2010). The retention of the *C4A*-*C4B* nucleotide differences is observed in great apes; hence, this specialization of duplicated *C4* genes confers evolutionary advantage and provides a potential explanation for the emergence of RCCX CNV (Kawaguchi et al. 1992; Innan and Kondrashov 2010). In addition, each *C4* gene contains a deletion CNV (0 or 1 haploid copy number [Conrad et al. 2010]) derived from the insertion of a human endogenous retrovirus K (HERV-K) sequence (Dangel et al. 1994; Tassabehji et al. 1994), and the prevalence of the insertion allele of this HERV-K (C4) CNV depends on the position of its harboring module in the RCCX (Blanchong et al. 2000). These variations in copy number and gene content result in a CNV with a highly complex structure, which is traditionally described by the copy number of RCCX modules, and, per module, by the deletion or insertion allele (the absence or presence of the insertion) of HERV-K CNV and the type of *C4* gene (Yu et al. 2003), even though these features embody genetic polymorphisms that differ in nature and size. In this article, a haplotypic RCCX module is abbreviated with two letters, the first represents the alleles of the HERV-K CNV (L—long allele [insertion allele] or S—short allele [deletion allele], the use of L and S abbreviation follows the tradition of published works on RCCX CNV) and the second symbolizes the types of *C4* gene (A or B). The multiplication of these two letters indicates the bi- and trimodular structures (see fig. 1 for some examples).

NAHR contributes substantially to the genetic diversity of RCCX. On the one hand, unequal crossover generates copy number changes (Yang et al. 1999; Blanchong et al. 2000) and very rare RCCX CNV alleles such as quadrimodular structures (Chung et al. 2002a; Koppens et al. 2002b), chromosomes with more than one active *CYP21* gene (*CYP21A2*) (Koppens et al. 2002b), chromosomes with only *CYP21* pseudogenes (*CYP21A1P*) (Koppens et al. 2003), and structures with chimeras of *CYP21* genes (Tusie-Luna and White 1995; Koppens et al. 2002a; Lee 2004). On the other hand, the deleterious mutations of the *CYP21A1P* gene, such as an 8-bp deletion in exon 3 and four related substitutions in exon 6, can be transferred by nonallelic gene conversion, causing the majority of the point mutations in *CYP21A2* (Collier et al. 1993; Tusie-Luna and White 1995; Concolino et al. 2010). *CYP21A2* deficiency is by far the most common cause of congenital adrenal hyperplasia (CAH), the inherited inability to synthesize cortisol and aldosterone (White and Speiser 2000).

We assumed that RCCX structures were related to particular *CYP21A2* alleles, and the primary aim of this study was to unravel the intraspecific evolution of RCCX CNV by means of the whole-gene haplotypes (the term haplotype was used to indicate "gene-based functional" haplotype [Hoehe 2003] in this study) of the polymorphic internal gene, *CYP21A2*. A molecular haplotyping technique has been developed for RCCX CNV based on the concept of allele-specific long-range polymerase chain reaction (ASLR-PCR) (Michalatos-Beloin et al. 1996), and full-length *CYP21A2* haplotypes have been determined from the haplotypic products of ASLR-PCR in many cases. Bioinformatic haplotype reconstruction has followed the experimental work to resolve the experimentally indeterminable haplotypes from genotypic *CYP21A2* sequences. Besides the intraspecific evolution, we also attempted to trace the NAHR events of RCCX CNV by the genealogical haplotype network. The characteristics of RCCX structure-*CYP21A2* haplotype variants and NAHR events forming RCCX CNV described here have provided reasonable expectations for the intraspecific evolution of complex, multiallelic, tandem CNVs in humans.

Material and Methods

Subjects

Unrelated European subjects from Hungary who participated in a previous study on full-length *CYP21A2* gene sequences (Blasko et al. 2009) were investigated initially, but original subjects with three copies of *CYP21A2* (see later for the method of determination) and those who did not have

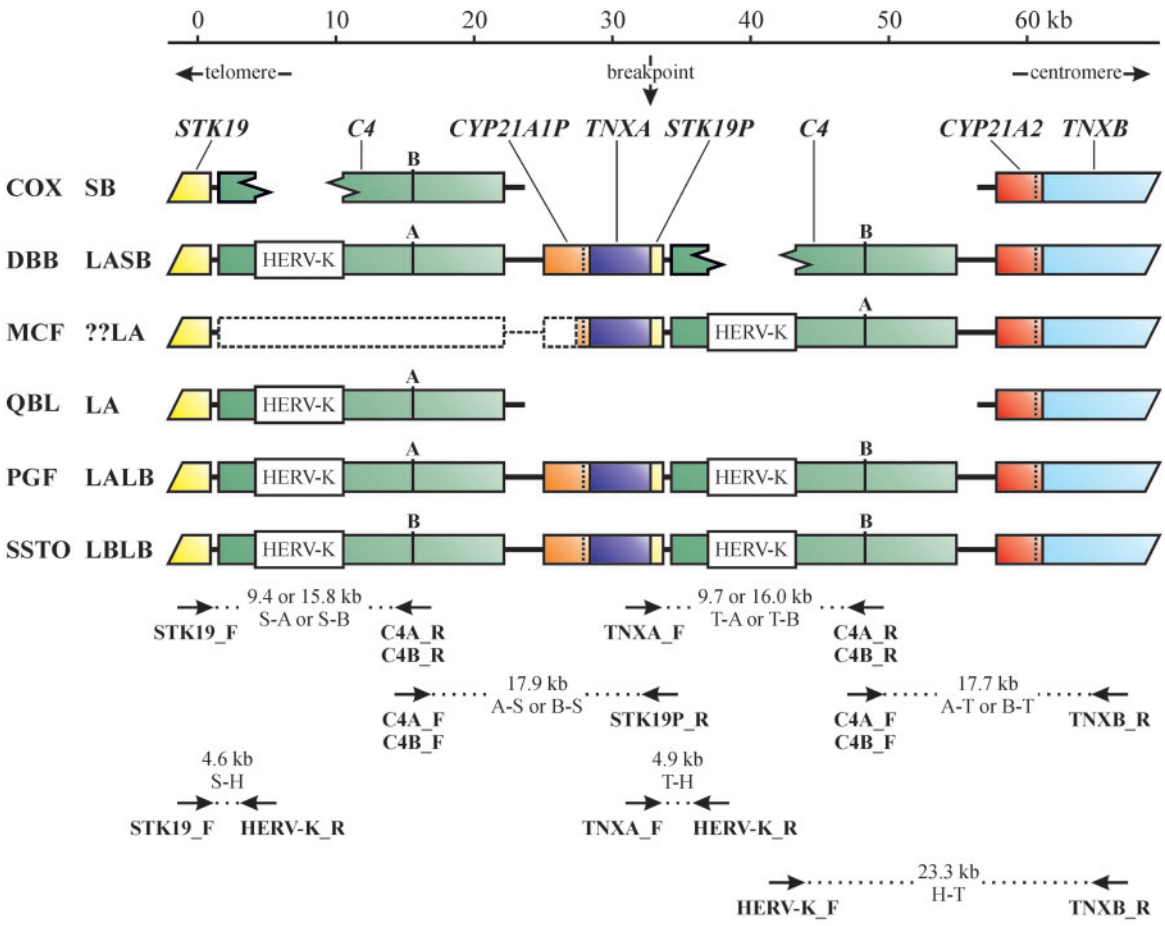

Fig. 1.—Scale representation of the alignment of the RCCX variable region sequences from the external database and the localizations of the developed ASLR-PCRs. The names of cell lines and the schematic abbreviation of RCCX structures are indicated on the left side. A module is abbreviated with two letters, the first represents the alleles of HERV-K CNV (L—the long allele or S—short allele), and the second symbolizes the type of *C4* gene (A or B). The duplication of these two letters indicates the bimodular structure. The alignment of the RCCX variable region has been generated from six MHC haplotype sequences of HLA-homozygous cell lines (NG_005163.2, NT_007592.15, NT_167245.1, NT_167247.1, NT_167248.1, and NT_167249.1). The alignment spans from the telomeric end of exon 4 of *STK19* to the centomeric end of exon 28 of *TNXB*. Dashed line indicates sequence absent from the MCF cell line. The RCCX structures of cell lines are monomodular and bimodular. The variable region of bimodular RCCX contains two pairs of complete genes, complement component 4 (*C4A* and *C4B*), steroid 21-hydroxylase (*CYP21A1P* and *CYP21A2*), and two pairs of partial genes, serine/threonine kinase 19 (*STK19* and *STK19P*) and tenascin-X (*TNXA* and *TNXB*). The CNV of the HERV-K virus sequence is located in the *C4* genes. The module breakpoint of bimodular structures and the direction of the ends of chromosome 6 are indicated under the scale bar. The positions and names of ASLR-PCR primers and the length of PCR products are shown at the bottom. The names of ASLR-PCRs are abbreviated by the first letter or the *C4* gene type of forward and reverse primers.

sufficient quality (fragmented DNA is inappropriate for long-range PCR) or enough DNA for *CYP21A2* resequencing from haplotypic products were excluded, resulting in 72 study subjects (A summary of experimental design and work flow can be found in supplementary fig. S1, Supplementary Material online). At the second stage, RCCX structure was investigated (see later) in 244 unrelated Hungarian subjects with European ancestry, and 40 unrelated subjects with two copies of *CYP21A2* were included in such a way as to represent a sufficient amount of all kinds of known RCCX structure (Blanchong et al. 2000) and to be suitable for the molecular haplotyping of *CYP21A2*. This sorting strategy enabled us to maximize the coverage of *CYP21A2* haplotype space and the discovery of rare haplotypes (these are miscalled when a statistical inference approach is applied [Tishkoff et al. 2000]) related to rare RCCX structures. However, the utilization of allele frequencies had to be rejected (the allele frequencies were not used in the subsequent bioinformatic analyses) because the subjects were sorted. Overall, 112 unrelated Hungarian subjects were included and fully investigated. The subjects gave informed consent, the study was approved by the Hungarian National Ethical Committee, and was executed according to the principles of the Declaration of Helsinki.

Molecular Haplotyping and Determination of RCCX Structures

Haplotypic RCCX structures and the suitable diploid RCCX structure combinations for the molecular haplotyping of *CYP21A2* (one *C4A* and one *C4B* gene next to *TNXB* gene, or one *C4* with the insertion allele of HERV-K CNV and one *C4* with the deletion allele next to *TNXB*) were determined using a set of ASLR-PCRs (fig. 1) and the copy number analyses of *C4* genes and HERV-K CNV.

The ASLR-PCRs principally relied on *C4A* and *C4B* allele-specific forward and reverse primers (C4A_F, C4B_F, C4A_R, and C4B_R) complementary to the discriminating nucleotide substitutions of *C4* genes in exon 26. In addition to the *C4* type-specific primers, *STK19* and *TNXB* gene-specific primers (STK19_F and TNXB_R), that matched only the active genes, and the *STK19P* and *TNXA* primers (STK19P_R and TNXA_F), which adhered to both active and pseudogene of *STK19* and *TNX*, but they were able to generate PCR products only from the pseudogenes with the *C4* allele-specific primers, were applied. Finally, HERV-K-specific primers (HERV-K_F and HERV-K_R) were also used, fitting only to *C4* genes with the insertion allele of HERV-K CNV (fig. 1 and supplementary table S1, Supplementary Material online). The following 10 primer pairs were applied from the possible combinations of these ASLR-PCR primers: STK19_F or TNXA_F with C4A_R or C4B_R (four pairs), C4A_F or C4B_F with STK19P_R or TNXB_R (four pairs), HERV-K_F with TNXB_R (one pair), and TNXA_F with HERV-K_R (one pair). An additional primer pair, STK19_F-HERV-K_R, was set up and described, but it was not needed in this study. The ASLR-PCRs were performed (supplementary table S2, Supplementary Material online) using LongAmp *Taq* DNA polymerase (New England Biolabs) according to the manufacturer's protocol with some modifications. PCRs were carried out in 10 μl total volume containing 1 U LongAmp *Taq* DNA polymerase, 1× LongAmp *Taq* reaction buffer, 300 μM of each dNTP, 0.4 μM of each primer, and 10–100 ng genomic DNA depending on DNA quality.

The copy numbers of the *C4A* and *C4B* genes, as well as the number of *C4* genes with the insertion and deletion alleles of HERV-K CNV, were determined by quantitative PCR (qPCR) as described previously (Szilagyi et al. 2006; Wu et al. 2007) with some modifications. *C4*-specific Taqman probes (Applied Biosystems) were labeled with the fluorescent dye 6-FAM, whereas *RPPH1*, used as an endogenous reference (RNase P reference assay, Applied Biosystems) in multiplex reactions, was labeled with the dye VIC.

Amplification and Sequencing of the *CYP21A2* Gene

To study the *CYP21A2* gene, the *CYP21A1P* pseudogene and their possible chimeric forms, two allele-specific nested PCRs corresponding to 5′- and 3′-parts of the *CYP21* genes (supplementary fig. S2, Supplementary Material online) were performed from the genotypic or haplotypic products of ASLR-PCRs generated by C4A_F or C4B_F with STK19P_R or TNXB_R, and HERV-K_F with TNXB_R primers. The subjects whose *CYP21A2*-specific nested PCR product was amplified from ASLR-PCR products by C4A_F or C4B_F with STK19P_R primers were considered as subjects with three *CYP21A2* copies and were excluded. The nested PCRs of the 5′-part were achieved by primers adhered to the allele-specific nucleotide substitutions in exon 6 (CYP21A1P_R or CYP21A2_R), with nonspecific primer matched to the 5′-flanking region (CYP21_F). The nested PCRs of 3′-part were accomplished by *CYP21A1P* and *CYP21A2* allele-specific primers complementary to the 8 bp indel difference in exon 3 (CYP21A1P_F or CYP21A2_F), with the nonspecific primer matched to the 3′-flanking region (CYP21_R). Each reaction (15 μl total volume) contained 1 U GoTaq DNA polymerase (Promega), 1× GoTaq colorless Flexi buffer, 1.5 mM $MgCl_2$, 200 μM of each dNTP, 133 nM of each primer, and ~4 ng ASLR-PCR product directly from the ASLR-PCR mix. The cycle conditions were 95°C for 5 min, 15 cycles of 95°C for 10 s, 64°C for 5 s, and 72°C for 90 s (CYP21_F with CYP21A1P_R or CYP21A2_R) or 150 s (CYP21A1P_F or CYP21A2_F with CYP21_R), finishing with extension at 72°C for 5 min.

The full-length *CYP21A2* was capillary sequenced following the allele-specific nested PCR. Nested PCR products were treated with exonuclease I (New England Biolabs) and rAPid alkaline phosphatase (Roche), then directly sequenced on both strands by 7–7 primers (supplementary table S1, Supplementary Material online) using the BigDye Terminator Sequencing Kit v3.1 (Applied Biosystems) and run on an ABI 3100 Genetic Analyzer (Applied Biosystems).

Bioinformatic Sequence and Haplotype Analyses

The sequences of RCCX CNV and *CYP21* genes from GenBank were used (supplementary table S3, Supplementary Material online). The 491 expressed sequence tags (EST) sequences of *CYP21* genes and 6 MHC haplotype sequences of HLA-homozygous cell lines (Horton et al. 2008) from *STK19* to *TNXB* were aligned using ClustalX2 v2.0.5 (Larkin et al. 2007). The start of the *CYP21A2* gene was defined at 8 bp upstream from the start of the coding region (Higashi et al. 1986) by 5′-EST analysis (supplementary fig. S3, Supplementary Material online) (Nagaraj et al. 2007). The sequence calls of *CYP21A2* were assembled with CLC DNA Workbench v5.7.1 (CLC bio) and inspected manually by two different operators. RCCX structures and *CYP21A2* haplotypes, which could not be determined experimentally, were inferred with PHASE v2.1.1 (Stephens et al. 2001; Stephens and Donnelly 2003) (Specialized phasing tools for CNVs [Kato et al. 2008; Su et al. 2010] could not been used because of the lack of ability to handle the known phase information from individual to individual.). To input RCCX structure data into PHASE, HERV-K CNV and the type of *C4* gene in each module were treated as

independent loci. Because the trimodular RCCX structures are relatively prevalent, but quadrimodular structure is extremely rare, six loci represented the three modules. First four loci represented the 5′-modules, and zero allele indicated the lack of the particular 5′-module(s) in bimodular and monomodular RCCX structures. Because the deletion of the full RCCX region has not yet been observed, zero alleles could not be present on the last two loci representing the 3′-module. The known phasing information from the experiments was input with the phasing option (-k) of PHASE (an example is given in supplementary fig. S4, Supplementary Material online). For the connection of unconnected RCCX structure-*CYP21A2* haplotypes, the experimental RCCX structure data were inferred together with the correctly resolved (experimentally determined or above 0.99 confidence probability threshold) *CYP21A2* haplotypes coded as known phase. To check the correctness of both coding of the RCCX structure and the connecting of RCCX structures and *CYP21A2* haplotypes, a simulated RCCX CNV data set was generated by the random connection of the haplotypic RCCX structures from a recent family study not relying on any bioinformatic inferences (Banlaki, Doleschall, et al. 2012) and the resolved *CYP21A2* haplotypes. Median-joining networks were constructed from the experimental and inferred haplotypes with Network v4.6.1.0 (Bandelt et al. 1999). A chimpanzee (*Pan troglodytes*) *CYP21A2* sequence was applied as an outgroup (root) for the network building.

Results

Experimental Determination of RCCX Structures and *CYP21A2* Haplotypes

The organization of RCCX structures with respect to the number of RCCX modules, *C4* gene types, and HERV-K CNV was investigated by a set of ASLR-PCRs (fig. 1), allele-specific nested PCRs (supplementary fig. S2, Supplementary Material online) and qPCRs. Furthermore, full-length *CYP21A2* haplotypes determined by allele-specific nested PCR and sequencing were assigned to these RCCX structures in many cases. Although ASLR-PCR was capable of determining the *C4* gene types in conjunction with the alleles of HERV-K CNV within a haplotypic module, the number of modules and relationship between modules on a chromosome were deduced from the results of ASLR-PCRs and qPCRs.

First, the deduction method is exemplified by the alignment of RCCX structures of six HLA-homozygous cell lines (Horton et al. 2008) (fig. 1). Two monomodular RCCX structures of COX and QBL in diploid state do not produce PCR fragments from *TNXA* and *STK19P*, but STK19_F and C4A_R primers result in a 15.8 kb fragment corresponding to QBL, and STK19_F and C4B_R primers generate a 9.4 kb fragment corresponding to COX. The TNXB_R primer with C4A_F and C4B_F primers (in two separate tubes) results in two 17.7 kb haplotypic products. Both *CYP21A2* haplotypes can be amplified and sequenced from the two separated 17.7 kb ASLR-PCR products, and thus *CYP21A2* haplotypes can be assigned to the corresponding RCCX structures. In addition to the PCR product of C4A_F and TNXB_R primers, the *CYP21A2* haplotype related to QBL can be determined from the 23.3 kb product of HERV-K_F and TNXB_R primers, because this ASLR-PCR product is not generated from COX. Therefore, the diploid combination of monomodular COX and QBL is abbreviated to LA/SB.

Henceforth, the deduction is demonstrated by the results of ASLR-PCRs, allele-specific nested PCRs, and qPCRs in four samples (fig. 2). The organization of two identical monomodular RCCX structures such as LA/LA was also deduced as described earlier, but the ASLR-PCR products of two chromosomes by C4A_F-TNXB_R (A-T) or HERV-K_F-TNXB_R (H-T) primer pairs could not be separated (sample 1). Therefore, the *CYP21A2* alleles of the two chromosomes could only be genotyped after the *CYP21A2*-specific nested PCR. Besides the products by TNXB_R (A-T, B-T, H-T) and by STK19_F-C4A_R (S-A) primer pair from the 5′- and 3′-ends of RCCX CNV, the products by C4A_F-STK19P_R (A-S) and TNXA_F-C4B_R (T-B) primer pairs were amplified in sample 2, verifying the presence of at least one multimodular RCCX structure. The presence of A-T product and the absence of product by TNXA_F-C4A_R (T-A) primer pair indicated that there was a monomodular LA RCCX structure on one chromosome, and the absence of T-A product also implied that there was no trimodular RCCX structure with the *C4A* gene in the middle module on the other chromosome. The absence of product by C4B_F-STK19P_R (B-S) and STK19_F-C4B_R (S-B) primer pairs indicated that there was a bimodular RCCX structure with *C4A* in the 5′-end and *C4B* in the 3′-end on the second chromosome. The size of S-A and T-B products verified that *C4A* genes were found together with the L allele of HERV-K CNV in a module, and the *C4B* gene was together with the S allele. Therefore, the diploid combination of haplotypic RCCX structures was LA/LASB in sample 2, which was concordant with the copy numbers of *C4A* and *C4B* genes (CN A and B) and the alleles of the HERV-K CNV (CN L and S). The *CYP21A1P*-specific products were amplified by 5′- and 3′-nested PCR from AS ASLR-PCR product as template, and *CYP21A2*-specific products were generated from the product of C4B_F-TNXB_R (B-T) primer pair. In the case of A-T and H-T products (these were amplified from the same LA RCCX structure), a *CYP21A2*-specific 5′-nested PCR product and a *CYP21A1P*-specific 3′-nested PCR product were generated, indicating that LA RCCX structure harbored a chimeric *CYP21* gene.

The LALB/LASB RCCX structures of sample 3 were unambiguously determined similarly to that of sample 2, but only a short (9.7 kb) T-B fragment could be detected in spite of the presence of H-T product. The shorter fragment may be preferred to the longer one in PCR, thus sample 3 was checked by a TNXA_F-HERV-K_R (T-H) primer pair to avoid this potential

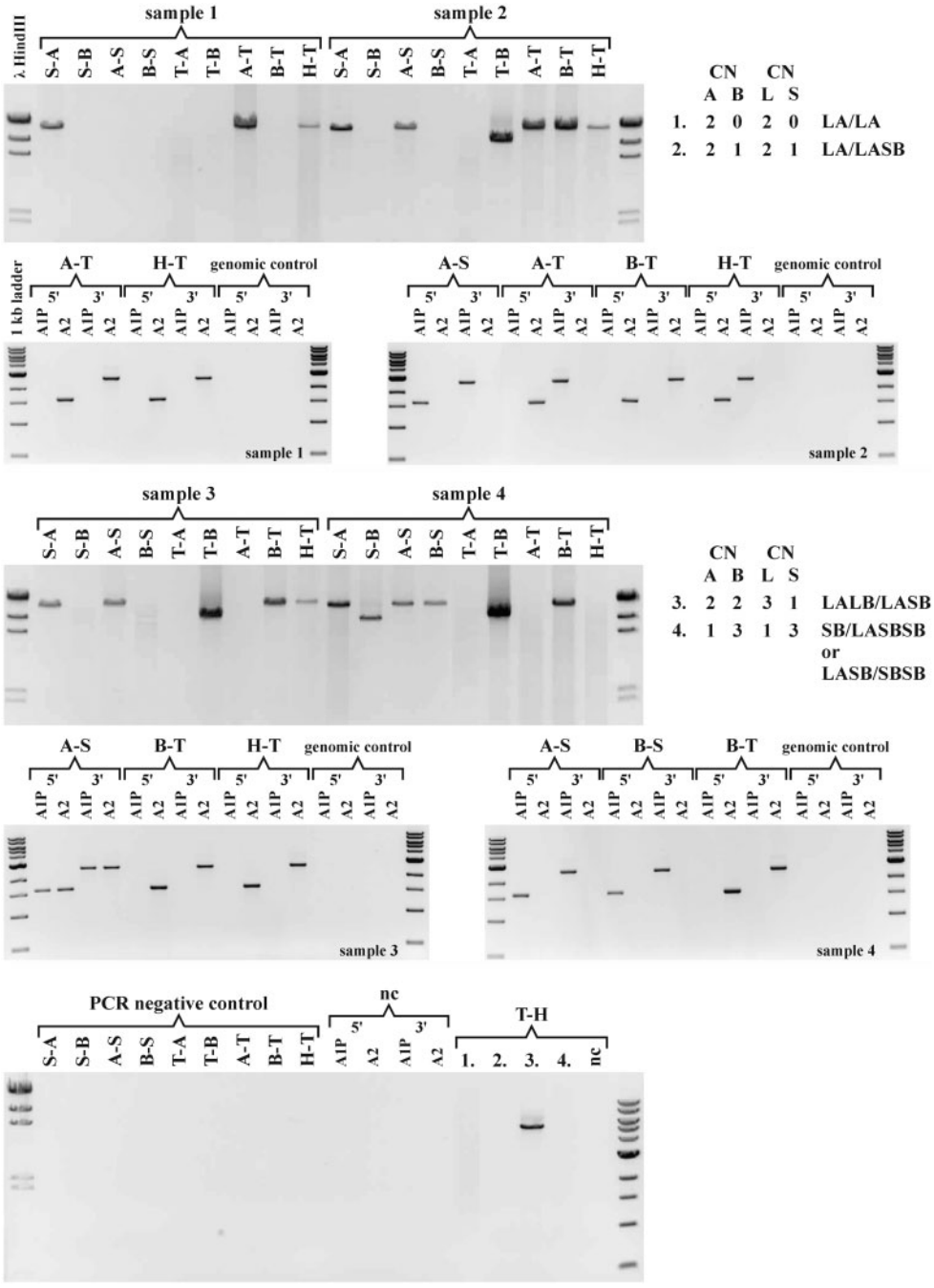

Fig. 2.—Results of ASLR-PCRs, allele-specific nested PCRs and qPCRs demonstrated in four samples (samples 1–4). The names of ASLR-PCRs are abbreviated by the first letter or the *C4* gene type of forward and reverse primers, in alphabetical order: C4A_F-STK19P_R (A-S), C4A_F-TNXB_R (A-T), C4B_F-STK19P_R (B-S), C4B_F-TNXB_R (B-T), HERV-K_F-TNXB_R (H-T), STK19_F-C4A_R (S-A), STK19_F-C4B_R (S-B), TNXA_F-C4A_R (T-A), TNXA _F-C4B_R (T-B), and TNXA_F-HERV-K_R (T-H). The names of *CYP21A1P*- and *CYP21A2*-specific nested PCRs are abbreviated by the specific tag of *CYP21* genes (A1P and A2) and the corresponding half of the gene from where the products can be amplified (5′ and 3′). The copy numbers (CN) of *C4* genes and the alleles of HERV-K CNV determined by qPCRs are abbreviated by the types of *C4* (A or B) and the long and short CNV alleles of HERV-K (L or S). Haplotypic RCCX module is abbreviated with two letters, the first represents the alleles of HERV-K CNV (L or S) and the second symbolizes the types of *C4* gene (A or B). The multiplication of the two letters in a structure indicates the module number. For *CYP21A1P*- and *CYP21A2*-specific nested PCRs, a portion of ASLR-PCR mix containing ~4 ng ASLR-PCR product was used as template. Genomic control confirms that the nested PCR products could not be amplified from the same amount of genomic DNA of the particular sample as the amount being included in a mix with ASLR-PCR product. PCR-negative control (nc) signifies the traditional control of PCR (complete PCR mix without DNA). HindIII digested lambda DNA (New England Biolabs) and 1 kb DNA ladder (New England Biolabs) were used as markers.

error. The presence of T-H products verified that the long (16 kb) T-B fragment became undetectable (usually, the long products of S-A, S-B, T-A, and T-B could be detected in our hands). The A-S and B-T products were derived from both chromosomes, hence the haplotypic fragments could not be separated from each other, but the H-T product was haplotypic. Therefore, only a genotypic *CYP21A2*-specific nested-PCR product could be acquired from the B-T ASLR-PCR product and a haplotypic product from the H-T product. Intriguingly, both *CYP21A1P*- and *CYP21A2*-specific nested PCR products were amplified from the A-S product, indicating that one of the two chromosomes harbored a *CYP21A2* gene in the 5′-module. Those subjects whose *CYP21A2*-specific nested PCR product was amplified from the A-S or B-S product were considered as subjects with three *CYP21A2* copies and were therefore excluded from the study because three gene copies from a diploid subject would have severely complicated the subsequent bioinformatic haplotype reconstruction. However, it should be noted that *CYP21* haplotypes in 5′-modules can also be examined using the ASLR- and nested PCRs, which may be proven helpful for a recent research area (Tsai et al. 2011).

The diploid RCCX combinations of the first three samples could be unambiguously determined only from the pattern of ASLR-PCRs. In many subjects, RCCX combinations were unambiguous merely based on the set of ASLR-PCRs, taking all conceivable RCCX haplotypes into account. In the cases when the determination of the copy numbers by ASLR-PCR and qPCR was redundant, a perfect concordance was observed between the data of the two assays, demonstrating the reliability of both. In addition, the sequencing of haplotypic and genotypic PCR products also confirmed the reliability and accuracy of the methods. In spite of the deduction from the redundant results of different type of assays, the deduction was made unambiguously only in a proportion of diploid combinations of haplotypic RCCX structures. Some of the combinations showed the same ASLR-PCR pattern, but copy numbers could distinguish them from each other. Some combinations could not be deciphered from experimental results, hence we were only able to narrow down the number of possible combinations. For example, the presence of S-A, S-B, and T-B products and the absence of A-T products in sample 4 indicated that there were an LA and an SB module in the 5′-end of RCCX structures and two *C4B* genes in the 3′-end modules. However, the B-S product might be derived from a bimodular RCCX structure with a 5′-end *C4B* gene or from a trimodular RCCX structure with a *C4B* gene in the middle module. Therefore, it was not possible to determine whether SB/LASBSB or LASB/SBSB was the real RCCX combination.

Overall, haplotypic RCCX structures were experimentally determined on 110 (49%) of 224 chromosomes of the 112 subjects and full-length *CYP21A2* haplotypes on 64 (29%) chromosomes. Molecular haplotyping (the experimental determination) revealed 8 different kinds of haplotypic RCCX structure variant and 35 different *CYP21A2* haplotype variants (GenBank: JN034382–JN034411 and JQ993310–JQ993314). Moreover, 23 of these *CYP21A2* haplotype variants were unambiguously assigned to haplotypic RCCX structures (supplementary table S4, Supplementary Material online). In addition, one chimeric *CYP21A1P-CYP21A2* gene harbored by LA RCCX structure also occurred (sample 2).

Bioinformatic Reconstruction of Haplotypic RCCX Structures and *CYP21A2* Haplotypes

The RCCX structures and *CYP21A2* alleles undetermined by molecular haplotyping were inferred using bioinformatic haplotype reconstruction by PHASE software. The RCCX polymorphism (module copy number, *C4* gene type, and HERV-K CNV in each module) data set and the *CYP21A2* polymorphism data set were separately analyzed, taking account of the experimentally determined haplotypic structures or haplotypes (known phases). Haplotypic RCCX structures and *CYP21A2* haplotypes were considered as resolved above the confidence probability threshold of 0.99 (this value is much stricter than those of most published works [Garrick et al. 2010]). From the 224 chromosomes, 148 (66%) haplotypic RCCX structures belonging to 10 kinds of RCCX structure variant were above the 0.99 confidence threshold, and 213 (95%) *CYP21A2* haplotypes were above the 0.99 threshold (supplementary table S5, Supplementary Material online). When the 5′-parts of resolved *CYP21A2* haplotypes were compared with the 377 filtered *CYP21A2* 5′-ESTs of ADRGL2 data set from dbEST, *CYP21A2* sequences proved to be highly concordant (supplementary table S6, Supplementary Material online). The *CYP21A2* haplotypes below the 0.99 confidence limit and the chimeric *CYP21* haplotype were excluded from the subsequent analyses to remove ambiguous structures and haplotypes. Furthermore, the RCCX structures were assigned to the resolved *CYP21A2* haplotypes by PHASE. From the 213 resolved *CYP21A2* haplotypes, 199 (93%) *CYP21A2* haplotypes were connected to RCCX structures above the 0.99 threshold, and only 14 (7%) could not be unambiguously assigned (supplementary table S4, Supplementary Material online). In addition, both RCCX and *CYP21A2* polymorphism data sets were analyzed without known phases (treated as genotype data) to evaluate the contribution of the haplotypic information to the efficiency of haplotype reconstruction (supplementary table S5, Supplementary Material online), and a simulated RCCX CNV data set was also analyzed to check the correctness of both coding of the RCCX structures and the connecting of RCCX structures and *CYP21A2* haplotypes (supplementary table S7, Supplementary Material online).

Altogether, the 213 experimental and inferred *CYP21A2* haplotypes represented 61 different variants (supplementary table S4, Supplementary Material online) containing 51 segregating sites in total (supplementary fig. S2 and table S8,

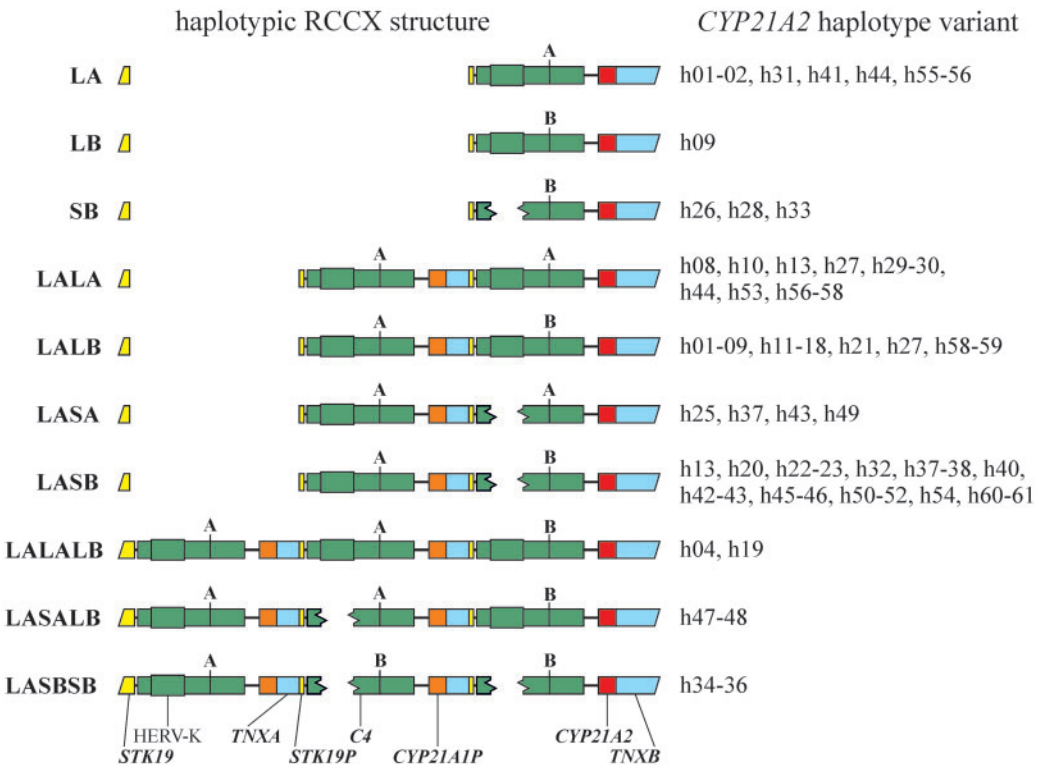

Fig. 3.—Graphic representation of resolved 71 haplotypic RCCX structure-*CYP21A2* haplotype variants. Haplotypic RCCX structures on the left side are abbreviated with the multiplication of the two letters of a module. In a module, the first represents the alleles of HERV-K CNV (L or S) and the second symbolizes the types of *C4* gene (A or B). *CYP21A2* haplotype variants are on the right side, and the names of the genes at the bottom. Eleven *CYP21A2* haplotype variants are connected to more than one RCCX structure. The segregating sites of *CYP21A2* haplotype variants with the related haplotypic RCCX structures can be also found in supplementary table S4, Supplementary Material online (In the figure, the *CYP21A2* haplotype variants are grouped by the haplotypic RCCX structure, and in supplementary table S4, haplotypic RCCX structures are grouped by the *CYP21A2* haplotype variant.).

Supplementary Material online). The number of singleton haplotype variants was 26 (43%), implying that a considerable degree of mutation events occurred in the recent past. The combined molecular and inferred haplotyping approach finally resulted in 71 different RCCX structure-*CYP21A2* haplotype variants (fig. 3), and only 2 (3%) of the *CYP21A2* haplotype variants could not be assigned to a haplotypic RCCX structure. Given the genetic linkages between the haplotypic RCCX structures and the harbored *CYP21A2* haplotypes, one *CYP21A2* haplotype variant was related to only one RCCX structure in 48 (79%) cases, but one haplotype variant related to more RCCX structures was also observed in 11 (18%) cases. To confirm the RCCX structure-*CYP21A2* haplotype variants, they were compared with the RCCX structures and harbored *CYP21A2* sequences of HLA-homozygous cell lines: The structures and sequences were highly concordant (supplementary table S4, Supplementary Material online).

Genealogical Network of *CYP21A2* Haplotype Variants

To construct a haplotype network that allows for the unique characteristics of intraspecific level such as persistent ancestral nodes, multifurcations, and reticulations (Posada and Crandall 2001), the median-joining method was applied. *CYP21A2* haplotype variants were free from the marks of crossover, presumably owing to the shortness of the gene and the low rate of meiotic (equal) crossover in the RCCX region (Cullen et al. 2002), and thus the prerequisite for the applicability of the median-joining algorithm was realized. To give an evolutionary direction to the network, it was rooted by a chimpanzee *CYP21A2* orthologue. The network showed a tree-like structure with some reticulations and intensive multifurcations (fig. 4*A* and supplementary fig. S5, Supplementary Material online). The h58 haplotype variant was connected to the root, implying that the h58 haplotype variant was the most ancestral haplotype in the network. When haplotypic RCCX structures were projected onto the corresponding *CYP21A2* haplotype variants, tight grouping related to RCCX structures was evident (fig. 4*B*). The haplotypic RCCX structures were not taken into consideration for the construction of the network, therefore, the *CYP21A2* haplotype variants as "complex genetic markers" (Hoehe 2003) independently reflected the genealogy of the entire RCCX CNV. To test the stability of the network and the effect of inferred *CYP21A2* haplotype

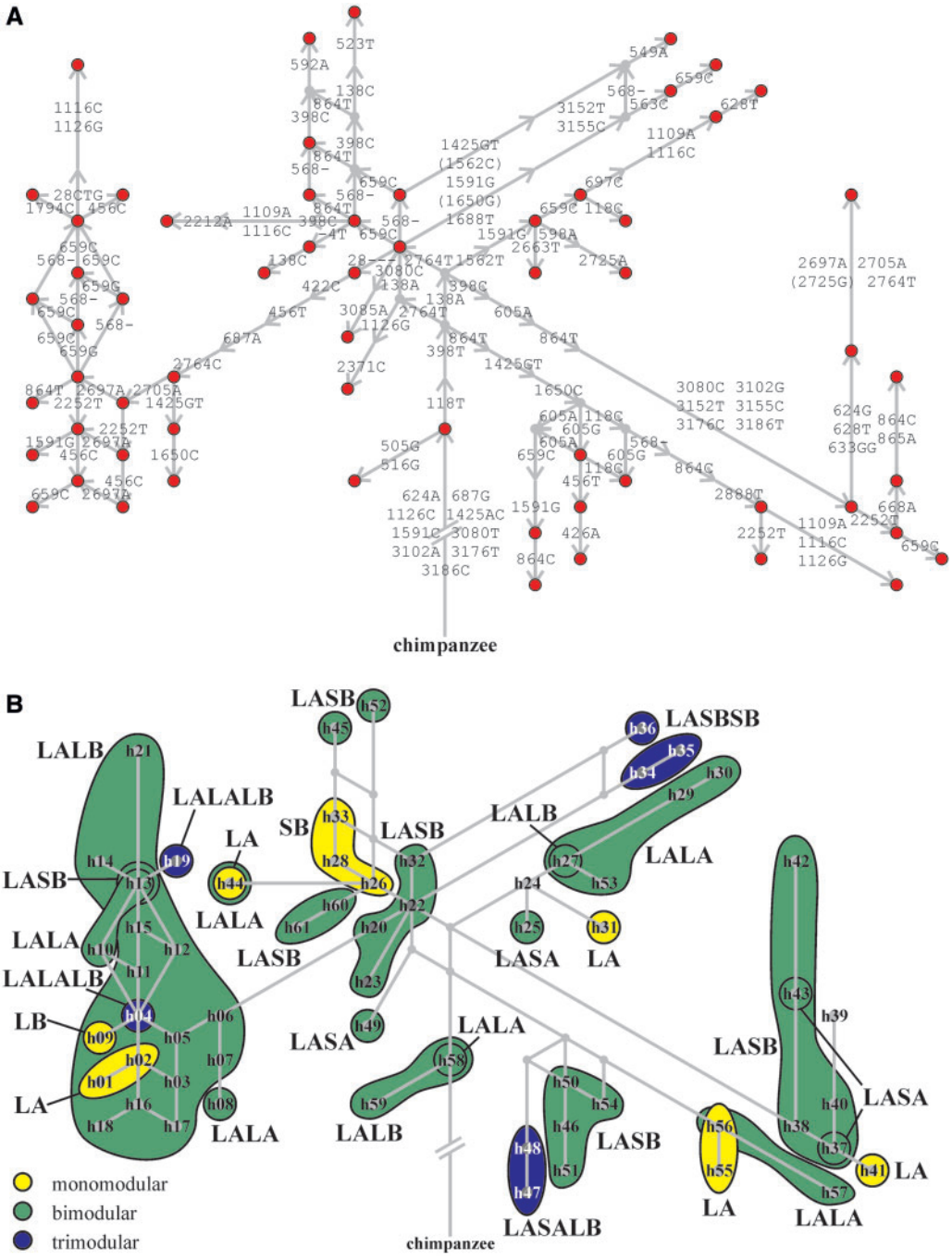

Fig. 4.—Genealogical haplotype networks of *CYP21A2* haplotype variants (the root is abridged.). (*A*) Haplotype network constructed from *CYP21A2* haplotype variants. Red circles indicate the (sampled) *CYP21A2* haplotype variants, light gray circles show the missing intermediates, and light gray arrows symbolize the allele-state changes (character-state changes) with the positions of segregating sites and the arisen allele. Two or more allele changes belonging to adjacent segregating sites have been considered as unambiguous gene conversion events: These allele-state changes are indicated together. (*B*) Haplotype network with projected RCCX structures constructed from *CYP21A2* haplotype variants. Light gray circles indicate the *CYP21A2* haplotype variants with their names. Monomodular CNV alleles (copy number on a chromosome) are indicated by yellow, bimodular by green, and trimodular by blue. Haplotypic RCCX module is abbreviated with two letters, the first represents the alleles of HERV-K CNV (L—long allele or S—short allele), and the second symbolizes the types of *C4* gene (A or B). The multiplication of the two letters in a structure indicates the module number.

variants on the network, an additional network was reconstructed using only the 35 experimentally determined *CYP21A2* haplotype variants (supplementary fig. S6, Supplementary Material online). The position of the root was identical (connected to h58 haplotype variant), and the *CYP21A2* haplotype variants remained in the original groups.

The h58 haplotype variant harbored by LALA or LALB RCCX structure was directly connected to the root, but there were no haplotype variants connected directly to the h58 variant. There were eight variants (h22, h24, h38, h48, h49, h50, h54, and h56) with quite different allele composition related indirectly to h58 variant through six missing intermediates (median vectors). Noticeably, none of these eight haplotype variants were carried by LALA or LALB structure. The deviations in the allele composition of h58-related haplotype-variants, the missing intermediates, and the different RCCX structures harboring these variants suggested larger evolutionary distances among h58 and its indirectly connected haplotype variants than inside the well-separated groups of directly connected haplotypes. These well-separated groups of directly connected haplotypes with branch-like junctions and with reticulation were found toward the tips of the network. With respect to CNV alleles (copy number on a chromosome), the RCCX structure groups of each allele (mono-, bi-, and trimodular) were widely scattered in the network, supporting their polyphyletic origin. Many groups with identical RCCX structure also occurred several times in the network. For instance, LA structure could be found in five distinct locations, which also indicated that they were polyphyletic groups. In contrast, some RCCX groups such as SB and LASALB occurred only once in the network, implying the monophyletic origin of these RCCX structures.

Unequal Crossovers in RCCX CNV

It is commonly understood that the mono- and trimodular RCCX structures are generated from bimodular structures by unequal crossover (fig. 5*A*) in humans (Yang et al. 1999; Blanchong et al. 2000). We therefore attempted to trace the origin of mono- and trimodular structures by unequal crossover based on the haplotype network. A mono- or trimodular RCCX structure group was considered as the product of an unequal crossover event if it was embedded in a group of a bimodular RCCX structure containing an identical *CYP21A2* haplotype variant or if it had a direct connection to an adjacent bimodular group. Thus one parental bimodular and one resultant mono- or trimodular (recombinant) structure could be examined by means of the haplotype network from the two parental and the two resultant chromosomes of an unequal crossover (The unequal crossover of monomodular structures cannot lead to the copy number change, and the trimodular groups of the network were in tip positions, or were embedded, and consequently these structures were not regarded as parental structures).

Corresponding to the aforementioned definition, eight unequal crossover events were observed in the network (fig. 5*B*), and six monomodular and two trimodular structures arose from them. Moreover, four events resulted in independent LA structures, and two events led to independent LALALB structures, and hence, these unequal crossover events were recurrent with respect to the particular RCCX structure. In the case of some monomodular structures such as the h44 haplotype variant harbored by LA, the breakpoint in front of the 5′-end of *CYP21A2* on the parental bimodular structure and at the back of the 5′-LA part on the other parental structure and the breakpoint between the module boundary of parental bimodular structure and in front of the 5′-C4 gene on the other chromosome could create the monomodular structure (fig. 5*B*). In effect, other breakpoints between these two breakpoints can be also envisioned, and therefore, LA-h44 structure could arise by a breakpoint located from the 3′-end of *CYP21A1P* to the 5′-end of *CYP21A2* of the parental bimodular structure harboring h44. For the generation of SB structure, the contribution of a breakpoint between the 5′-end of the HERV-K CNV and the *CYP21A2* gene could be excluded because the 5′-end of the *C4B* gene with an S allele of the HERV-K CNV (5′-SB part) did not occur except in itself, the arising SB structure. Therefore, the haplotype variants of SB structure presumably arose from a breakpoint around the module boundary of LASB-h22 and three subsequent, consecutive allele-state changes. In this scenario, the SB structure did not change during the allele-state changes. This is further supported by the fact that the SB structure is totally absent in a European CAH population of a previous study, because SB structure probably reduces the unequal crossover events due to a greater degree of dissimilarity compared to other RCCX structures (Blanchong et al. 2000). Similar to the LA-h44 structure, LALALB structures could be generated by breakpoints with different locations and by different RCCX structures along with the parental LALB structure, including bi- and trimodular structures.

It should be noted that the other two trimodular groups excluded by virtue of their connection to the missing intermediates were probably generated by unequal crossover as well. LASBSB might be created by two LASB structures, and LASALB might be born from a LASB and a LALB structure, as described previously (Chung et al. 2002a). In addition to unequal crossover, some events of (equal) crossover or conversions affecting *C4* type-specific nucleotides (Braun et al. 1990; Jaatinen et al. 2002) were also observed in the network.

Gene Conversions in the *CYP21A2* Gene

Two or more nonconsecutive allele changes belonging to adjacent segregating sites are considered as unambiguous gene conversion events (Chen et al. 2007). In addition to this criterion, an event was regarded as a gene conversion only if the allele combination of change was present in the *CYP21A2*

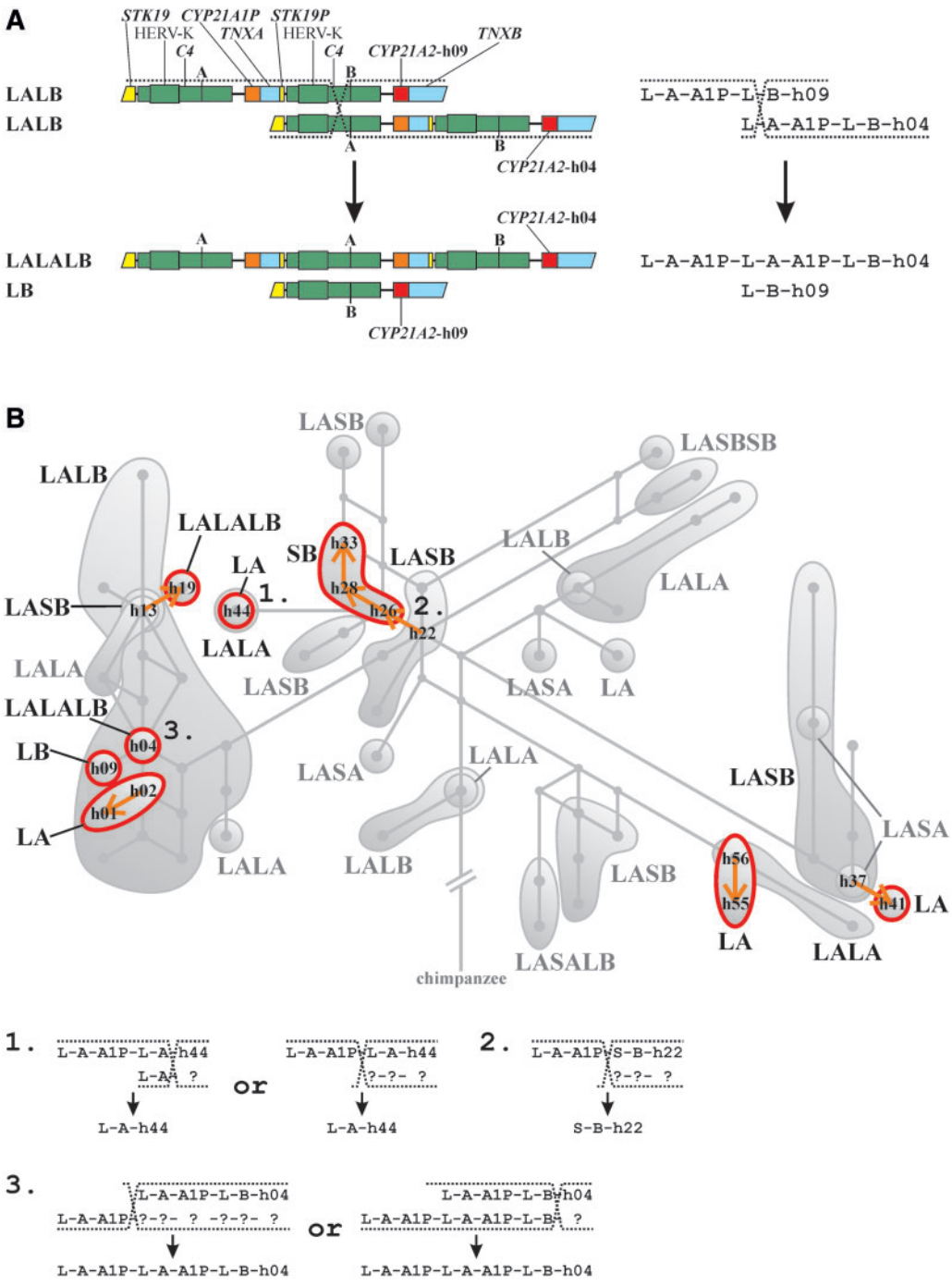

Fig. 5.—Unequal crossover events in RCCX CNV. (*A*) Scale and schematic representations of a hypothetical unequal crossover event in RCCX CNV (although both haplotypic RCCX structures with corresponding *CYP21A2* haplotypes exist, it is improbable that they have been generated exactly by this unequal crossover event). The name of genes in the RCCX CNV and the insertion allele of HERV-K CNV are indicated at the top. Haplotypic RCCX module is abbreviated with two letters, the first represents the alleles of HERV-K CNV (L—long allele or S—short allele), and the second symbolizes the types of *C4* gene (A or B). The multiplication of the two letters in a structure indicates the module number. (*B*) Unequal crossover of haplotypic RCCX structures on *CYP21A2* haplotype networks. Red circles indicate the unequal crossover events, orange arrows indicate the unequal crossover or consecutive mutational events, and the numbers preceded by h represent the CYP21A2 haplotypes. Numbers are assigned to particular unequal crossover events, the detailed explanations of which can be seen at the bottom. The question marks indicate the unknown RCCX polymorphisms of the parental structures. If two possibilities are present at the generation of a particular structure, the two possible breakpoints of unequal crossover are on the border of the range where the breakpoint can be located.

Table 1
Gene conversions in the *CYP21A2* haplotype network

Position of Segregating Sites	Arisen Allele	Origin	Tract Length	
			Min	Max
505–516	GG	*CYP21A1P*	12	37
624–633	GTGG	*CYP21A1P*	30	107
864–865	CA	*CYP21A1P*	2	658
1109–1116	AC		8	
1109–1116	AC		8	
1109–1126	ACG	*CYP21A1P*	18	658
1116–1126	CG		11	
1425–1688	GTCGGT	*CYP21A1P*	264	377
2697–2764	AAGT	*CYP21A1P*	68	429
3080–3186	CGTCCT	*CYP21A1P*	107	1,433+
3152–3155	TC		4	1,433+

NOTE.—Segregating sites are denoted by their position numbered from the start of the *CYP21A2* coding region on the PGF sequence. The minimum tract length was measured between the first and last nucleotide of allele combination of the conversion. The maximum tract length could be examined only at nonallelic conversions because it was measured from the first nucleotide difference between all *CYP21A2* sequences and all *CYP21A1P* sequences in the 5′-direction of allele combination to the first nucleotide difference between all *CYP21A2* sequences and all *CYP21A1P* sequences in the 3′-direction of allele combination. "+" indicates that the 3′-end of conversion tract was open (the tract was not closed where the sequenced section of the *CYP21A2* gene was finished), and the maximum tract length was larger than the value in the table.

haplotype variants or haplotypic *CYP21A1P* sequences of the utilized external sequences (supplementary table S3, Supplementary Material online). A conversion event was considered as a nonallelic event if its allele combination occurred only in one part (branch) of the network and if it contained a CYP21A1P-specific allele or alleles (if an event meets the definition, it will be nonallelic with high probability). Overall, 11 such conversion events could be observed in the *CYP21A2* network (fig. 4*A* and table 1). From these conversions, seven events could be regarded as nonallelic conversions. Altogether nine conversions occurred only once, but the 1109 A and 1116 C allele changes appeared twice and were located in different parts of the network, indicating that the same conversion could be recurrent, and its allele changes might belong to haplotypes with different intraspecific origins. The minimum tract lengths ranged from 4 to 264 bp (mean: 48.36 bp, median: 12 bp), and the maximum tract lengths as defined according to a previous article (Chen et al. 2007) spanned from 37 to 658 bp (mean: 377.7 bp, median: 403 bp). These values matched with the values in both the human genome and *CYP21A2* gene (Chen et al. 2007).

Discussion

In this study, the intraspecific evolution of the complex, multiallelic, tandem RCCX CNV has been traced by whole-gene *CYP21A2* haplotype variants, which were applied as complex genetic markers. To summarize the theoretical significance, the known genetic phenomena of human RCCX CNV such as the frequent variations in copy number and in the content of *C4* gene and HERV-K CNV (Yang et al. 1999; Blanchong et al. 2000), the transfer of sequence tracts by gene conversion (Tusie-Luna and White 1995; Concolino et al. 2010), and the generation of monomodular variants and trimodular variants by unequal crossover (Tusie-Luna and White 1995; Chung et al. 2002a) have been encompassed by one evolutionary framework. The studied subjects were Hungarians. The genome-wide polymorphisms of Europeans from Hungary deviate negligibly from those of the European reference population (CEU) and other European populations (Semino et al. 2000; Tomory et al. 2007; Heath et al. 2008), and the same applies to the MHC region of these populations (de Bakker et al. 2006; Szilagyi et al. 2010), therefore, the results of this study can be extrapolated for other European populations.

The delineation of intraspecific evolution of RCCX CNV mainly relied on haplotypic information obtained by a set of ASLR-PCRs, which enabled us to haplotype RCCX CNV alleles and structures in many cases that genome-wide platforms for CNV discovery have not yet resolved (Alkan et al. 2011). In contrast to genome-wide, high-throughput methods, ASLR-PCR is only feasible for particular genomic regions because of their specific primers. For a future perspective, our approach can be extended, because the DNA products of ASLR-PCR can serve as templates for high-throughput methods such as next-generation sequencing (Mamanova et al. 2010). Therefore, the advantages (haplotypic and high-throughput) of the two approaches can be merged, eliciting the full-length haplotypic sequences of large, complex, and multiallelic CNVs.

To unravel the complex structures of a duplicated region in diploid subjects, not only must the information of homologous chromosomes be separated from each other but also the duplicated modules on a chromosome. ASLR-PCR can span the module-specific parts of a CNV on a chromosome, enabling the separation of the duplicated modules. Because the centromeric modules of RCCX CNV regularly contain *CYP21A2* genes on both chromosomes, the allele-specific nested *CYP21A2* PCR can inherit the allele specificity only from ASLR-PCR, and its own allele specificity may seem to be unnecessary. Actually, the allele specificity of nested PCRs can prove to be rewarding, as the orthologous modules may comprise different *CYP21* genes, as seen in the case of the chimeric *CYP21* gene. Besides the traditional Southern-based restriction fragment length polymorphism (RFLP) (Yang et al. 1999), many methods such as pulsed-field gel electrophoresis (Chung et al. 2002b), long-range PCR (Kristjansdottir and Steinsson 2004), long-range PCR with *C4* type-specific RFLP (Chung et al. 2002a), ASLR-PCR (Lee et al. 2006), qPCR (Szilagyi et al. 2006; Wu et al. 2007), multiplex ligation-dependent probe amplification (Concolino et al. 2009; Wouters et al. 2009), paralog ratio test (Fernando et al. 2010), and long-range PCR coupled to nested PCR and

sequencing (Tsai et al. 2011) have been developed for the investigation of particular features of RCCX, but the lack of allele specificity and/or module separation often limit their performance. However, ASLR-PCR also has a limitation, because the separation of haplotypic modules cannot be achieved in all diploid combinations of RCCX structures.

The combined molecular and inferred haplotyping approach enabled us to construct a dense genealogical *CYP21A2* haplotype network, shedding new light on the evolution of structure in a complex, multiallelic, tandem CNV. The history of RCCX dates from one primigenial duplication event, which exhibits the sign of prevalent breakpoint microhomology of genome-wide structural variations (Kawaguchi and Klein 1992; Horiuchi et al. 1993; Mills et al. 2011). This duplication occurred in early mammals or the ancestor of mammals, and it must have already existed for at least 90 million years (Hedges 2002). The common genetic features of closely related species, such as the HERV-K CNV in great apes and the 8 bp deletion of exon 3 of *CYP21A1P* in chimpanzee and human, are also considered to originate from one event (Kawaguchi and Klein 1992; Dangel et al. 1995). The CNV status of RCCX has been reliably proven in chimpanzee (Perry et al. 2008b), and in all probability, this status in human and chimpanzee has continuously existed since the common ancestor (Marques-Bonet et al. 2009). Although extensive RCCX polymorphism or *CYP21A2* haplotype data from chimpanzee or other great ape populations has not yet been made available, some signatures have been presented by the haplotype network for the coalescence of RCCX structures. The two RCCX structures of the h58 haplotype variant were not identical to any of the eight indirectly connected RCCX structures, which were located far from each other and represented four different structures. It is hard to imagine that all the eight indirectly connected RCCX structures arose from several different recombination events. Therefore, the h58 haplotype variant and its RCCX structures should not considered as the one and only ancestral haplotype and structure, but rather one of several ancestral RCCX structure-*CYP21A2* haplotype variants still extant.

Diverse and sometimes contradictory selection forces keeping the balance of various RCCX structures may underlie the continuance of the RCCX CNV. For instance, the SB RCCX structure is advantageous with a view to deleterious nonallelic conversion and unequal crossover (Chung et al. 2002a; Lee et al. 2006), but disadvantageous in terms of the retention of the *C4A* gene (Kawaguchi et al. 1992), and the two forces therefore attenuate the effects of each other. The haplotypes harboring the 1688 T allele (h34, h35, and h36) are also intriguing from this viewpoint, because this allele causes CAH (Rumsby et al. 1998), and should be under the effect of purifying selection. Contrary to this, more related haplotypes with the 1668 T allele were observed, implying that these haplotypes have already existed in the long term. The contradiction may be resolved by a presumed positive selection force that compensates the deleterious effects of the allele and may be generated by a genetic feature of the common LASBSB structure of these haplotypes, such as the increased *C4* copy number (Yang et al. 2007). In addition to *C4* copy number, an elevated cortisol response has been actually found in the heterozygous carriers of *CYP21A2* CAH mutations, which may also provide greater fitness (Witchel et al. 1997). Moreover, the elevated cortisol response and the changes of other hormone levels in association with *C4B* copy number have recently been described (Banlaki et al. 2012), hence an advantageous phenotype determined by a subset of particular RCCX structures or *CYP21A2* haplotypes may be realistic.

The cumulative effects of potential selection forces are hard to assess by virtue of the difficulty in the quantitative analysis of selection forces, and the picture is further complicated by the fact that a particular CNV allele is not, of necessity, balanced by selection. If the cumulative effect of selection forces is quite small and negative on a particular CNV allele, then the CNV allele will have existed for a while but not in the long term (Innan and Kondrashov 2010). If this CNV allele is generated (again by unequal crossover) as frequently as it is removed by selection, then it will be continuously present among CNV alleles. We speculate that the recurrent unequal crossover of a particular CNV allele, which was observed in the case of the polyphyletic LA structure, may lead to the repeated generation and removal of the particular CNV allele. Therefore, recurrent unequal crossover events may maintain the polymorphic state of a complex, multiallelic, tandem CNV.

The haplotype network has also provided some further insight into the NAHR events shaping the RCCX CNV. Besides the observed gene conversion events, recurrent unequal crossovers generating the same RCCX structure occurred several times. Although the haplotypic frequencies were not followed in this study, the consequences of unequal crossover were apparent enough for clarifying the relationship of a CNV allele and the SNPs inside the CNV. The recurrent unequal crossover events could result in the same RCCX structures: Therefore, more *CYP21A2* haplotypes with rather different SNP allele patterns belonged to the same RCCX structure. However, only one recurrent gene conversion event was observed in the study, but the same can apply to the effect of gene conversions on the genetic linkage. Even if the CNV alleles were correctly inferred, the strong linkage could be hampered between CNV alleles and their harbored SNP alleles by recurrent NAHR. Therefore, the recurrent NAHR may be one of the causes for the poor tagging of multiallelic CNV alleles by SNPs (Conrad et al. 2010; Campbell et al. 2011).

Supplementary Material

Supplementary figures S1–S6 and tables S1–S8 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Mark Eyre for English proofreading and László Cervenak for critical reading of the manuscript. They also thank Balázs Gereben for his helpful advice, Andrásné Dóczy for help with sequencing, and Anikó Bíró and Anikó Páy for help with the running of the sequencing reactions. This work was supported by the Hungarian Scientific Research Fund (OTKA, CK8842 to G.F. and A.S., and PD100648 to A.P.).

Literature Cited

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet. 12:363–376.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.

Banlaki Z, Doleschall M, Rajczy K, Fust G, Szilagyi A. 2012. Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. Genes Immun. 13:530–535.

Banlaki Z, et al. 2012. ACTH-induced cortisol release is related to the copy number of the *C4B* gene encoding the fourth component of complement in patients with non-functional adrenal incidentaloma. Clin Endocrinol. 76:478–484.

Blanchong CA, et al. 2000. 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 major histocompatibility complex-associated disease. J Exp Med. 191:2183–2196.

Blasko B, et al. 2009. Linkage analysis of the *C4A*/*C4B* copy number variation and polymorphisms of the adjacent steroid 21-hydroxylase gene in a healthy population. Mol Immunol. 46:2623–2629.

Braun L, Schneider PM, Giles CM, Bertrams J, Rittner C. 1990. Null alleles of human complement C4. Evidence for pseudogenes at the *C4A* locus and for gene conversion at the *C4B* locus. J Exp Med. 171: 129–140.

Campbell CD, et al. 2011. Population-genetic properties of differentiated human copy-number polymorphisms. Am J Hum Genet. 88:317–332.

Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet. 8:762–775.

Chung EK, et al. 2002a. Genetic sophistication of human complement components *C4A* and *C4B* and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex. Am J Hum Genet. 71:823–837.

Chung EK, et al. 2002b. Determining the one, two, three, or four long and short loci of human complement C4 in a major histocompatibility complex haplotype encoding C4A or C4B proteins. Am J Hum Genet. 71:810–822.

Collier S, Tassabehji M, Sinnott P, Strachan T. 1993. A de novo pathological point mutation at the 21-hydroxylase locus: implications for gene conversion in the human genome. Nat Genet. 3:260–265.

Concolino P, Mello E, Zuppi C, Capoluongo E. 2010. Molecular diagnosis of congenital adrenal hyperplasia due to 21-hydroxylase deficiency: an update of new *CYP21A2* mutations. Clin Chem Lab Med. 48: 1057–1062.

Concolino P, et al. 2009. Multiplex ligation-dependent probe amplification (MLPA) assay for the detection of *CYP21A2* gene deletions/duplications in congenital adrenal hyperplasia: first technical report. Clin Chim Acta. 402:164–170.

Conrad DF, et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704–712.

Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. 2002. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet. 71:759–776.

Dangel AW, Baker BJ, Mendoza AR, Yu CY. 1995. 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.

Dangel AW, et al. 1994. The dichotomous size variation of human complement *C4* genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates. Immunogenetics 40:425–436.

de Bakker PI, et al. 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 38:1166–1172.

Fernando MM, et al. 2010. Assessment of complement *C4* gene copy number using the paralog ratio test. Hum Mutat. 31:866–874.

Garrick RC, Sunnucks P, Dyer RJ. 2010. Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. BMC Evol Biol. 10:118.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nat Rev Genet. 10:551–564.

Heath SC, et al. 2008. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet. 16:1413–1429.

Hedges SB. 2002. The origin and evolution of model organisms. Nat Rev Genet. 3:838–849.

Higashi Y, Yoshioka H, Yamane M, Gotoh O, Fujii-Kuriyama Y. 1986. Complete nucleotide sequence of two steroid 21-hydroxylase genes tandemly arranged in human chromosome: a pseudogene and a genuine gene. Proc Natl Acad Sci U S A. 83:2841–2845.

Hoehe MR. 2003. Haplotypes and the systematic analysis of genetic variation in genes and genomes. Pharmacogenomics 4:547–570.

Horiuchi Y, Kawaguchi H, Figueroa F, O'HUigin C, Klein J. 1993. Dating the primigenial *C4-CYP21* duplication in primates. Genetics 134: 331–339.

Horton R, et al. 2004. Gene map of the extended human MHC. Nat Rev Genet. 5:889–899.

Horton R, et al. 2008. Variation analysis and gene annotation of eight MHC haplotypes: the MHC haplotype project. Immunogenetics 60: 1–18.

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. Trends Genet. 24:238–245.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 11:97–108.

Jaatinen T, Eholuoto M, Laitinen T, Lokki ML. 2002. Characterization of a de novo conversion in human complement *C4* gene producing a C4B5-like protein. J Immunol. 168:5652–5658.

Kato M, Nakamura Y, Tsunoda T. 2008. An algorithm for inferring complex haplotypes in a region of copy-number variation. Am J Hum Genet. 83:157–169.

Kato M, et al. 2010. Population-genetic nature of copy number variations in the human genome. Hum Mol Genet. 19:761–773.

Kawaguchi H, Klein J. 1992. Organization of *C4* and *CYP21* loci in gorilla and orangutan. Hum Immunol. 33:153–162.

Kawaguchi H, Zaleska-Rutczynska Z, Figueroa F, O'HUigin C, Klein J. 1992. *C4* genes of the chimpanzee, gorilla, and orang-utan: evidence for extensive homogenization. Immunogenetics 35:16–23.

Kidd JM, et al. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell 143: 837–847.

Koppens PF, Hoogenboezem T, Degenhart HJ. 2002a. Carriership of a defective tenascin-X gene in steroid 21-hydroxylase deficiency patients: *TNXB-TNXA* hybrids in apparent large-scale gene conversions. Hum Mol Genet. 11:2581–2590.

Koppens PF, Hoogenboezem T, Degenhart HJ. 2002b. Duplication of the *CYP21A2* gene complicates mutation analysis of steroid 21-hydroxylase deficiency: characteristics of three unusual haplotypes. Hum Genet. 111:405–410.
Koppens PF, Smeets HJ, de Wijs IJ, Degenhart HJ. 2003. Mapping of a de novo unequal crossover causing a deletion of the steroid 21-hydroxylase (*CYP21A2*) gene and a non-functional hybrid tenascin-X (*TNXB*) gene. J Med Genet. 40:e53.
Kristjansdottir H, Steinsson K. 2004. A study of the genetic basis of C4A protein deficiency. Detection of *C4A* gene deletion by long-range PCR and its associated haplotypes. Scand J Rheumatol. 33:417–422.
Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.
Lee HH. 2004. The chimeric *CYP21P*/*CYP21* gene and 21-hydroxylase deficiency. J Hum Genet. 49:65–72.
Lee HH, Chang SF, Tseng YT, Lee YJ. 2006. Identification of the size and antigenic determinants of the human *C4* gene by a polymerase chain-reaction-based amplification method. Anal Biochem. 357:122–127.
Mamanova L, et al. 2010. Target-enrichment strategies for next-generation sequencing. Nat Methods. 7:111–118.
Marques-Bonet T, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457:877–881.
Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G. 1996. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res. 24:4841–4843.
Mills RE, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65.
Nagaraj SH, Gasser RB, Ranganathan S. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform. 8:6–21.
Perry GH, et al. 2008a. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet. 82:685–695.
Perry GH, et al. 2008b. Copy number variation and evolution in humans and chimpanzees. Genome Res. 18:1698–1710.
Posada D, Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol. 16:37–45.
Redon R, et al. 2006. Global variation in copy number in the human genome. Nature 444:444–454.
Rumsby G, Avey CJ, Conway GS, Honour JW. 1998. Genotype-phenotype analysis in late onset 21-hydroxylase deficiency in comparison to the classical forms. Clin Endocrinol. 48:707–711.
Semino O, et al. 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290:1155–1159.
Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet. 18:74–82.
Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 73:1162–1169.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68: 978–989.
Stranger BE, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853.
Su SY, et al. 2010. Inferring combined CNV/SNP haplotypes from genotype data. Bioinformatics 26:1437–1445.
Szilagyi A, Doleschall M, Fust G. 2010. Complement genes in the central region of the MHC. In: Mehra NK, editor. The HLA complex in biology and medicine: a resource book. New Delhi (India): Jaypee Brothers Medical Publishers. p. 135–158.
Szilagyi A, et al. 2006. Real-time PCR quantification of human complement *C4A* and *C4B* genes. BMC Genet. 7:1.
Szilagyi A, et al. 2010. Frequent occurrence of conserved extended haplotypes (CEHs) in two Caucasian populations. Mol Immunol. 47: 1899–1904.
Tassabehji M, et al. 1994. Identification of a novel family of human endogenous retroviruses and characterization of one family member, HERV-K(*C4*), located in the complement *C4* gene cluster. Nucleic Acids Res. 22:5211–5217.
Teshima KM, Innan H. 2012. The coalescent with selection on copy number variants. Genetics 190:1077–1086.
Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK. 2000. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet. 67:518–522.
Tomory G, et al. 2007. Comparison of maternal lineage and biogeographic analyses of ancient and modern Hungarian populations. Am J Phys Anthropol. 134:354–368.
Tsai LP, Cheng CF, Chuang SH, Lee HH. 2011. Analysis of the *CYP21A1P* pseudogene: indication of mutational diversity and *CYP21A2*-like and duplicated *CYP21A2* genes. Anal Biochem. 413:133–141.
Tusie-Luna MT, White PC. 1995. Gene conversions and unequal crossovers between *CYP21* (steroid 21-hydroxylase gene) and *CYP21P* involve different mechanisms. Proc Natl Acad Sci U S A. 92: 10796–10800.
Tuzun E, et al. 2005. Fine-scale structural variation of the human genome. Nat Genet. 37:727–732.
Vatay A, et al. 2003. Relationship between complement components *C4A* and *C4B* diversities and two *TNFA* promoter polymorphisms in two healthy Caucasian populations. Hum Immunol. 64:543–552.
White PC, Speiser PW. 2000. Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. Endocr Rev. 21:245–291.
Witchel SF, Lee PA, Suda-Hartman M, Trucco M, Hoffman EP. 1997. Evidence for a heterozygote advantage in congenital adrenal hyperplasia due to 21-hydroxylase deficiency. J Clin Endocrinol Metab. 82: 2097–2101.
Wouters D, et al. 2009. High-throughput analysis of the *C4* polymorphism by a combination of MLPA and isotype-specific ELISA's. Mol Immunol. 46:592–600.
Wu YL, et al. 2007. Sensitive and specific real-time polymerase chain reaction 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.
Yang Y, et al. 2007. Gene copy-number variation and associated polymorphisms of complement component *C4* in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet. 80:1037–1054.
Yang Z, Mendoza AR, Welch TR, Zipf WB, Yu CY. 1999. Modular variations of the human major histocompatibility complex class III genes for serine/threonine kinase *RP*, complement component *C4*, steroid 21-hydroxylase *CYP21*, and tenascin *TNX* (the RCCX module). A mechanism for gene deletions and disease associations. J Biol Chem. 274:12147–12156.
Yu CY, et al. 2003. Dancing with complement C4 and the RP-C4-CYP21-TNX (RCCX) modules of the major histocompatibility complex. Prog Nucleic Acid Res Mol Biol. 75:217–292.
Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 10:451–481.

Associate editor: B. Venkatesh