ARTICLE

A unique haplotype of RCCX copy number variation: from the clinics of congenital adrenal hyperplasia to evolutionary genetics

Márton Doleschall[*,1,2,3], Andrea Luczay[4], Klára Koncz[5], Kinga Hadzsiev[6], Éva Erhardt[7], Ágnes Szilágyi[3], Zoltán Doleschall[8], Krisztina Németh[4], Dóra Török[4], Zoltán Prohászka[3], Balázs Gereben[9], György Fekete[4], Edit Gláz[5], Péter Igaz[5], Márta Korbonits[2], Miklós Tóth[5], Károly Rácz[1,5] and Attila Patócs[1,10]

There is a difficulty in the molecular diagnosis of congenital adrenal hyperplasia (CAH) due to the c.955C > T (p.(Q319*), formerly Q318X, rs7755898) variant of the *CYP21A2* gene. Therefore, a systematic assessment of the genetic and evolutionary relationships between c.955C > T, *CYP21A2* haplotypes and the RCCX copy number variation (CNV) structures, which harbor *CYP21A2*, was performed. In total, 389 unrelated Hungarian individuals with European ancestry (164 healthy subjects, 125 patients with non-functioning adrenal incidentaloma and 100 patients with classical CAH) as well as 34 adrenocortical tumor specimens were studied using a set of experimental and bioinformatic methods. A unique, moderately frequent (2%) haplotypic RCCX CNV structure with three repeated segments, abbreviated to LBSASB, harboring a *CYP21A2* with a c.955C > T variant in the 3′-segment, and a second *CYP21A2* with a specific c.*12C > T (rs150697472) variant in the middle segment occurred in all c.955C > T carriers with normal steroid levels. The second *CYP21A2* was free of CAH-causing mutations and produced mRNA in the adrenal gland, confirming its functionality and ability to rescue the carriers from CAH. Neither LBSASB nor c.*12C > T occurred in classical CAH patients. However, CAH-causing *CYP21A2* haplotypes with c.955C > T could be derived from the 3′-segment of LBSASB after the loss of functional *CYP21A2* from the middle segment. The c.*12C > T indicated a functional *CYP21A2* and could distinguish between non-pathogenic and pathogenic genomic contexts of the c.955C > T variant in the studied European population. Therefore, c.*12C > T may be suitable as a marker to avoid this genetic confound and improve the diagnosis of CAH.

INTRODUCTION

RCCX copy number variation (CNV) is a complex (it contains another CNV), multiallelic (there are more than two frequent CNV alleles in a population) and tandem (CNV segments are next to each other) CNV.[1–3] Two full-length genes, complement component 4 (*C4*) and steroid 21-hydroxylase (*CYP21*), are located in RCCX CNV. *C4* has two types with slightly different immune functions,[4] *C4A* and *C4B*, and it contains an additional CNV, HERV-K(C4) CNV, in its intron 9 (Figure 1), derived from an ancient insertion of human endogenous retrovirus. RCCX CNV in one chromosome is referred as a haplotypic structure, and is traditionally described by the copy number of the segment, and, per segment, by the types of *C4* and the alleles of HERV-K(C4) CNV (see the explanation of more specialized genetic terms in Table 1). A haplotypic RCCX CNV structure usually contains one functional *CYP21* gene (*CYP21A2*) in the 3′-segment, and zero, one or two disabled pseudogenes (*CYP21A1P*) in the segments towards the 5′-direction (Figure 1). *CYP21A2* expresses the steroid 21-hydroxylase enzyme (cytochrome *P*450c21) uniquely in adrenal cortex, which is one of the enzymes responsible for the biosynthesis of the two principal steroid hormones, aldosterone and cortisol.[5] Aldosterone regulates blood pressure by acting on kidney functions, and cortisol suppresses immune functions, responds to stress and takes part in the intermediary metabolism of different tissues. Two chromosomal copies of disabled *CYP21A2* causes congenital adrenal hyperplasia (CAH; we use this broader, but more prevalent, term instead of 21-hydroxylase deficiency), an autosomal recessive disorder accompanied by a partial or complete lack of aldosterone and cortisol.[6] Special genetic rearrangements generally featured by multi-allelic CNVs, such as non-allelic gene conversion and unequal crossover,[7] are responsible for the vast majority of *CYP21A2* mutations leading to CAH;[8] therefore, the complex genetic structure of RCCX CNV indirectly manifests in CAH.

Difficulties in the molecular diagnosis of CAH due to the complicated genetic nature of RCCX CNV have been recognized for a long time,[9] and the best-known case for this is an LSS haplotypic RCCX CNV structure with three segments harboring two copies of

[1]Molecular Medicine Research Group, Hungarian Academy of Sciences and Semmelweis University, Budapest, Hungary; [2]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK; [3]3rd Department of Internal Medicine, Semmelweis University, Budapest, Hungary; [4]2nd Department of Pediatrics, Semmelweis University, Budapest, Hungary; [5]2nd Department of Internal Medicine, Semmelweis University, Budapest, Hungary; [6]Department of Medical Genetics, University of Pecs, Pecs, Hungary; [7]Department of Pediatrics, University of Pecs, Pecs, Hungary; [8]Department of Pathogenetics, National Institute of Oncology, Budapest, Hungary; [9]Department of Endocrine Neurobiology, Institute of Experimental Medicine, Hungarian Academy of Sciences, Budapest, Hungary; [10]'Lendület' Hereditary Endocrine Tumors Research Group, Hungarian Academy of Sciences and Semmelweis University, Budapest, Hungary
*Correspondence: Dr M Doleschall, Molecular Medicine Research Group, Hungarian Academy of Sciences and Semmelweis University, 46 Szentkiralyi Street, Budapest 1088, Hungary. Tel: +36 305364111; Fax: +36 1 2660816; E-mail: doleschall.marton@med.semmelweis-univ.hu
Received 10 June 2016; revised 8 February 2017; accepted 14 February 2017; published online 12 April 2017
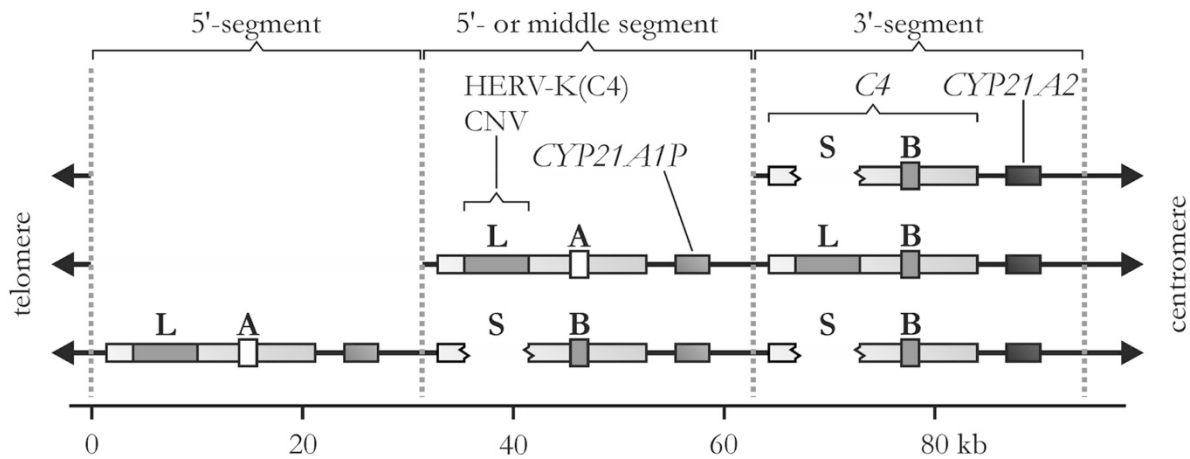
Figure 1 Scaled representation of the organization of human RCCX CNV depicted by RCCX structures with one, two and three segments. Each segment (repeat) is abbreviated with two letters, the first represents the alleles of the HERV-K(C4) CNV (**L** – long (insertion) allele or **S** – short (deletion) allele), and the second symbolizes the types of *C4* gene (**A** or **B**). The duplications of these two letters indicate RCCX structures with more than one segment. Dotted lines indicate the segment boundaries. The variable region of the RCCX with two segments contains two pairs of full-length genes, complement component 4 (*C4A* and *C4B*) and steroid 21-hydroxylase (*CYP21A1P* and *CYP21A2*). Other two gene pairs partially residing in RCCX CNV, *TNXA-TNXB* and *STK19* and *STK19P*, are not illustrated.

Table 1 Glossary of genetic abbreviations and key terms used in the text of the current study

CNV	Copy number variation	A form of structural variations where a longer DNA segment occurs with different repeat numbers in the different copies of the same homolog chromosome; thus, the copy number of the DNA segment will vary among chromosomes and individuals.
	CNV allele	A particular number of repeated segments in one chromosome.
	CYP21A2 haplotype	A particular combination of the alleles of all variations (SNPs and mutations) in *CYP21A2* of one chromosome. It is identical to a Mendelian allele of the full-length *CYP21A2*.
	Haplotype	A particular combination of the alleles of adjacent genetic variants in one chromosome, which are not independent of each other, and are frequently inherited together.
	Haplotypic RCCX CNV structure (its short form is RCCX structure)	A complex variant in one chromosome which is traditionally described by the copy number of the repeated segment of RCCX CNV (CNV allele), and, per segment, by the alleles of HERV-K (C4) CNV and the type of *C4* gene.
HERV-K(C4)	Human endogenous retrovirus-K	An ancient genomic remnant of a human endogenous retrovirus-K in the intron 9 of human *C4* genes.
HERV-K(C4) CNV		A CNV with a deletion or insertion CNV allele (null or one segment) in a *C4* gene of one chromosome derived from a human endogenous retrovirus-K.
LASA	LASA RCCX structure	A RCCX structure with two segments containing a *C4A* with HERV-K(C4) insertion in the 5′-segment and a *C4A* without HERV-K(C4) in the 3′-segment.
LBSASB	LBSASB RCCX structure	A RCCX structure with three segments containing a *C4B* with HERV-K(C4) insertion in the 5′-segment, a *C4A* without HERV-K(C4) in the middle segment and a *C4A* without HERV-K(C4) in the 3′-segment.
LD	Linkage disequilibrium	A statistical term where two or more alleles of adjacent variants tend to be associated with each other. It is a statistical representation of the haplotypes within a particular population.
LSS	LSS RCCX structure	A RCCX structure with three segments, where the types of *C4* are not determined, contains a HERV-K(C4) insertion only in the 5′-segment.
RCCX CNV		A CNV of chromosome 6, which was named after the traditional name of its gene contents.
SB	SB RCCX structure	A RCCX structure with one segment containing a *C4B* without HERV-K(C4) insertion in its single segment.

Abbreviations: CNV, copy number variation; SNP, single-nucleotide polymorphism.

CYP21A2 and the c.955C > T (p.(Q319*), rs7755898) variant. (For numbering the genetic variants, we use the most recent human reference gene and the corresponding mRNA sequence (RefSeqGene NG_007941.3 and NM_000500.7), which contains the insertion (CTG) allele of the c.28_30delCTG (rs61338903) variant. Therefore, our numbering differs from the studies using a *CYP21A2* sequence (eg, former RefSeqGene NG_007941.2) with the deletion allele of c.28_30delCTG. The p.(Q319*) variant name is used with parentheses at the protein level according to the nomenclature recommendation of

Human Genome Variation Society (http://varnomen.hgvs.org/), which considers a protein variant without experimental evidence at RNA or protein level as a predicted one; however, the segregation patterns of classical (severe form) CAH families[8,10–12] (especially patients from consanguineous families with family history having homozygous c.955T genotypes without any other CAH mutations in exons and exon/intron boundaries and without reported duplicated *CYP21A2*[13]) have provided clinical evidences for the pathogenicity since the initial report of c.955C > T.[14]) This unique RCCX structure was found for

Table 2 Published literature on the topic of the c.955C>T variant and two CYP21A2 gene copies

First author, year	Healthy	CAH	Experimental method	Haplotypic RCCX structure	Genetic variants CYP21A2	c.955C>T	Other CAH mutation	Haplotyping Familial	Molecular	Population Nationality	Allele frequency
Wedell, 1994	—	✓	AS-PCR, RFLP-SD, sequencing	LSS	2	✓	c.293−13G	✓	—	Swedish?	—[a]
Koppens, 2000	✓	SW	AS-PCR, (PFGE), RFLP-SD	Three segments	2	✓	c.293−13G	✓	✓	Dutch	1/150 (0.67%)
Koppens, 2002	✓	—	AS-PCR, PFGE, RFLP-SD	LSS	2	✓	—	—	✓	Dutch	2/282 (0.71%)
Baumgartner-Parzer, 2003	✓	—	AS-PCR, RFLP-SD, sequencing	—	2	✓	—	✓	—	Austrian?	—[a]
Baumgartner-Parzer, 2006[b]	✓	—	AS-PCR, RFLP-SD, sequencing	Three segments	2	✓	—	✓	—	Austrian?	—[a]
Parajes, 2008	✓	—	AS-PCR, qPCR, sequencing	—	2	✓	—	—	—	Spanish?	8/288 (2.77%)
Kleinle, 2009	✓	✓	AS-PCR, MLPA, RFLP-SD, sequencing	Three segments[c]	2	✓	—	✓	—	Austrian?	—[a]
Kharrat, 2011	✓	—	MLPA, qPCR, RFLP-PCR	—	2	✓	—	—	—	Tunisian	18/272 (6.62%)
Tsai, 2011[d]	✓	—	ACRS, AS-PCR, sequencing	Three segments[c]	2	?	Yes	—	—	Taiwanese	5/200 (2.50%)
Concolino, 2013	✓	—	AS-PCR, MLPA, RFLP-PCR, sequencing	Three segments[c]	2	✓	—	—	—	Italian	18/750 (2.40%)
Lekarev, 2013	—	✓	AS-PCR, sequencing	—	2	✓	—	✓	—	Mixed European	—[a]
Leccese, 2014	✓	—	AS-PCR, MLPA, RFLP-PCR, sequencing	Three segments	2	✓	—	✓	—	Italian?	—[a]

Abbreviations: ACRS, amplification-created restriction site; AS, allele-specific; CAH, congenital adrenal hyperplasia; LSS, RCCX structure with 5′-long, middle short, and 3′-short C4 genes in its three segments; MLPA, multiplex ligation-dependent probe amplification; PCR, polymerase chain reaction; PFGE, pulsed-field gel electrophoresis; qPCR, quantitative polymerase chain reaction; RFLP, restriction fragment length polymorphism; SD, Southern blot; SW, salt wasting. RefSeqGene NG_007941.3 and NM_000500.7 mRNA sequence were used for the nomenclature of variants.
[a]Case study.
[b]The examined family was the same as in Baumgartner-Parzer et al.[37]
[c]It was probably assumed based on previous studies, but the experimental data on RCCX structure were not documented in the study.
[d]The primary objective of the study was not the examination of RCCX structure with duplicated CYP21A2 and c.955C>T; therefore, neither variants in CYP21A2 locus (next to TNXB) nor RCCX structure were investigated.

the first time in three patients with CAH, and the c.293−13C>R (rs6467) CAH mutation accompanied the c.955C>T variant.[10] The unique RCCX structure with c.955C>T and c.293−13C>R has been confirmed in a Dutch CAH population,[15] but the same RCCX structure without the c.293−13C>R has been observed in the healthy subjects of a Dutch population as well.[16] After the initial findings, several studies on the topic have been published over time. However, healthy and CAH populations have rarely been examined simultaneously, and the determination of underlying RCCX structures has often been incomplete or has not been documented (Table 2). Nevertheless, new information on the unique haplotypic RCCX structure has been accumulated. In a healthy Spanish population, two CYP21A2 gene copies in one chromosome are associated with three genetic variants in complete linkage disequilibrium (LD): c.955C>T, c.293−79G>A (rs114414746) and c.*12C>T (rs150697472) variants. They occur with a moderate allele frequency (2.77%), and c.955C>T is located in the 3′-segment of CYP21A2, whereas a wild-type CYP21A2 is in the middle segment.[17] An incomplete association among c.955C>T, c.293−79G>A and c.*12C>T has also been found in a later study,[18] but, unfortunately, the diagnoses of CAH in the studied individuals have not been clearly documented, and therefore it cannot be known if the presence or absence of the association has been assigned to patients with CAH or healthy subjects.

In the current study, a healthy and a CAH population with the same European ancestry were used to assess the occurrences and association of the haplotypic RCCX structure with three segments, two CYP21A2 in the same RCCX structure, c.955C>T, c.293−79G>A and c.*12C>T, and to offer a solution for the diagnostic problem. A recent set of experimental and bioinformatic methods[1,19] were applied to decipher the exact haplotypes of RCCX structure and CYP21A2, and to translate the evolutionary origin of the studied genetic variants into clinically useful information. Furthermore, patients with non-functioning adrenal incidentaloma (NFAI), which is a hormonally inactive adrenal mass unintentionally discovered by medical imaging,[20] and the adrenocortical tumor specimens of NFAI patients, were used to compare blood hormone levels and to examine the mRNA levels. The rationale for studying NFAI patients was: (i) the baseline hormone profiles of NFAI subjects do not differ from those of healthy subjects, (ii) the NFAI subjects had a broader hormone profile and (iii) their tumor specimens are available.

MATERIALS AND METHODS
Subjects and clinical data
For the subjects with the c.955C>T variant, 96 healthy subjects (50% women) from a family study,[21] 68 healthy subjects (65.2% women) and 125 patients (76.0% women) with NFAI from a recent study,[22] and 100 patients (58.0% women) with classical CAH due to 21-hydroxylase deficiency from the Second Department of Internal Medicine, Semmelweis University (Budapest, Hungary) and the Second Department of Pediatrics, Semmelweis University (Budapest, Hungary) were screened. In addition, one patient with supposed classical CAH was referred to us by the Department of Pediatrics, University of Pecs (Pecs, Hungary) for reexamination of the original genetic diagnosis. All subjects were genetically unrelated and Hungarian with European ancestry. The exclusion criteria for NFAI patients were described recently.[22] The diagnosis was confirmed in all CAH patients by hormonal and genetic testing (except for the one who was reexamined) as described.[23,24] Furthermore, 34 adrenocortical tumor specimens of NFAI patients were obtained as described,[25] and screened for the c.955C>T variant. All subjects gave written informed consent. The study protocol was approved by the National Scientific and Ethical Committee, Medical Research Council of Hungary (TUKEB, ETT), and was executed according to the Declaration of Helsinki principles.
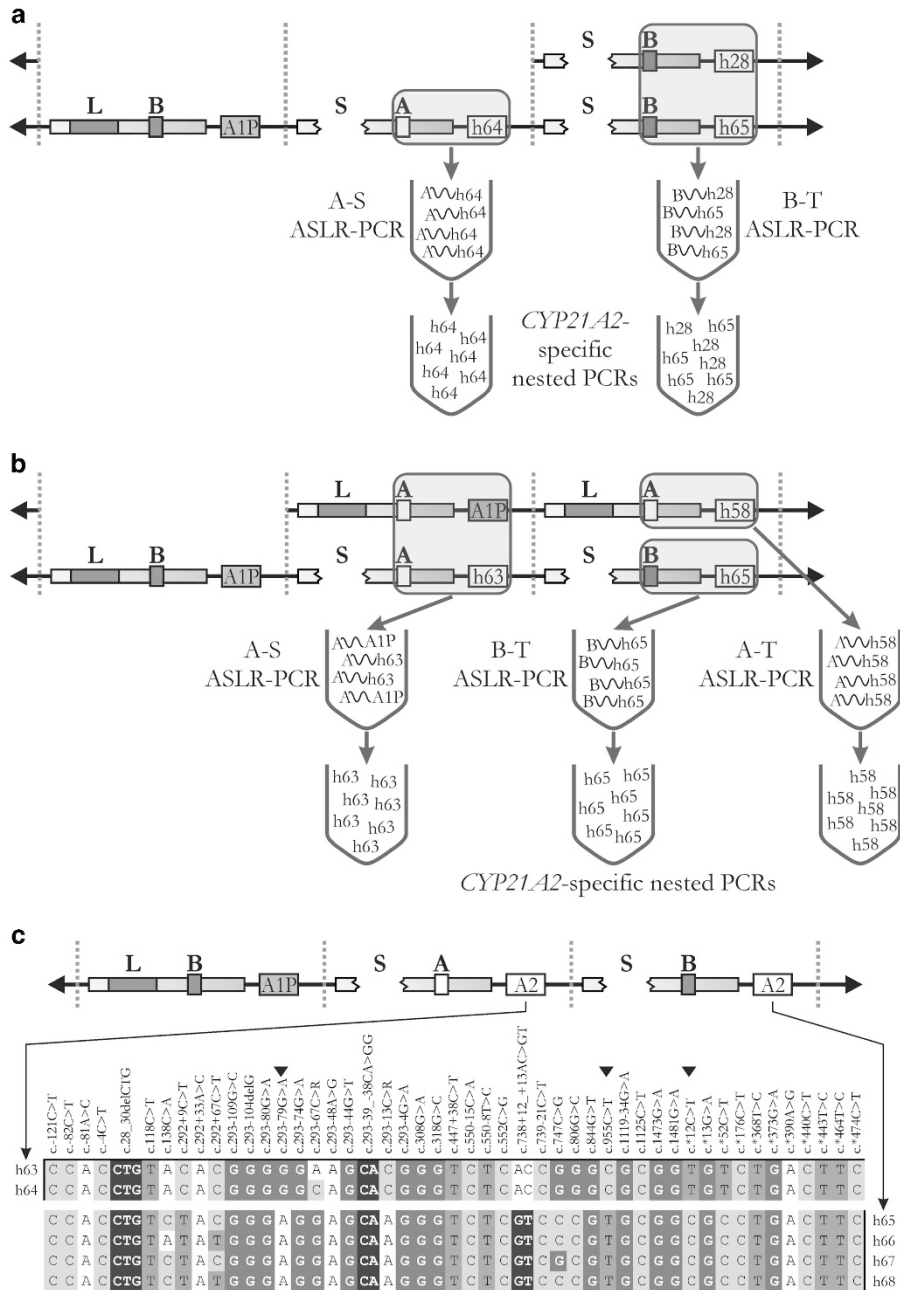
Figure 2 The experimental determination (molecular haplotyping) of haplotypic RCCX CNV structures, *CYP21A2* haplotypes and the resultant *CYP21A2* haplotypes of LBSASB RCCX structure. A-S ASLR-PCR indicates allele-specific long-range PCR only from the *C4A* gene in the 5′-segment or the middle segment, A-T ASLR-PCR indicates ASLR-PCR only from the *C4A* gene in the 3′-segment, and B-T ASLR-PCR indicates ASLR-PCR only from the *C4B* gene in the 3′-segment. Each haplotypic RCCX segment is abbreviated with two letters, the first represents the alleles of HERV-K(C4) CNV (**L** – long (insertion) allele or **S** – short (deletion) allele), and the second symbolizes the types of *C4* gene (**A** or **B**). The multiplication of the two letters in a structure indicates the segment number. Other two gene pairs partially residing in RCCX CNV, TNXA-TNXB and STK19 and STK19P, are not illustrated. (**a**) The haplotypic RCCX structures and the *CYP21A2* haplotype in the middle segment can be determined from the depicted diploid combination of the haplotypic RCCX structures, but the *CYP21A2* haplotype in the 3′-segment cannot. (**b**) All *CYP21A2* haplotypes can be determined from the depicted diploid combination of the haplotypic RCCX structures, because ASLR-PCR can separate the two 3′-segments from the different haploid copies of chromosome 6. (**c**) *CYP21A2* haplotypes of the LBSASB RCCX structure. Only variants with at least moderate frequencies according to a recent study[19] are presented. Specific genetic variants of these haplotypes are indicated by black triangles.

Genetic analyses of RCCX CNV and *CYP21* genes

The c.955T allele, full haplotypic RCCX CNV structures, and full-length *CYP21A1P* and *CYP21A2* haplotypes were determined as described previously.[1,19,22–24] Briefly, a combined molecular and inferred haplotyping approach was used with two stages. *C4* type-specific quantitative PCRs

(qPCRs), *CYP21* gene-specific qPCR, HERV-K(C4) CNV allele-specific qPCR, a set of *CYP21* gene-specific nested polymerase chain reactions (PCRs) resting on a set of allele-specific long-range (ASLR) PCRs, and the Sanger sequencing of the nested PCR products were carried out in the first, experimental stage. One smaller modification was set up using a new *CYP21A1P*-specific nested

PCR primer for one individual, CYP21A1P_2R (Supplementary Table 1), because a mutation of *CYP21A1P* decreased the binding efficacy of the original primer, CYP21A1P_R.[1] The experimentally determined new *CYP21A1P* and *CYP21A2* haplotypes are available in GenBank (www.ncbi.nlm.nih.gov/genbank/) under KU302771–KU302776 IDs, and variants (IDs in Supplementary Table 2) in dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/). The above-mentioned haplotypic genetic features of RCCX CNV and *CYP21* genes, which could not be determined in some subjects, were inferred by bioinformatic haplotype reconstruction in the second stage. To map the origin of haplotypic RCCX structures and *CYP21A2* haplotypes, a genealogical haplotype network was constructed using the new and published *CYP21A2* haplotypes (Supplementary Table 2 and 3) as described.[1]

Functional studies on CYP21A2

Total RNA was isolated from adrenocortical tumor specimens by RNeasy Lipid Tissue Kit (Qiagen, Hilden, Germany), cDNA was synthesized by ProtoScript II Kit (New England Biolabs, Ipswich, MA, USA) and *CYP21A2* mRNA level was determined by qPCR using a combined approach as described.[19,26] Briefly, the forward primer was designed on the boundary of exons 1 and 2 of the *CYP21* genes, the reverse primer was located on c.332_339delGAGACTAC variant distinguishing *CYP21A1P* and *CYP21A2* and the Taqman minor groove binder probe (Applied Biosystems, Foster City, CA, USA) on the boundary of exons 2 and 3 was labeled with the fluorescent dye 6-FAM (Supplementary Table 1). Reactions were carried out in a 7500 Fast Real-Time PCR equipment by TaqMan Fast Advanced PCR master mix (Applied Biosystems). The human *OAZ1* gene was used as an internal control.[27] Full-length *CYP21* cDNA was amplified and cloned with primers designed for mRNA according to a previous study.[28] Briefly, *CYP21A2* cDNA was amplified by Phusion DNA polymerase (New England Biolabs), the PCR product was directionally cloned into a pTZ57R vector and subjected to Sanger sequencing. The methods of hormone measurements for NFAI patients and the software for statistics were recently described.[22]

RESULTS

A unique haplotypic RCCX CNV structure with three segments, LBSASB, harboring two CYP21A2 gene copies and the c.955C>T variant

The c.955C>T variant was detected in 7 healthy subjects out of 164 (allele frequency is 2.1%), in 5 NFAI patients out of 125 (2.0%), and in 6 classical CAH patients out of 100 patients (3.0%). To find out which haplotypic RCCX CNV structure and *CYP21A2* haplotype harbor the c.955T allele, a set of molecular haplotyping methods[1] were used for the experimental dissection in the first stage. Then, the haplotypic information based on molecular haplotyping from the first stage and the genotypic information of all studied healthy subjects and patients with NFAI were used for bioinformatic haplotype reconstruction in the second stage (see detailed description of how the experimental and bioinformatic methods work elsewhere[1]). The first, experimental stage revealed the following genetic features: (i) All the healthy subjects and all the NFAI patients who had the c.955T allele (12 individuals) carried a haplotypic RCCX CNV structure on one chromosome harboring two *CYP21A2* gene copies (they had three *CYP21A2* gene copies as a whole). (ii) The extra *CYP21A2* gene copy in the 5′-segment or the middle segment of RCCX CNV was always related to a *C4A* gene in the 5′-direction, and the *CYP21A2* haplotype of these segments could be determined in all 13 individuals (Figure 2a). Accordingly, two almost identical and unique (not observed in former studies[1,19]) *CYP21A2* haplotypes were located in the 5′-segment or the middle segment, the h63 (in 1 individual) and h64 (in 11 individuals). (iii) From these 11 individuals with the h64 haplotype, the middle segment of an LBSASB RCCX structure harbored h64 in two individuals. (iv) Two individuals with c.955C>T carried a suitable RCCX structure combination for the *CYP21A2* haplotype determination in the 3′-segment (Figure 2b), and they had a unique, undescribed h65 *CYP21A2* haplotype there. In addition, both *CYP21A2* haplotypes in one haplotypic RCCX structure were also determined in a 1-year-old CAH patient with c.955C>T, whose DNA sample was received to reexamine the original genetic diagnosis based on parental genotypes, because the required dose of steroid substitution was unchanged in spite of weight gain. In this putative CAH patient, the 3′-segment also harbored a h65 haplotype, and a h64 haplotype was carried as a second *CYP21A2* copy.

The second, bioinformatic stage resolved a unique LBSASB RCCX structure harboring very similar h63 or h64 *CYP21A2* haplotypes in the middle segment and very similar h65, h66, h67 or h68 haplotypes in the 3′-segment above a 0.99 confidence probability threshold in all 12 individuals and the reexamined CAH patient (Figure 2c). A specific allele, c.*12T, occurred only in h63 and h64 haplotypes of the middle segment and not in any other haplotypes of healthy subjects and patients with NFAI, and a specific allele, c.293−79A, occurred only in h65–h68 haplotypes of the 3′-segment (Supplementary Table 2).

Table 3 Haplotypes and genotypes of classical CAH patients with the c.955C>T mutation

Patient ID	RCCX structure	5′-Segment	3′-Segment	CYP21 genes Classical CAH mutation in 3′-segment	Haplotype	c.293−79A	c.*12T
Patient 1	LALB	A1P	A2	c.332_339delGAGACTAC	Det	N	N
	LALB	A1P	5′-A2 and 3′-A1P (intron 6)	c.955C>T	h69	N	N
Patient 2	LASB	A1P	A2	c.293−13C>A,G	—	Y (het)	N
	LASB			c.955C>T			
Patient 3	LASB	A1P	A2	c.293−13C>A,G	—	N	N
	LA?B			c.955C>T			
Patient 4	LA	—	A1P	—	Det	N	N
	LASB	A1P	A2	c.955C>T	h70	N	N
Patient 5	?BSA	A1P	A2	c.955C>T	—	Y (het)	N
	??LA		5′-A2 and 3′-A1P (intron 4?)	c.955C>T			
Patient 6	LASA	A1P	A2	c.955C>T	h65	Y	N
	??LB	?	A2	c.332_339delGAGACTAC	Det	N	N

Abbreviations: Det, determined haplotypes; het, in heterozygous form; N, it does not contain the particular allele; Y, it contains the particular allele.
Each haplotypic RCCX segment is abbreviated with two letters, the first represents the alleles of HERV-K(C4) 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 segment number. Question mark (?) indicates uncertainty in the particular genetic feature. RefSeqGene NG_007941.3 and NM_000500.7 mRNA sequence were used for the nomenclature of variants.

Consequently, these two variants and c.955C>T have the same allele frequency as the LBSASB RCCX structure harboring them, they were in complete LD ($r^2 = 1$), and, furthermore, there was no significant difference between the allele frequencies in healthy subjects and patients with NFAI (Fisher's exact test, $P = 1.0000$).

CYP21A2 haplotypes of classical CAH patients with c.955C>T

In total, seven CYP21A2 haplotypes with the c.955T allele were found in six patients out of the studied 100 patients with classical CAH (Table 3). The c.*12T allele did not occur in these six CAH patients, but the c.293−79A allele was observed in three patients. From the seven haplotypes with the c.955T allele, three CYP21A2 haplotypes were experimentally determined; a h65 haplotype harbored by an LASA RCCX CNV structure, a haplotype (h69), which had an extended CYP21A1P sequence approximately from intron 6 to the 3′ end of the gene including c.955T, and a haplotype (h70), which did not harbor any alleles characterizing CYP21A1P apart from c.955T. Bioinformatic haplotype reconstruction can accurately resolve haplotypes when they recurrently occur in a studied population, but CYP21A2 haplotypes with CAH variants not determined by experiments did not satisfy this requirement. However, it was highly probable that a CYP21A2 haplotype (in patient 2) was identical with the h67 haplotype of LBSASB RCCX structure, and a haplotype (in patient 5) also had an extended CYP21A1P sequence at the 3′ end of the gene. Furthermore, the CYP21A2 haplotypes of CAH patients were not assigned to LBSASB or a chromosome harboring a second copy of CYP21A2, although the presence of the c.293−79A allele in classical CAH patients implied that the CYP21A2 haplotypes harboring c.293−79A had derived from the CYP21A2 in the 3′-segment of LBSASB.

Origin of CYP21A2 haplotypes with c.955C>T in healthy and CAH individuals

To assess the evolutionary origin of CYP21A2 haplotypes, a genealogical (evolutionary) haplotype network was applied as described recently.[1] Briefly, haplotypes that are more similar to each other are located close to each other in the haplotype network, and are considered to be closer relatives.[29]

The CYP21A2 haplotypes (h63 and h64) in the middle segment of the LBSASB RCCX structure clustered together and were far from the group of the haplotypes (h65–h68) in the 3′-segment of LBSASB on the haplotype network (Figure 3a). This supports the idea that the haplotypes in each group were closely linked, but a distant relationship was observed between the two groups. The distant relationship implies that an unequal crossover event between two distantly related RCCX structures generated the LBSASB structure. However, the SB and LASB RCCX structure groups next to the RCCX structure assigned to h63–h64 and h65–h68 groups do not suggest a one-step mechanism for this structure generation. Nevertheless, a potential multistep generation with unequal crossover or an alternative explanation of multiple non-allelic gene conversion events equally indicates a relatively ancient origin of the LBSASB structure. Not only the multistep mechanism but the presence of LBSASB-specific variants also confirms this.

One experimentally determined CYP21A2 haplotype with the c.955T allele from classical CAH patients was in the haplotype group of the 3′-segment of the LBSASB, indicating a common descent. The LASA RCCX structure harboring the LBSASB-identical h65 haplotype was generated by simple (equal) or unequal crossover[1,30] between LBSASB and existing RCCX structures in the studied population.[1,21] In the case of (equal) crossover, the breakpoint would have been after the C4A in the 3′-segments of the LASA structure and between the

C4B and CYP21A2 in the 3′-segment of the LBSASB harboring the h65 haplotype (Figure 3b), while it would have been after the C4A in the middle segments of the LASALB structure and in the same place of the LBSASB structure in the case of unequal crossover. In both cases, the functional CYP21A2 in the middle segment of LBSASB vanished from the emerged LASA RCCX structure.

Two other experimentally determined CAH-causing CYP21A2 haplotypes with the c.955T allele (h69 and h70) were located separately from either haplotypes in LBSASB and from each other, supporting their independent origins from the LBSASB RCCX structure (probably by non-allelic gene conversions from the CYP21A1P of the same RCCX structure where c.955T occurs). The connections of these two CYP21A2 haplotypes to the stem of the haplotype network were confirmed by the harboring RCCX structures of the connected haplotypes (h69 harbored by LALB RCCX structure and h70 harbored by LASB RCCX structure).

Gene expression and function of CYP21A2 in the carriers of LBSASB RCCX structure with the c.955C>T variant

LBSASB RCCX structure with the c.955T allele was observed in 1 out of 34 adrenocortical tumor specimens of patients with NFAI. The CYP21A2 mRNA level of this c.955C>T carrier patient did not significantly differ (t-test, $P = 0.3312$) from those of non-carrier patients (Figure 4a). Furthermore, this c.955C>T carrier patient had an h64 CYP21A2 haplotype in the middle segment of LBSASB, an h65 haplotype in the 3′-segment, and an h28 haplotype in the SB RCCX structure of the other chromosome (the exact RCCX structures of this patient is actually shown in Figure 2a). The transcripts of these three CYP21A2 haplotypes harbored distinctive alleles on c.*12C>T and c.*52C>T variants; CC alleles for h65, CT alleles for h28 and TT alleles for h64 (Figure 4b). To ascertain the frequencies of the different transcripts derived from these three different CYP21A2 haplotypes, cDNA was generated from total mRNA, and the CYP21A2 transcripts were cloned and sequenced in five independent experiments. The h64 transcript was present, indicating that CYP21A2 in the middle segment was transcriptionally active. The h65 transcript was not detected, whereas the frequency of the h28 transcript from the other chromosome was significantly lower than that of the h64 transcript (Figure 4c). The hormone profiles of the carriers of LBSASB with c.955C>T ($N = 5$) and non-carriers ($N = 102$) of NFAI patients from our recent study[22] were compared, but no significant difference was detected in morning cortisol, midnight cortisol, adrenocorticotropic hormone (ACTH)-induced cortisol, morning aldosterone, ACTH-induced aldosterone, morning 17-OH-progesterone, morning corticosterone, ACTH-induced corticosterone, metyrapone-blocked 11-deoxycortisol, morning dehydroepiandrosterone sulfate, ACTH, or metyrapone-blocked ACTH levels (Supplementary Table 4).

DISCUSSION

A moderately frequent, unique haplotypic RCCX CNV structure with three segments harboring two CYP21A2 gene copies, abbreviated to LBSASB, was determined in 13 individuals (seven healthy subjects, five patients with NFAI and one reexamined CAH patient) with normal steroid levels by a set of experimental and bioinformatic methods published recently.[1,19] This set of methods is capable of not only a more exact determination of RCCX structure than some current methods but also the segment-specific sequencing of CYP21A2: thus the haplotypes of the two harbored CYP21A2 in the middle and the 3′-segments of the RCCX CNV were revealed. The CYP21A2 haplotypes in the middle segment harbored the c.*12T allele, whereas the CYP21A2 haplotypes in the 3′-segment harbored the c.293-79A and
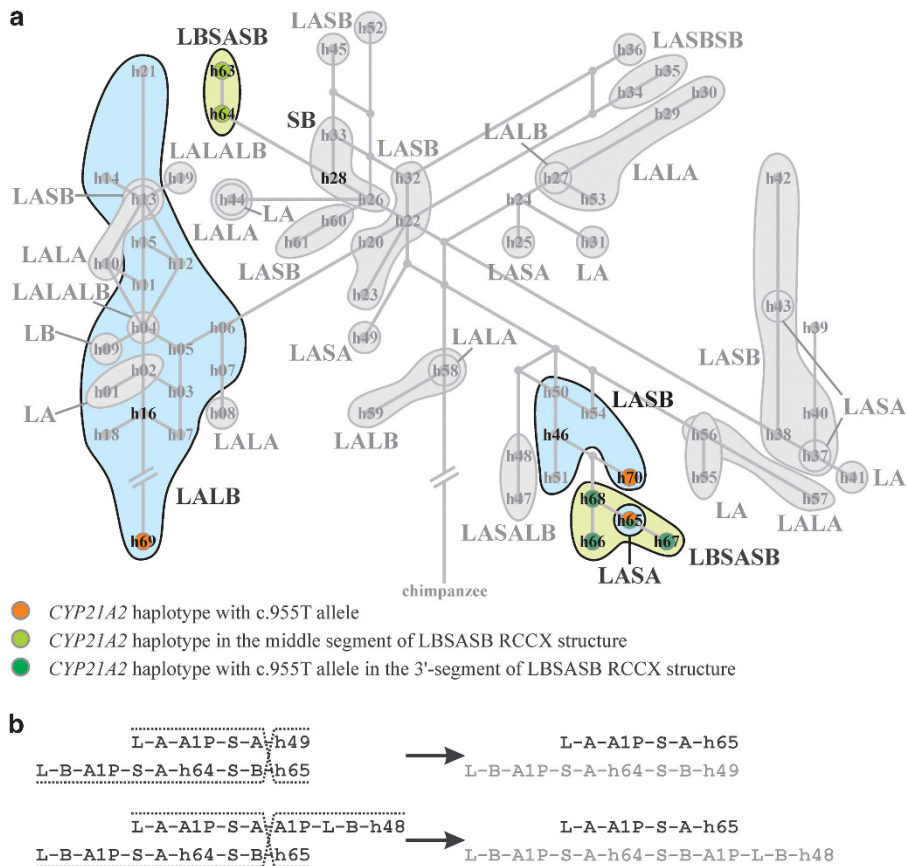
Figure 3 Evolutionary processes of RCXX CNV structure harboring the c.955C>T variant in *CYP21A2*. Each haplotypic RCCX segment is abbreviated with two letters, the first represents the alleles of HERV-K(C4) CNV (**L** – long (insertion) allele or **S** – short (deletion) allele), and the second symbolizes the types of *C4* gene (**A** or **B**). The multiplication of the two letters in a structure indicates the segment number. (**a**) Genealogical haplotype networks of *CYP21A2* haplotypes with projected haplotypic RCCX CNV structures. Numbers in gray circles indicate the known *CYP21A2* haplotype, and numbers in orange, green, and dark green circles represent the *CYP21A2* haplotype published here for the first time. Orange circles indicate *CYP21A2* haplotypes with the c.955C>T variant in classical CAH patients, green circles represent *CYP21A2* haplotypes in the middle segment of the LBSASB RCCX structure, and dark green circles indicate *CYP21A2* haplotypes with the c.955C>T in the 3′-segment of LBSASB. The length of gray lines between the *CYP21A2* haplotypes is proportional with the nucleotide (allele) identity: greater length implies more different alleles between haplotypes and lower identity. (**b**) Both simple (equal) crossover between LASA and LBSASB RCCX structures and unequal crossover between LASALB and LBSASB structures could generate the LASA structure harboring a h65 *CYP21A2* haplotype in the 3′-segment.

c.955T alleles. Both c.293−79G>A and c.*12C>T variants are in complete LD with the c.955C>T variant in the healthy subjects of two European (Spanish and Italian) populations,[17,31] suggesting a straightforward usage of these variants to distinguish between pathogenic and non-pathogenic genomic contexts of the c.955C>T variant in the genetic diagnosis of CAH, but unfortunately there is no available information from c.293−79G>A and c.*12C>T in the CAH patients of these populations. Our findings on the healthy subjects and the patients with NFAI and CAH of the Hungarian population indicated that only the c.*12T allele is unique in the individuals with normal steroid hormone levels. Therefore, the genetic examination of c.*12C>T allows us to avoid the genetic misinterpretation of p.(Q319*).

The ability of the c.*12C>T to discriminate between *CYP21A2* copies with c.955C>T in non-pathogenic and pathogenic genomic contexts can be readily built into molecular diagnosis approaches,[32,33] which seems to be especially important for populations with high c.955C>T frequency from North Africa and Asia.[34,35] Before its inclusion in the routine genetic diagnosis of CAH, the occurrence pattern of c.*12C>T should be investigated in both healthy subjects

and CAH patients belonging to the target population. The pioneer study on c.955C>T in RCCX structure with three segments and two *CYP21A2*[10] obviously illustrates the importance of validation in other populations: it is probable that the c.*12T and c.293−13G CAH alleles are simultaneously harbored in the middle segment in three CAH patients. However, the nationality and ancestry of these patients have not been documented, and thus a conclusion with regard to the origin of this RCCX structure unfortunately cannot be drawn.

Nevertheless, the three CAH-causing mutations in one patient harboring c.*12C>T implied that one of the mutations could occur in the *CYP21A2* of the middle segment, disabling its functionality. Our evolutionary analysis provided a new insight into the underlying mechanism that generates such a clinical case. The *CYP21A2* haplotypes with the c.955T allele are derived from the 3′-segment of the LBSASB RCCX structure or from *CYP21A1P*, according to the genealogical haplotype network. However, the second, functional copy of *CYP21A2* in a chromosome was never carried by classical CAH patients with c.955C>T, independent of whether the c.955T allele of *CYP21A2* haplotypes was derived from the 3′-segment of LBSASB or elsewhere. This putative rearrangement mechanism of *CYP21A2*
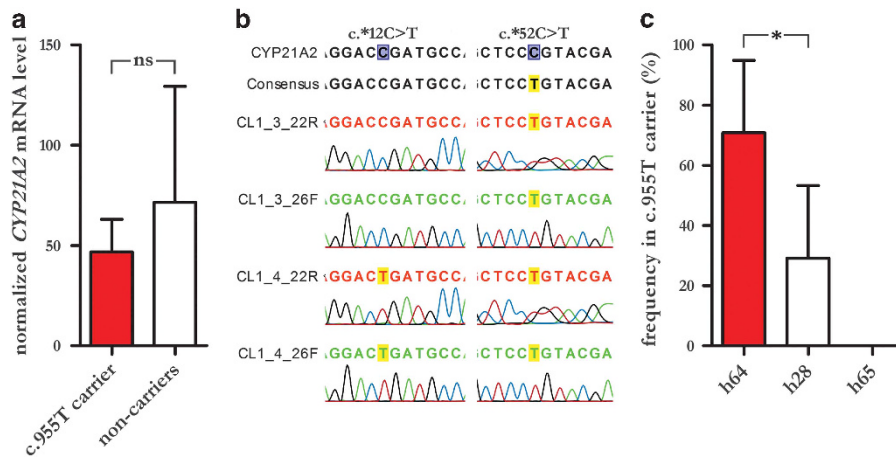
Figure 4 The results of the functional studies on adrenocortical specimens of NFAI patients. One asterisk indicates significant difference at $P<0.05$, whereas ns represents a nonsignificant difference, as calculated by t-test. (**a**) *CYP21A2* mRNA level in the adrenal specimens of c.955T allele (on LBSASB RCCX structure) carrier NFAI patient ($N=1$) and non-carrier patients ($N=3$). (**b**) The forward and reverse chromatogram of h28 and h64 transcripts from the c.955T allele carrier NFAI patient, where the distinctive alleles on c.*12C>T and c.*52C>T variants are indicated (CT alleles for h28 and TT alleles for h64). (**c**) The abundance of transcripts expressed from three *CYP21A2* haplotypes (h28, h64, and h65) of the NFAI patient with c.955T allele on LBSASB RCCX structure.

haplotypes with the c.955T derived from the LBSASB RCCX structure sheds light on why the c.*12T allele did not occur in Hungarian CAH patients carrying a *CYP21A2* haplotype with c.293-79A and c.955T; if a rearranged RCCX structure loses the functional *CYP21A2* from the middle segment of LBSASB, it will become pathogenic, and if a rearranged RCCX structure retains the functionally intact *CYP21A2* with c.*12T from the middle segment, it will rescue its carrier from CAH. Nevertheless, it cannot be excluded that a rearrangement such as non-allelic gene conversion deletes or transfers the c.*12T allele from its original genomic context, or transfers a pathogenic mutation to *CYP21A2* in the middle segment (the frequency of such an event will be very rare, because a rare event happens in the moderately frequent LBSASB). Therefore, careful consideration is always a basic requirement in the genetic diagnosis of CAH, and the sequencing of full-length *CYP21A2* from the segment-specific PCR product of the middle segment is needed after the finding of the c.*12T allele in the extremely rare case when more than two CAH mutations (including c.955C>T) are detected in one individual.

The significance of the c.955C>T variant in the molecular diagnosis of CAH is testified by a recent case study on a misdiagnosed c.955C>T variant.[36] This kind of case is not extremely rare, and there may be no better example of how true that is than one of our cases. The original genetic diagnosis of a 1-year-old patient with classical CAH was reexamined, because the required dose of hormone substitution was unchanged in spite of weight gain. The father had a known history of classical CAH due to *CYP21A2* deficiency and the healthy mother carried a c.955T allele. The patient inherited the maternal c.955T allele besides a paternal RCCX structure without *CYP21A2*, and was therefore treated with mineralocorticoids and glucocorticoids from birth. The reexamination demonstrated that an intact *CYP21A2* was harbored by the middle segment of LBSASB next to the c.955C>T variant in the 3′-segment. The steroid supplementation had been gradually reduced, and then completely discontinued based on this finding, without any clinical signs of adrenocortical insufficiency.

Several studies based on the normal steroid levels of subjects with two carried CAH variants have found indirect evidence proving that one of the two *CYP21A2* associated with the c.955C>T variant in one chromosome expresses active steroid 21-hydroxylase enzyme.[36–39] However, direct evidence for the functional *CYP21A2* in the middle segment was provided by the current study for the first time. In the adrenal specimen of a patient with NFAI, the *CYP21A2* in the middle segment of the LBSASB RCCX structure produced an mRNA transcript differing from the transcripts of the *CYP21A2* copies in the 3′-segment and the other chromosome, and this middle segment-specific *CYP21A2* transcript was detected. The *CYP21A2* transcript from the middle segment was significantly more abundant than the other two transcripts, but neither *CYP21A2* mRNA expression levels nor hormone levels in blood significantly differed between the LBSASB carriers and the non-carriers in NFAI patients. It should be noted that the quantitative analysis of the abundance of different *CYP21A2* transcripts was not absolutely reliable because only one adrenal specimen was examined (however, we have no reason to question the existence of the *CYP21A2* transcript expressed from the middle segment of LBSASB), and the sample size of NFAI patients was not sufficient for the detection of subtle hormonal differences in blood. In fact, several hundreds of adrenal specimens or NFAI patients should be examined to collect enough LBSASB carriers for the more reliable analyses on the abundance of *CYP21A2* mRNAs or the subtle hormonal differences due to the moderate prevalence of LBSASB.

The current study is in tune with the results that have been presented so far (Table 2). In addition, the allele frequencies (~1–2%) and the almost complete LD between c.293−79G>A and c.*12C>T in the European populations of the 1000 Genomes project[40] also imply the presence of LBSASB RCCX structure. However, the next-generation sequencing (NGS)-based methods cannot reveal the exact internal content of multiallelic CNVs,[41] and the bioinformatic analysis of NGS data is still prone to errors;[42] hence, c.*12C>T is documented by dbSNP as a variant of *CYP21A2* in the 3′-segment. According to the current finding, the c.*12 nucleotide position is never polymorphic in the 3′-segment, but it is in the middle segment. Therefore, this phenomenon cannot be described by the traditional variant definition of non-duplicated genomic regions, and the usage of the variant categories for duplicated genes[19,43] should be considered.

1 Banlaki Z, Szabo JA, Szilagyi A *et al*: Intraspecific evolution of human RCCX copy number variation traced by haplotypes of the *CYP21A2* gene. *Genome Biol Evol* 2013; **5**: 98–112.

2 Handsaker RE, Van Doren V, Berman JR *et al*: Large multiallelic copy number variations in humans. *Nat Genet* 2015; **47**: 296–303.

3 Blanchong CA, Zhou B, Rupert KL *et al*: Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease. *J Exp Med* 2000; **191**: 2183–2196.

4 Szilagyi A, Doleschall M, Fust G: Complement genes in the central region of the MHC. In: Mehra NK (ed): *The HLA Complex in Biology and Medicine: A Resource Book*. New Delhi, India: Jaypee Brothers Medical Publishers, 2010, pp 135–158.

5 Miller WL, Auchus RJ: The molecular biology, biochemistry, and physiology of human steroidogenesis and its disorders. *Endocr Rev* 2011; **32**: 81–151.

6 White PC, Speiser PW: Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Endocr Rev* 2000; **21**: 245–291.

7 Cooper DN, Bacolla A, Ferec C, Vasquez KM, Kehrer-Sawatzki H, Chen JM: On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. *Hum Mutat* 2011; **32**: 1075–1099.

8 New MI, Abraham M, Gonzalez B *et al*: Genotype–phenotype correlation in 1,507 families with congenital adrenal hyperplasia owing to 21-hydroxylase deficiency. *Proc Natl Acad Sci USA* 2013; **110**: 2611–2616.

9 Wedell A: Molecular genetics of congenital adrenal hyperplasia (21-hydroxylase deficiency): implications for diagnosis, prognosis and treatment. *Acta Paediatr* 1998; **87**: 159–164.

10 Wedell A, Stengler B, Luthman H: Characterization of mutations on the rare duplicated C4/CYP21 haplotype in steroid 21-hydroxylase deficiency. *Hum Genet* 1994; **94**: 50–54.

11 Witchel SS, Lee PA, Trucco M: Who is a carrier? Detection of unsuspected mutations in 21-hydroxylase deficiency. *Am J Med Genet* 1996; **61**: 2–9.

12 Janner M, Pandey AV, Mullis PE, Fluck CE: Clinical and biochemical description of a novel CYP21A2 gene mutation 962_963insA using a new 3D model for the P450c21 protein. *Eur J Endocrinol/Eur Fed Endocr Soc* 2006; **155**: 143–151.

13 Mohamed S, El-Kholy S, Al-Juryyan N, Al-Nemri AM, Abu-Amero KK: A CYP21A2 gene mutation in patients with congenital adrenal hyperplasia. Molecular genetics report from Saudi Arabia. *Saudi Med J* 2015; **36**: 113–116.

14 Globerman H, Amor M, Parker KL, New MI, White PC: Nonsense mutation causing steroid 21-hydroxylase deficiency. *J Clin Invest* 1988; **82**: 139–144.

15 Koppens PF, Hoogenboezem T, Degenhart HJ: CYP21 and CYP21P variability in steroid 21-hydroxylase deficiency patients and in the general population in the Netherlands. *Eur J Hum Genet* 2000; **8**: 827–836.

16 Koppens PF, Hoogenboezem T, Degenhart HJ: Duplication of the CYP21A2 gene complicates mutation analysis of steroid 21-hydroxylase deficiency: characteristics of three unusual haplotypes. *Hum Genet* 2002; **111**: 405–410.

17 Parajes S, Quinteiro C, Dominguez F, Loidi L: High frequency of copy number variations and sequence variants at CYP21A2 locus: implication for the genetic diagnosis of 21-hydroxylase deficiency. *PLoS ONE* 2008; **3**: e2138.

18 Kleinle S, Lang R, Fischer GF *et al*: Duplications of the functional CYP21A2 gene are primarily restricted to Q318X alleles: evidence for a founder effect. *J Clin Endocrinol Metab* 2009; **94**: 3954–3958.

19 Szabo JA, Szilagyi A, Doleschall Z *et al*: Both positive and negative selection pressures contribute to the polymorphism pattern of the duplicated human CYP21A2 gene. *PLoS ONE* 2013; **8**: e81977.

20 Young WFJr: Clinical practice. The incidentally discovered adrenal mass. *N Engl J Med* 2007; **356**: 601–610.

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

22 Doleschall M, Szabo JA, Pazmandi J *et al*: Common genetic variants of the human steroid 21-hydroxylase gene (CYP21A2) are related to differences in circulating hormone levels. *PLoS ONE* 2014; **9**: e107244.

23 Luczay A, Torok D, Ferenczi A, Majnik J, Solyom J, Fekete G: Potential advantage of N363S glucocorticoid receptor polymorphism in 21-hydroxylase deficiency. *Eur J Endocrinol/Eur Fed Endocr Soc* 2006; **154**: 859–864.

24 Patocs A, Toth M, Barta C *et al*: Hormonal evaluation and mutation screening for steroid 21-hydroxylase deficiency in patients with unilateral and bilateral adrenal incidentalomas. *Eur J Endocrinol/Eur Fed Endocr Soc* 2002; **147**: 349–355.

25 Boyle B, Butz H, Liko I *et al*: Expression of glucocorticoid receptor isoforms in human adrenocortical adenomas. *Steroids* 2010; **75**: 695–700.

26 Cervenak J, Doleschall M, Bender B *et al*: NFkappaB induces overexpression of bovine FcRn: a novel mechanism that further contributes to the enhanced immune response in genetically modified animals carrying extra copies of FcRn. *mAbs* 2013; **5**: 860–871.

27 de Jonge HJ, Fehrmann RS, de Bont ES *et al*: Evidence based selection of housekeeping genes. *PLoS ONE* 2007; **2**: e898.

28 Doleschall M, Mayer B, Cervenak J, Cervenak L, Kacskovics I: Cloning, expression and characterization of the bovine p65 subunit of NFkappaB. *Dev Comp Immunol* 2007; **31**: 945–961.

29 Posada D, Crandall KA: Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol Evol* 2001; **16**: 37–45.

30 Chung EK, Yang Y, Rennebohm RM *et al*: 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* 2002; **71**: 823–837.

31 Concolino P, Mello E, Minucci A, Giardina B, Capoluongo E: Genes, pseudogenes and like genes: the case of 21-hydroxylase in Italian population. *Clin Chim Acta* 2013; **424**: 85–89.

32 Tsai LP, Cheng CF, Chuang SH, Lee HH: Analysis of the CYP21A1P pseudogene: indication of mutational diversity and CYP21A2-like and duplicated CYP21A2 genes. *Anal Biochem* 2011; **413**: 133–141.

33 Xu Z, Chen W, Merke DP, McDonnell NB: Comprehensive mutation analysis of the CYP21A2 gene: an efficient multistep approach to the molecular diagnosis of congenital adrenal hyperplasia. *J Mol Diagn* 2013; **15**: 745–753.

34 Kharrat M, Tardy V, M'Rad R *et al*: Molecular genetic analysis of Tunisian patients with a classic form of 21-hydroxylase deficiency: identification of four novel mutations and high prevalence of Q318X mutation. *J Clin Endocrinol Metab* 2004; **89**: 368–374.

35 Marumudi E, Sharma A, Kulshreshtha B, Khadgawat R, Khurana ML, Ammini AC: Molecular genetic analysis of CYP21A2 gene in patients with congenital adrenal hyperplasia. *Indian J Endocrinol Metab* 2012; **16**: 384–388.

36 Lekarev O, Tafuri K, Lane AH *et al*: Erroneous prenatal diagnosis of congenital adrenal hyperplasia owing to a duplication of the CYP21A2 gene. *J Perinatol* 2013; **33**: 76–78.

37 Baumgartner-Parzer SM, Nowotny P, Waldhausl W, Vierhapper H: A rare duplicated 21-hydroxylase haplotype and a *de novo* mutation: a family analysis. *J Clin Endocrinol Metab* 2003; **88**: 2794–2796.

38 Kharrat M, Riahi A, Maazoul F, M'Rad R, Chaabouni H: Detection of a frequent duplicated CYP21A2 gene carrying a Q318X mutation in a general population with quantitative PCR methods. *Diagn Mol Pathol Part B* 2011; **20**: 123–127.

39 Leccese A, Longo V, Dimatteo C *et al*: Lack of genotypephenotype correlation in congenital adrenal hyperplasia due to a CYP21A2-like gene. *Clin Chim Acta* 2014; **437**: 48–51.

40 Auton A, Brooks LD, Durbin RM *et al*: A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

41 Nuttle X, Huddleston J, O'Roak BJ *et al*: Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nat Methods* 2013; **10**: 903–909.

42 Li H: Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 2014; **30**: 2843–2851.

43 Innan H: Population genetic models of duplicated genes. *Genetica* 2009; **137**: 19–37.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)