Relation between HLA and copy number variation of steroid 21-hydroxylase in a Swedish cohort of patients with autoimmune Addison’s disease

Short title: 21-hydroxylase copy number variation

Christian Lundtoft1, Daniel Eriksson2,3,4,#, Matteo Bianchi5,#, Maribel Aranda-Guillén2, Nils Landegren2,6, Solbritt Rantapää-Dahlqvist7, Peter Söderkvist8, Jennifer R. S. Meadows5, the DISSECT consortium†, the ImmunoArray consortium†, The Swedish Addison Registry Study Group†, Sophie Bensing9,10, Gerli Rosengren Pielberg5, Kerstin Lindblad-Toh5,11, Lars Rönnblom1, Olle Kämpe2,9,12,13,

1Department of Medical Sciences, Uppsala University, Uppsala, Sweden
2Center for Molecular Medicine, Department of Medicine (Solna), Karolinska Institutet, Stockholm, Sweden
3Department of Clinical Genetics, Uppsala University Hospital, Uppsala, Sweden
4Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden
5Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden
6Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden
7Department of Public Health and Clinical Medicine/Rheumatology, Umeå University, Umeå, Sweden
8Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden
9Department of Endocrinology, Karolinska University Hospital, Stockholm, Sweden
10Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden
11Broad Institute, MIT and Harvard, Cambridge, MA, United States
12Department of Clinical Science, University of Bergen, Bergen, Norway
13K.G. Jebsen Center for Autoimmune Diseases, University of Bergen, Bergen, Norway

#These authors contributed equally
†Consortia members are listed in Supplementary Information.
‡Corresponding author: Olle Kämpe, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden, olle.kampe@ki.se

Keywords
Addison’s disease, Autoimmune Disease, Autoantibodies, Genetics.

© The Author(s) 2023. Published by Oxford University Press on behalf of European Society of Endocrinology. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abbreviations

AAD, autoimmune Addison’s disease; CI, confidence interval; HLA, human leukocyte antigen; RCCX module, RP/STK-C4-CYP21-TNX gene module; SNV, single-nucleotide variant.

Word count

2,082 words
Abstract

Objective. Autoantibodies against the adrenal enzyme 21-hydroxylase is a hallmark manifestation in autoimmune Addison’s disease (AAD). Steroid 21-hydroxylase is encoded by \textit{CYP21A2}, which is located in the HLA region together with the highly similar pseudogene \textit{CYP21A1P}. A high level of copy number variation is seen for the two genes, and therefore we asked if genetic variation of the \textit{CYP21} genes is associated with AAD.

Methods. Using next-generation DNA sequencing, we estimated the copy number of \textit{CYP21A2} and \textit{CYP21A1P} together with HLA alleles in 479 Swedish patients with AAD and autoantibodies against 21-hydroxylase, and in 1,393 healthy controls.

Results. With 95\% of individuals carrying two functional 21-hydroxylase genes, no difference in \textit{CYP21A2} copy number was found when comparing patients and controls. In contrast, we discovered a lower copy number of the pseudogene \textit{CYP21A1P} among AAD patients ($p = 5 \times 10^{-44}$) together with associations of additional nucleotide variants in the \textit{CYP21} region. However, the strongest association was found for \textit{HLA-DQB1*02:01} ($p = 9 \times 10^{-63}$), which in combination with the \textit{DRB1*04:04–DQB1*03:02} haplotype imposed the greatest risk of AAD.

Conclusions. We identified strong associations between copy number variants in the \textit{CYP21} region and risk of AAD, although these associations most likely are due to linkage disequilibrium with disease-associated HLA class II alleles.

Significance statement

The majority of patients with autoimmune Addison’s disease present with autoantibodies against the enzyme steroid 21-hydroxylase that is encoded by \textit{CYP21A2}. The gene is located in the HLA region together with the similar pseudogene \textit{CYP21A1P}. Based on a focused analysis of the complex genetic region, we demonstrate extensive copy number variation of \textit{CYP21A2} and \textit{CYP21A1P} in patients with autoimmune Addison’s disease and healthy controls. Although copy numbers differ strongly between patients and controls, HLA risk alleles remain as the strongest risk factor for autoimmune Addison’s disease.
Introduction

Autoimmune Addison’s disease (AAD) is characterised by an autoimmune destruction of the adrenal glands, resulting in inadequate synthesis of cortisol and aldosterone\(^1,2\). The disease has a strong genetic component as shown in twin studies\(^3\), and the HLA region harbours the main genetic risk factors associated with the disease, including the haplotypes DRB1*03:01–DQB1*02:01 and DRB1*04:04–DQB1*03:02\(^4,5\). In more than 85% of adult patients, the autoimmune aetiology is confirmed by the presence of autoantibodies against the enzyme 21-hydroxylase, which is specifically expressed in the adrenal cortex and essential for the synthesis of cortisol and aldosterone\(^4,6,7\).

Steroid 21-hydroxylase is encoded by the gene CYP21A2 that is located between the HLA class I and class II regions on chromosome 6\(^8\). CYP21A2 is part of the RCCX module that harbours extensive copy number variation in the general population, generally ranging between 1-4 copies on each chromosome (Figure 1A)\(^9\). In addition to the functional 21-hydroxylase gene, a non-functional pseudogene CYP21A1P with ~98% sequence similarity is located in the same region\(^10,11\).

In addition to the CYP21 genes, the genes C4A and C4B encoding complement C4 are also located in the RCCX module (Figure 1A). Previous studies have shown a strong link between DRB1*03:01 and low copy number of C4A in populations of European descent\(^12,13\). Although a few studies exist\(^14-16\), a systematic evaluation of the CYP21 locus in relation to AAD has been restrained by the high level of copy number variation in combination with the high sequence similarity between the functional and non-functional genes.

Here we performed a focused investigation of genetic variation in the CYP21 region among AAD patients with 21-hydroxylase autoantibodies in comparison to healthy controls. Furthermore, HLA alleles were analysed in order to evaluate the relative contribution of CYP21 copy number variation and HLA haplotypes to the association with AAD.

Materials and Methods

Study participants

The 479 AAD patients, all with autoantibodies against 21-hydroxylase, and 1,394 healthy controls have been described in detail previously\(^17-20\). Study subjects had been recruited in Sweden, and patients...
fulfilled diagnostic criteria for primary adrenal insufficiency, whereas patients with other suspected causes of adrenal failure or anti-cytokine autoantibodies had been excluded. In addition, genetic quality control had been performed to exclude samples of low genotyping quality and population outliers. We further excluded one control from the current study that had been included in duplicate, leaving 1,393 healthy controls in the analysis. The study was performed in accordance with the Declaration of Helsinki, and all study participants gave informed consent. The study was approved by the Swedish regional ethics committees (Stockholm Dnr 2008/296-31/2; Uppsala Dnr 2009/013; Linköping 2010/182-3).

**DNA sequencing and analysis of copy number and HLA**

Sequence capture was performed using a custom array targeting exons and regulatory regions of 1,853 genes selected for their role in regulation of the immune system or in autoimmune diseases. DNA sequencing was performed with paired-end reads (100 bp) using an Illumina HiSeq 2500 system, followed by mapping to GRCh37, genotyping using GATK HaplotypeCaller and genetic quality control, as described previously.

A method for analysis of copy number variation of complement C4 described earlier was extended to include copy number variation of the neighbouring CYP21 genes that are also part of the RCCX module. The total copy number of CYP21 and complement C4 was analysed using GATK GermlineCNVCaller as described previously, with few modifications detailed in Supplementary Information. Copy number of the C4A/C4B and CYP21A2/CYP21A1P paralogs were determined using the relative read depth of paralog-specific variants. Genetic variants in the RCCX module were analysed using GATK HaplotypeCaller while accounting for total copy number of C4/CYP21. The method for CYP21 copy number variation was tested on WES data from 90 samples in the 1000 Genomes Project and compared to CYP21 copy number estimates generated from WGS data from the same individuals using GenomeSTRiP. As described in Supplementary Information, a match in total copy number of CYP21 was found for 97.8% of the 90 samples, while an agreement in CYP21A1P and CYP21A2 copy number calls was seen for 95.6% of the samples.
HLA alleles of the genes HLA-A, -B, -C, -DPB1, -DQB1 and -DRB1 were typed from sequencing reads at 2-field (i.e. 4-digit) resolution using the tool xHLA as described previously.

**Statistical analysis**

R version 4.0.4 was used for statistical analyses, and the models and covariates included are described in the text or in the figure legends. Two-tailed p-values < 0.05 were considered significant, while a conservative Bonferroni-adjusted GWAS threshold at \( p < 5 \times 10^{-8} \) was used for variant analysis.

**Results**

**Copy number of CYP21A2 and the pseudogene CYP21A1P**

We determined the copy numbers of CYP21A2 together with the pseudogene CYP21A1P in 479 AAD patients with 21-hydroxylase autoantibodies and 1,393 healthy controls. In order to understand the relation to copy number variation in the RCCX modules (Figure 1A), we also determined copy numbers of complement C4A and C4B in parallel.

Analysis of CYP21A2 revealed that 95% of both patients and controls carried 2 functional 21-hydroxylase genes, while few individuals had 1, 3 or even 4 CYP21A2 genes (Figure 1B, left plot). However, no difference in copy number was observed between cases and controls (\( p_{\text{CYP21A2}} = 0.05 \)). In contrast, low copy number of the pseudogene CYP21A1P was more common among AAD patients when compared to controls (\( p_{\text{CYP21A1P}} = 5 \times 10^{-44}; \ OR = 3.44 \ [\text{CI}_{95\%}: 2.90-4.10] \) for each decrease in copy number) (Figure 1B, right plot).

To understand the levels of linkage disequilibrium between CYP21 and the neighbouring genes C4A and C4B, we also analysed the copy number of the two complement genes. The copy number of C4A generally correlated with the copy number of CYP21A1P (\( r_{\text{Spearman}} = 0.76 \)), whereas weaker linkage was seen between copy numbers of C4B and CYP21A2 (\( r_{\text{Spearman}} = 0.13 \)).

Association analysis showed that low copy number of C4A was associated with AAD to a higher extent than a low copy number of CYP21A1P (\( p_{\text{C4A}} = 3 \times 10^{-56}; \ OR = 3.82 \ [\text{CI}_{95\%}: 3.24-4.53] \) for each decrease in copy number; Figure 1C, right plot). In comparison, the copy number of C4B showed a slightly increased risk of AAD (\( p_{\text{C4B}} = 2 \times 10^{-10}; \) Figure 2C, left plot).
With a stronger association between C4A copy number and AAD, we concluded that copy numbers of CYP21A2 and CYP21A1P were likely not associated with the risk of AAD.

**Genetic variants in RCCX module**

We continued by evaluating individual nucleotide variants in the RCCX module and their association with AAD. Due to high sequence homology of the C4/CYP21 paralogs (≥ 98% identical), nucleotide variants in the two genes are generally excluded during genotyping, and the high level of copy number variation further challenges the assumption of diploidy in autosomal genes.

By remapping sequencing reads to a genomic reference in which RCCX module 1 had been masked while at the same time accounting for the copy number of C4/CYP21, we were able to genotype genetic variants in the RCCX module. However, these variants could not be assigned unambiguously to the individual gene paralogs, and we instead analysed the total number of C4/CYP21 copies carrying a reference allele and an alternative allele, respectively (Supplementary Information).

In total, we identified 317 SNVs and INDELs distributed across the RCCX module (Supplemental Data). In addition to the nucleotide variants in exon 26 of C4 that are used in defining the C4A and C4B paralogs, we observed two variants associated with AAD to a similar extent as the copy number of C4A (Figure 2): a Gly1073Asp missense variant in C4 (chr6:31996297G/A; p = 5 x 10^{-57}) and a variant located in the promoter region of CYP21 (chr6:32005267C/T; p = 1 x 10^{-56}). With multiple genetic variants in the RCCX module showing strong associations with AAD, we continued by evaluating genetic variants in C4/CYP21 in a broader context comprising the entire HLA region.

**Analysis of the HLA region**

We continued by analysing the entire HLA region focusing on variants in the RCCX module and HLA alleles. The strongest association was seen for DQB1*02:01 (p = 9 x 10^{-63}; Figure 3A, upper plot), which is in high linkage disequilibrium with B*08:01 and DRB1*03:01. Therefore, the overall association of HLA was stronger than the copy number of CYP21A1P and other variants in the RCCX module. Conditional analysis on DQB1*02:01 revealed an additional association with the
21-hydroxylase copy number variation

DRB1*04:04 allele (p = 2 x 10^-31; Figure 3A, lower plot), which is in strong linkage disequilibrium with DQB1*03:02.

Overall, the B*08:01–DRB1*03:01–DQB1*02:01 together with the DRB1*04:04–DQB1*03:02 risk haplotypes showed the strongest association with AAD when compared to genetic variants in the RCCX module. Combined analysis of DQB1*02:01 and DRB1*04:04 identified a statistical interaction for the heterozygous combination of the two HLA class II alleles, indicating that the heterozygous combination imposes a significantly greater risk than the summed risk of the two (OR = 53.8 [CI95%: 33.0-91.1], pinteraction = 0.03; Figure 3B). In total, 114 AAD patients (23.8%) carried the heterozygous combination of DQB1*02:01 and DRB1*04:04 compared to 23 controls (1.7%).

In summary, these results indicate that AAD mainly is associated with the HLA risk haplotypes DRB1*03:01–DQB1*02:01 and DRB1*04:04–DQB1*03:02, and the associations found for copy number of CYP21 and variation in the RCCX module is to a high degree due to linkage disequilibrium with disease-associated HLA alleles.

Discussion

Here, we presented an extensive genetic analysis of the copy number variation of the coding gene and pseudogene for steroid 21-hydroxylase genes in relation to the surrounding HLA region in patients with AAD and healthy controls. Special attention was paid to the RCCX module that harbour the functional gene for 21-hydroxylase together with a non-functional pseudogene. Despite the potential relevance in AAD, genetic variation in the RCCX module is generally omitted from genetic analyses due to the high sequence similarity combined with copy number variation of the CYP21 genes. While most individuals carry 2 functional 21-hydroxylase genes, we detected 3 copies of CYP21A2 among 4% of study participants, whereas 1% had a single copy of the 21-hydroxylase gene. Nevertheless, we found no difference in CYP21A2 copy number when comparing patients and controls. Common copy number variation of CYP21A2 has been described in other populations, with complete deficiency of 21-hydroxylase being a major cause of congenital adrenal hyperplasia. For the pseudogene CYP21A1P, extensive copy number variation was detected in the study participants, generally ranging between 0-5 copies with lower copy number for AAD patients.
However, the association between low \textit{CYP21A1P} copy number and AAD was to a higher extent mediated through linkage between \textit{CYP21A1P} copy number, \textit{C4A} copy number and \textit{HLA-DRB1*03:01}, which are in high linkage disequilibrium in populations of European descent\textsuperscript{12}. Similarly, while several nucleotide variants in the RCCX module were strongly associated with AAD, the strongest association was found for \textit{DRB1*03:01–DQB1*02:01}, which is in line with a previous study on selected \textit{CYP21A2} variants and HLA alleles in 381 AAD patients and 340 healthy controls\textsuperscript{16}. As shown previously, the \textit{DRB1*03:01–DQB1*02:01} alleles were especially associated with risk of AAD when present in combination with the \textit{DRB1*04:04–DQB1*03:02} alleles\textsuperscript{4,5}. By accounting for copy number variation while genotyping variants in the RCCX module, we were able to analyse genetic variation in this complex region. Considering the intrinsic limitations of short-read sequencing used in the current study, alternative methods such as long-read sequencing would be better suited for capturing variation of the individual \textit{C4/CYP21} paralogs in their complete genetic context. Further, analysis of patients from other populations with different linkage disequilibrium structures between HLA alleles and variants in the \textit{C4/CYP21} region may help elucidating the genetic factors associated with development of autoantibodies against 21-hydroxylase and AAD. We note that the method for \textit{CYP21} copy number estimation described here will not suffice e.g. in clinical evaluation of congenital adrenal hyperplasia. Instead we refer to the best practice guidelines for genetic testing of 21-hydroxylase deficiency that comprises techniques such as multiplex ligation dependent probe amplification (MLPA) and Sanger sequencing\textsuperscript{30}. Despite these limitations, the current method is scalable and allows analysis of a complex genetic region using next generation sequencing data, which is increasingly being used in the genetic analysis of cohorts of various sizes. In summary, we performed a comprehensive analysis of variation in the coding gene and pseudogene for 21-hydroxylase, a protein that the majority of AAD patients elicit an autoimmune response to. Despite associations between copy number variation of the \textit{CYP21} region and risk of AAD, stronger associations were identified for the \textit{DRB1*03:01–DQB1*02:01} and \textit{DRB1*04:04–DQB1*03:02} haplotypes, with HLA remaining as the main risk factor associated with 21-hydroxylase autoantibodies and AAD.
**Declaration of interest**

LR has received honorarium for lectures from AstraZeneca and participated in advisory board for UCB. OK is board member of Navinci Diagnostics AB and member of the Scientific Advisory Board for the Leo Foundation Skin Immunology Research Center, University of Copenhagen.

**Funding**

This work was supported by the Swedish Research Council for Medicine and Health, King Gustav V’s 80-year Foundation, the Knut and Wallenberg Foundation, the Novo Nordisk Foundation, and the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet.

**Acknowledgements**

DNA sequencing and genotyping was performed at the SNP&SEQ Technology Platform in Uppsala, part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The computations were enabled by resources in projects sens2017142, sens2017513, sens2018526 and sens2020577 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX. Raw sequencing data from SweAS controls was provided following data agreements, from the Uppsala and Linköping Ankylosing Spondylitis Research Group (Department of Medical Biochemistry and Microbiology, Uppsala University, Sweden and Department of Biomedical and Clinical Sciences, Linköping University, Sweden). The authors thank Västerbotten biobank for providing DNA samples from control individuals from Northern Sweden.

**Data availability**

Raw data for individual figures is available in Supplemental Data. Genotype data at the individual level is not publicly available since it contains information that could compromise research participant privacy and consent. Scripts for calling CYP21/C4 copy number in GATK GermlineCNVCaller are available upon request.
References


8. White PC, Grossberger D, Onufer BJ, Chaplin DD, New MI, Dupont B & Strominger JL. Two genes encoding steroid 21-hydroxylase are located near the genes encoding the fourth component of complement in man. Proceedings of the National Academy of Sciences 1985 82 1089.


21-hydroxylase copy number variation


Figure legends

Figure 1 Copy number variation of the 21-hydroxylase genes in Addison’s disease
A Genetic structure of the RCCX modules (STK19 [formerly RP], C4, CYP21, TNX) in the reference genome (GRCh37). A majority of healthy individuals carry two RCCX modules on each chromosome although it may vary between 1-4 modules. Pseudogenes have been marked with an asterisk. B Copy number of the functional 21-hydroxylase gene CYP21A2 (left plot) and the pseudogene CYP21A1P (right plot) in patients with autoimmune Addison’s disease (AAD) and controls. C Copy number of complement C4B (left plot) and complement C4A (right plot) in AAD patients and controls. B-C Logistic regression of copy numbers from 479 AAD patients and 1,393 controls adjusting for sex and population structure (principal component 1 (PC1)-PC3). Bars indicate the relative proportion of individuals with the indicated copy number. Controls with five gene copies are not shown in plots (n_cyp21a1p = 1, n_c4b = 1, n_c4a = 5). Abbreviations: DXO decapping and exoribonuclease protein, STK19 serine/threonine-protein kinase 19, TNX tenascin X.

Figure 2 Genetic variants in the RCCX module
Logistic regression analysis of C4/CYP21 copy numbers and nucleotide variants in the RCCX module for 479 patients with autoimmune Addison’s disease and 1,393 controls. Variants in the RCCX module were analysed by accounting for total copy number of C4/CYP21 followed by logistic regression using the number of genes carrying the reference allele and the alternative allele, respectively, while adjusting for sex and PC1-PC3. Variants present among ≥ 5 patients and controls were included in the analysis. Genomic position refers to the RCCX2 module in the GRCh37 reference, and dashed line indicate GWAS significance threshold (p < 5 x 10^-8).

Figure 3 Association analysis of HLA region
A Association of HLA alleles, SNPs, C4/CYP21 copy number, and nucleotide variants in the RCCX module with autoimmune Addison’s disease (n = 479) as compared to controls (n = 1,393). HLA alleles for six genes are shown (from left: HLA-A, -C, -B, -DRB1, -DQB1 and -DPB1). Genomic position refers to GRCh37, and dashed line indicate GWAS significance threshold (p < 5 x 10^-8).
B Interaction analysis between HLA-DQB1*02:01 and DRB1*04:04 and the association with autoimmune Addison’s disease compared to controls. The size of the points indicates the number of individuals with each combination of HLA alleles.
A-B Analysed by logistic regression adjusting for sex and PC1-PC3.
21-hydroxylase copy number variation

Figure 1

127x147 mm (x DPI)
21-hydroxylase copy number variation

Figure 2
122x80 mm (x DPI)
21-hydroxylase copy number variation

Figure 3

169x104 mm (x DPI)