scientific reports

Check for updates

OPEN

High precision characterization of RCCX rearrangements in a 21-hydroxylase deficiency Latin American cohort using oxford nanopore long read sequencing

Aldana Claps[1,6], Jorge E. Kolomenski[2,6], Franco Fernández[3], Natalia Macchiaroli[2], Marina L. Ingravidi[2], Marisol Delea[4], Cecilia Fernández[5], Tania Castro[1], Julieta Laiseca[1], Laura Kamenetzky[2], Melisa Taboas[1] & Liliana Dain[1,2]✉

The *CYP21A2* gene, mapped to the RCCX module in 6p21.3, is responsible for 21-hydroxylase deficiency (21-HD). In this work, we leveraged Oxford Nanopore Technology (ONT) Long Read sequencing (LRS) to analyze samples from an Argentinian cohort of 21-HD. A total of 34 samples were sequenced in 2 amplicons of 8.5 Kb covering the centromeric and telomeric RCCX modules. The number of variants found varied between 3 and 106 and all expected pathogenic variants and new ones were obtained with the LR sequencing workflow developed. We defined with higher accuracy the breakpoints of the rearrangements allowing the reclassification of chimeras and/or converted genes in 18.75% of the samples, some of them with clinical implications. By addressing the study of the telomeric RCCX module/s in depth, we found 19 genetic variants (GVs) for *CYP21A1P* and 29 GVs for *TNXA* not previously described in Latin American populations. This study may represent the first application of ONT LRS for clinical evaluation of Latin American subjects, highlighting the importance of LRS as a high-resolution method of diagnosis. It would allow a better understanding of the diversity of the RCCX modules and improve our knowledge of the variation of genetic mechanisms behind the disease.

Keywords RCCX modules, *CYP21A2*, Long read sequencing, 21-hydroxylase deficiency

Congenital adrenal hyperplasia (CAH) due to 21-hydroxylase deficiency (21-HD) accounts for 90–95% of CAH cases. This autosomal recessive disorder, the most frequent inborn error of metabolism, has a broad spectrum of clinical forms, ranging from severe or classical (CL), including the salt-wasting (SW) and simple virilizing (SV) forms, to a mild late onset form, better known as nonclassical (NC) CAH[1].

The gene encoding 21-hydroxylase, *CYP21A2*, is mapped to the short arm of chromosome 6 (6p21.3), within the human leukocyte antigen complex, in the 30 kb RCCX (*STK19*(***R***P)-***C***4-***C***YP21-TN***X***) module. The module is duplicated in tandem and arranged in a centromeric and a telomeric copy (Fig. 1). The centromeric copy includes the pseudogene for the serine-threonine nuclear protein kinase 19 B (formerly *RP2)*, and the active genes from the complement factor 4 B (*C4B*), *CYP21A2* and tenascin B (*TNXB*). The telomeric copy contains the serine-threonine nuclear protein kinase 19 (formerly *RP1*), the C4A and the pseudogenes *CYP21A1P* and tenascin A (*TNXA*). In particular for *CYP21A2*, it shares 98% sequence identity with its pseudogene, *CYP21A1P*[2–4]. Due to high sequence identity, most pathogenic variants in *CYP21A2* are known to originate from the pseudogene *CYP21A1P* through unequal crossing over and gene conversion. Around 20–30% of the

[1]Centro Nacional de Genética Médica, Administración Nacional de Laboratorios e Institutos de Salud (ANLIS) "Dr. Carlos G Malbrán", Buenos Aires, Argentina. [2]Instituto de Biociencias, Biotecnología y Biología Traslacional (iB3), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina. [3]Instituto de Patología Vegetal (IPAVE-CIAP-INTA), Córdoba, Argentina. [4]Unidad de Conocimiento Traslacional Hospitalaria Patagónica, Hospital de Alta Complejidad El Calafate SAMIC, Santa Cruz, Argentina. [5]Laboratorio Novagen, Buenos Aires, Argentina. [6]Aldana Claps and Jorge E. Kolomenski contributed equally to this work. ✉email: ldain@fbmc.fcen.uba.ar

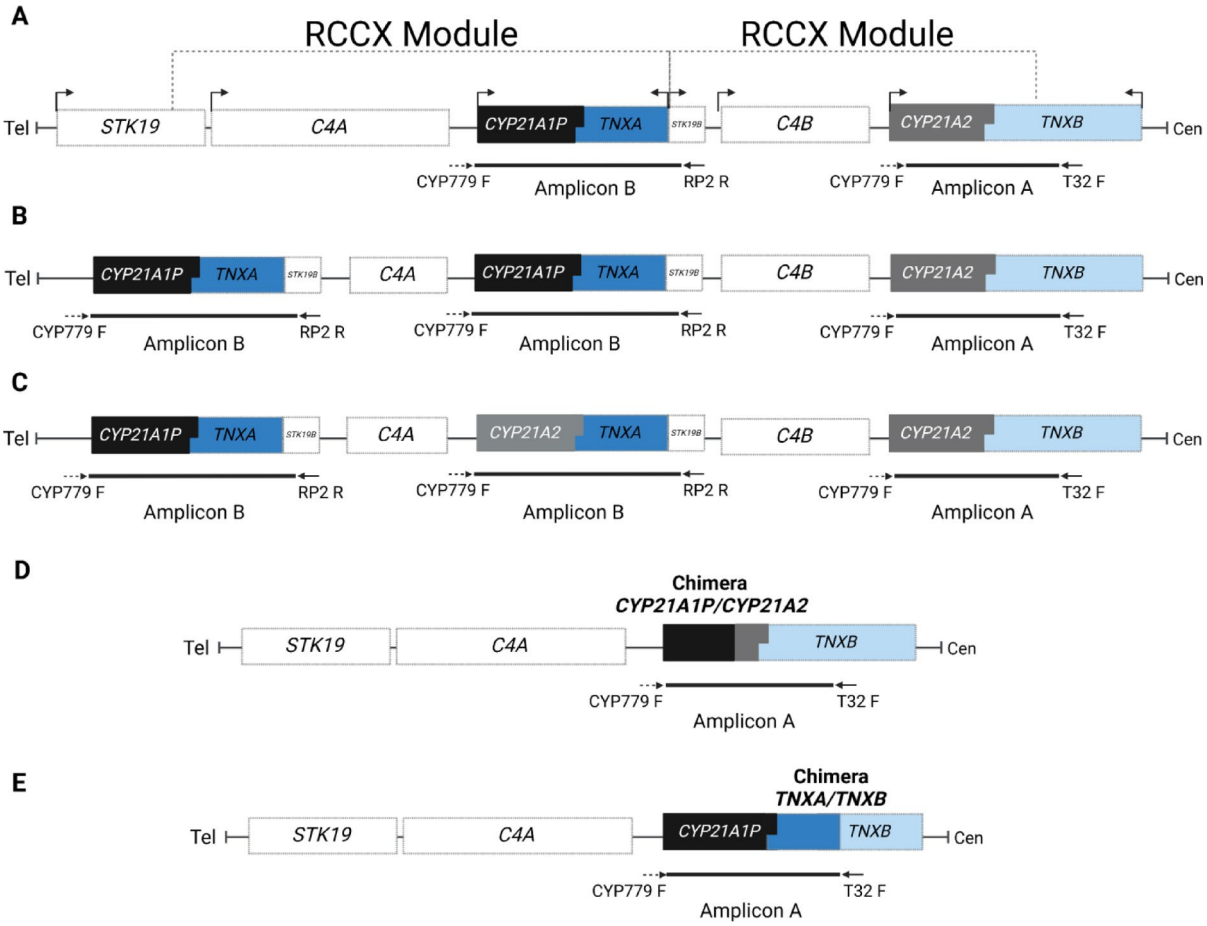

Fig. 1. Structure of different arrangements of the RCCX modules. (**A**) Bimodular rearrangement. (**B**) Trimodular rearrangement containing a duplication of *CYP21A1P*. (**C**) Trimodular rearrangement containing a duplication of *CYP21A2*. (**D**) and (**E**) Structure of the RCCX module when the module containing *CYP21A2* is deleted. In D the product of the unequal crossover is a *CYP21A1P*/*CYP21A2* chimera and, in E, a *TNXA*/*TNXB* chimera. Boxes represent the genes: *TNX*: Tenascin X (B active gene, A: pseudogene). *C4 A* and *B*: genes of the complement factor 4. *CYP21A2* and *CYP21A1P*: gene and pseudogene, respectively: *STK19* and *STK19B*: active gene and pseudogene, respectively. For clarity, in B and C, the box of *C4A* is shorter and *STK19* is omitted (indicated with a slanted line). Arrows above the boxes indicate orientation of transcription. Arrows below the boxes indicate the annealing position of the primers CYP779F, RP2R and Tena32F to generate the 8.5 Kb amplicon A and amplicon B.

alleles presented a 30 kb deletion resulting in chimeric genes with variable junction sites due to unequal crossing over at the RCCX module. Pseudogene-derived pathogenic variants due to gene conversion represent 70–75% of the pathogenic variants. Lastly, rare, naturally occurring pathogenic variants unrelated to the pseudogene account for approximately ~ 5% of the disease-causing alleles[1,5]. To date, 11 different *CYP21A1P*/*CYP21A2* chimeras have been described, with chimeras 1 to 9 being the most frequently observed[6–8]. In addition, 5 different types of *TNXA*/*TNXB* chimeras have been reported[9,10]. The *TNXA*/*TNXB* chimeras are associated with Congenital Adrenal Hyperplasia and Ehlers-Danlos Syndrome (CAH-X) and present joint hypermobility and a spectrum of other comorbidities associated with their connective tissue disorder, including chronic arthralgia, joint subluxations, hernias and cardiac defects[11].

In a previous work, we performed the genetic characterization of a large cohort of 21-HD patients from Argentina. We used differential amplification of the *CYP21A2* gene and Sanger sequencing together with Multiplex Ligation-dependent Probe Amplification (MLPA) to elucidate the presence of pathogenic variants, as well as some of the possible genetic rearrangements in the RCCX region[12]. Nevertheless, these approaches are very time-consuming and do not reveal all the possible rearrangements of this complex genomic locus. They also do not include the search for putative pathogenic variants in the *TNXB* gene.

With this in mind, in this work we aim to leverage long read third-generation sequencing (LRS) using Oxford Nanopore Technology (ONT) to analyze genetic variants (GVs) and rearrangements in the RCCX region. LRS has taken a more prominent role in recent years, especially by allowing for better analysis of genome rearrangements, repetitive sequences and regions with high sequence identity, such as discriminating genuine genes from pseudogenes as well as to better elucidate *cis*/*trans* location of GVs[13]. Other methods based on short-

read sequencing do not assemble it efficiently, so the analysis of these regions is difficult[14]. We selected a group of samples previously studied in our laboratory as well as new samples recruited for this work and sequenced both their centromeric and telomeric RCCX modules. Our goal was to gain additional information about the GVs, the organization of the RCCX modules involved in the rearrangements and to obtain higher resolution for the breakpoints of converted and chimeric alleles. We also aimed to retrieve new data of the *TNXA* and *CYP21A1P* genes that may contribute to the knowledge of the putative genetic diversity of the RCCX locus in different populations and therefore to a better understanding of the results found in clinical cases.

Results

We used third-generation ONT LRS to analyze 34 samples from an Argentinian cohort of 21-HD. Twenty-eight samples were previously studied in our lab by Sanger sequencing and MLPA and the remaining 6 were studied for the first time in this work. To this end, we amplified and sequenced 2 amplicons of 8.5 kb spanning the entire *CYP21A2 and* exon 32–44 of *TNXB* genes (amplicon A) and *CYP21A1P-TNXA* pseudogenes or a duplicated *CYP21A2-TNXA* (amplicon B), respectively. For each amplicon, we analyzed the presence of pathogenic and non-pathogenic GVs, indels, and their zygosity after haplotyping and phasing, and defined gene conversions or chimeric alleles (Fig. 2).

We successfully sequenced both amplicons in 32/34 samples. The remaining 2 samples were genotyped as having both alleles with chimeric genes and, as expected, only amplicon A showed amplification (see below). The read depth was > 400X for all the amplicons and the estimated N50 was 8.5 kb for all sequence output.

The quality cutoff of reads was > Q5 for Flow Cell R9.4.1 chemistry with the number of GVs found varied between 3 and 106. For Flow Cell R10.4.1 chemistry, the cutoff of reads was > Q7 with a number of GVs varied between 12 and 80. Excluding large rearrangements, we found a total of 248 single nucleotide variants (SNVs), 3 insertions, 6 deletions and 11 duplications (Supplementary Table S1 and S2). Particularly, for the coding regions of *CYP21A2* and the analyzed *TNXB* region (exons 32–44), we retrieved 17 synonymous (9 in *CYP21A2* and 8 in *TNXB*), 28 missense (15 in *CYP21A2* and13 in *TNXB*), 5 frameshift (4 in *CYP21A2* and1 in *TNXB*), 1 in frame and 2 nonsense variants (only in *CYP21A2*). In addition, we found 2 duplications and 3 deletions in *CYP21A2* and 1 duplication and 1 deletion in *TNXB*. Finally, we observed 39 SNVs in *CYP21A1P* and 54 in *TNXA*, and 2 duplications, 1 insertion and 1 deletion in *CYP21A1P*.

The results of genotyping by ONT LRS are summarized in Table 1, along with the results obtained with Sanger sequencing and MLPA. As shown, all the pathogenic GVs found by Sanger sequencing in *CYP21A2* were indeed observed after ONT LRS in all the analyzed samples. Common non-pathogenic variants were also concordant using both approaches (data not shown). LRS, however, allowed us to precisely phase *cis*/*trans* locations of GVs (see Fig. 3 for some examples), and to complete the list of GVs involved in the rearrangements (Table 1).

After analyzing the results, we found that some of the common *CYP21A1P* pseudogene GVs were absent in the rearrangements. Among the 21 alleles that had a converted or chimeric gene that included the promoter of

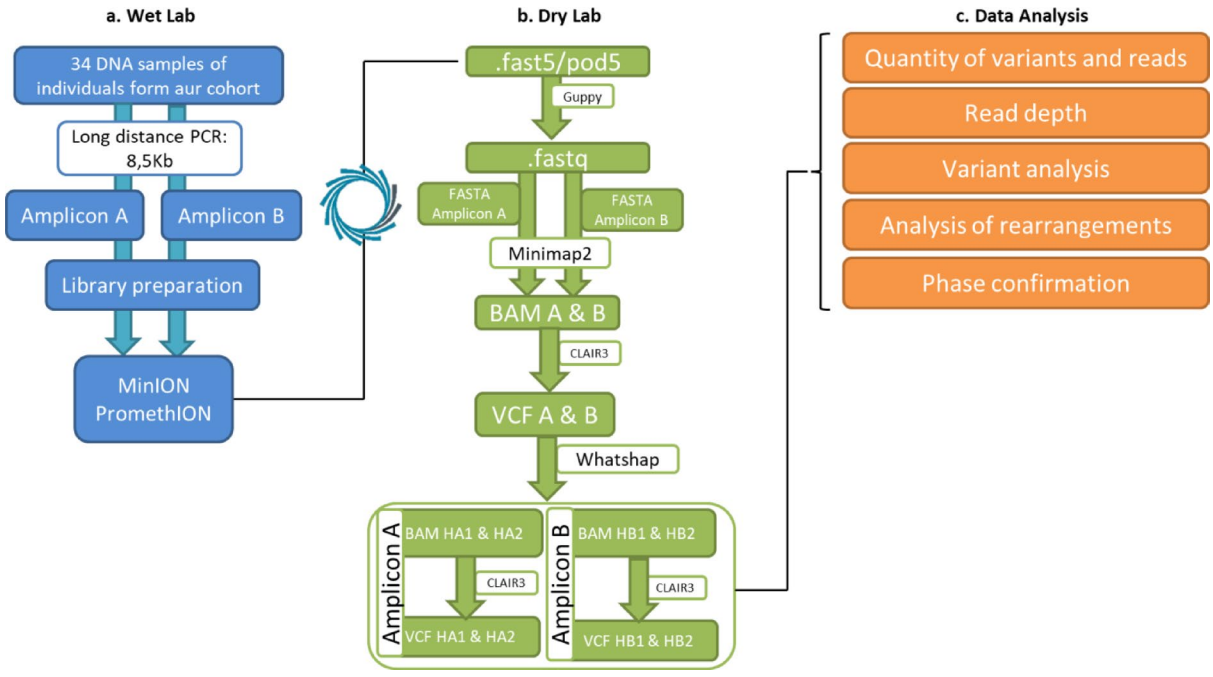

Fig. 2. Summary of the workflow. (**a**) Wet Lab optimization. Both amplicons of the RCCX module were amplified by PCR and sequenced using a MinION or PromethION device. (**b**) Dry Lab optimization. Fast5 or pod5 files were processed using Guppy, Minimap2, Clair3 and Whatshap in order to obtain bam and vcf files. (**c**) Data analysis. Analysis of ONT sequencing output. HA1 and HA2: Haplotype 1 and haplotype 2 for amplicon A, respectively. HB1 and HB2: Haplotype 1 and haplotype 2 for amplicon B, respectively.

ID	Condition	Presumptive phenotype	Previous results			LRS results				
						AMPLICON A		AMPLICON B		
						Pathogenic GVs		Absence of typical pseudogene GVs/duplicated 21A2 pathogenic GVs		
			Pathogenic GVs by Sanger sequencing	RCCX arrangement by MLPA	Deduced genotype	HA1	HA2	HB1	HB2	Deduced genotype
1	Control	N/A	None	[B]; [B]	[WT]; [WT]	None	None	***21A1P***: p.V282L; p.R357W. ***TNXA***: c.12,174 C > G	-	[WT]; [WT]
2	Patient	NC	p.V282L	[T] (*21A1P* dup); [M] (CH-5)	[p.V282L]; [CH-5]	***21A2***: p.V282L	***21A2***: c.-126 C > T; c.-113 G > A; c.-110T > C; c.-103 A > G; p.P31L; c.293–13 A > G; p.G111Vfs*21; p.I173N; p.I237N-p.V238E-p.M240K; p.L308Ffs*6	***21A1P***: p.Q319*; p.R357W ***TNXA***: c.12524G > A; c.12514G > A; c.12,174 C > G	***21A1P***: p.V282L; p.R357W	[p.V282L]; [Conv] (CH-5 like)
3	Patient	NC	c.293–13 A > AG; p.G111Vfs*21; p.282 L	[T] (*21A1P* dup); [B] (Conv CH-1 like)	[p.V282L]; [Conv] (CH-1 like)	***21A2***: p.V282L	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c.293–13 A > G; p.G111Vfs*21	***TNXA***: c.12524G > A[1]	***21A1P***: p.V282L; p.Q319*; p.R357W.***TNXA***: c.12,174 C > G	[p.V282L]; [Conv] (CH-1 like)
4	Patient	CL (SW)	c.293–13 C > G; p.G111Vfs*21	[M] (CAH-X CH-1); [T] (Conv CH-1 like & 21A1P dup) or [B] (Conv CAH-X CH-1 like); [B] (Conv CH-1 like)	[CAH-X CH-1]; [Conv] (CH-1 like) or [Conv] (CAH-X CH-1 like); [Conv] (CH-1 like)	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c.293–13 C > G; p. G111Vfs*21; p.I173N; p.I237N-p.V238E-p.M240K; p.V282L; p. L308Ffs*6; p.R357W ***TNXB***: c.12524G > A; c.12218G > A; c.12,174 C > G c.11435_11524 + 30del	***21A2***: c.-126 C > T; c.-113 G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c. 293–13 C > G; p.G111Vfs*21	***21A1P***: p.Q319* ***TNXA***: c.12514G > A	None	[CAH-X CH-1]; [Conv] (CH-1 like) or [Conv] (CAH-X CH-1 like); [Conv] (CH-1 like)
5	Patient	CL (SW)	c.293–13 C > G; p.G111Vfs*21	[M] (CH-1);[M] (CH-1)	[CH-1]; [CH-1]	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c.293–13 C > G; p.G111Vfs*21	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c.293–13 C > G; p.G111Vfs*21	No amplification		[CH-1]; [CH-1]
6	Partner	N/A	None	[B]; [M] (*21A1P* del)	[WT]; [WT]	None	None	None	-	[WT]; [WT]
7	Partner	N/A	c.*13G > A; p.Q319*	[M] (*21A1P* del); [T] (*21A2* dup)	[c.*13G > A]; [p.Q319*;WT]	***21A2***: c.*13G > A	***21A2***: p.Q319* ***TNXB***: c.12,463 + 2T > C	***21A1P***: p.P31L; p.I173N; p.R357W ***TNXA***: c.12218G > A	***21A2*WT**	[c.*13G > A]; [p.Q319*; c.12463 + 2T > C;WT]
8	Patient	CL (SW)	p.G111Vfs*21	[M] (CH-5);[B]	[CH-5]; [p.G111Vfs*21]	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; p.P31L; c. 293–13 C > G; p.G111Vfs*21; p.I173N; p.I237N-p.V238E-p.M240K; p.L308Ffs*6	***21A2***: p.G111Vfs*21	None	-	[CH-5]; [p.G111Vfs*21] or [Conv] (CH-5 like); [p.G111Vfs*21]
9	Patient	CL (SW)	c.293–13 C > G; p.G111Vfs*21	[T] (*21A1P* dup); [B] (Conv CH-1 like) or [B]; [T] (Conv CH-1 like & *21A1P* dup)	[c.293–13 C > G]; [Conv] (CH-1 like)	***21A2***: c.293–13 C > G	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c.293–13 C > G; p.G111Vfs*21	None[1]	***21A1P***: p.V282L; p.R357W ***TNXA***: c.12,174 C > G	[c.293–13 C > G]; [Conv] (CH-1 like)
10	Patient	CL (SW)	p.L308Ffs*6	[M] (CH-5);[B]	[CH-5]; [p.L308Ffs*6]	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G;c.-4 C > T; p. P31L; c.293–13 C > G; p. G111Vfs*21; p.I173N; p.I237N-p.V238E-p.M240K; p.V282L; p. L308Ffs*6; p.Q319*; p.R357W ***TNXB***: c.12514G > A; c.12218G > A; c.12,174 C > G	***21A2***: p.L308Ffs*6	***TNXA***: c.12524G > A; c.12514G > A; c.12218G > A	-	[CAH-X CH-2]; [p.L308Ffs*6] or [Conv] (CAH-X CH-2 like); [p.L308Ffs*6]

ID	Condition	Presumptive phenotype	Previous results			LRS results				
						AMPLICON A		AMPLICON B		
			Pathogenic GVs by Sanger sequencing	RCCX arrangement by MLPA	Deduced genotype	Pathogenic GVs		Absence of typical pseudogene GVs/duplicated 21A2 pathogenic GVs		Deduced genotype
						HA1	HA2	HB1	HB2	
11	Patient	NC	None	[B] (Conv CH-5 like); [B] or [M] (CH-5);[T] (*21A1P* dup)	[Conv] (CH-5 like); [WT] or [CH-5]; [WT]	***21A2***: c.-126 C > T; c. -113G > A; c.-110T > C; c. -103 A > G;c.-4 C > T; p. P31L; p.H63L; c.293–13 C > G; p.G111Vfs*21; p.I173N; p.I237N-p. V238E-p.M240K; p. V282L; p.L308Ffs*6; p. Q319*	None	None	***21A1P*²**: p.R357WWT	[Conv] (CH-3 like);] or [CH-3]; [WT]
12	Patient	NC	p.V282L	[T] (*21A1P* dup); [B]	[p.V282L]; [p.V282L]	***21A2***: p.V282L	-	***21A1P***: p.V282L; p. R357W	***21A1P***: p.Q319* ***TNXA***: c.12524G > A; c. 12514G > A; c. 12,174 C > G	[p.V282L]; [p.V282L]
13	Patient	CL (SW)	c.293–13 C > G; p. G111Vfs*21	[M] (CH-1);[M] (CH-1)	[CH-1]; [CH-1]	***21A2***: c.-126 C > T; c. -113G > A; c.-110T > C; c. -103 A > G; c.-4 C > T; p. P31L; c.293–13 C > G; p. G111Vfs*21	***21A2***: c.-126 C > T; c.-113 G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c. 293–13 C > G; p.G111Vfs*21	No amplification		[CH-1]; [CH-1]
14	Patient	CL (SW)	c.293–13 C > G; p. G111Vfs*21; p. R445*	[B]; [M] (CH-1)	[p.R445*]; [CH-1]	***21A2***: p.R445*	***21A2***: c.-126 C > T; c.-113 G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c. 293–13 C > G; p.G111Vfs*21	***21A1P***: p.V282L; p. R357W	-	[p.R445*]; [CH-1] or [p.R445*]; [Conv] (CH-1 like)
15	Patient	NC	c.-126 C > T;c.-113 G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L	[B]; [B] (Conv CH-4 or CH-9 like)	[p.P31L]; [Conv] (CH-4 like)	***21A2***: c.-126 C > T; c. -113G > A; c.-110T > C; c. -103 A > G; c.-4 C > T; p. P31L	***21A2***: p.P31L	None	***TNXA***: c.12524G > A; c. 12514G > A; c. 12218G > A	[p.P31L]; [Conv] (CH-4 like)
16	Patient	NC	p.V282L	[T] (*21A1P* dup); [T] (21A1P dup)	[p.V282L]; [p.V282L]	***21A2***: p.V282L	-	***21A1P***: p.Q319*; p. R357W ***TNXA***: c.12524G > A; c. 12514G > A; c. 12,174 C > G	***21A1P***: p.V282L; p. R357W	[p.V282L]; [p.V282L]

Continued

ID	Condition	Presumptive phenotype	Previous results			LRS results				
						AMPLICON A		AMPLICON B		
						Pathogenic GVs		Absence of typical pseudogene GVs/duplicated 21A2 pathogenic GVs		
			Pathogenic GVs by Sanger sequencing	RCCX arrangement by MLPA	Deduced genotype	HA1	HA2	HB1	HB2	Deduced genotype
17	Patient	NC	p.Q319*	[T] (*21A2* dup); [M] (*21A1P* del)	[p.Q319*;WT]; [WT]	None	***21A2***: p.Q319* ***TNXB***: c.12,463 + 2T > C	***21A1P***: p.P31L; p. I173N; p.R357W ***TNXA***: c.12218G > A	**21A2 WT**	[WT]; [p.Q319*; c. 12463 + 2T > C;WT]
18	Patient	CL (SW)	c.293–13 C > G; p. G111Vfs*21	[B]; [B] (Conv CH-1 like)	[c.293–13 C > G]; [Conv] (CH-1 like)	***21A2***: c.293–13 C > G	***21A2***: c.-126 C > T; c.-113G > A; c. -110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c.293–13 C > G; p.G111Vfs*21	None	-	[c.293–13 C > G]; [CH-1] or [c.293–13 C > G]; [Conv] (CH-1 like)
19	Patient	CL (SV)	c.293–13 C > G; p. V282L; p. L308Ffs*6; p. Q319*; p.R357W	[B]; [B] (3' Conv)	[c.293–13 C > G]; [3' Conv]	***21A2***: c.293–13 C > G	***21A2***: p.V282L; p.L308Ffs*6; p. Q319*; p.R357W	***21A1P*²**: c.-4 C > T; p.P31L ***TNXA***: c.12524G > A	***21A1P***: p.R357W ***TNXA***: c. 12524G > A; c. 12514G > A; c. 12,174 C > G	[c.293–13 C > G]; [3' Conv]
20	Patient	NC	p.I237N-p. V238E-p.M240K; p. V282L; p. L308Ffs*6	[T] (dup *21A1P*);[B] (3' Conv)	[p.V282L]; [3' Conv]	***21A2***: p.V282L	***21A2***: p.I237N-p.V238E-p. M240K; p.V282L; p. L308Ffs*6	***21A1P***: p.Q319*; p.R357W	***21A1P*¹**: p.V282L; p. R357W ***TNXA***: c.12,174 C > G	[p.V282L]; [3' Conv]
21	Patient	NC	p.V282L	[T]; [M] (CH-5)	[p.V282L]; [CH-5]	***21A2***: p.V282L	***21A2***: c.-126 C > T; c.-113G > A; c. -110T > C; c.-103 A > G; p. P31L; c.293–13 A > G; p. G111Vfs*21; p.I173N; p. I237N-p.V238E-p. M240K; p.L308Ffs*6	***21A1P***: p.Q319*; p. R357W ***TNXA***: c.12524G > A	***21A1P***: p.V282L; p. R357W	[p.V282L]; [Conv] (CH-5 like)
22	Patient	CL (SV)	p.I173N	[M] (CAH-X CH-1);[B]	[p.I173N]; [CAH-X CH-1]	***21A2***: p.I173N	***21A2***: c.-126 C > T; c.-113 G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c. 293–13 A > G; p. G111Vfs*21; p.I173N; p. I237N-p. V238E-p.M240K; p.V282L; p.L308Ffs*6; p. Q319*; p.R357W ***TNXB***: c.12524G > A; c. 12514G > A; c.12218G > A; c. 12,174 C > G; c. 11435_11524 + 30del	None	-	[p.I173N]; [CAH-X CH-1] or [p.I173N]; [Conv] (CAH-X CH-1 like)

Continued

ID	Condition	Presumptive phenotype	Previous results			LRS results				
						AMPLICON A		AMPLICON B		
						Pathogenic GVs		Absence of typical pseudogene GVs/duplicated 21A2 pathogenic GVs		
			Pathogenic GVs by Sanger sequencing	RCCX arrangement by MLPA	Deduced genotype	HA1	HA2	HB1	HB2	Deduced genotype
23	Relative	N/A	c.293–13 C > G; p.G111Vfs*21; p.Q319*	[T] (*21A2* dup); [B] (Conv CH-1 like)	[p.Q319*; WT]; [Conv] (CH-1 like)	***21A2***: p.Q319* ***TNXB***: c.12,463 + 2T > C	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; c.293–13 C > G; p.G111Vfs*21	None	**21A2** WT	[p.Q319*; c.12463 + 2T > C;WT]; [CH-1] or [p.Q319*; c.12463 + 2T > C;WT]; [Conv](CH-1 like)
24	Patient	CL (SV)	c.293–13 C > G; p.H403Rfs*5	[B]; [M](*21A1P* del)	[c.293–13 C > G]; [p.H403Rfs*5]	***21A2***: c.293–13 C > G	***21A2***: p.H403Rfs*5	None	-	[c.293–13 C > G]; [p.H403Rfs*5]
25	Relative	N/A	c.293–13 C > G; p.Q319*	[T] (*21A2* dup); [B]	[c.293–13 C > G; p.Q319*]; [WT]	None	***21A2***: p.Q319* ***TNXB***: c.12,463 + 2T > C	***21A1P***[1]: p.I173N; p.V282L; p.R357W ***TNXA***: c.12,174 C > G	***21A2***: c.293-13 C > G	[WT]; [p.Q319*; c.12463 + 2T > C;c.293-13 C > G]; [WT]
26	Patient	CL (SW)	No amplification	[B] (Conv CH-5 like); [M] (CH-5)	[Conv] (CH-5 like); [CH-5]	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p.P31L; p.H63L; c.293–13 A > G; p. G111Vfs*21; p.I173N; p.I237N-p.V238E-p.M240K; p.V282L; p.L308Ffs*6; p.Q319*	-	***21A1P***[2]: p.R357W	-	[Conv] (CH-3 like); [CH-3]
27	Patient	NC	p.I173N; p.I237N-p.V238E-p.M240K; p.V282L; p.L308Ffs*6; p.Q319*	[B]; [B] (Conv CAH-X CH-1 like) or [T] (*21A1P* dup); [M] (CAH-X CH-1)	[CAH-X CH-1]; [WT] or [Conv] (CAH-X CH-1 like); [WT]	***21A2***: c.-126 C > T; c.-113G > A; c.-110T > C; c.-103 A > G; c.-4 C > T; p. P31L; c.293–13 C > G; p. I173N; p.I237N-p.V238E-p.M240K; p. V282L; p.L308Ffs*6; p.Q319* ***TNXB***: c.12524G > A; c.12218G > A; c.12,174 C > G; c.11435_11524 + 30del	None	***21A1P***: p.G111fs; p.R357W ***TNXA***: c.12,174 C > G	None	[CAH-X CH-1]; [WT] or [Conv] (CAH-X CH-1 like); [WT]
28	Patient	NC	p.V282L	[T] (*21A1P* dup); [T] (21A1P dup)	[p.V282L]; [p.V282L]	***21A2***: p.V282L	-	***21A1P***: p.Q319*; p. R357W ***TNXA***: c.12524G > A; c. 12514G > A; c.12,174 C > G	***21A1P***: p.V282L; p. R357W	[p.V282L]; [p.V282L]

Continued

ID	Condition	Presumptive phenotype	Previous results			LRS results				
						AMPLICON A		AMPLICON B		
						Pathogenic GVs		Absence of typical pseudogene GVs/duplicated 21A2 pathogenic GVs		
			Pathogenic GVs by Sanger sequencing	RCCX arrangement by MLPA	Deduced genotype	HA1	HA2	HB1	HB2	Deduced genotype
29	Patient	NC	p.R484P; p.V282L	[B]; [T] (*21A1P* dup)	[p.R484P]; [p.V282L]	***21A2***: p.R484P	***21A2***: p.V282L	None	***21A1P***: p.V282L; p.Q319*; p.R357W ***TNXA***: c.12,174 C > G	[p.R484P]; [p.V282L]
30	Patient	NC	p.I237N-p.V238E-p.M240K; p. V282L; p.L308Ffs*6	[B] (3' Conv); [T] (*21A1P* dup)	[3' Conv]; [p.V282L]	***21A2***: p.I237N-p.V238E-p. M240K; p.V282L; p.L308Ffs*6	***21A2***: p.V282L	***21A1P***[1]: p.Q319; p.R357W	***21A1P***: p.V282L; p.Q319; p.R357W ***TNXA***: c.12,174 C > G	[3' Conv]; [p.V282L]
31	Patient	NC	p.V282L	[T] (*21A1P* dup); [T] (*21A1P* dup)	[p.V282L]; [p.V282L]	***21A2***: p.V282L	-	***21A1P***: p.Q319*; p.R357W ***TNXA***: c.12524G > A; c.12514G > A; c.12,174 C > G	***21A1P***: p.V282L; p.R357W	[p.V282L]; [p.V282L]
32	Patient	NC	p.V282L; p.I173N	[B]; [T] (*21A1P* dup)	[p.I173N]; [p.V282L]	***21A2***: p.I173N	***21A2***: p.V282L	None[1]	***21A1P***: p.Q319*; p.R357W ***TNXA***: c.12524G > A; c.12514G > A; c.12,174 C > G	[p.I173N]; [p.V282L]

Continued

ID	Condition	Presumptive phenotype	Previous results			LRS results					
						AMPLICON A			AMPLICON B		
						Pathogenic GVs			Absence of typical pseudogene GVs/duplicated 21A2 pathogenic GVs		
			Pathogenic GVs by Sanger sequencing	RCCX arrangement by MLPA	Deduced genotype	HA1	HA2		HB1	HB2	Deduced genotype
33	Patient	NC	p.V282L	[T] (*21A1P* dup); [T] (*21A1P* dup)	[p.V282L]; [p.V282L]	***21A2***: p.V282L		-	***21A1P***: p.Q319*; p.R357W ***TNXA***: c.12524G > A; c.12514G > A; c.12,174 C > G	***21A1P***: p.V282L; p. R357W	[p.V282L]; [p.V282L]
34	Patient	CL (SV)	c.293–13 C > G	[B]; [B]	[c.293–13 C > G]; [c. 293–13 C > G]	***21A2***: c.293–13 C > G		***21A2***: c.293–13 C > G	None	-	[c.293–13 C > G]; [c. 293–13 C > G]

Table 1. Summary of the results obtained by LRS in the 34 selected samples and its comparison using Sanger. CAH: Congenital adrenal hyperplasia. CL: Classical, SW: salt-wasting, SV: simple virilizing, NC: Nonclassical. GVs: Genetic Variants. LRS: Long Reads sequencing. MLPA: Multiplex Ligation-dependent Probe Amplification. HA1 and HA2: Haplotype 1 and haplotype 2 for amplicon A; HB1 and HB2: Haplotypes 1 and 2 for amplicon B. WT: Wild Type. B. Bimodular, M: Monomodular, T: Trimodular. del: deletion; dup: duplication. *21A2*: *CYP21A2*. *21A1P*: *CYP21A1P*. Conv: conversion. 3' Conv: *CYP21A2-CYP21AP* conversion. CH-#: Chimera *CYP21A1P-CYP21A2*. CAH-X CH-#: Chimera *TNXA-TNXB*. N/A: Not Applicable. None: No variants found. A dash (-) denotes that only one haplotype could be phased, indicating that this patient is homozygous or hemizygous for the haplotype. 1: Heterozygous GVs in HB1. 2:*CYP21A1P pre*sented the p.63 L GV. The GVs are referenced to *CYP21A2*: RefSeq for cDNA: NM_000500.9; protein: NP_000491.4; *TNXB*: RefSeq for cDNA: NM_019105 and Ensembl transcript: ENST00000375244.3; protein: NP_001352205.1; 21A1P: Ensembl for non-coding transcript ENST00000354927; *TNXA*: Ensembl for non-coding transcript ENST00000507684.1.

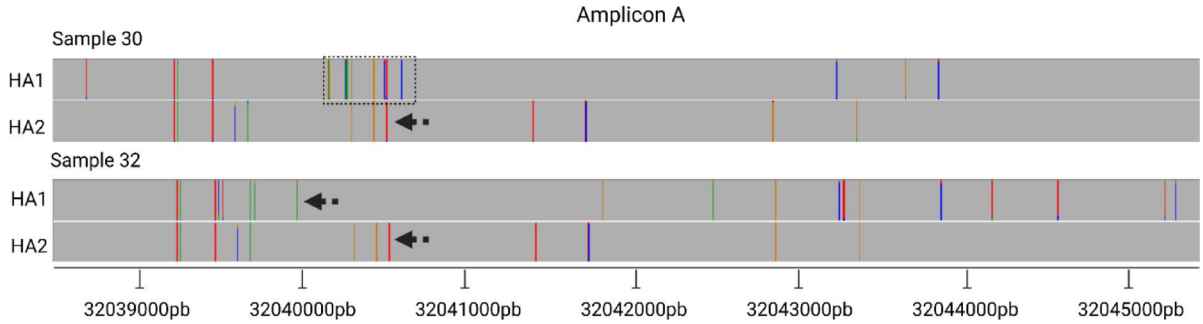

Fig. 3. Representative IGV plots showing *cis*/*trans* location of variants. Above: a 3' conversion *(CYP21A2/CYP21A1P* conversion) in HA1 (rectangle) and p.V282L in HA2 (arrow) for sample 30. Below: p.I173N in HA1 and p.V282L in HA2 (arrows) for sample 30. Genomic locations in hg38 reference sequence are also included below the plots.

the *CYP21A1P* and different lengths of the coding region, 3 alleles with a breakpoint at least at the p.L308Ffs*6 GV, lacked the expected c.-4 C > T and p.V282L GVs (samples 2, 8 and 21). In 1 allele (sample 4), the chimera reaches the *TNXB* gene, although the p.Q319* GV was not observed. Likewise, we observed a *TNXA/TNXB* converted allele lacking the p.G111Vfs*21GV as does the pseudogene of the same sample (sample 27). Similarly, some common GVs in *TNXA* are absent in the converted or chimeric alleles (Table 1).

By analyzing the entire region in a single read, the genomic region involved in the rearrangements could be more accurately defined (Fig. 4). For a more accurate definition of the breakpoints of the chimeras and converted genes, we also took advantage of other GVs differentially described in population databases compared to the active genes. This adds to the definition made commonly with the most frequent pathogenic GVs in the pseudogene to narrow the sequences involved in the rearrangements. In that sense, the start point of the breakpoint for CH-1 genes was narrowed to c.342C > T, for CH-5 to c.939 + 11G > C, and for the allele in sample 15 (a conversion CH-4 like) to c.292 + 33A > C. On the other hand, CAH-X CH-1 genes had a start and stop points at c.11417A > G and c.11387-9T, respectively, and in CAH-X CH-2, at c.11616G > A and c.11548 C. This analysis also allowed us to accurately define the breakpoints involved in alleles with a 3' conversion (*CYP21A2/CYP21A1P* rearrangements) and to determine that all the breakpoints were conserved for the different types of recombinant alleles, either chimeras or conversions.

It should be noted that after LRS the classification of the chimeras or the extent of the converted alleles had changed in some samples. One patient (sample 10) had been previously classified as a carrier of CH-5, but by LRS the rearrangement extends up to the *TNXB* reclassifying it as CAH-X CH-2. These new findings change the genetic counseling for the patient and their family related to the presence of a segregated allele for CAH-X. Likewise, samples 11 and 26 have been previously genotyped as having a converted allele and a chimera, both of them involving a *CYP21A1P*/*CYP21A2* rearrangement up to p.L308Ffs*6 GV as tested by MLPA. Once again, LRS allowed us to extend the breakpoints to p.Q319* and reclassified it to CH-3 instead of CH-5. Additionally, we could uncover that all the GVs from amplicon A in this sample are in homozygosity, despite the fact that one allele presented a conversion and the homologous allele a chimera. Although consanguinity is not evident for this family, we also found the less frequent p.H63L GV in both alleles and in the only one copy of the pseudogene.

We have previously observed that in many patients with a p.V282L/p.V282L genotype, the entire gene showed homozygosity for common non-pathogenic variants. After analyzing LR sequencing of some samples with this genotype (samples 12, 16, 28, 31, and 33), we conclude that the entire amplicon A presented homozygosity and thus phasing and haplotyping suggested the existence of only one allele. Amplicon B, on the other hand, could be separated into 2 different haplotypes in all the analyzed samples, in accordance with the putative arrangement of 2 RCCX modules per chromosome. In addition, we confirm that in an RCCX module conformation with a duplicated *CYP21A2*, either wild type or having the c.293–13 C > G GV, the p.Q319* GV in *CYP21A2* and the c.12,463 + 2T > C GV in *TNXB* are always in the centromeric copy.

Some discordant results were apparently found in the conformation of amplicon B for some previously genotyped samples (Table 1). For example, previous MLPA results of sample 1 showed a bimodular conformation of the RCCX module for both chromosomes. However, LRS output suggested that one of the alleles had a monomodular arrangement lacking *CYP21A1P*. Note, however, that the results may also be interpreted as a bimodular arrangement and homozygosity for amplicon B. Likewise, it is well documented that p.V282L is part of a known conserved haplotype with a pseudogene duplication[15–19]. Although MLPA results confirm this rearrangement for some of the samples studied, LRS showed mostly a bimodular arrangement. Nevertheless, pseudogene duplication can be inferred, in some instances, by the presence of heterozygous GVs in one of the haplotypes after phasing amplicon B.

By including the analysis of amplicon B, we also retrieved the frequencies of the different GVs in *CYP21A1P* and *TNXA* (Supplementary Table S2). Of note, 19 GVs for *CYP21A1P* and 28 GVs for *TNXA* were not previously described in Latino American populations with similar ethnic origin than in our country. We also found 1 GV in *TNXA*, n.322-82G > C, in 11 alleles that have not been reported in any of the consulted databases. After analyzing the most frequent pseudogene variants in the samples, all of their *CYP21A1P* pseudogenes have c.-126T; c.-113 A; c.-110 C; and c.-103G in the promoter region, c.293-13G in intron 2, p.237 N-p.238E-p.240 K cluster

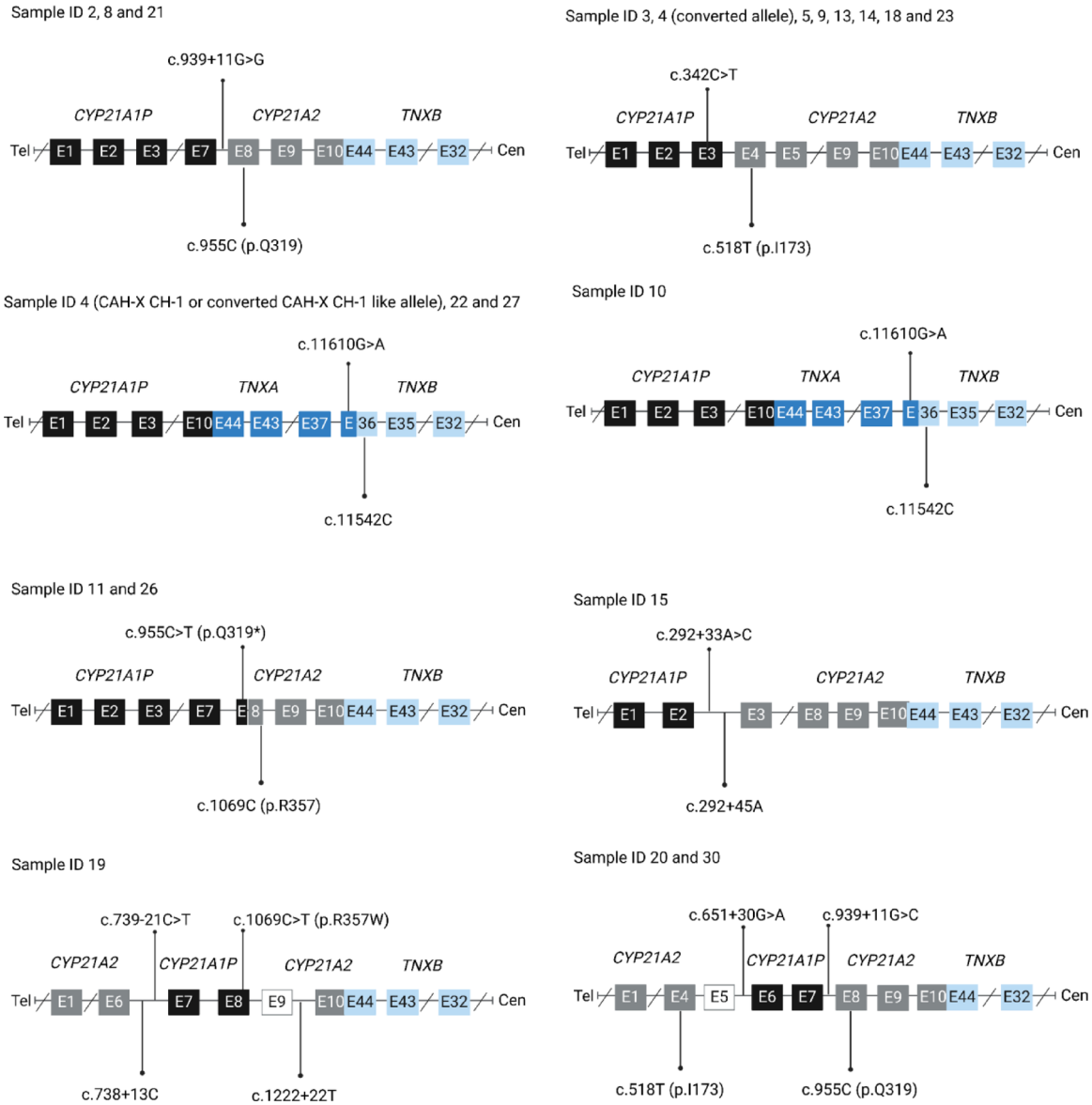

Fig. 4. Schematic representation of the breakpoints of the different chimeras and converted genes in the analyzed samples for Amplicon A. *CYP21A1P* exons are shown in black boxes, *CYP21A2* exons in gray boxes, *TNXA* exons in blue boxes and *TNXB* exons in light blue boxes. White boxes represent the uncertain origin of the sequence. Above the boxes are the GVs from *CYP21A1P* or *TNXA* that limit the breakpoints, and below the GVs from *CYP21A2* or *TNXB*. E: Exon. Tel: telomere. Cen: centromere.

(exon 6) and c.923dup (ins T in exon 7). Nevertheless, for the remaining common pseudogene GVs, we noticed a considerable degree of variability by means of their absence as part of the genomic sequences (Table 1). Besides the above mentioned allele lacking the p.G111Vfs*21 GV (n.331_332insGAGACTAC), 3 alleles lack the p.31 L GV (n.92T > C), 3 the p.173 N GV (n.510 A > T), 15 the p.282 L GV (n.836T > G), 13 the p.319* GV (n.948T > C) and 30 lack the p.357 W GV (n.1062T > C) GV. On the other hand, none of the *CYP21A1P* presented the p.454 S GV (n.1353 C > T), 3 alleles had the p.63 L GV (n.188 A > G), and the p.492 S GV (n.1474 A > G) variant was found in 21 alleles. In addition, all the *TNXA* genes have the 121 bp deletion, but similarly to the results for *CYP21A1P*, not all the alleles had the most common pathogenic variants. Indeed, 12 alleles lack the frequent c.12,524 A GV, 5 the c.12,218 A GV, 10 the c.12,514 A GV, and the c.12174G GV is absent in 16 alleles.

Finally, we are describing a novel GV: NM_000500.9:c.1208_1209del, p.H403Rfs*5 in a SV patient (sample 24). The NM_000500.9:c.1208_1209del GV has not been reported previously in neither population databases nor the literature. The GV was classified as pathogenic according to the ACMG guidelines as it introduces a change in the reading frame and the appearance of a premature stop codon 5 amino acids ahead (See Supplementary Text S1 and Supplementary Figure S1 for more details).

Discussion

In this work, we analyzed the utility of ONT third-generation sequencing for the study of GVs and structural rearrangements of the RCCX modules implicated in the CAH due to 21-HD and in CAH-X. We included samples previously studied in our laboratory along new ones, encompassing SW, SV and NC clinical forms. Although up to date a number of studies have already presented data using LRS in CAH patients[8,20–29], to our knowledge, this work represents the first study analyzing patients from Latin America, and the second one using ONT sequencing[8]. Our study also includes a detailed analysis of the RCCX module containing the *CYP21A1P* and *TNXA* pseudogenes, which contributes to elucidate the arrangement of this complex locus and adds valuable information of the distribution and frequencies of the GVs in Latin American populations, barely reported in clinical and population databases and in previous reports.

The human RCCX module represents a complex genomic region with a high variability in the number of copies per allele. Although 2/3 of the chromosomes analyzed have a duplicated RCCX module, monomeric arrangements and up to four copies of the module have been described[30]. In addition, pseudogene-derived variants are also very frequently found in disease-causing alleles due to gene conversion events. In this scenario, the presence of more than one pathogenic GV in *cis* is very frequent in 21-HD[12]. In conventional molecular methodologies, most of the *cis*/*trans* location of GVs requires the study of segregated parents, especially for NC patients, for accurate genetic counseling, and in many instances, they are not available. LRS has the advantage over classical molecular methodologies in that the study of segregated parents is not mandatory, as the *cis*/*trans* configuration of GVs can be detected in a single read, and therefore the genotype can be defined by analyzing only the proband.

By comparing ONT LRS outcomes with our previous results, we confirmed the presence of already characterized variants. Nevertheless, LRS adds information of GVs involved in the rearrangements that were not detected previously and/or were otherwise assumed. For example, we observed that common pseudogene-derived GVs were absent in a considerable number of the alleles. These observations point to the diversity of the sequences involved in accordance with the variability observed when the *CYP21A1P* gene was analyzed in our samples and in samples from other populations[31–33]. In line with these observations, we described a converted *TNXA/TNXB* allele lacking the p.G111Vfs*21GV in a patient with a NC phenotype (sample 27). This GV was also absent in the *CYP21A1P* gene from the same sample. Although this is a very rare haplotype (0.0024% individuals in GnomAD), the importance of this finding relies on the fact that the majority of the molecular diagnosis methods -including ours- take advantage of the absence of this 8 bp deletion to differentially amplify the *CYP21A2* gene. LRS overcomes this limitation, avoiding the amplification of a pseudogene lacking this GV and a false positive result. Indeed, we have previously genotyped this sample as having p.I173N; p.I237N-p.V238E-p.M240K; p.V282L; p.L308Ffs*6; and p.Q319* GVs, and therefore classified it as a 3' conversion (*CYP21A2/CYP21A1P* rearrangement). In contrast, MLPA showed a conversion up to *TNXB* with a conflicting result for the p.G111Vfs*21 GV suggesting a rare converted allele. LRS allowed us to verify a CAH-X CH-1 chimera and to conclude that the results found by Sanger sequencing may represent the amplification of the 3' fragment of the converted allele, but may also include the pseudogene lacking the p.G111Vfs*21 GV.

As it was mentioned above, chimeras and gene conversions represent a significant number of disease-causing alleles in 21-HD and in CAH-X as well[30,34]. Currently, most of these arrangements are detected by the application of MLPA in combination with Sanger sequencing to define the genotype of affected patients. Nevertheless, in this work and by applying LRS, we were able to define with a higher accuracy the breakpoints of the rearrangements adding to our understanding of gene conversion mechanisms. It also allows the (re)classification of the chimeras and/or converted genes in 3 out of 16 (18.75%) of the samples (Table 1). These results are in accordance with a recent report comparing standard molecular analysis and targeted LRS in CAH patients, improving the definition and classification of chimeras in 41.46% of 82 probands[21]. Similarly, a new chimera has recently been described using ONT LRS of 21-HD patients[8].

One important clinical result of our study is the finding of a previously CH-5 allele in a patient that was reclassified to a CAH-X CH-2 (sample 10). It should be noted that only CAH-X CH-1 can be detected using SALSA MLPA Probemix P050 CAH. Hence, patients with other *TNXA/TNXB* chimeras are missing by applying this method, triggering the search of pathogenic GVs in *TNXB* or other types of rearrangements only in the presence of clinical manifestations of CAH-X.

The most frequent pathogenic GV associated with alleles in NC patients is p.V282L[1]. Thus, homozygosity for this GV is quite common in this clinical form of 21-HD. We have previously observed that this homozygosity involved the entire *CYP21A2* gene in a considerable number of patients, and after LRS we confirm that it extends at least to the neighboring region of the *TNXB* gene that was analyzed in this study (exon 32–44) but not to the RCCX module containing the pseudogenes. These observations were possible by including the analysis of the telomeric module/s and the GVs in the *TNXB* gene, not routinely studied in molecular settings, and suggest a conserved haplotype for the active genes. Once again, the analysis of the telomeric module/s by LRS makes the application of MLPA unnecessary for exclusion of hemizygosity for these samples. A detailed analysis of additional samples using LRS could contribute to a more precise characterization of disease-causing alleles and to the knowledge of putative founder events. To add to the importance of the inclusion of LRS of amplicon B, the detection of duplicated *CYP21A2* is also possible considering that any duplicated module is expected to lie next to *TNXA* (Fig. 1) and, again, application of MLPA is not mandatory.

One of the limitations is that LRS of amplicons cannot separate haplotypes if both copies are homozygous and, therefore, homozygosity may be indistinguishable from a putative deletion of the region. Likewise, if an identical pseudogene is duplicated in the same allele, we cannot differentiate them. These reasons may account for some of the apparent discordant results found when comparing the use of Sanger sequencing and MLPA in defining the number of RCCX modules in the analyzed samples. Nevertheless, and although we do not phase more than 2 haplotypes for each amplicon, the presence of GVs in heterozygosis in one haplotype allowed us to posit that the

RCCX module containing the pseudogene is indeed duplicated in some samples. This represents an advantage of the inclusion of the analysis of the telomeric module/s (amplicon B) in our algorithm compared to other works analyzing only the module containing the active *CYP21A2* gene[8,24]. Even though the study of the duplicated pseudogene is not essential for a clinical analysis, LRS of both amplicons can be used to diagnose samples where MLPA testing is not performed. Nevertheless, and for a more accurate definition of the configuration of the modules, adaptive sampling developed by ONT may be a more suitable option[35,36]. Its drawback is that it allows the analysis of a reduced number of samples/runs in contrast to the utilization of amplicons.

Taking into account all of the above considerations, the utilization of LRS on 2 amplicons may represent not only an advantage in terms of accuracy in genotyping results but, also, it streamlines laboratory workflows, reducing staff time and overall operational demands. For our workflow in particular, LRS replaces the amplification of six different fragments, eight Sanger sequencing runs, the utilization of MLPA to detect chimeras and *CYP21A2* duplications and the need to perform PCRs and Sanger sequencing for segregation studies.

We also aimed at analyzing in detail the GVs in the *CYP21A1P* and *TNXA* pseudogenes, as its variability is mainly limited to the information deposited in population databases from whole exome or genome initiatives. Indeed, most of the other reports that may include LRS of the RCCX module containing the pseudogenes do not present data of their GVs[20–29]. In that sense, we have identified some GVs not previously described for Latin American populations and a variant not reported in any of the consulted databases. Although the number of alleles sequenced may not be high enough to give a more precise estimation of its frequencies, this data may add to the characterization of sequence diversity in different populations and should be considered as the basis for further analysis for the study of allele frequencies in individuals from our region and for the set-up of molecular studies.

Conclusions

To our knowledge, this is the first work in Latin America that uses LRS for analyzing a specific genomic region associated with human disease. We were able to confirm all the GVs previously observed by other methods, adding valuable information of the GVs involved, narrowing genetic breakpoints, and redefining genomic regions involved in rearrangements. It was also possible to determine with high confidence the phase of each allele, allowing for the first time for the analysis of GVs in neighboring genes with putative clinical implications in a single read. In addition, we present data of the frequencies of GVs in the RCCX module containing the pseudogenes, contributing to the characterization of this complex genomic region.

Materials and methods

Ethical approval

All the procedures performed in this study were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Written informed consent was obtained from all individuals recruited for this study. The protocol was approved by the ethics committee of the Administración Nacional de Laboratorios e Institutos de Salud (A.N.L.I.S.), Buenos Aires, Argentina.

Subjects and genotyping

We selected a total of 34 DNA samples from our cohort for the study (Table 1). Twenty-eight (samples 1–28) were previously genotyped in our laboratory. Briefly, DNA samples were initially amplified in 2 differentially and overlapping gene fragments from the promoter region to the p.I237N-p.V238E-p.M240K variants (cluster in exon 6) and from the p.G111Vfs*21 variant (8 bp deletion) in exon 3 to the 3' UTR in exon 10, respectively. These 2 fragments were further used as templates for a new round of specific gene amplification in 4 overlapping fragments and subject to Sanger sequencing in the forward and reverse orientation. When necessary, MLPA (SALSA MLPA Probemix P050 CAH MRC-Holland BV, Amsterdam, Holland) was also applied for selected samples to elucidate hemizygosity, homozygosity, large gene conversions, and/or duplications[12].

The selection of retrospective samples was based on their diversity in the RCCX module rearrangements and pathogenic GVs and comprised 12 samples from CL patients (3 SV, 9 SW), 11 NC, and 5 samples of relatives/partners/controls. In addition, we included 6 new DNA samples (samples 29–34) from 21-HD patients (1 SV and 5 NC) that were analyzed in a double-blind setting using ONT LRS (see below) in parallel with the methodologies described above.

ONT long read sequencing

Figure 2 summarizes the workflow applied for LRS and data analysis.

Long range PCR

For each sample, 2 different fragments (Amplicon A and B) of 8.5 kb were amplified using primers CYP779F and TENA32F or CYP779F and RP2R[37,38] (Fig. 1). Amplicon A includes the module containing *CYP21A2* and part of *TNXB* (exon 32–44 of *TNXB*) while amplicon B includes the module/s containing *CYP21A1P-TNXA-STK19B(RP2)* and/or *CYP21A2-TNXA-STK19B*(*RP2*). Long range PCRs were performed using 3 µL of 300 ng of genomic DNA, 3.75 UI of Expand Long Polymerase (Expand Long Template PCR System Kit, Roche®), 1.75 mM of Expand Long Template buffer 3 (Expand Long Template PCR System Kit, Roche®), 350 µM of dNTPs, 300 nM of each primer and PCR grade water up to 50 µL. Cycling conditions were: 94 °C denaturation for 2 min followed by 10 cycles of 94 °C denaturation for 10 s, 60 °C annealing for 30 s, 68 °C elongation for 9 min, followed by 25 cycles of 94 °C denaturation for 15 s, 60 °C annealing for 30 s, 68 °C elongation for 9 min adding 20 s cycle elongation for each successive cycle and a final elongation of 7 min at 68 °C. All of the PCR products were detected by electrophoresis on a 1.0% agarose gel and ethidium bromide stain.

Library preparation and sequencing

After amplification, each amplicon was purified using 1X AgencourtAMPure XP beads (Beckman Coulter®) and quantified using the Qubit dsDNA High Sensitivity Reagent (ThermoFisher Scientific®) while quality was evaluated using NanoDrop™ (ThermoFisher Scientific®). DNA repair and end-prep were performed using 200 fmol of each amplicon and NEBNext Ultra II End repair/dA-tailing Module (E7546-New England Biolabs®), Native barcode ligation was performed using NEB Blunt/TA Ligase Master Mix (M0367- New England Biolabs®).

Library preparation, barcoding and sequencing were performed with Native barcoding amplicons kit (EXP-NBD104, EXP-NBD114, and SQK-LSK109 versionNBA_9093_v109_revC_12Nov2019) using Nanopore Flow Cell R9.4.1 and Native barcoding Kit 24 V14 (SQK-NBD114.24) using Nanopore Flow Cell R10.4.1 following ONT recommendations. Amplicon A and B from each sample used different barcodes and libraries were sequenced using MinION or PromethION devices.

Data analysis

The analysis of the sequencing data was done with custom scripts based on the recommended ONT pipelines for the analysis of LRS. The data was basecalled and barcoded using Guppy version 6.5.7 (Oxford Nanopore Technologies Ltd., 2000)[39]. The obtained reads were aligned to their corresponding reference sequences using Minimap2 version 2.17[40] and Samtools version 1.3.1[41]. The generated BAM files were analyzed with Clair3 version 1.0.3[42] and Whatshap version 2.0[43] for variant calling and haplotyping. The called variants were filtered by their corresponding quality scores (Q-score) and depth (DP) before defining the haplotypes.

The Q-score (used for variant calling and base calling) is Phred-scaled, which means Q=-10*log10(p), where "p" is the estimated probability of the variant call or base call being wrong.

The reads were aligned to both reference sequences of each amplicon and the human genome (hg38). When aligning to the GRCh38 human genome reference sequence (hg38), amplicon A corresponds to positions chr6:32037620 to chr6:32046165, whereas amplicon B corresponds to positions chr6:32004884 to chr6:32013454. Variants were confirmed using the Integrative Genomics Viewer (IGV) based on BAM files, and any novel variant found was confirmed by Sanger sequencing.

Nomenclature of the GVs was done following the recommendations of the Human Genome Variation Society (HGVS)[44] using NC_000006.12 as a reference sequence, corresponding to the 6th chromosome of the GRCh38.p14 human genome reference (*CYP21A2*: RefSeq for cDNA: NM_000500.9 and Ensembl transcript: ENST00000418967; protein: NP_000491.4; *TNXB*: RefSeq for cDNA: NM_019105.8 and Ensembl transcript: ENST00000375244.3; protein: NP_001352205.1; *CYP21A1P*: Ensembl for non-coding transcript ENST00000354927; *TNXA*: Ensembl for non-coding transcript ENST00000507684.1; *STK19B*: GenBank assembly GCF_000001405.40). *CYP21A1P/CYP21A2* chimeras or converted genes were classified according to the different breakpoints involved in the junction[6,10]. *CYP21A2/CYP21A1P* rearrangements (5′ region of *CYP21A2* and the 3′ region of the *CYP21A1P* pseudogene) were designated 3' conversions.

The following population databases were consulted to retrieve information and frequencies for the different GVs: GnomAD (https://gnomad.broadinstitute.org/ version 2.1.1[45], 1000 genome project (https://www.internationalgenome.org/)[46], dbSNP (https://www.ncbi.nlm.nih.gov/snp/)[47] including dbGaP dataset derivate from Allele Frequency Aggregator (ALFA) project[48]. In addition, ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)[49], ClinGen (https://www.clinicalgenome.org/)[50], DECIPHER (https://www.deciphergenomics.org/)[51] and PDB (Protein Data Bank: https://www.rcsb.org/) were also used to analyse the GVs using the 4Y8W structure as a template. Novel variants found in active genes were deposited in ClinVar (submission ID: SUB14707012), with their classification done following the recommendations of the American College of Medical Genetics (ACMG) guideline[52].

Data availability

Data Availability StatementThe original contributions presented in the study are included in the article/supplementary material. The sequencing data and analysis scripts used in this study are available on GitHub at https://github.com/EmilioKolo/nanopore_ib3. Raw data was deposited at NCBI, Sequence Read Archive (SRA), accession number PRJNA1242071. Further inquiries can be directed to the corresponding author.

Received: 21 March 2025; Accepted: 22 May 2025
Published online: 10 July 2025

References

1. Claahsen-van derGrinten, H. L. et al. Congenital adrenal hyperplasia-current insights in pathophysiology, diagnostics, and management. *Endocr. Rev.* **43**, 91–159 (2022).
2. White, P. C., New, M. I. & Dupont, B. Structure of human steroid 21-hydroxylase genes. *Proc. Natl. Acad. Sci. U S A.* **83**, 5111–5115 (1986).
3. Donohoue, P. A. et al. Gene conversion in salt-losing congenital adrenal hyperplasia with absent complement C4B protein. *J. Clin. Endocrinol. Metab.* **62**, 995–1002 (1986).
4. Higashi, Y., Tanae, A., Inoue, H. & Fujii-Kuriyama, Y. Evidence for frequent gene conversion in the steroid 21-hydroxylase P-450(C21) gene: implications for steroid 21-hydroxylase deficiency. *Am. J. Hum. Genet.* **42**, 17–25 (1988).
5. Simonetti, L. et al. CYP21A2 mutation update: comprehensive analysis of databases and published genetic variants. *Hum. Mutat.* **39**, 5–22 (2018).
6. Chen, W. et al. Junction site analysis of chimeric CYP21A1P/CYP21A2 genes in 21-hydroxylase deficiency. *Clin. Chem.* **58**, 421–430 (2012).
7. Lao, Q., Burkardt, D. D., Kollender, S., Faucz, F. R. & Merke, D. P. Congenital adrenal hyperplasia due to two rare CYP21A2 variant alleles, including a novel attenuated CYP21A1P/CYP21A2 chimera. *Mol. Genet. Genomic Med.* **11**, e2195 (2023).
8. Adachi, E. et al. A MinION-based Long-Read sequencing application with one-step PCR for the genetic diagnosis of 21-Hydroxylase deficiency. *J. Clin. Endocrinol. Metab.* **109**, 750–760 (2024).

9. Lee, H. H., Lee, Y. J. & Chao, M. C. Comparing the Southern blot method and polymerase chain reaction product analysis for chimeric RCCX detection in CYP21A2 deficiency. *Anal. Biochem.* **399**, 293–298 (2010).
10. Carrozza, C., Foca, L., De Paolis, E. & Concolino, P. Genes and pseudogenes: complexity of the RCCX locus and disease. *Front. Endocrinol.* **12**, 709758 (2021).
11. Miller, W. L. & Merke, D. P. Tenascin-X, congenital adrenal hyperplasia, and the CAH-X syndrome. *Horm. Res. Paediatr.* **89**, 352–361 (2018).
12. Fernández, C. S. et al. Genetic characterization of a large cohort of Argentine 21-hydroxylase deficiency. *Clin. Endocrinol.* **93**, 19–27 (2020).
13. Tyson, J. R. et al. MinION-based long-read sequencing and assembly extends the reference genome. *Genome Res.* **28**, 266–274 (2018).
14. Karaoğlan, M., Nacarkahya, G., Aytaç, E. H. & Keskin, M. Challenges of CYP21A2 genotyping in children with 21-hydroxylase deficiency: determination of genotype-phenotype correlation using next generation sequencing in southeastern Anatolia. *J. Endocrinol. Invest.* **44**, 2395–2405 (2021).
15. Laron, Z. et al. Late onset 21-hydroxylase deficiency and HLA in the Ashkenazi population: a new allele at the 21-hydroxylase locus. *Hum. Immunol.* **1**, 55–66 (1980).
16. Pollack, M. et al. (ed, S.) HLA linkage and B14, DR1, BfS haplotype association with the genes for late onset and cryptic 21-hydroxylase deficiency. *Am. J. Hum. Genet.* **33** 540–550 (1981).
17. Kohn, B. et al. Late-onset steroid 21-hydroxylase deficiency: a variant of classical congenital adrenal hyperplasia. *J. Clin. Endocrinol. Metab.* **55**, 817–827 (1982).
18. Werkmeister, J. W., New, M. I., Dupont, B. & White, P. C. Frequent deletion and duplication of the steroid 21-hydroxylase genes. *Am. J. Hum. Genet.* **39**, 461–469 (1986).
19. Garlepp, M. J., Wilton, A. N., Dawkins, R. L. & White, P. C. Rearrangement of 21-hydroxylase genes in disease-associated MHC supratypes. *Immunogenetics* **23**, 100–105 (1986).
20. Zhang, R. et al. Evaluating the efficacy of a long-read sequencing-based approach in the clinical diagnosis of neonatal congenital adrenocortical hyperplasia. *Clin. Chim. Acta*. **555**, 117820 (2024).
21. Wang, Y. et al. High clinical utility of long-read sequencing for precise diagnosis of congenital adrenal hyperplasia in 322 probands. *Hum. Genomics*. **19**, 3 (2025).
22. Lan, T. et al. Comparison of long-read sequencing and MLPA combined with long-PCR sequencing of mutations in patients with 21-OHD. *Front. Genet.* **15**, 1472516 (2024).
23. Wang, R. et al. Long-Read sequencing solves complex structure of CYP21A2 in a large 21-hydroxylase deficiency cohort. *J. Clin. Endocrinol. Metab.* **110**, 406–416 (2025).
24. Tantirukdham, N. et al. Long-read amplicon sequencing of the CYP21A2 in 48 Thai patients with steroid 21-Hydroxylase deficiency. *J. Clin. Endocrinol. Metab.* **107**, 1939–1947 (2022).
25. Zhang, X. et al. Targeted long-read sequencing for comprehensive detection of CYP21A2 mutations in patients with 21-hydroxylase deficiency. *J. Endocrinol. Invest.* **47**, 833–841 (2024).
26. Li, H. et al. Long-read sequencing: an effective method for genetic analysis of CYP21A2 variation in congenital adrenal hyperplasia. *Clin. Chim. Acta*. **547**, 117419 (2023).
27. Liu, Y. et al. Comprehensive analysis of congenital adrenal hyperplasia using long-read sequencing. *Clin. Chem.* **68**, 927–939 (2022).
28. Yuan, D. et al. Improved genetic characterization of congenital adrenal hyperplasia by long-read sequencing compared with multiplex ligation-dependent probe amplification plus Sanger sequencing. *J. Mol. Diagn.* **26**, 770–780 (2024).
29. Zhang, X. et al. Chimeric CYP21A1P/CYP21A2 genes in 21-Hydroxylase deficiency detected by Long-Read sequencing and phenotypes correlation. *J. Clin. Endocrinol. Metab.* https://doi.org/10.1210/clinem/dgae819 (2025).
30. Chen, W. et al. Complement component 4 copy number variation and CYP21A2 genotype associations in patients with congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Hum. Genet.* **131**, 1889–1894 (2012).
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*. **424**, 85–89 (2013).
32. Tsai, L. P., Cheng, C. F., Chuang, S. H. & Lee, H. H. Analysis of the CYP21A1P pseudogene: indication of mutational diversity and CYP21A2-like and duplicated CYP21A2 genes. *Anal. Biochem.* **413**, 133–141 (2011).
33. Cantürk, C. et al. Sequence analysis of CYP21A1P in a German population to aid in the molecular biological diagnosis of congenital adrenal hyperplasia. *Clin. Chem.* **57**, 511–517 (2011).
34. Koppens, P. F. J., Hoogenboezem, T. & Degenhart, H. J. 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 (2002).
35. Miller, D. E. et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).
36. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods.* **13**, 751–754 (2016).
37. Lee, H. H., Lee, Y. J. & Lin, C. Y. PCR-based detection of the CYP21 deletion and TNXA/TNXB hybrid in the RCCX module. *Genomics* **83**, 944–950 (2004).
38. Parajes, S., Quinteiro, C., Domínguez, 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.* **3**, e2138 (2008).
39. Welcome to Oxford Nanopore Technologies. *Oxford Nanopore Technologies* https://nanoporetech.com/.
40. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
41. Releases · samtools/samtools. *GitHub* https://github.com/samtools/samtools/releases
42. Zheng, Z. et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* **2**, 797–803 (2022).
43. Martin, M. et al. WhatsHap: fast and accurate read-based phasing. *BioRxiv* **085050** https://doi.org/10.1101/085050 (2016).
44. den Dunnen, J. T. et al. HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mutat.* **37**, 564–569 (2016).
45. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2023).
46. Dunham, I., Kulesha, E., Iotchkova, V., Morganella, S. & Birney, E. FORGE: A tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions. *F1000Research* **4**, 18 (2015).
47. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
48. ALFA. Allele Frequency Aggregator. www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/.
49. Landrum, M. J. et al. .ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
50. Rehm, H. L. et al. ClinGen–the clinical genome resource. *N Engl. J. Med.* **372**, 2235–2242 (2015).
51. Firth, H. V. et al. Database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
52. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.* **17**, 405–424 (2015).
53. Revista Medicina. pdf. *Google Docs* https://drive.google.com/file/d/1tU1D5vlPFssFHMoCF9n7NMk66tofo_G4/view?usp = embed_facebook (2024).

54. Claps, A. et al. High precision characterization of RCCX rearrangements in 21-hydroxylase Argentine patients using Oxford nanopore long read sequencing. *Revista Med. Vol.* **83**, 64 (2023).
55. Claps, A. et al. High precision characterization of Rccx rearrangements in a 21-hydroxylase deficiency Latin American cohort using Oxford nanopore long read sequencing. *MedRxiv* https://doi.org/10.1101/2024.11.14.24317161 (2024).

Acknowledgements
We thank Dr Verónica Ferreiro for her helpful discussion of the results. AC was a fellow of the Salud Investiga Program, Ministerio de Salud de la Nación. JEK and MLI are Ph.D fellows from the National Research Council (CONICET). NM, LK and LD are researchers from the National Research Council (CONICET). AC, TC, JL, MT and LD are professional staff from the Administración Nacional de Laboratorios e Institutos de Salud (ANLIS).

Author contributions
Conceptualization: LD, LK, FF and MT; Performed the experiments: AC, JEK, MT, CF, TC, JL, and MD; Data analysis and interpretation: AC, JEK, MT, and LD; Bioinformatics: LK, JEK, MLI, and NM; Writing—original draft preparation: AC, JEK, MT and LD; Writing—review and editing: All the authors.

Funding
This work was supported by grants from the University of Buenos Aires (PIDAE 2020 and PIDAE 2022), from the National Agency for the Promotion of Science and Technology (ANPCyT, PICT-2021-CAT-II-00018), and from the Administración Nacional de Laboratorios e Institutos de Salud (FOCANLIS 2022, NRU:2022-5). Partial results were presented as a conference abstract[53,54]. In addition, a preprint has been published[55].

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03799-7.

Correspondence and requests for materials should be addressed to L.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.