NR5A1 c.991-1G>C splice-site variant causes familial 46,XY partial gonadal dysgenesis with incomplete penetrance

Running title: Recurrent *NR5A1* splice-site variant and PGD

Maris Laan[1]*, Laura Kasak[1], Kęstutis Timinskas[2], Marina Grigorova[1], Česlovas Venclovas[2], Alexandre Renaux[3,4,5], Tom Lenaerts[3,4,5], Margus Punab[6,7]

[1] Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia

[2] Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania

[3] Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles - Vrije Universiteit Brussel, Brussels, Belgium

[4] Machine Learning Group, Université libre de Bruxelles, Brussels, Belgium

[5] Artificial Intelligence lab, Vrije Universiteit Brussel, Brussels, Belgium

[6] Andrology Center, Tartu University Hospital, Tartu, Estonia

[7] Institute of Clinical Medicine, University of Tartu, Tartu, Estonia

*** Corresponding author:**

Maris Laan, Institute of Biomedicine and Translational Medicine, University of Tartu

Ravila St. 19, 50411 Tartu, Estonia; tel: +372-7375008; +372-53495258, email: maris.laan@ut.ee

Manuscript length (Introduction to Discussion): 3,211 words

Abstract: 235 words

Tables: 3

Figures: 3

Acknowledgments

Eveliis Koppel and Eve Laasik are thanked for technical assistance in DNA extractions and Sanger sequencing. Andrew Sinclair is thanked for critical reading and commenting an early version of the manuscript. This study was supported by Estonian Research Council (grant IUT34-12 to ML), Research Council of Lithuania grant (09.3.3-LMT-K-712-01-0080 to ČV), the European Regional Development Fund (ERDF), the Brussels-Capital Region-Innoviris within the framework of the Operational Program 2014-2020 through the ERDF-2020 project ICITY-RDI.BRU (27.002.53.01.4524) and the Fondation de la Recherche Scientifique (FNRS-F.R.S) through the research credit 35276964 (to TL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

A Conflict of Interest Statement

The authors declare no conflict of interest.

A Data Availability Statement

The authors declare that all the data supporting the findings of this study are available within the paper and its supporting information files, excluding identifiable patient data.

SUMMARY

Objective: The study aimed to identify the genetic basis of partial gonadal dysgenesis (PGD) in a non-consanguineous family from Estonia. **Patients:** Cousins P(proband)1 (12 years; 46,XY) and P2 (18 years; 46,XY) presented bilateral cryptorchidism, severe penoscrotal hypospadias, low bitesticular volume; and azoospermia in P2. Their distant relative, P3 (30 years; 46,XY) presented bilateral cryptorchidism and cryptozoospermia. **Design:** Exome sequencing was targeted to P1-P3 and five unaffected family members. **Results:** P1-P2 were identified as heterozygous carriers of *NR5A1* c.991-1G>C. *NR5A1* encodes the steroidogenic factor-1 essential in gonadal development and specifically expressed in adrenal, spleen, pituitary and testes. Together with a previous PGD case from Belgium (Robevska et al. 2018), c.991-1G>C represents the first recurrent *NR5A1* splice-site mutation identified in patients. The majority of previous reports on *NR5A1* mutation carriers have not included phenotype-genotype data of the family members. Segregation analysis across three generations showed incomplete penetrance (<50%) and phenotypic variability among the carriers of *NR5A1* c.991-1G>C. The variant pathogenicity was possibly modulated by rare heterozygous variants inherited from the other parent, *OTX2* p.P134R (P1) or *PROP1* c.301_302delAG (P2). For P3, the pedigree structure supported a distinct genetic cause. He carries an undescribed likely pathogenic variant SOS1 p.Y136H. *SOS1*, critical in Ras/MAPK signaling and fetal development, is a strong novel candidate gene for cryptorchidism. **Conclusions:** Detailed genetic profiling facilitates counseling and clinical management of the probands and unaffected mutation carriers in the family for their reproductive decision-making.

Keywords from the *Clinical Endocrinology* webpage: clinical medicine, clinical endocrinology, genetics, testis, pediatrics, pituitary

Specific keywords: partial gonadal dysgenesis, exome sequencing, *NR5A1*, splice-site variant, incomplete penetrance, *SOS1*, patient management

INTRODUCTION

Genital anomalies of newborns with a 46,XY karyotype are prevalent developmental disorders, including hypospadias (0.2-0.5% of males), cryptorchidism (~2-9%), and other testicular abnormalities.[1, 2] Extreme cases may present ambiguous genitalia or even complete sex reversal. So far, around 30 confident genes implicated in genital anomalies have been identified. Known genetic factors causing genital anomalies are mutations in loci implicated in early gonadal development, such as specific transcription factors (e.g. *SRY, NR5A1, WT1*), signaling and endocrine regulators (e.g. *AR, SRD5A2, HSD17B3*).[3] The diagnostic yield of targeted gene panel sequencing applied to 46,XY patients with genital anomalies has been reported ~40%.[4, 5] There is accumulated knowledge that a large proportion of these cases are caused by *de novo* pathogenic variants, present incomplete penetrance and variable phenotype in families.[6, 7] Availability of advanced sequencing technologies has not only expanded the knowledge of genetics underlying genital dysgenesis, but also provided emerging evidence that the penetrance of a primary pathogenic variant may depend on the modulatory effect of other genes.[8] The largest study on 46,XY patients with genital anomalies (n=278) identified a likely genetic diagnosis in 118 patients (43%), whereas every tenth case carried two or more contributing variants (n=13; 4.7%).[5] Others have predicted the proportion of cases with oligogenic causes to reach even 50% or more.[4, 9]

In clinical practice, determination of the genetic cause in each patient is critical not only for optimal lifelong management but also to identify variant carriers among close relatives and to counsel them in reproductive and general health-related issues. Accurate estimation of variant penetrance has been challenging as the majority of the reported cases are sporadic or lack proper clinical anamnesis and genetic testing among relatives. There are only limited genetic studies on large pedigrees with several non-sibling cases affected with genital dysgenesis.[6, 10, 11] We aimed to dissect the genetic cause(s) in a three-generation non-consanguineous Estonian family with a complex history of partial gonadal dysgenesis (PGD) and other reproductive phenotypes.

MATERIALS AND METHODS

Ethics approval and participants' consent

The study was approved by the Ethics Review Committee of Human Research of the University of Tartu, Estonia (approval date 254/M-17, 21.12.2015). Written informed consent for evaluation and use of their clinical data for scientific purposes was obtained from each person prior to recruitment. The study was carried out in compliance with the Helsinki Declaration.

Clinical profiling of the probands and their family members

The three probands, P1 (aged 12 years at recruitment), P2 (18), and P3 (30), representing first cousins and their joint first cousin once removed (III-1, III-2, II-1, respectively; **Figure 1**) were managed by M.P. at the Andrology Centre, Tartu University Hospital (AC-TUH), Estonia. The primary cause of the subjects to seek andrological consultation were genital anomalies (**Table 1)**. The applied routine andrological pipeline at the AC-TUH to document the epidemiological, laboratory and clinical examination data of patients has been detailed in[12] and **Supporting Information**. Andrological workup of the probands excluded chromosomal abnormalities, *AZF* microdeletions, hypogonadotropic hypogonadism, testicular diseases, androgen abuse, severe trauma or operation in the genital area, chemo- and radiotherapy.

The recruited family members were sampled for venous blood from the cubital vein for genetic testing and analysis of circulating hormone levels (**Table 1, Table S1**). Nuclear family members of P1-P2 and their grandmother consented to genetic testing. The data and DNA of the parents and brother of P3 were unavailable.

Exome sequencing, data analysis, variant prioritization and validation

Genomic DNAs from the three probands P1, P2, P3 (III-1, III-2, II-1; **Figure 1**) and five available immediate family members (II-4, II-5, II-6, II-7, III-3) were subjected to exome sequencing (ES). The applied ES data generation and analysis pipeline is detailed in[13] and in **Supporting Information**. Briefly, wet-lab processing, base calling of the raw sequencing data (Illumina HiSeq 2500; San Diego, CA), primary sequence analysis and variant calling was performed at the Institute for Molecular Medicine Finland (FIMM) NGS Service (**Supporting Information**).

Population sampling probability (PSAP) pipeline[14] was applied in order to prioritize potential causative variants from the ES data. It is a model-based framework to evaluate the significance of genotypes ascertained from a single case by determining the by-chance probability of sampling the detected genotypes in the unaffected population based on the pathogenicity scores and observed frequencies of variants (see details in **Supporting Information**). Implementation of the PSAP pipeline for the filtering of the Variant Call Format (VCF) data files resulted in 11,763 - 13,198 (median 12,958) variants for each ES dataset (**Table S2**). The variants in the PSAP output were further sequentially prioritized to satisfy the criteria: (i) the PSAP statistical significance value, termed as popScore $P<0.005$ (162-185 retained variants/exome); (ii) high confidence undescribed or rare variants with minor allele frequency (MAF) in general population <0.0005 (55-72); (iii) omitting synonymous variants (39-54). After applying the Combined Annotation Dependent Depletion (CADD) tool[15] to rank the deleteriousness of variants (C-score >20), the final list of prioritized variants in the probands (P1 and P2, n=35; P3, n=32) was inspected manually using scientific literature and genome databases (**Tables S3-5**). Based on the family history, the following disease inheritance models were considered: autosomal dominant with incomplete penetrance or autosomal recessive (homozygous, compound heterozygous). Classification of prioritized variants as 'pathogenic', 'likely pathogenic', 'variant of uncertain significance' (VUS), 'likely benign', 'benign' and subsequent ranking followed the recommendations of the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology.[16] For the pathogenicity assessment of the *SOS1* variant, ClinGen's RASopathy Expert Panel Consensus Methods for Variant Interpretation were applied.[17]

Variants highlighted in the exomes of index patients were experimentally confirmed using Sanger sequencing. PCR and sequencing primers are presented in **Table S6**. The analysis included eight family members initially assessed by ES, plus the grandmother of P1 and P2.

In silico protein sequence and structure analysis of *SOS1* p.Y136H

Full description of the *in silico* assessment of the protein sequence and structure of SOS1 along with detailed methodological references is provided in **Supporting Information**. Briefly, homologs of the human SOS1 protein (Uniprot id Q07889) were identified in Uniclust90 database using HMMER, a

software for biosequence analysis using profile hidden Markov models (profile HMMs).[18] Multiple protein sequence alignments were produced with MAFFT (l-INS-i) programme.[19] Visualization and analysis of structures and surface properties of SOS1, as well as modeling of p.Y136H variant was performed using UCSF Chimera platform.[20] Molecular interactions in experimentally determined structures of macromolecular complexes were searched and analyzed using the PPI3D server.[21]

Statistical testing the hypothesized digenic variant combinations using the ORVAL platform

The ORVAL (Oligogenic Resource for Variant AnaLysis) platform was implemented for computational modelling and estimating of the probability of a hypothesized combined pathogenic effect of the variant pairs *NR5A1* c.991-1G>C and *OTX2* p.P134R in P1 and *NR5A1* c.991-1G>C and *PROP1* c.301_302delAG in P2.[22] This *in silico* prediction tool quantifies the potential pathogenicity of targeted digenic variant combination(s) and provides various exploratory bioinformatics tools to interpret the outcome. The pathogenicity scores of tested bi-locus variant combinations were computed by VarCoPP with a threshold >95% confidence zone.[23] Details of ORVAL are provided in **Supporting Information**.

RESULTS

Family medical history

The analyzed clinical case represents a three-generation family including several members diagnosed with various forms of partial gonadal dysgenesis (PGD). Proband 1 (P1, III-1; **Figure 1A**), aged 12 years at the first clinical assessment, and P2 (III-2) aged 18 years, are first cousins with similar clinical symptoms of congenital genital anomalies. They presented bilateral cryptorchidism (operated), severe penoscrotal hypospadias (operated), bitesticular volume ≤8 ml, extremely elevated FSH and LH, as well as confirmed azoospermia in P2 (**Table 1**, **Figure 1A**). P1 was born at 36 weeks' gestation from a medically induced delivery due to complicated pregnancy characterized by severe premature placental ageing and fetal growth restriction. He was also diagnosed with asthma in childhood. General postnatal development and puberty progression of both patients has been normal. Their joint first cousin once removed, P3 (II-1) aged 30 years, also presented bilateral cryptorchidism, but with less severe testicular dysfunction – cryptozoospermia (sperm count $0.2x10^{-6}$ per ejaculate). Unlike P1 and P2, he had neither

hypospadias nor extremely low testes size (bi-testicular volume 30 ml). No apparent signs of genital dysgenesis were present in other male family members available for andrological phenotyping (brother of P2, fathers of P1 and P2; **Table S1**).

Medical history of the family across the three generations included other reproductive complications, such as an unexplained stillbirth (sibling of the grandmother of P1 and P2) and an induced termination of a 2nd trimester pregnancy due to medical indications (unborn sibling of P2; **Figure 1A**). The grandmother of P1-P2 reported a family history of sub/infertile members in four preceding generations. Notably, her first child had been born with an ambiguous sex and diagnosis of purpura fulminans and had died at the age of 2.

NR5A1 c.991-1G>C splice-site variant exhibits variable penetrance in three generations

A heterozygous splice-site variant (c.991-1G>C) in a well-known gene *nuclear receptor subfamily 5 group A member 1* (*NR5A1*) implicated in genital dysgenesis was identified by exome sequencing (ES) as the primary causative variant in cases P1 and P2 (**Table 2; Tables S3-4, S7**). *NR5A1* encodes a transcription factor steroidogenic factor-1 (SF-1) that is critical for gonadal development in both sexes, acting in dose-dependent manner.[24] In adults, it is specifically expressed in adrenal gland, spleen, pituitary and testes. *NR5A1* c.991-1G>C has not been reported in human genome databases, but was described in a heterozygous state in a Belgian PGD patient (46,XY karyotype, but reared as female), presenting clitoromegaly, abdominal bilateral testes and no internal female structures.[25] It is predicted to affect the intron 5 splice acceptor site and potentially lead to the usage of a new acceptor 6 bp downstream in exon 6 and an in-frame deletion p.Gly330_Ser331del in the ligand-binding and dimerization domain of SF-1 (**Figure 1B**).

P1 inherited the variant from his mother (II-5) and P2 from his father (II-6), asymptomatic for the genital phenotypes (**Figure 1A**). The variant originates from the grandmother (I-3) of P1 and P2, and is also carried by the brother of P2. At the age of 17 years he (III-3) presented no apparent clinical phenotype, but his bitesticular volume (28 ml; **Table S1**) was lower than typical for young men of his age (median 50.0, range 35.0–70.0 ml).[26] As his sperm analysis was not available, sub/infertility in adulthood cannot be fully excluded. In the literature, *NR5A1* mutations have been also reported in cases with severe

spermatogenic failure without genital dysgenesis.[27] Notably, he and the mother of P1 (II-5) have been diagnosed with depression and anxiety disorder, stress-induced panic attacks and dizziness, consistent with a previous observation on asymptomatic carriers of *NR5A1* pathogenic variants.[28] In addition, the mother of P1 (II-5) had experienced late menarche at 15 years of age. Even assuming that the family member presenting with ambiguous sex at birth (II-3; died at the age of 2 and unavailable for testing) was a highly likely carrier of the *NR5A1* c.991-1G>C variant (**Figure 1A**), the penetrance of the variant is estimated to be less than 50%.

So far, only eight clinical cases with different *NR5A1* splicing variants have been reported that were either sporadic (*de novo*) or with unavailable family data (**Table 3**). None of these variants have been reported in the general population. All these subjects exhibited 46,XY PGD with variably positioned bilateral undescended testes and other severe abnormalities of external genitalia, mostly without any detected internal female structures. The variant c.991-1G>C detected in P1 and P2 is the first recurrent *NR5A1* splice variant, identified in three PGD cases. It is also the first *NR5A1* splice variant shown to segregate through three generations due to its incomplete penetrance and variable expressivity.

Hypothesized co-contributing variants in genes encoding transcription factors upstream of *NR5A1*

P1 carried an additional, paternally inherited rare heterozygous variant in the gene *orthodenticle homeobox 2* (*OTX2*; c.425C>G, p.P134R; rs199761861; gnomAD MAF=2.10×10^{-4}) (**Table 2, Figure 1A, Figure S1**). This mutation has been previously reported in a Dutch pediatric male patient with pituitary malformation and an underdeveloped left optic nerve, also inherited from an asymptomatic father.[29] OTX2 acts as a pleiotropic transcription factor in the development of brain, eyes, craniofacial structures and male genitalia, and heterozygous *Otx2*-null male mice have impaired fertility and significantly reduced testes weight.[30] The variant p.P134R was shown to inhibit *in vitro* the expression of OTX2 target genes in a dominant negative fashion.[29] Variable penetrance and expressivity of *OTX2* mutations is known.[31] For P1, no craniofacial or ocular phenotypes were detected; however, his postnatal growth and weight at both visits (12 and 15 years old) were in the lowest 10th percentile based on national

growth charts.[32] By the age of 15, his height was only 164 cm and weight 47 kg (lowest 3rd percentile; **Table 1**).

P2 carries an additional pathogenic variant in the gene *PROP paired-like homeobox 1* (*PROP1*), c.301_302delAG (p.L102Cfs*8; rs193922688; MAF=$1.81x10^{-4}$; **Table 2)** inherited from his mother. *PROP1* encodes a transcription factor that specifically guides the differentiation of pituitary cell types. A typical phenotype characteristic to *PROP1* biallelic mutation carriers is congenital pituitary hormone deficiency (CPHD).[33] This was not observed in P2 and his mother (**Table 1, Table S1**). On the contrary, the levels of gonadotropins FSH and LH in P2 were extremely elevated and rising with age, reflecting progressive testicular failure (**Table 1**). We hypothesize that 50% reduction of the functioning PROP1 protein due to the carriership of the heterozygous inactivating variant[33] may modulate the penetrance of *NR5A1* c.991-1G>C in P2 through a cascade effect and lead to insufficient dosage of SF-1 transcription factor during critical developmental time windows (**Figure 2A**).

Computational modelling of the hypothesized digenic contributions in P1-P2 were implemented at the ORVAL platform.[22] Both tested digenic effects, *NR5A1* c.991-1G>C and *OTX2* p.P134R, *NR5A1* c.991-1G>C and *PROP1* c.301_302delAG reached statistically significant predictions as candidate disease-causing with >95% confidence (**Figure 2B**).

Patient 3 carries a novel variant in *SOS1*, a candidate gene implicated in testicular development

The pedigree structure supported a distinct genetic cause for cryptorchidism and cryptozoospermia in case P3 (**Figure 1A**). A novel heterozygous variant in the developmental gene *SOS1* (c.406T>C, p.Y136H) was identified as a primary candidate variant responsible for his clinical condition (**Table 2; Table S5, S7**). *SOS1* gain-of-function dominant missense variants alter Ras/MAPK signaling and account for a significant proportion of Noonan syndrome cases (NS, MIM: 610733).[34] Known pathogenic variants in NS genes, including *SOS1*, have been reported in isolated cryptorchidism cases without typical characteristics of NS.[35] Re-assessment of the P3 medical history revealed repeated visits to a dermatologist due to multiple nevi across the body, and erythema on his face and chest. In literature, NS-causing *SOS1* variants have been associated with a high prevalence of ectodermal abnormalities.[34]

Residue Y136 is located in the N-terminal histone-like domain of SOS1 that is homologous to the H2A/H2B histone complex (**Figure 3A-B**). This position is highly conserved in SOS1, SOS2 and histone H2A across all vertebrates (**Figure 1B, Figure S2**). Y136 is positioned structurally very close to the PH-Rem linker helix, a hotspot for known pathogenic NS variants, and it is important for the maintenance of intramolecular contacts to the helical linker and for protein-protein interactions (**Figure 3C-D; Figures S3-4**). Compared to Y136 the mutated H136 residue lacks the hydroxyl group, is smaller and therefore less exposed to the domain surface, predicted to weaken molecular interactions.

DISCUSSION

NR5A1 represents the 2nd most frequently affected gene in 46,XY cases with gonadal dysgenesis and one of the few well-established loci linked to non-obstructive azoospermia.[5, 27] In total, 188 different pathogenic *NR5A1* mutations have been detected in either 46,XY or 46,XX patients and the vast majority of these represent single cases and/or *de novo* variants.[7] This is the first report on a recurrent *NR5A1* splicing variant in patients with PGD (**Table 3**). The variant c.991-1G>C has been initially reported in a Belgian 46,XY case reared as a female and presenting clitoromegaly, abdominal testes, but no internal female structures.[25] Our study detected the same variant in two affected cousins with bilateral cryptorchidism, severe hypospadias, very low testicular volume and azoospermia. Although unable to confirm, the deceased family member with ambiguous sex was also a highly likely carrier (sibling of P1's mother and P2's father). As a critical observation for the clinical practice, the patient phenotypes in the two studies reporting c.991-1G>C were quite different. This is consistent with the observed complex phenotypic expressivity and variable penetrance of *NR5A1* pathogenic variants (including asymptomatic cases), challenging clinical counseling in the affected families.[6, 11, 25, 36]

So far, only eight splice-site variants have been reported in single PGD cases and the lack of proper family-based assessments has limited the estimation of their penetrance and genotype-phenotype correlation. Analysis of affected and non-affected family members showed that *NR5A1* c.991-1G>C variant has been segregating in at least three generations with variable penetrance (<50%) and phenotypic consequences. Among 41 previously described cases with maternal inheritance of other types of *NR5A1* mutations, one in four carrier mothers has been diagnosed with premature ovarian

insufficiency.[7] In this study, mother of P1 had experienced late menarche (age 15 years) and a single severely complicated pregnancy (36 years) with the anamnesis of premature placental aging. Like the case of P2, most reported paternal mutations were inherited from asymptomatic fathers.[7]

Notably, ES-based studies have revealed that a substantial group of 46,XY patients with *NR5A1* mutations carry also pathogenic variants in other genes that are functionally linked to SF-1 action.[4, 8, 9, 25] In cases P1 and P2, pathogenicity of *NR5A1* c.991-1G>C was possibly modulated by co-contributing rare variants inherited from the other parent, *OTX2* p.P134R and *PROP1* c.301_302delAG (**Figure 1, Table 2**). These transcription factors act upstream of SF-1 in fetal development (**Figure 2**).

Proband P3 presenting less severe genital phenotype and cryptozoospermia was identified as a carrier of an undescribed heterozygous variant p.Y136H in the Noonan syndrome (NS) related gene *SOS1*. In addition to reproductive phenotype, P3 had various forms of dermatological concerns that have been often noticed in *SOS1*-related NS patients. Cryptorchidism and hypospermatogenesis are observed in 60-80% of male NS cases[37] and patients with isolated cryptorchidism have been reported to carry known pathogenic *SOS1* variants that have been identified in NS patients.[35] In addition, cryptorchid testis has been shown to display significantly reduced *SOS1* gene expression.[38] The SOS1 p.Y136H substitution in a highly conserved residue that is structurally close to the hotspot region for known NS mutations (**Figure 3**). Among these, position R552 is a target for a range of (likely) pathogenic substitutions (p.R552S, p.R552K, p.R552T, p.R552M, p.R552W, p.R552G) and is altered in 30% of NS patients with the *SOS1* defect.[34] The gathered evidence supports p.Y136H as a likely pathogenic variant and *SOS1* as a strong candidate contributor to cryptorchidism and consequently, spermatogenic failure. Interestingly, *de novo* variants dysregulating the RAS/MAPK pathway, including mutations in *SOS1*, were shown to be under positive selection leading to clonal expansion in the aging male germline.[39] It may be speculated that *de novo* mutations arising during spermatogenesis in spermatogonial stem cells in RAS/MAPK pathway genes represent so far undescribed etiology behind sporadic cryptorchidism cases.

In summary, our study reports the first recurrent *NR5A1* splice-site variant linked to PGD and highlights genetic heterogeneity causing genital dysgenesis in a single family. Detailed genetic profiling facilitates high quality counseling and clinical management of the index patients and also unaffected carriers in the family for their reproductive decision-making.

REFERENCES

1. Boisen KA, Kaleva M, Main KM, et al. Difference in prevalence of congenital cryptorchidism in infants between two Nordic countries. *Lancet*. 2004;363(9417):1264-9.
2. Nordenvall AS, Frisen L, Nordenstrom A, Lichtenstein P, Nordenskjold A. Population based nationwide study of hypospadias in Sweden, 1973 to 2009: incidence and risk factors. *J Urol*. 2014;191(3):783-9.
3. Baetens D, Verdin H, De Baere E, Cools M. Update on the genetics of differences of sex development (DSD). *Best Pract Res Clin Endocrinol Metab*. 2019;33(3):101271.
4. Wang H, Zhang L, Wang N, et al. Next-generation sequencing reveals genetic landscape in 46, XY disorders of sexual development patients with variable phenotypes. *Hum Genet*. 2018;137(3):265-277.
5. Eggers S, Sadedin S, van den Bergen JA, et al. Disorders of sex development: insights from targeted gene sequencing of a large international patient cohort. *Genome Biol*. 2016;17(1):243.
6. Brauner R, Picard-Dieval F, Lottmann H, et al. Familial forms of disorders of sex development may be common if infertility is considered a comorbidity. *BMC Pediatr*. 2016;16(1):195.
7. Fabbri-Scallet H, de Sousa LM, Maciel-Guerra AT, Guerra-Junior G, de Mello MP. Mutation update for the NR5A1 gene involved in DSD and infertility. *Hum Mutat*. 2020;41(1):58-68.
8. Mazen I, Abdel-Hamid M, Mekkawy M, et al. Identification of NR5A1 Mutations and Possible Digenic Inheritance in 46,XY Gonadal Dysgenesis. *Sex Dev*. 2016;10(3):147-51.
9. Camats N, Fernandez-Cancio M, Audi L, Schaller A, Fluck CE. Broad phenotypes in heterozygous NR5A1 46,XY patients with a disorder of sex development: an oligogenic origin? *Eur J Hum Genet*. 2018;26(9):1329-1338.
10. Ayers K, Kumar R, Robevska G, et al. Familial bilateral cryptorchidism is caused by recessive variants in RXFP2. *J Med Genet*. 2019;56(11):727-733.
11. Eggers S, Smith KR, Bahlo M, et al. Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1. *Eur J Hum Genet*. 2015;23(4):486-93.
12. Punab M, Poolamets O, Paju P, et al. Causes of male infertility: a 9-year prospective monocentre study on 1737 patients with reduced total sperm counts. *Hum Reprod*. 2017;32(1):18-31.
13. Kasak L, Punab M, Nagirnaja L, et al. Bi-allelic Recessive Loss-of-Function Variants in FANCM Cause Non-obstructive Azoospermia. *Am J Hum Genet*. 2018;103(2):200-212.
14. Wilfert AB, Chao KR, Kaushal M, et al. Genome-wide significance testing of variation from single case exomes. *Nat Genet*. 2016;48(12):1455-1461.
15. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019;47(D1):D886-D894.
16. Richards S, Aziz N, Bale 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*. 2015;17(5):405-24.
17. Gelb BD, Cave H, Dillon MW, et al. ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genet Med*. 2018;20(11):1334-1345.
18. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol*. 2011;7(10):e1002195.
19. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-80.
20. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-12.
21. Dapkunas J, Timinskas A, Olechnovic K, Margelevicius M, Diciunas R, Venclovas C. The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures. *Bioinformatics*. 2017;33(6):935-937.
22. Renaux A, Papadimitriou S, Versbraegen N, et al. ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res*. 2019;47(W1):W93-W98.
23. Papadimitriou S, Gazzo A, Versbraegen N, et al. Predicting disease-causing variant combinations. *Proc Natl Acad Sci U S A*. 2019;116(24):11878-11887.

24. Suntharalingham JP, Buonocore F, Duncan AJ, Achermann JC. DAX-1 (NR0B1) and steroidogenic factor-1 (SF-1, NR5A1) in human disease. *Best Pract Res Clin Endocrinol Metab*. 2015;29(4):607-19.
25. Robevska G, van den Bergen JA, Ohnesorg T, et al. Functional characterization of novel NR5A1 variants reveals multiple complex roles in disorders of sex development. *Hum Mutat*. 2018;39(1):124-139.
26. Grigorova M, Punab M, Poolamets O, Adler M, Vihljajev V, Laan M. Genetics of Sex Hormone-Binding Globulin and Testosterone Levels in Fertile and Infertile Men of Reproductive Age. *J Endocr Soc*. 2017;1(6):560-576.
27. Kasak L, Laan M. Monogenic causes of non-obstructive azoospermia: challenges, established knowledge, limitations and perspectives. *Hum Genet*. 2020;
28. Suwanai AS, Ishii T, Haruna H, et al. A report of two novel NR5A1 mutation families: possible clinical phenotype of psychiatric symptoms of anxiety and/or depression. *Clin Endocrinol (Oxf)*. 2013;78(6):957-65.
29. Gorbenko Del Blanco D, Romero CJ, Diaczok D, de Graaff LC, Radovick S, Hokken-Koelega AC. A novel OTX2 mutation in a patient with combined pituitary hormone deficiency, pituitary malformation, and an underdeveloped left optic nerve. *Eur J Endocrinol*. 2012;167(3):441-52.
30. Larder R, Kimura I, Meadows J, Clark DD, Mayo S, Mellon PL. Gene dosage of Otx2 is important for fertility in male mice. *Mol Cell Endocrinol*. 2013;377(1-2):16-22.
31. Tajima T, Ishizu K, Nakamura A. Molecular and Clinical Findings in Patients with LHX4 and OTX2 Mutations. *Clin Pediatr Endocrinol*. 2013;22(2):15-23.
32. Salm E, Käärik E, Kaarma H. The growth charts of Estonian schoolchildren. Comparative analysis. *Papers on Anthropology*. 1970;22(0):171-183.
33. Wu W, Cogan JD, Pfaffle RW, et al. Mutations in PROP1 cause familial combined pituitary hormone deficiency. *Nat Genet*. 1998;18(2):147-9.
34. Lepri F, De Luca A, Stella L, et al. SOS1 mutations in Noonan syndrome: molecular spectrum, structural insights on pathogenic effects, and genotype-phenotype correlations. *Hum Mutat*. 2011;32(7):760-72.
35. Rodriguez F, Vallejos C, Ponce D, et al. Study of Ras/MAPK pathway gene variants in Chilean patients with Cryptorchidism. *Andrology*. 2018;6(4):579-584.
36. Bashamboo A, Donohoue PA, Vilain E, et al. A recurrent p.Arg92Trp variant in steroidogenic factor-1 (NR5A1) can act as a molecular switch in human sex development. *Hum Mol Genet*. 2016;25(23):5286.
37. Marcus KA, Sweep CG, van der Burgt I, Noordam C. Impaired Sertoli cell function in males diagnosed with Noonan syndrome. *J Pediatr Endocrinol Metab*. 2008;21(11):1079-84.
38. Hadziselimovic NO, de Geyter C, Demougin P, Oakeley EJ, Hadziselimovic F. Decreased expression of FGFR1, SOS1, RAF1 genes in cryptorchidism. *Urol Int*. 2010;84(3):353-61.
39. Maher GJ, Ralph HK, Ding Z, et al. Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Res*. 2018;28(12):1779-1790.
40. Avbelj Stefanija M, Kotnik P, Bratanic N, et al. Novel Mutations in HESX1 and PROP1 Genes in Combined Pituitary Hormone Deficiency. *Horm Res Paediatr*. 2015;84(3):153-158.
41. Kohler B, Lin L, Mazen I, et al. The spectrum of phenotypes associated with mutations in steroidogenic factor 1 (SF-1, NR5A1, Ad4BP) includes severe penoscrotal hypospadias in 46,XY males without adrenal insufficiency. *Eur J Endocrinol*. 2009;161(2):237-42.
42. Song Y, Fan L, Gong C. Phenotype and Molecular Characterizations of 30 Children From China With NR5A1 Mutations. *Front Pharmacol*. 2018;9:1224.
43. Nishina-Uchida N, Fukuzawa R, Numakura C, Suwanai AS, Hasegawa T, Hasegawa Y. Characteristic testicular histology is useful for the identification of NR5A1 gene mutations in prepubertal 46,XY patients. *Horm Res Paediatr*. 2013;80(2):119-28.
44. Rehkamper J, Tewes AC, Horvath J, Scherer G, Wieacker P, Ledig S. Four Novel NR5A1 Mutations in 46,XY Gonadal Dysgenesis Patients Including Frameshift Mutations with Altered Subcellular SF-1 Localization. *Sex Dev*. 2017;11(5-6):248-253.
45. Fabbri HC, Ribeiro de Andrade JG, Maciel-Guerra AT, Guerra-Junior G, de Mello MP. NR5A1 Loss-of-Function Mutations Lead to 46,XY Partial Gonadal Dysgenesis Phenotype: Report of Three Novel Mutations. *Sex Dev*. 2016;10(4):191-199.

46. Gergics P. Pituitary Transcription Factor Mutations Leading to Hypopituitarism. *Exp Suppl*. 2019;111:263-298.
47. Patti G, Guzzeti C, Di Iorgi N, et al. Central adrenal insufficiency in children and adolescents. *Best Pract Res Clin Endocrinol Metab*. 2018;32(4):425-444.

WEB REFERENCES

NCBI ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/
The Human Protein Atlas: https://www.proteinatlas.org
gnomAD, https://gnomad.broadinstitute.org
OMIM, https://www.omim.org
Uniclust90, https://uniclust.mmseqs.com
Ensembl: https://www.ensembl.org/
UCSC Genome Browser: https://genome.ucsc.edu
Mouse Genome Informatics: http://www.informatics.jax.org
ORVAL https://orval.ibsquare.be

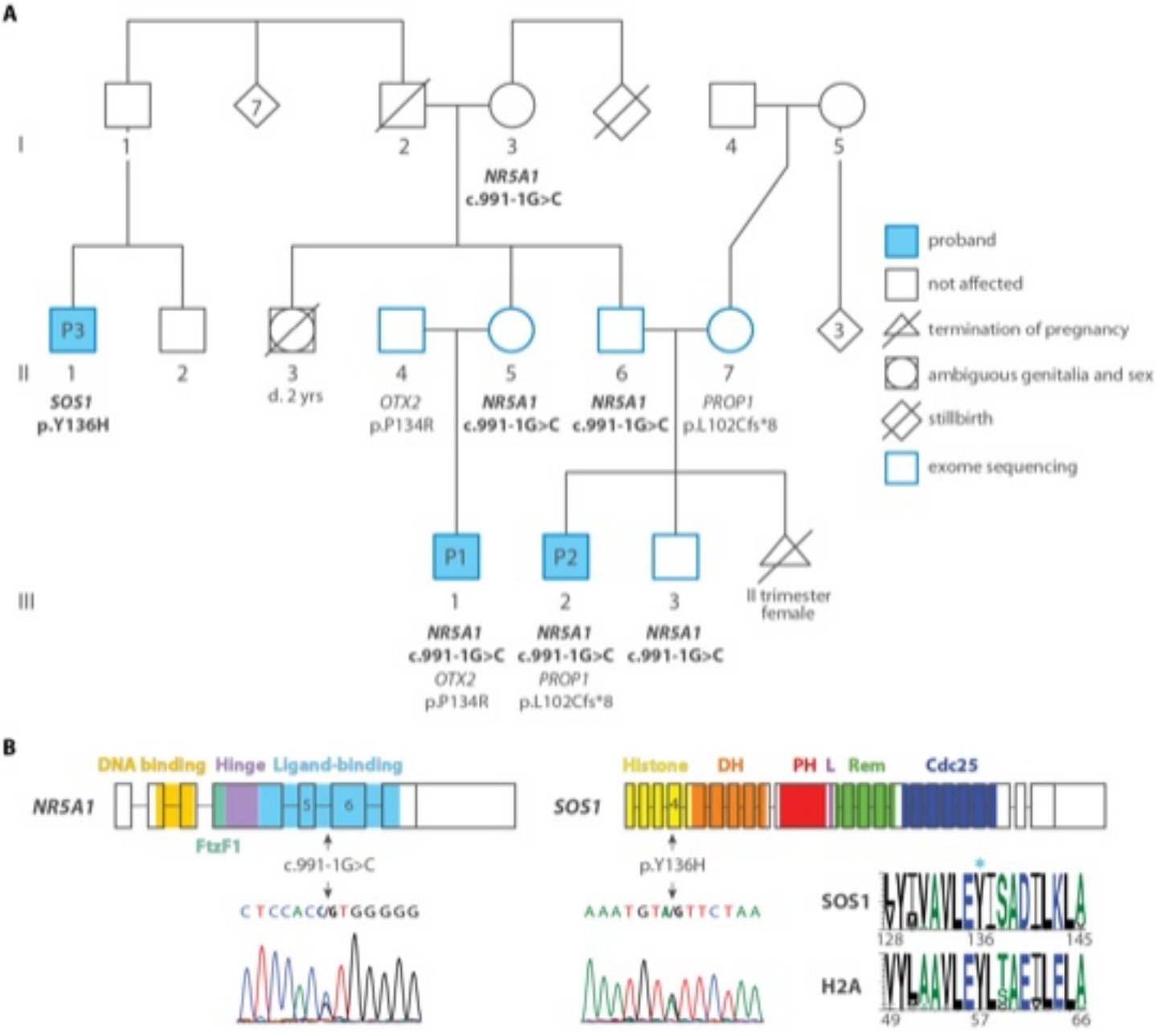
Figure Legends

Figure 1 Primary causative and modulatory heterozygous variants in the index family with partial gonadal dysgenesis. A, Pedigree of the index family. Prioritized rare genetic variants are listed under tested family members with the primary causative variants for their phenotype highlighted in bold. B, Schematic presentation of the exon-intron structure of *NR5A1* and *SOS1*, highlighting the encoded protein domains and the position of the primary pathogenic variant. Variant validation by Sanger sequencing is shown relative to the sense strand. Residue Y136 in the SOS1 histone-like domain (light blue star) represents a highly conservative position not only across vertebrate SOS homologs, but also in the corresponding position Y57 of H2A histone proteins. cdc25, catalytic GEF domain; DH, Dbl homology; Histone, histone-like domain; L, linker domain; PH, pleckstrin homology; Rem, Ras exchanger motif

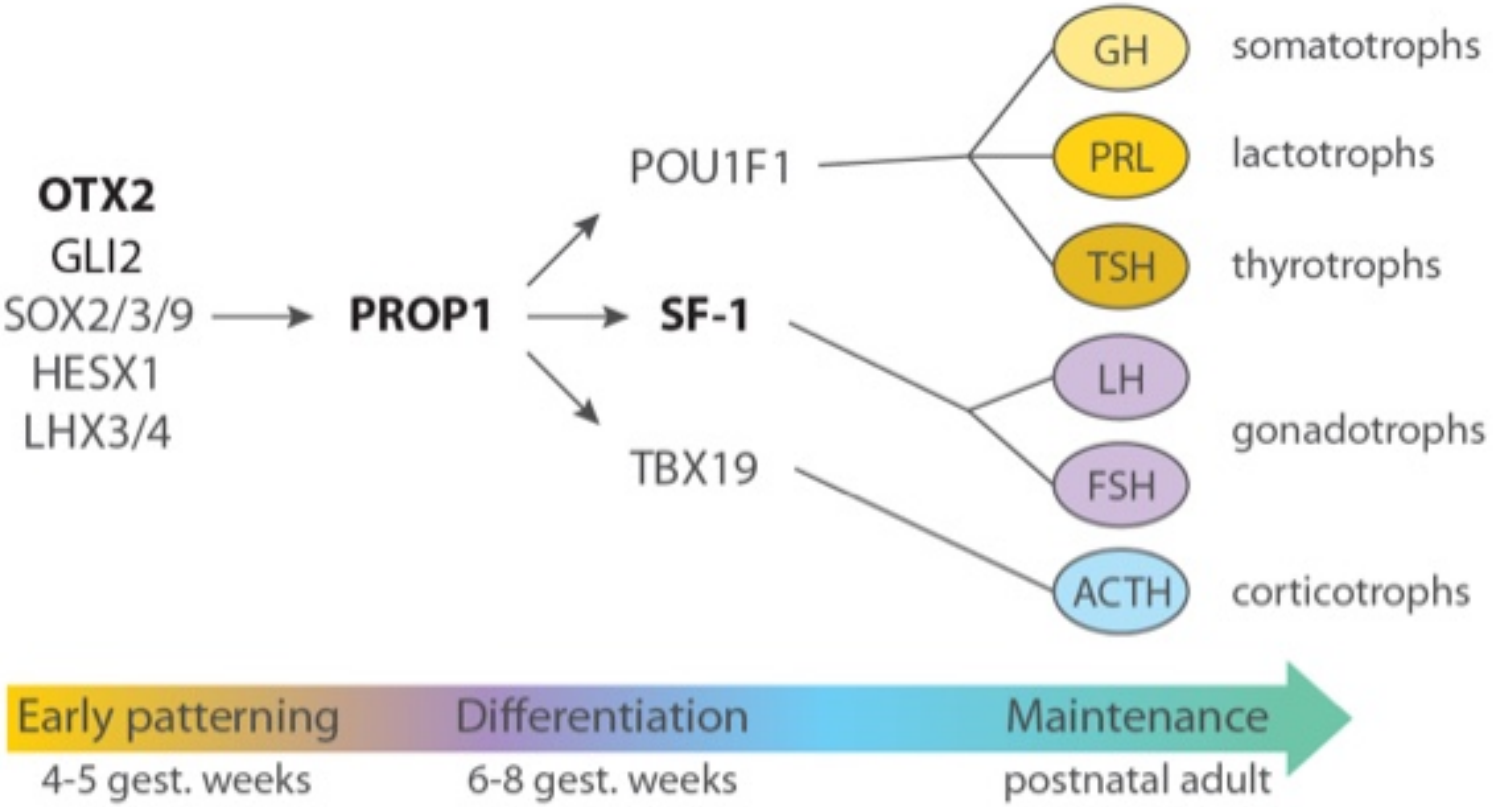
Figure 2 Hypothesized scenario of co-contributing genetic effects modulating the penetrance of *NR5A1* pathogenic variants. A, Simplified schematic representation of a critical transcription factor cascade in the anterior pituitary development in humans (Refs.[46, 47]). Genes *OTX2, PROP1* and *NR5A1* (encoding SF-1) are functioning sequentially and interdependently in early development, whereas SF-1 acts in dose-dependent manner.[24] B, Statistical support for the pathogenicity of hypothesized disease-causing digenic variant combinations. The probability of a combined pathogenic effect of the variant pairs was modelled and estimated using the ORVAL platform.[22]

Figure 3 *In silico* structural analysis of SOS1 p.Y136H variant. A, The structure of the human hSOS1 protein (hSOS1; PDB id: 3KSY) colored by structural domains. B, SOS1 histone-like domain is structurally similar to the H2A/H2B histone complex. The full human nucleosome (PDB id 2CV5) and hSOS1 (PDB id 3KSY) structures (bottom) are shown along with their highlighted and superimposed histone domains (above). Amino acid Y136 in hSOS1 corresponds to position Y57 in human H2A. C, Known Noonan syndrome (NS) causing variants shown on the hSOS1 structure using the respective color code of the location domain. The variant annotation was derived from Uniprot 'Natural variant'

section for the hSOS1 protein (Uniprot id Q07889). The Y136 residue is located in the N-terminal histone-like domain but is structurally close to the PH-Rem linker helix (violet). The latter is enriched in reported pathogenic NS causing variants involving residues 548-558, whereas NS mutations in the SOS1 histone-like domain are structurally distant from residue Y136. D, Structural comparison of native hSOS1 (PDB id: 3KSY) and modeled p.Y136H substitution in histone-like domain suggests a smaller surface area for the latter, resulting in altered contacts to the linker region. Although residue Y136 has a direct contact to p.V556 in the linker helix, p.Y136H may not form a similar interaction. Partial surfaces are shown for the main (dark blue) and side chains (light blue) of Y136 and H136 residues.

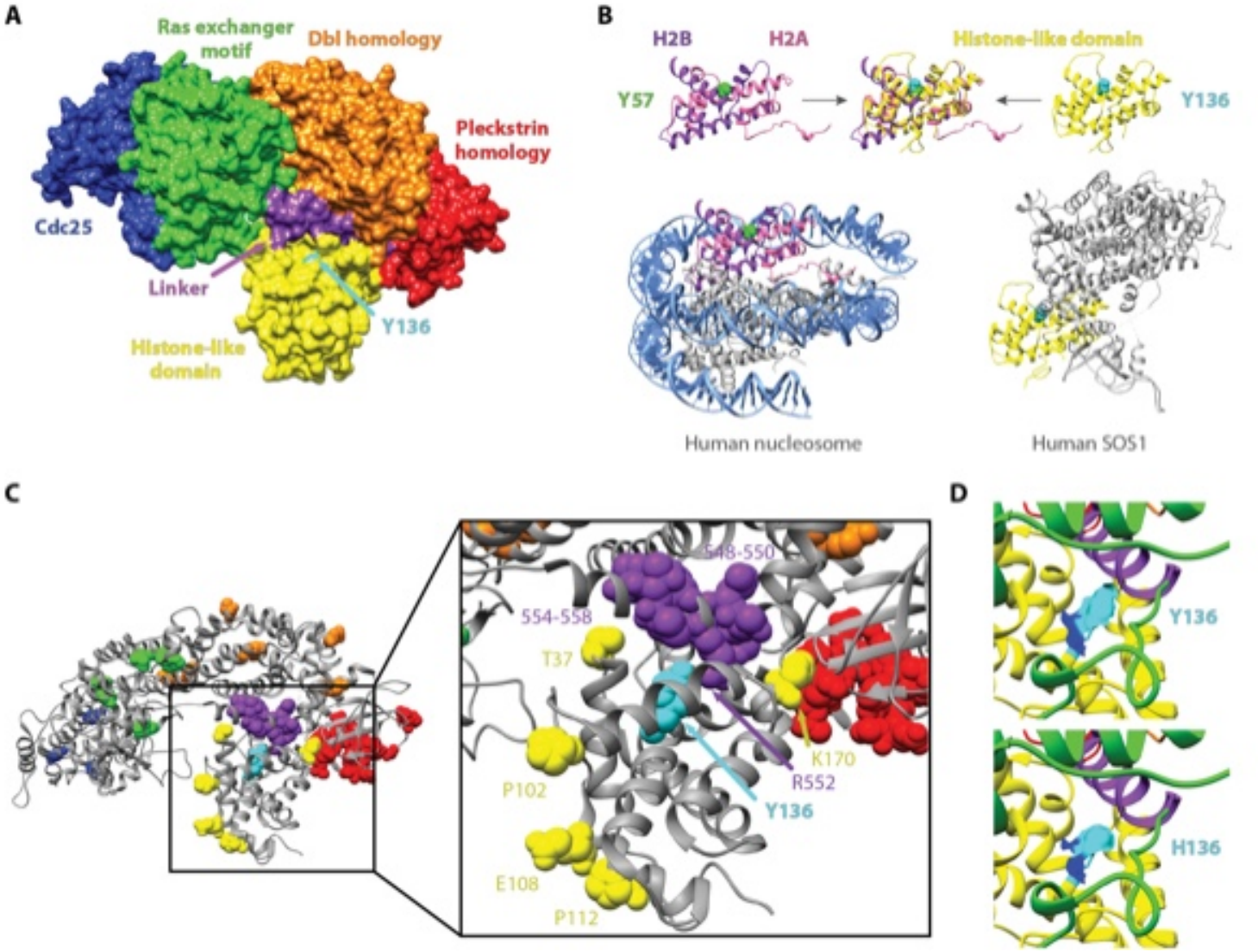

A
I
II
III
1
2
3
4
5
6
7
P3
P1
P2
NR5A1
c.991-1G>C
SOS1
p.Y136H
d. 2 yrs
OTX2
p.P134R
PROP1
p.L102Cfs*8
II trimester
female
proband
not affected
termination of pregnancy
ambiguous genitalia and sex
stillbirth
exome sequencing
B
DNA binding
Hinge
Ligand-binding
NR5A1
FtzF1
c.991-1G>C
C T C C A C G/A G T G G G G G
Histone
DH
PH
L
Rem
Cdc25
SOS1
p.Y136H
A A A T G T A/G T T C T A A
SOS1
H2A
128
136
145
49
57
66

A



B

Proband	Digenic variant combinations Variant A	Variant B	ORVAL pathogenicity effect >95% confidence zone
P1	*NR5A1* c.991-1G>C	*OTX2* c.425C>G	disease-causing
P2	*NR5A1* c.991-1G>C	*PROP1* c.301_302delAG	disease-causing


A
Ras exchanger motif
Dbl homology
Pleckstrin homology
Cdc25
Linker
Y136
Histone-like domain
B
H2B
H2A
Histone-like domain
Y57
Y136
Human nucleosome
Human SOS1
C
548-550
554-558
T37
K170
R552
P102
Y136
E108
P112
D
Y136
H136

SUPPLEMENTARY INFORMATION

NR5A1 c.991-1G>C splice-site variant causes familial 46,XY partial gonadal dysgenesis with incomplete penetrance

Running title: Recurrent *NR5A1* splice-site variant and PGD

Maris Laan[1]*, Laura Kasak[1], Kęstutis Timinskas[2], Marina Grigorova[1], Česlovas Venclovas[2], Alexandre Renaux[3,4,5], Tom Lenaerts[3,4,5], Margus Punab[6,7]

[1] Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia

[2] Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania

[3] Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles - Vrije Universiteit Brussel, Brussels, Belgium

[4] Machine Learning Group, Université libre de Bruxelles, Brussels, Belgium

[5] Artificial Intelligence lab, Vrije Universiteit Brussel, Brussels, Belgium

[6] Andrology Center, Tartu University Hospital, Tartu, Estonia

[7] Institute of Clinical Medicine, University of Tartu, Estonia

*** Corresponding author:**

Maris Laan, Institute of Biomedicine and Translational Medicine, University of Tartu

Ravila St. 19, 50411 Tartu, Estonia; tel: +372-7375008; +372-53495258, email: maris.laan@ut.ee

SUPPLEMENTARY METHODS

Additional details on the clinical assessment of the patients

Probands and other male members of the recruited pedigree were examined by specialist andrologist (M.P.).[1] Physical examination for the assessment of genital pathology and testicular size (orchidometer; made of birch wood, Pharmacia & Upjohn, Denmark) was performed with the patients in standing position. The total testes volume is the sum of right and left testicles. The position of the testicles in the scrotum, pathologies of the genital ducts (epididymis and ductus deference) and the penis, urethra, presence and if applicable grade of varicocele were registered for each subject. Physical development was staged according to Tanner's classification.[2]

Genomic DNA was extracted from EDTA-blood. After blood draw in the morning, serum and plasma fractions were separated immediately for hormone measurements (FSH, LH, testosterone, SHBG, estradiol, TSH, ACTH, PRL, GH). All laboratory analyses and routine genetic testing (karyotyping, Y-chromosomal microdeletions) were performed at the United Laboratories of Tartu University Hospital according to the established clinical laboratory guidelines. Free testosterone levels were calculated from measured testosterone, SHBG and fixed albumin levels (43 g/L) using the Vermeulen's equation.[3]

Semen samples were obtained by patient masturbation and semen analysis was performed in accordance with the World Health Organization recommendations.[4] In brief, after ejaculation, the semen was incubated at 37°C for 30–40 min for liquefaction. Semen volume was estimated by weighing the collection tube with the semen sample and subsequently subtracting the predetermined weight of the empty tube assuming 1 g = 1 mL. For assessment of the spermatozoa concentration, the samples were diluted in a solution of 0.6 mol,L NaHCO3 and 0.4% (v,v) formaldehyde in distilled water. The spermatozoa concentration was assessed using the improved Neubauer haemocytometers.

Anthropometric parameters were documented during clinical examination upon patients' consent. Family history of reproductive disorders and general health disturbances were collected by the managing clinician from the probands and their close relatives, if available. All recruited family members are native Estonians and the anamnesis excluded consanguinity.

Whole-exome sequencing (WES) and data analysis

WES data generation and primary processing

Wet-lab processing, base calling of the raw sequencing data, primary sequence analysis and variant calling was performed at the Institute for Molecular Medicine Finland (FIMM; Helsinki, Finland) Next Generation Sequencing Service using the established pipeline and has been described previously.[5] Whole exome enrichment was undertaken with the SeqCap EZ MedExome Target Enrichment Kit (Roche NimbleGen, Madison, WI, US) following the manufacturer's protocol. Sequencing was

performed on Illumina HiSeq 2500 sequencing system (San Diego, CA) using paired-end 101-bp read length. Sequencing resulted in 75-132 million reads per exome.

Primary sequence analysis and variant calling was performed using the Variant Calling Pipeline (VCP3.7) described in.[6] In brief, Illumina paired-end reads were trimmed with Trimmomatic (version 0.36) and aligned against human genome build hg19 with the Burrows-Wheeler Aligner (BWA, version 0.6.2).[7] After the alignment, PCR duplicates were removed with Picard MarkDuplicates (version 2.9.0). Across all samples, 92-97% of targeted bases were covered at least 30X and the mean target coverage was 147X. SNVs and small indels were called with SAMtools (version 1.4)[8] and Pindel (version 0.2.5b8),[9] respectively. The resulting eight Variant Call Format (VCF) files contained from 127,942 to 168,760 (median 135,460) variants per exome dataset (**Table S1**). Subsequent filtering of the VCF files, prioritization and QC of the retained variants was performed separately for the SNVs and indels.

Data analysis and variant prioritization

Population sampling probability (PSAP) pipeline[5,10] was applied in order to prioritize potential causative variants from the WES data. Prior to variant annotation and testing for the genotype sampling probability, the PSAP pipeline applies the following filtering of the VCF dataset: (i) removal of genes with poor genotype calling; (ii) exclusion of variants that are not located within the coding sequences; (iii) removal of variants exhibiting major discrepancies for their minor allele frequencies (MAF) among the alternative human genome sequencing databases – The Exome Aggregation Consortium (ExAC), n=61,000 exomes; Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, n=6,500 exomes; 1000 Genomes Project (1000G), n=2,500 exomes; (iv) removal of variants with missing Combined Annotation Dependent Depletion (CADD) scores,[11,12] except for indels.

A PSAP option 'individual analysis pipeline' was applied. The pipeline resulted in 11,763 - 13,198 (median 12,958) variants for each of the eight analyzed WES datasets. The variants in the PSAP output were further sequentially prioritized to satisfy the following criteria: (i) the PSAP statistical significance value, popScore <0.005; (ii) previously undescribed or rare variants with minor allele frequency (MAF) <0.0005 in ExAC/gnomAD and 1000 Genomes databases; (iii) exclusion of synonymous variants; (iv) applying Combined Annotation Dependent Depletion (CADD) to rank the deleteriousness of the variants (C-score >20).[5] Variants prioritized by the PSAP, but missing allele frequency information in the ExAC/gnomAD,[13] ESP6500 and 1000GP were double-checked in dbSNP (build 151) before classifying as previously unreported. If needed, visual inspection of the quality of sequencing reads was performed using The Integrative Genomics Viewer (IGV) software.[14] Low confidence variants due to unreliable reads, a position within a long mononucleotide track, location in genes reported susceptible to sequencing errors, and/or rare variants detected in multiple unrelated pedigree members were discarded. The final list of prioritized variants in the probands (P1 and P2, n=35; P3, n=32) were inspected manually using scientific literature and genome databases (**Tables S3-5**). Based on the family

history the following disease inheritance models were considered: autosomal dominant with incomplete penetrance or autosomal recessive (homozygous, compound heterozygous).
Prioritized variants were classified as (likely) pathogenic, (likely) benign or with uncertain significance (VUS) based on the American College of Medical Genetics and Genomics (ACMG) guidelines.[15] For the pathogenicity assessment of the novel *SOS1* variant, ClinGen's RASopathy Expert Panel Consensus Methods for Variant Interpretation were applied.[16]

Experimental assessment of prioritized variants by Sanger sequencing

Variants highlighted in the exomes of the three probands (P1, P2, P3) were experimentally confirmed using Sanger sequencing. The analysis included eight family members initially assessed by WES and an additional subject II-3 (grandmother of P1 and P2). For P1 and P2, the parent-of-origin of all variants was determined. For P3, parental DNA samples were not available. Primers for amplification and sequencing (**Table S3**) were designed in Primer3,[17] tested by NCBI Primer-BLAST. DNA fragments were amplified by standard PCR, sequenced with the BigDye Terminator v.3.1 Cycle Sequencing Kit and run on an ABI 3730 DNA Analyzer (both Applied Biosystems, Carlsbad, CA, USA). Sequences were analyzed with BioEdit 7.2.5.

Utilized annotated protein sequences and structures for SOS1 p.Y136H *in silico* analysis

The reviewed annotations of human SOS1 protein are available in Uniprot (Uniprot id Q07889) and the basic structural annotations at PDB, respectively (https://www.uniprot.org/uniprot/Q07889; http://www.rcsb.org/pdb/protein/Q07889). There are two available structures that include the SOS1 histone-like domain: 1Q9C (only histone-like domain) and 3KSY (nearly complete protein missing only the Grb2 binding C-terminal domain). Since the histone-like domain in the two structures shows only negligible differences, the more complete 3KSY structure was used for all modeling and structural analysis.

Sequence conservation analysis of SOS1 p.Y136H

Homologs for human SOS1 (hSOS1) protein were identified in Uniclust90 database[18] using HMMER[19] (1 iteration, 1e-4 E-value cutoff). Only the N-terminal sequence of hSOS1 (1-180 aa) was used to limit the search for homologs of the hSOS1 histone-like domain. Full length sequences of all collected homologs were clustered according to their global sequence similarities using CLANS,[20] SOS1 and H2A groups were separated at p-value=1e-60. Multiple sequence alignments (MSAs) were produced for these groups with MAFFT (l-INS-i).[21] Using Jalview,[22] the alignments were visually inspected, sequence fragments and apparent alternative splicing variants (compared to human sequence) were discarded. The alignments used for sequence logo figures were further filtered by annotated taxonomy to retain only vertebrate sequences. The sequence logos were created using the WebLogo 3 server.[23]

Modeling and structural analysis of SOS1 p.Y136H variant

The model for the SOS1 p.Y136H substitution was constructed by direct replacement of the tyrosine side chain with a selected histidine rotamer using tools implemented in UCSF Chimera.[24] UCSF Chimera was also used for the analysis of surface properties of SOS1 and the variant p.Y136H. The quality of structural models was evaluated both by visual inspection in UCSF Chimera and by automatic assessment with VoroMQA.[25] PPI3D[26] server was used to identify structurally known interactions with human SOS1 or other homologous histone proteins. Only the histone-like domain sequence (1-190) of human SOS1 was used in the search (standard parameters). UCSF Chimera[24] was used for superposition and visualization of identified homologous complexes.

Prediction of disease-causing oligogenic variant combinations using ORVAL platform

Computational methods for the prediction of disease-causing oligogenic variant combinations and their effects were assessed using the ORVAL platform (https://orval.ibsquare.be).[27] ORVAL tests bi-locus variant combinations, annotates and analyzes them using two machine-learning based methods trained on digenic cases contained in the Digenic Disease Database (DIDA)[28] using features at the variant, gene and combination level. As a result, the pathogenicity of each tested bi-locus variant combination is predicted using VarCoPP, the Variant Combination Pathogenicity Predictor.[29] For *PROP1* c.301_302delAG, the current version of VarCoPP software misses the P(rec) score, the estimated probability that a gene is a recessive disease gene. Therefore, to the probability of digenetic effect for *PROP1* c.301_302delAG + *NR5A1* c.991-1G>C variant pair manual encoding P(rec) = 0.9 was applied.

SUPPLEMENTARY FIGURES

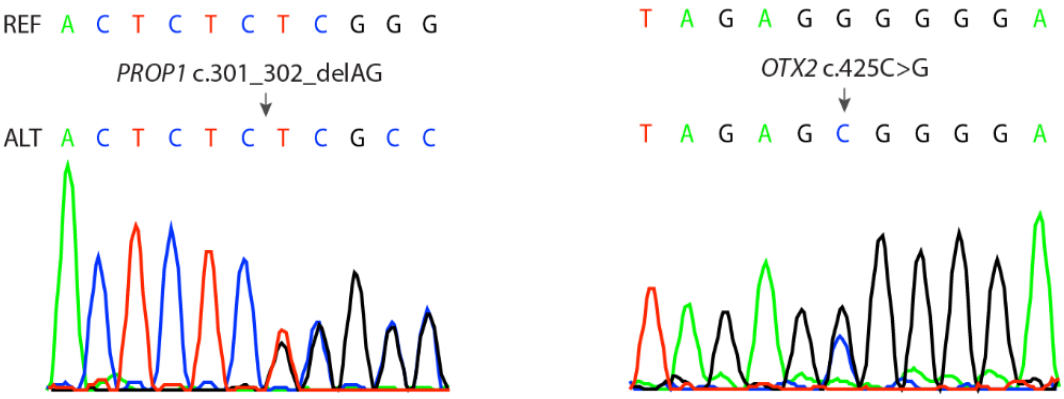


Figure S1. Sanger sequencing validation of additional variants.

Figure S2. Conservation of vertebrate SOS1 residues structurally close to the human hSOS1 Y136 residue. Position numbering corresponds to that of hSOS1. Grey stars mark positions that have side chains no further than 5Å from the side chain of Y136 (based on hSOS1 structure, PDB id: 3KSY). The 547-564 motif is part of the helical linker, that links PH and Rem domains. The 77-92 motif corresponds to a motif in H2B histone in the homologous H2A/H2B histone complex.

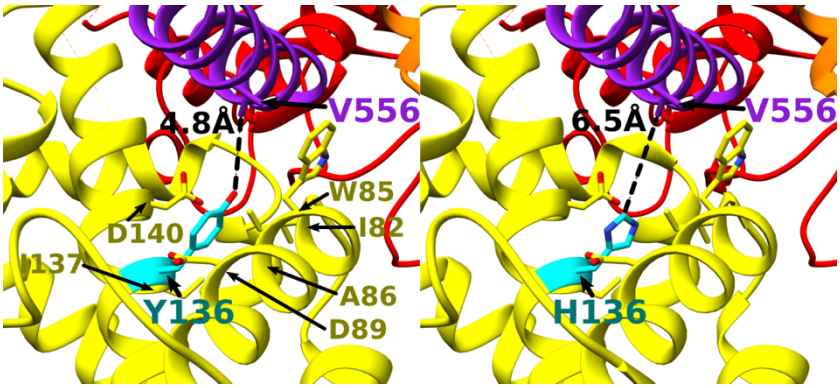


Figure S3. Structural comparison of native hSOS1 (PDB id: 3KSY) and a modelled substitution. The wild-type Y136 and the substitution variant H136 residues are exposed to the surface of the protein between histone-like and linker domains. The Y136 residue has a direct contact to V556 in the helical linker (distance 4.8 Å), whereas the p.Y136H substitution is unlikely to form a similar interaction (distance to V556 is 6.5 Å). All other residues close to the Y136 (marked with grey stars in online **Figure S2** above) are also indicated in the left part of the figure.

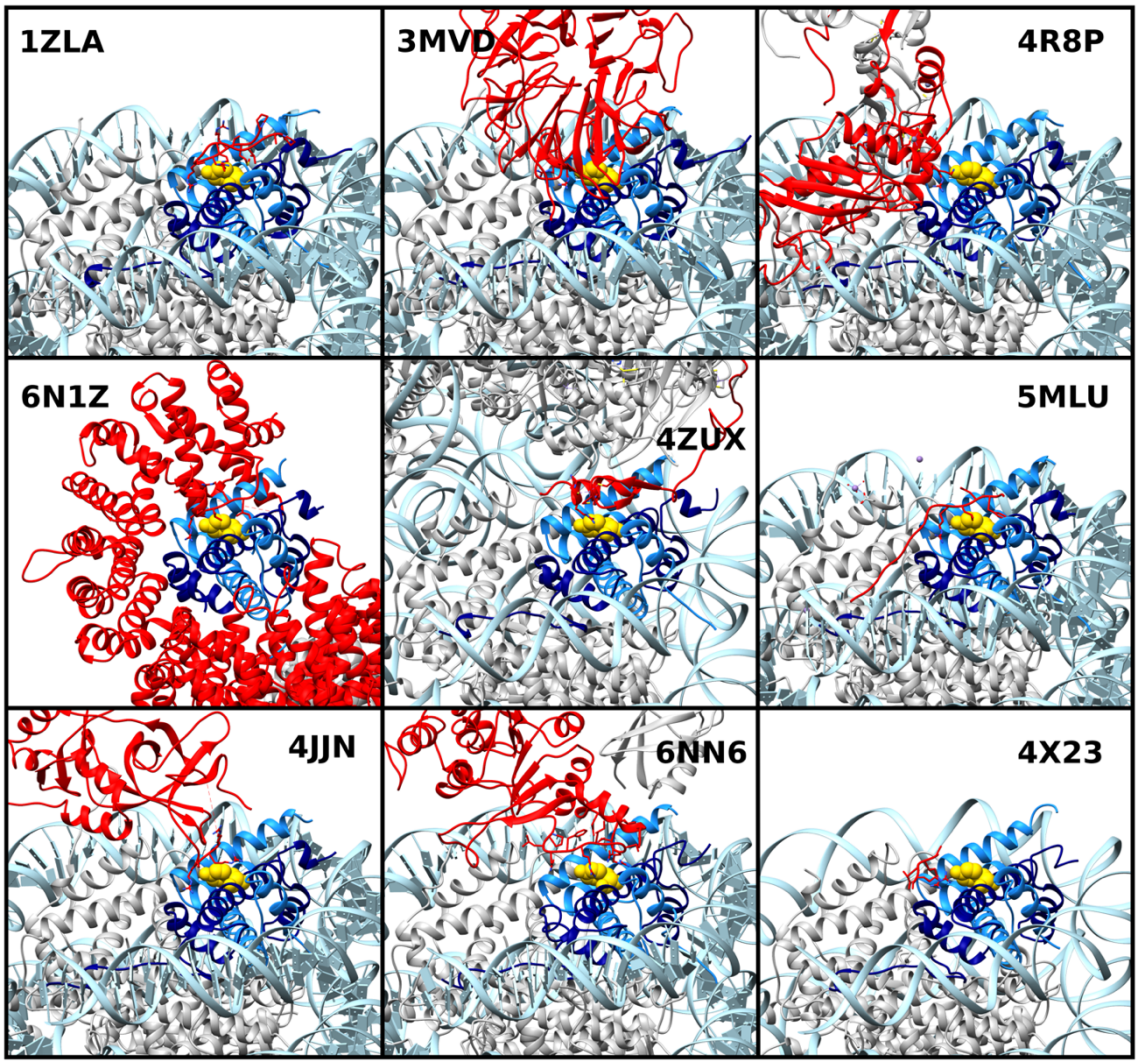


Figure S4. Examples of structures of known nucleosome complexes in which histone H2A residue p.Y57 is involved. There are no solved structures for protein complexes involving SOS1. The histone H2A/H2B complex has been extensively investigated and position Y57 of histone H2A (corresponding to Y136 in SOS1) has been identified as part of the interaction interface in critical functional complexes

Color code: **red**, interacting peptide/protein chain; **yellow**, p.Y57 (or a corresponding) position; **dark blue**/**blue**, H2A/H2B; **light blue**, DNA; **grey**, other protein chains.

SOS1 histone-like domain is similar to the H2A/H2B histone complex (blue) and the corresponding position to hSOS1 p.Y136 is p.Y57 hH2A (yellow), conserved in all H2A homologs. The surface including p.Y57 is involved in multiple nucleosome interactions to other proteins/peptides (red):

- Kaposi's sarcoma-associated herpesvirus (KSHV) latency-associated nuclear antigen (LANA), that mediates viral genome attachment to mitotic chromosomes (PDB id 1ZLA);
- RCC1 (regulator of chromosome condensation) protein (PDB id 3MVD);
- Polycomb repressive complex 1 (PRC1), that uniquitylates histone H2A K119 and is known for repressing the expression of developmentally regulated genes in eukaryotes (PDB id 4R8P);
- Histone chaperone Importin-9 to H2A/H2B (PDB id 6N1Z);
- Spt-Ada-Gcn5 acetyltransferase (SAGA)-associated factor 11 (PDB id 4ZUX), involved in H2B deubiquitination;
- Chromatin-binding sequence of the GAG peptide of prototype foamy virus (PFV) (PDB id 5MLU);
- Heterochromatin protein Sir3 with nucleosome (PDB id 4JJN);
- H3K79 methyltransferase Dot1L (PDB id 6NN6);
- centromere protein CENP-C (PDB id 4X23).

REFERENCES TO SUPPLEMENTARY INFORMATION

1. Carlsen E, Andersen AG, Buchreitz L, et al. Inter-observer variation in the results of the clinical andrological examination including estimation of testicular size. *Int J Androl.* 2000;23(4):248-253.
2. Tanner JM, Whitehouse RH, Takaishi M. Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. II. *Arch Dis Child.* 1966;41(220):613-635.
3. Vermeulen A, Verdonck L, Kaufman JM. A critical evaluation of simple methods for the estimation of free testosterone in serum. *J Clin Endocrinol Metab.* 1999;84(10):3666-3672.
4. Lu JC, Huang YF, Lu NQ. [WHO Laboratory Manual for the Examination and Processing of Human Semen: its applicability to andrology laboratories in China]. *Zhonghua Nan Ke Xue.* 2010;16(10):867-871.
5. Kasak L, Punab M, Nagirnaja L, et al. Bi-allelic Recessive Loss-of-Function Variants in FANCM Cause Non-obstructive Azoospermia. *Am J Hum Genet.* 2018;103(2):200-212.
6. Sulonen AM, Ellonen P, Almusa H, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* 2011;12(9):R94.
7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120.
8. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987-2993.
9. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics.* 2009;25(21):2865-2871.
10. Wilfert AB, Chao KR, Kaushal M, et al. Genome-wide significance testing of variation from single case exomes. *Nat Genet.* 2016;48(12):1455-1461.
11. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-315.
12. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47(D1):D886-D894.
13. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv.* 2019:531210.
14. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24-26.
15. Richards S, Aziz N, Bale 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.* 2015;17(5):405-424.
16. Gelb BD, Cave H, Dillon MW, et al. ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genet Med.* 2018;20(11):1334-1345.
17. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3--new capabilities and interfaces. *Nucleic Acids Res.* 2012;40(15):e115.
18. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* 2017;45(D1):D170-D176.

19. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol.* 2011;7(10):e1002195.
20. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics.* 2004;20(18):3702-3704.
21. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772-780.
22. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics.* 2009;25(9):1189-1191.
23. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004;14(6):1188-1190.
24. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* 2004;25(13):1605-1612.
25. Olechnovic K, Venclovas C. VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins.* 2017;85(6):1131-1145.
26. Dapkunas J, Timinskas A, Olechnovic K, Margelevicius M, Diciunas R, Venclovas C. The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures. *Bioinformatics.* 2017;33(6):935-937.
27. Renaux A, Papadimitriou S, Versbraegen N, et al. ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res.* 2019;47(W1):W93-W98.
28. Gazzo A, Raimondi D, Daneels D, et al. Understanding mutational effects in digenic diseases. *Nucleic Acids Res.* 2017;45(15):e140.
29. Papadimitriou S, Gazzo A, Versbraegen N, et al. Predicting disease-causing variant combinations. *Proc Natl Acad Sci U S A.* 2019;116(24):11878-11887.

WEB RESOURCES

NCBI ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/
The Human Protein Atlas: https://www.proteinatlas.org
gnomAD, https://gnomad.broadinstitute.org
OMIM, https://www.omim.org
Uniclust90, https://uniclust.mmseqs.com
Ensembl: https://www.ensembl.org/
UCSC Genome Browser: https://genome.ucsc.edu
Mouse Genome Informatics: http://www.informatics.jax.org
ORVAL https://orval.ibsquare.be