Biochemistry and Biophysics Reports 19 (2019) 100659

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

journal homepage: www.elsevier.com/locate/bbrep

Missense mutations in the C-terminal portion of the *B4GALNT2*-encoded glycosyltransferase underlying the Sd(a−) phenotype

Linn Stenfelt[a], Åsa Hellberg[b], Mattias Möller[a], Nicole Thornton[c], Göran Larson[d,e], Martin L. Olsson[a,b,*]

[a] Hematology and Transfusion Medicine, Department of Laboratory Medicine, Lund University, BMC C14, Sölvegatan 19, SE-22184, Lund, Sweden
[b] Department of Clinical Immunology and Transfusion Medicine, Laboratory Medicine, Office of Medical Service, F-blocket, Klinikgatan 21, SE-22185, Lund, Sweden
[c] International Blood Group Reference Laboratory, NHS Blood and Transplant, 500, North Bristol Park, Filton, Bristol, BS34 7QH, United Kingdom
[d] Department of Laboratory Medicine, Institute of Biomedicine, Sahlgrenska Academy at the University of Gothenburg, Bruna Stråket 16, SE-41345, Gothenburg, Sweden
[e] Laboratory of Clinical Chemistry, Sahlgrenska University Hospital, Bruna Stråket 16, SE-41345, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Sd[a] histo-blood group antigen
Sd(a–) phenotype
B4GALNT2
Glycosyltransferase
Red blood cell

ABSTRACT

Sd[a] is a high-frequency carbohydrate histo-blood group antigen, GalNAcβ1-4(NeuAcα2-3)Galβ, implicated in pathogen invasion, cancer, xenotransplantation and transfusion medicine. Complete lack of this glycan epitope results in the Sd(a−) phenotype observed in 4% of individuals who may produce anti-Sd[a]. A candidate gene (*B4GALNT2*), encoding a Sd[a]-synthesizing β-1,4-*N*-acetylgalactosaminyltransferase (β4GalNAc-T2), was cloned in 2003 but the genetic basis of human Sd[a] deficiency was never elucidated. Experimental and bioinformatic approaches were used to identify and characterize *B4GALNT2* variants in nine Sd(a−) individuals. Homozygosity for rs7224888:T > C dominated the cohort (n = 6) and causes p.Cys466Arg, which targets a highly conserved residue located in the enzymatically active domain and is judged deleterious to β4GalNAc-T2. Its allele frequency was 0.10–0.12 in different cohorts. A Sd(a−) compound heterozygote combined rs7224888:T > C with a splice-site mutation, rs72835417:G > A, predicted to alter splicing and occurred at a frequency of 0.11–0.12. Another compound heterozygote had two rare nonsynonymous variants, rs148441237:A > G (p.Gln436Arg) and rs61743617:C > T (p.Arg523Trp), *in trans*. One sample displayed no differences compared to Sd(a+). When investigating linkage disequilibrium between *B4GALNT2* variants, we noted a 32-kb block spanning intron 9 to the intergenic region downstream of *B4GALNT2*. This block includes *RP11-708H21.4*, a long non-coding RNA recently reported to promote tumorigenesis and poor prognosis in colon cancer. The expression patterns of *B4GALNT2* and *RP11-708H21.4* correlated extremely well in > 1000 cancer cell lines. In summary, we identified a connection between variants of the cancer-associated *B4GALNT2* gene and Sd[a], thereby establishing a new blood group system and opening up for the possibility to predict Sd(a+) and Sd(a–) phenotypes by genotyping.

1. Introduction

The Sd[a] (also known as Sid) histo-blood group antigen was described in 1967 and Sd(a+) was reported as a high-frequency red blood cell (RBC) phenotype [1,2]. Whilst ~90% of individuals tested were Sd(a+), 96% had Sd[a] substance in secretions [3,4]. Thus, 4% are truly Sd(a−), and can make anti-Sd[a]. Dialyzed urine from humans or guinea pigs is used diagnostically in reference laboratories to neutralize anti-Sd[a] in plasma [3]. The chemical basis of this enigmatic antigen was defined as GalNAcβ1-4(NeuAcα2-3)Galβ, a terminal trisaccharide on ganglioside or RBC glycoproteins [5], or on Tamm-Horsfall glycoprotein in urine [6]. The latter prevents adherence of pathogenic bacteria to urothelial cells [3]. Already in 1996 a cDNA fragment with the partial sequence of a putative Sd[a]-synthesizing enzyme was isolated [7]. In 2003, two groups independently cloned and characterized the gene, *B4GALNT2* at 17q21.32, encoding β-1,4-*N*-acetylgalactosaminyltransferase (β4GalNAc-T2) able to synthesize the terminal β1-4GalNAc linkage in Sd[a] [8,9]. However, this gene does not seem to be expressed in erythroid cells [10] but in analogy with Lewis blood group glycolipids, small amounts of Sd[a] are found in human serum [4]. In addition, a Sd[a]-synthesizing β4GalNAc-T in plasma

* Corresponding author. Division of Hematology and Transfusion Medicine, Department of Laboratory Medicine, Lund University, BMC C14, Sölvegatan 19, SE-22184, Lund, Sweden.
E-mail addresses: linn.stenfelt@med.lu.se (L. Stenfelt), asa.hellberg@skane.se (Å. Hellberg), mattias.moller@med.lu.se (M. Möller), nicole.thornton@nhsbt.nhs.uk (N. Thornton), goran.larson@clinchem.gu.se (G. Larson), Martin_L.Olsson@med.lu.se (M.L. Olsson).

https://doi.org/10.1016/j.bbrep.2019.100659
Received 20 May 2019; Received in revised form 16 June 2019; Accepted 19 June 2019
Available online 17 July 2019
2405-5808/ © 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

was described already in 1987 [11]. Both Lewis [3] and Sd^a [3,12] antigens are progressively lost from human RBCs during pregnancy. Lewis antigens are mainly produced in gastrointestinal cells but adsorbed onto RBCs from plasma [3]. Sd^a-active glycans are also made in the gastrointestinal tract [13], where they may interfere with *E.coli* binding [14]. Interestingly, Sd^a inhibits invasion of malaria parasites into RBCs [15], and *B4GALNT2* was recently identified as the key inhibitory factor for avian influenza A [16]. Furthermore, Sd^a has been proposed as a tumor marker [17], may be involved in ovine fecundity [18] and constitutes part of the porcine xenotransplantation barrier [19]. The antigen and underlying glycosyltransferase were comprehensively reviewed in 2014 [20].

Despite all these clinically important implications, human Sd^a deficiency has remained genetically unexplained, thereby preventing formation of a new blood group system as defined by the International Society of Blood Transfusion (ISBT). We aimed to resolve the genetic basis underlying the Sd(a−) phenotype, hypothesizing that alterations in *B4GALNT2* would be found, which could enable genotyping for prediction of Sd(a+) and Sd(a−) phenotypes.

2. Materials and methods

2.1. Samples

Anonymized samples of blood- (n = 6) or plasma-derived (n = 3) DNA were obtained from Sd(a−) individuals with anti-Sd^a characterized at the International Blood Group Reference Laboratory (IBGRL, Bristol, UK), Hoxworth Blood Center's Immunohematology Reference Laboratory (IRL, Cincinnati, OH, USA) and LifeShare Blood Centers (Shreveport, LA, USA). Eleven blood donors were phenotyped with in-house anti-Sd^a reagents (plasma from Sd(a−) donors with serologically defined polyclonal antibodies of anti-Sd^a specificity) and confirmed Sd(a+) using an antiglobulin tube test incubated in low ionic strength solution enhanced with 30% albumin for 30–60 min at room temperature and read microscopically. Polymorphism screening was performed on samples from apparently healthy anonymized random blood donors provided by the Department of Clinical Immunology and Transfusion Medicine, Laboratory Medicine, Office of Medical Services in Lund, Sweden, following ethical approval (LU333-99) and informed consent. A Thai cohort of donor samples was obtained from the Lampang Hospital and Saraburi Hospital, Thailand, also following ethical approval (UP-HEC 2/024/59) and informed consent.

2.2. Amplification and sequencing of B4GALNT2

DNA was prepared from whole blood with a simple salting-out method [21], or (for the Thai screening cohort) with QIAsymphony DSP DNA Mini Kit from QIAGEN (Venlo, Netherlands). DNA from plasma samples was also extracted with QIAsymphony but using the DSP virus/pathogen Midi Kit from QIAGEN according to the manufacturer's instructions. Amplification was done as previously described [22] using Expand high-fidelity PCR system (Roche, Basel, Switzerland), with some modifications. In brief, 40.5–100 ng DNA was added to a 20 μL reaction mix of 0.2 U Taq polymerase, 6.7 pmol of each primer (Table S1) and 2 nmol dNTP mix. Reactions in 35 thermal cycles (after 3 min, 95 °C): 95 °C (20 s), 58 °C (30 s), 72 °C (40 s or 3 min, depending on amplicon size). Amplicons were purified in 3% agarose gel and extracted in nuclease free H_2O with gel extraction kits QIAquick (QIAGEN) and GeneJET (Thermo Scientific, Vilnius, Lithuania). Sanger sequencing was executed by Eurofins Genomics (Ebersberg, Germany), or in house as previously described [22] with primers stated in Table S1.

2.3. Allelic discrimination assay

Allelic discrimination (AD) was performed for genetic variant screening of Swedish and Thai blood donor cohorts. TaqMan SNP genotyping assays C_25755236_10 and C_25757464_10 (Thermo Fisher Scientific, Waltham, MA, USA) were performed as follows: DNA (20 ng) was used in a reaction volume of 10 μL according to the manufacturer's protocol. The reaction was run on a QuantStudio 3 Real-Time PCR instrument and the results analyzed in QuantStudio design and analysis software c1.4.1 (Thermo Fisher Scientific).

2.4. Compilation of variants in B4GALNT2

The dataset with genomes from 2504 individuals from the 1000 Genomes Project phase 3 [23] was downloaded and variants for the *B4GALNT2* gene were imported and processed in the *Erythrogene* database [24]. A total of 1886 variants were found in the region 10 kb upstream to 10 kb downstream from the long transcript (ENST00000300404). All variants predicted to alter the amino acid sequence with a frequency of at least 0.05% were selected for further analysis. This included eighteen missense variants, four splice region variants and one stop-gaining variant. Two larger structural variants affecting the coding parts of *B4GALNT2* were found among the structural variants in the 1000 Genomes dataset [25]. A 5829 bp duplication encompassing exon 1 (esv3640734) was found in three individuals in the East Asian superpopulation, and a 2676 bp deletion containing the entire exon 6 was detected in two individuals from the same superpopulation. However, due to their rarity and the unpredictable effects of these two variants, and because of previously identified shortcomings in regards to structural variants in this dataset [24], these were not included in Table S2. Variant effect predictor results for PolyPhen-2 [26] and SIFT [27] were extracted from Ensembl release 95 [28]. Variants for *B4GALNT2* were downloaded from the gnomAD database v2.1.1 [29] (gnomad.broadinstitute.org) and variant frequencies were extracted. No additional variants, with a frequency ≥0.1% and predicted to alter the amino acid sequence, were detected. Variant frequencies from the SweGen Variant Frequency Dataset [30] (swefreq.nbis.se) were manually extracted from the online graphical browser and included in Table S2.

2.5. Protein structure modelling

We searched the RCSB Protein Data Bank [31] (rcsb.org) for the term “B4GALNT2” but found no matches. Instead, a theoretical model was generated using the protein structure homology-modelling server SWISS-MODEL [32] resulting in a model based on chondroitin polymerase from *E. coli* (PDB ID: 2Z87) [33] established by X-ray diffraction as template. The resulting structure covers amino acids 319–524 of the long isoform of β4GalNAc-T2 (NP_703147) and has 20.1% identity with the template protein. This section of the enzyme is part of the C-terminal globular catalytic domain and contains the DXD motif and the amino acids affected by our three variants of interest. Finally, the model was visualized in NGL viewer v2.0.0 (nglviewer.org) [34].

2.6. Protein sequence homology

To evaluate the level of conservation in β4GalNAc-T2 in the region affected by rs7224888 (p.Cys466Arg), a Protein BLAST search [35] was performed in the reference protein database (refseq_protein) using the UniProt ID for the enzyme (Q8NHY0) as search parameter. The search was restricted to the region around rs7224888 (amino acids 451–481) and excluded models (XM/XP), non-redundant RefSeq proteins (WP) and uncultured/environmental sample sequences. The result showed thirteen sequences from eleven species producing significant alignments, after multiple isoforms of the same protein had been excluded.

2.7. Linkage disequilibrium and splice-site prediction

Gene coordinates were extracted from Ensembl release 95 [28], except for *RP11-708H21.4* where we used Ensembl release 78 since it

had been deprecated in later versions. Linkage disequilibrium for rs7224888 in all populations in the 1000 Genomes Project [23] was calculated using the LDproxy module in LDLink (ldlink.nci.nih.gov) [36]. The resulting dataset, containing all variants ± 500 kb of rs7224888 with a pairwise R^2 value greater than 0.01, was downloaded and variants were classified using the gene coordinates. A splice-site prediction tool, Human splicing finder (www.umd.be/HSF3/), was used to predict the effects of rs72835417, the SNP found close to the splice-site in the 5′-end of intron 8 [37].

2.8. Gene expression patterns in cell lines

Gene expression levels for 56,202 genes in 1019 cell lines as determined by RNA sequencing were downloaded from the Cancer Cell Line Encyclopedia (CCLE) (version 20180929) provided by the Broad Institute (portals.broadinstitute.org/ccle) [38]. The Pearson correlation coefficients (*r*) between mRNA levels for *B4GALNT2* and all other genes were calculated in R v3.5.3. The ten genes with the most similar expression patterns, i.e. highest correlation coefficients, were extracted and compiled in Table S5.

3. Results and discussion

All *B4GALNT2* exons (Fig. 1A) and ~2 kb upstream of the 5′-end were sequenced or analyzed by allelic discrimination assays in nine Sd(a−) individuals. Four SNPs of interest were identified (Fig. 1B). Strikingly, six Sd(a−) individuals were homozygous for a missense mutation, rs7224888 (c.1396T > C) in exon 10, causing p.Cys466Arg (Fig. 1C). One individual was heterozygous for rs7224888 and a splice-site mutation, rs72835417 (c.1134+5G > A) in intron 8. Another compound heterozygote harbored missense mutations in exons 10 (rs148441237, c.1307A > G, p.Gln436Arg) and 11 (rs61743617, c.1567C > T, p.Arg523Trp). In line with bioinformatic data, allele-specific amplification and sequencing confirmed these latter SNPs to be carried on different alleles. Finally, a single Sd(a−) sample did not deviate from consensus and remains unresolved for future study when regulatory elements have been characterized.

The SNPs rs7224888, rs148441237 and rs61743617 all lead to unorthodox protein changes involving the large and positively charged residue arginine in a region, C-terminal of the DXD motif, that typically interacts with substrate molecules in a glycosyltransferase [40]. A 3D-model of β4GalNAc-T2 (Fig. 2A) was developed to estimate the approximate location of affected residues (Fig. 2B). The p.Cys466 moiety is evolutionarily well-conserved (Fig. 3), which indicates its importance for enzymatic function [40]. A synonymous SNP in exon 11, rs16946912 (c.1590A > G), frequently accompanies rs7224888 with a linkage disequilibrium (LD) of $R^2 = 0.90$. To examine this haplotype further, we analyzed the LD with all surrounding SNPs (Fig. 4A) and found a ~32-kb block spanning intron 9 to the intergenic region downstream of *B4GALNT2*. Importantly, no other SNPs in this haplotypic block cause protein changes.

The other Sd(a−)-associated variants encountered, rs148441237 and rs61743617, are rare in all populations while rs7224888 occurs at 10.6–12% in Europeans (Table S2), which could explain why it dominates this Sd(a−) study. Its allele frequency in Swedish donors (n = 200) was 10% (Table S3), which agrees well with European datasets. However, compared to the previously predicted null-allele frequency at the (then unknown) Sd^a-responsible locus (19.7%) [4,41], these three missense mutations do not explain all Sd(a−) cases, even if limited sensitivity of Sd^a detection historically may have resulted in an

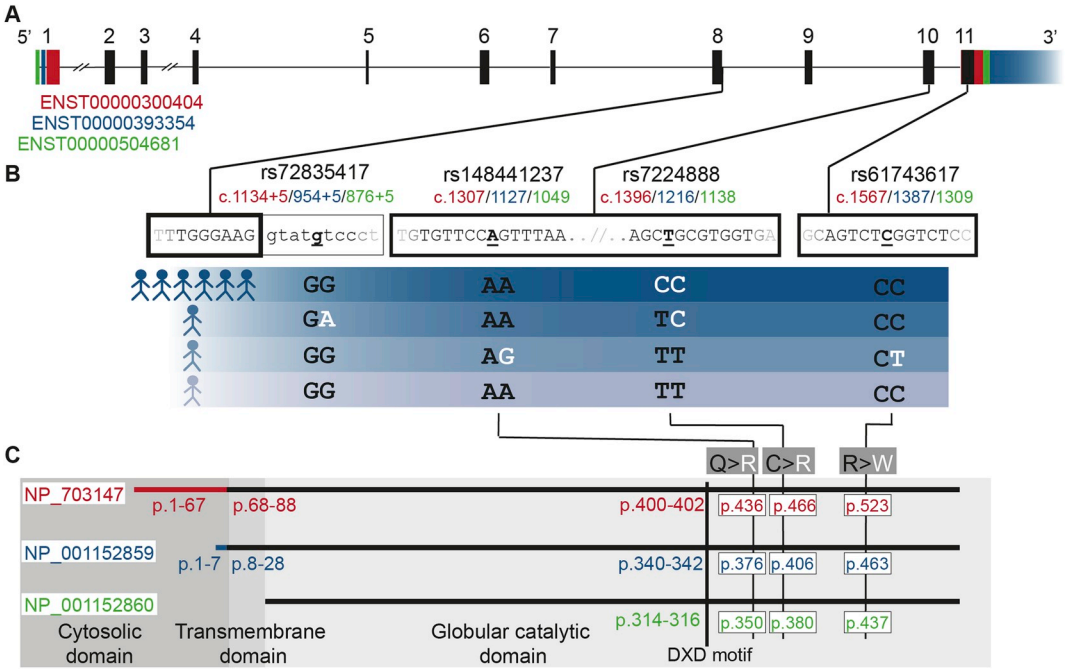

Fig. 1. The rs7224888 and additional SNPs that correlates with Sd(a−) phenotype. (A) Human *B4GALNT2* on chromosome 17 encodes three transcripts, differing in sequence due to differential use of exon 1; long (red box), short (blue box) or, a so far only theoretical, middle length exon 1 (green box) as stated in Ensembl release 96 [39]. All transcripts use the same coding sequence of exons 2–11 (black boxes), while the UTR of exon 11 differs (colored accordingly). Black horizontal lines in between the exons depict the introns. **(B)** The SNPs identified in nine individuals with the Sd(a−) phenotype, after sequencing the coding regions and the proposed promoter of the gene. Magnified sequences surrounding the SNPs (underlined) are shown at the stated nucleotide positions and below follows the SNP status of each subject symbolized by stick figures. The variants of interest are written in white. For one of the rs7224888 homozygotes, nucleotide status for rs148441237 was not established.
(C) Schematic sketch of the translated proteins from each transcript. Depending on which exon 1 is utilized the product is predicted to encode a transmembrane (NP_703147 and NP_001152859) or a soluble (NP_001152860) glycosyltransferase. The amino acid changes caused by the identified SNPs are found C-terminally of the DXD motif (in this case three consecutive aspartic acids, DDD) and are described in the dark grey boxes on the right and the amino acid positions are stated for each isoform.

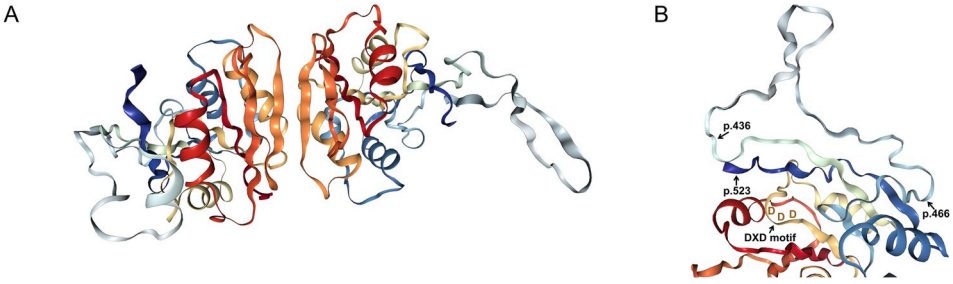

Fig. 2. The 3D protein structure of β-1,4-*N*-acetylgalactosaminyltransferase 2, based on homology modelling of the crystal structure of chondroitin synthase from *E. coli*, using SWISS-MODEL [32].
(A) The model consists of amino acids 319–524 in the catalytic domain and is colored by residue number from red in the N-terminal via yellow to blue in the C-terminal. The protein is predicted to form a dimer, as is common for glycosyltransferases.
(B) A close-up view of the structure detailing the locations of the DXD motif and the three SNPs of interest in relation to the Sd(a−) phenotype.

Glycosyltransferase	Species	Sequence identity	Accession number	Start	Sequence
β-1,4-*N*-acetylgalactosaminyltransferase 2	*Homo sapiens*	100.0%	NP_703147	458	Q P L D G F P S C V V T S G V V N F F L A H T E
	Ovis aries	83.9%	NP_001305005	398	G P L D G F P N C V V T S G V V N F F L A H T E
	Capra hircus	83.9%	NP_001301191	398	G P L D G F P N C V V T S G V V N F F L A H T E
	Sus scrofa	80.7%	NP_001231259	394	R P V D G F P D C V V T S G V V N F F L A H T E
	Mus musculus	77.4%	NP_032107	402	Q A L D G F P G C T L T S G V V N F F L A H T E
β-1,4-*N*-acetylgalactosaminyltransferase 1	*Homo sapiens*	66.7%	NP_001469	419	H E L V G F P G C V V T D G V V N F F L A R T D
	Mus musculus	66.7%	NP_032106	421	H E L V G F P S C V V T D G V V N F F L A R T D
	Xenopus tropicalis	65.5%	NP_001120168	415	H A I E G F P N C V V T D G V V N F F L A R T E
	Bos taurus	63.3%	NP_001017943	419	H E L V G F P G C V V T D G V V N F F L A R T D
	Rattus norvegicus	63.3%	NP_074051	421	H E L A G F P N C V V T D G V V N F F L A R T D
	Xenopus laevis	63.0%	NP_001079612	415	H A I E G F P N C V V T D G V I N F F L A R T E
	Salmo salar	61.3%	NP_001158803	419	H I I Q G F P N C V V T D G V I N F F L A R T D
	Danio rerio	58.1%	NP_001074039	420	H V I E G F P N C V V T D A V I N F F M A R T E

p.C466R

Fig. 3. The rs7224888 is located in a highly conserved region among different species.
Homologous proteins with significant amino acid alignments based on Protein BLAST of amino acids 451–481. Sequence identity represents similarity with human β-1,4-*N*-acetylgalactosaminyltransferase 2 (top row) in the analyzed region. Start refers to the position of the first amino acid shown on each line. The red frame shows the amino acid altered by rs7224888 (p.Cys466Arg), and the corresponding amino acids in the related proteins. Light blue, middle blue, and dark blue boxes symbolize identity between 9 and 10, 11–12, or all 13 of the compared sequences, respectively. The figure shows that a highly homologous region can be found in β-1,4-*N*-acetylgalactosaminyltransferase 1 in many species, and that the cysteine at position 466, as well as many surrounding residues, is conserved in all proteins.

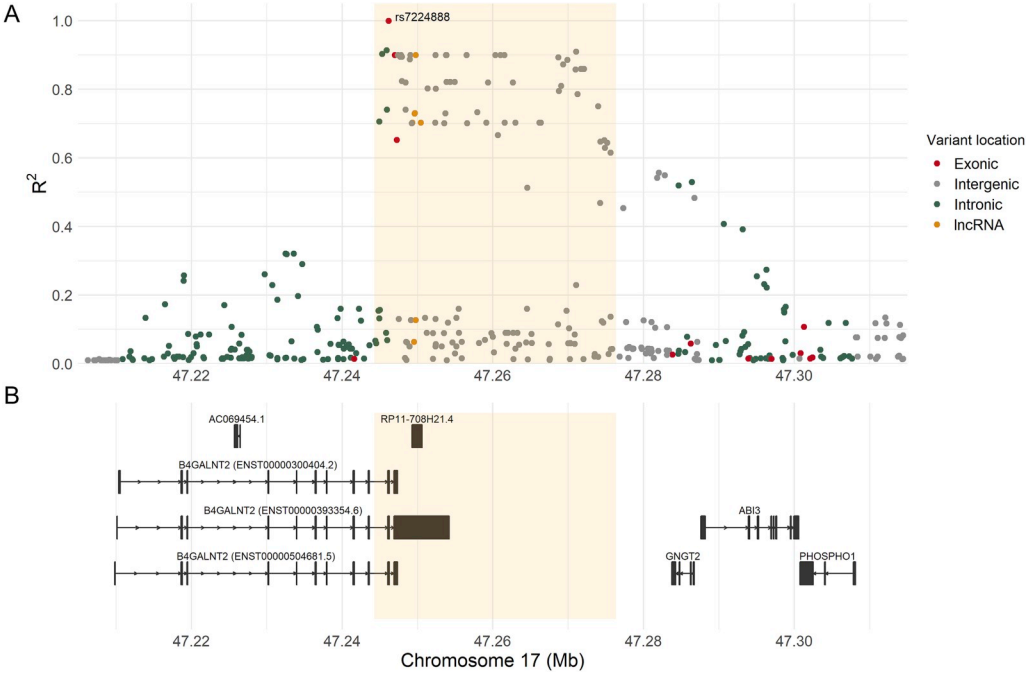

Fig. 4. The rs7224888 resides in a haplotype block of ~32 kb
(A) Linkage disequilibrium (LD) between rs7224888 and other variants in *B4GALNT2* and neighboring genes. Each dot represents a variant detected in the 1000 Genomes Project [23] and is color-coded according to its location. The long transcript (ENST00000300404) was used for color-coding variants in *B4GALNT2*. Variants in *AC069454.1* were coded as intronic. The x-axis shows the chromosomal location according to the GRCh37/hg19 human reference genome and the y-axis shows the level of LD with rs7224888 (R^2).
(B) The *B4GALNT2* transcripts and the canonical transcripts for neighboring genes *ABI3*, *GNGT2* and *PHOSPO1*. *AC069454.1* is a ribosomal protein S10 (*RPS10*) pseudogene and *RP11-708H21.4* is long non-coding RNA (lncRNA). The orange shaded area represents a presumed haplotype of ~32 kb, where 66 variants exhibit strong LD ($R^2 > 0.6$) with rs7224888. This haplotype consists of exon 10 and 11 in all *B4GALNT2* transcripts as well as *RP11-708H21.4*. The two other exonic variants in this haplotype are synonymous (rs16946912) or located in the 3′ UTR (rs28689968).

overestimated frequency. Our fourth finding, the splice-site mutation in intron 8 (rs72835417), occurred in 11% of Swedish donors' alleles (Table S3), comparing well to 11–12% in European cohorts (Table S2). Hence, it could potentially make up for the "missing" percentage of null alleles but given their similar frequencies, we would have expected a more balanced contribution of rs7224888 and rs72835417 towards the genetic background of Sd(a−). Instead, we observed an imbalance between rs7224888 (13 alleles) and rs72835417 (one allele) among nine Sd(a−) samples. Importantly, rs7224888 and rs72835417 constituted 14 of 18 *B4GALNT2* alleles in the Sd(a−) group whilst none of the eleven serologically phenotyped Sd(a+) control donors were homozygous for either allele, nor compound heterozygous (Table S3). The frequency of Sd(a−) in Sweden is unknown but screening of 200 random Swedish blood donors resulted in a total of 7.5% predicted Sd(a−) donors, i.e. 2% rs7224888 homozygotes, 1.5% rs72835417 and 4% compound heterozygotes (Table S3), which is somewhat higher than observed phenotypically in other Europeans [3,4].

In silico analysis of similar (+5G > A) splice-site alterations, predicts probable effects on splicing but further experimental work to verify this is required. Skipping of exon 8 would alter the reading frame, generating a β4GalNAc-T2 with truncated catalytic domain (p.Asn316GlyfsTer6). It can be speculated that some transcripts may escape loss of exon 8 and that homozygosity for rs72835417 could lead to very weak Sd^a expression and preclude production of anti-Sd^a, which was a selection criterium for this study. Due to the distance between rs7224888 and rs72835417 we were unable to confirm that these SNPs occur on different chromosomes in our compound heterozygous case. However, in-house screening and the 1000G dataset support that they are not linked but occur separately in the vast majority of cases.

To our knowledge, the Sd(a−) frequency has not been established outside of Europe, although a relatively high frequency of individuals with elevated Sd^a expression was reported among Thai donors [42]. This is interesting since this phenotype has been reported to inhibit malaria invasion [15]. We therefore examined SNP frequencies among Thai donors and found them low for both rs7224888 (3.3%) and rs72835417 (0.5%) with no predicted Sd(a−) donors (Table S4). In general, it is thought-provoking to note the unusually high variation in SNP frequency between the 1000G superpopulations, e.g. 2.2% in East Asians vs. 20% in South Asians and Africans for rs7224888 (Table S2).

Given the reported cancer association for Sd^a, we also investigated co-expression patterns between *B4GALNT2* and > 56,000 transcripts in the cancer cell line encyclopedia (CCLE) and found *RP11-708H21.4* to correlate exceptionally well (Table S5). Interestingly, this long non-coding (lnc) RNA is located just downstream of *B4GALNT2* (Fig. 4B) on the sense strand, and was recently reported to predict poor prognosis of colorectal cancer and promote tumorigenesis in a study unrelated to Sd^a [43]. Loss of Sd^a has been reported in colon cancer [7,13] so we speculate that *B4GALNT2* and *RP11-708H21.4* may be regulated by the same mechanism/element. Taken together, this may suggest that *RP11-708H21.4* expression may have a causative role whilst Sd^a serves as a marker.

In summary, we identified mutations in the cancer-associated histo-blood group candidate gene *B4GALNT2* in Sd(a−) individuals. Based on this, genotyping to predict Sd^a status is made possible and the Sd^a antigen can be moved from the ISBT series of high-frequency blood group antigens to form a new blood group system, which we propose to designate SID.

Note added at proof stage: Following a presentation of the above data on 22 June 2019, the ISBT Working Party on Red Cell Immunogenetics and Blood Group Terminology decided to acknowledge SID as a new blood group system with the official ISBT number 038.

Conflict of interest disclosures

The authors declare no competing financial interests.

Accession numbers

The following sequences identified during this study were deposited in GenBank:

MK765047 (rs7224888), MK765048 (rs148441237), MK765049 (rs61743617) and MK797056 (rs72835417).

Authorship contributions

LS and ÅH performed laboratory experiments, and interpreted data together with MM and MLO. MM performed bioinformatic analyses. MLO and GL conceived and designed the study which was coordinated by ÅH and MLO. NT serologically characterized samples for the study. LS, MM and MLO wrote the manuscript. All authors read, revised and approved the manuscript.

Acknowledgements

Gregory R. Halverson, the manager of the Immunohematology Reference Laboratory, LifeSouth Community Blood Centers, Atlanta, GA, USA, is thanked for providing archived plasma samples. Philaiphon Jongruamklang who collected the samples in the Thai cohort is gratefully acknowledged. We would like to honor the memory of Dr. Ammi Grahn who first identified the c.1396T > C allele in Sd(a−) samples when working on this project with us before her passing in 2016. We are also grateful to Dr. Elisabet Sjöberg Wester, who contributed to the early phases of this work, along with her deputy supervisor Assoc. Prof. Jill Storry, who is also thanked for technical assistance. The study was supported by the Knut and Alice Wallenberg Foundation (2014.0312) to MLO, the Swedish Research Council (2014-71X-14251) to MLO and governmental ALF grants to the university healthcare in Region Skåne, Sweden (ALFSKANE-446521) to MLO.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbrep.2019.100659.

Transparency document

Transparency document related to this article can be found online at https://doi.org/10.1016/j.bbrep.2019.100659

References

[1] P.H. Renton, P. Howell, E.W. Ikin, et al., Anti-Sd^a, a new blood group antibody, Vox Sang. 13 (1967) 493–501.
[2] S.I. Macvie, J.A. Morton, M.M. Pickles, The reactions and inheritance of a new blood group antigen, Sd^a, Vox Sang. 13 (1967) 485–492.
[3] M.E. Reid, C. Lomas-Francis, M.L. Olsson, The Blood Group Antigen FactsBook, third ed., Academic Press, London, UK, 2012.
[4] J.A. Morton, M.M. Pickles, A.M. Terry, The Sd^a blood group antigen in tissues and body fluids, Vox Sang. 19 (1970) 472–482.
[5] D. Blanchard, C. Capon, Y. Leroy, et al., Comparative study of glycophorin A derived O-glycans from human Cad, Sd(a+) and Sd(a-) erythrocytes, Biochem. J. 232 (1985) 813–818.
[6] A.S. Donald, A.D. Yates, C.P. Soh, et al., A blood group Sd^a-active pentasaccharide isolated from Tamm-Horsfall urinary glycoprotein, Biochem. Biophys. Res. Commun. 115 (1983) 625–631.
[7] T. Dohi, Y. Yuyama, Y. Natori, et al., Detection of *N*-acetylgalactosaminyl-transferase mRNA which determines expression of Sd^a blood group carbohydrate structure in human gastrointestinal mucosa and cancer, Int. J. Cancer 67 (1996) 626–631.
[8] L. Lo Presti, E. Cabuy, M. Chiricolo, et al., Molecular cloning of the human β1,4 *N*-acetylgalactosaminyltransferase responsible for the biosynthesis of the Sd^a histo-blood group antigen: the sequence predicts a very long cytoplasmic domain, J. Biochem. 134 (2003) 675–682.
[9] M.D. Montiel, M.A. Krzewinski-Recchi, P. Delannoy, et al., Molecular cloning, gene organization and expression of the human UDP-GalNAc:Neu5Acα2-3Galβ-R β1,4-*N*-acetylgalactosaminyltransferase responsible for the biosynthesis of the blood group Sd^a/Cad antigen: evidence for an unusual extended cytoplasmic domain, Biochem. J. 373 (2003) 369–379.

[10] M. Jöud, M. Möller, M.L. Olsson, Identification of human glycosyltransferase genes expressed in erythroid cells predicts potential carbohydrate blood group loci, Sci. Rep. 8 (2018) 6040.
[11] A. Takeya, O. Hosomi, T. Kogure, Identification and characterization of UDP-GalNAc: NeuAcα2 − 3Galβ1 − 4Glc(NAc) β1 − 4(GalNAc to Gal)*N*-acetylgalactosaminyltransferase in human blood plasma, J. Biochem. 101 (1987) 251–259.
[12] S. Spitalnik, M.T. Cox, J. Spennacchio, et al., The serology of Sd^a effects of transfusion and pregnancy, Vox Sang. 42 (1982) 308–312.
[13] N. Malagolini, F. Dall'Olio, G. Di Stefano, et al., Expression of UDP-GalNAc:NeuAcα2,3Galβ-R β1,4(GalNAc to Gal)*N*-acetylgalactosaminyltransferase involved in the synthesis of Sd^a antigen in human large intestine and colorectal carcinomas, Cancer Res. 49 (1989) 6466–6470.
[14] F. Serafini-Cessi, A. Monti, D. Cavallone, N-Glycans carried by Tamm-Horsfall glycoprotein have a crucial role in the defense against urinary tract diseases, Glycoconj. J. 22 (2005) 383–394.
[15] J.P. Cartron, O. Prou, M. Luilier, et al., Susceptibility to invasion by *Plasmodium falciparum* of some human erythrocytes carrying rare blood group antigens, Br. J. Haematol. 55 (1983) 639–647.
[16] B.E. Heaton, E.M. Kennedy, R.E. Dumm, et al., A CRISPR activation screen identifies a pan-avian influenza virus inhibitory host factor, Cell Rep. 20 (2017) 1503–1512.
[17] M. Tanaka-Okamoto, K. Hanzawa, M. Mukai, et al., Identification of internally sialylated carbohydrate tumor marker candidates, including Sd^a/CAD antigens, by focused glycomic analyses utilizing the substrate specificity of neuraminidase, Glycobiology 28 (2018) 247–260.
[18] X. Guo, X. Wang, B. Liang, et al., Molecular cloning of the *B4GALNT2* gene and its single nucleotide polymorphisms association with litter size in small tail han sheep, Animals 8 (2018) 160.
[19] G. Byrne, S. Ahmad-Villiers, Z. Du, et al., B4GALNT2 and xenotransplantation: a newly appreciated xenogeneic antigen, Xenotransplantation 25 (2018) e12394.
[20] F. Dall'Olio, N. Malagolini, M. Chiricolo, et al., The expanding roles of the Sd^a/Cad carbohydrate antigen and its cognate glycosyltransferase B4GALNT2, Biochim. Biophys. Acta 1840 (2014) 443–453.
[21] S.A. Miller, D.D. Dykes, H.F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells, Nucleic Acids Res. 16 (1988) 1215.
[22] Å. Hellberg, R. Steffensen, V. Yahalom, et al., Additional molecular bases of the clinically important p blood group phenotype, Transfusion 43 (2003) 899–907.
[23] The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (2015) 68–74.
[24] M. Möller, M. Jöud, J.R. Storry, et al., Erythrogene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project, Blood Adv 1 (2016) 240–249.
[25] P.H. Sudmant, T. Rausch, E.J. Gardner, et al., An integrated map of structural variation in 2,504 human genomes, Nature 526 (2015) 75–81.
[26] I.A. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[27] R. Vaser, S. Adusumalli, S.N. Leng, et al., SIFT missense predictions for genomes, Nat. Protoc. 11 (2016) 1–9.
[28] D.R. Zerbino, P. Achuthan, W. Akanni, et al., Ensembl 2018, Nucleic Acids Res. 46 (2018) D754–D761.
[29] M. Lek, K.J. Karczewski, E.V. Minikel, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291.
[30] A. Ameur, J. Dahlberg, P. Olason, et al., SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population, Eur. J. Hum. Genet. 25 (2017) 1253–1260.
[31] H.M. Berman, J. Westbrook, Z. Feng, et al., The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[32] A. Waterhouse, M. Bertoni, S. Bienert, et al., SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (2018) W296–W303.
[33] T. Osawa, N. Sugiura, H. Shimada, et al., Crystal structure of chondroitin polymerase from *Escherichia coli* K4, Biochem. Biophys. Res. Commun. 378 (2009) 10–14.
[34] A.S. Rose, A.R. Bradley, Y. Valasatava, et al., NGL viewer: web-based molecular graphics for large complexes, Bioinformatics 34 (2018) 3755–3758.
[35] S.F. Altschul, W. Gish, W. Miller, et al., Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[36] M.J. Machiela, S.J. Chanock, LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants, Bioinformatics 31 (2015) 3555–3557.
[37] F.O. Desmet, D. Hamroun, M. Lalande, et al., Human Splicing Finder: an online bioinformatics tool to predict splicing signals, Nucleic Acids Res. 37 (2009) e67.
[38] J. Barretina, G. Caponigro, N. Stransky, et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 (2012) 603–607.
[39] F. Cunningham, P. Achuthan, W. Akanni, et al., Ensembl 2019, Nucleic Acids Res. 47 (2019) D745–d751.
[40] C. Breton, L. Šnajdrová, C. Jeanneau, et al., Structures and mechanisms of glycosyltransferases, Glycobiology 16 (2006) 29r–37r.
[41] R.R. Race, R. Sanger, Blood groups in man, Oxford: Blackwell Scientific, 1975.
[42] S. Sringarm, P. Chiewsilp, J. Tubrod, Cad receptor in Thai blood donors, Vox Sang. 26 (1974) 462–466.
[43] L. Sun, C. Jiang, C. Xu, et al., Down-regulation of long non-coding RNA RP11-708H21.4 is associated with poor prognosis for colorectal cancer and promotes tumorigenesis through regulating AKT/mTOR pathway, Oncotarget 8 (2017) 27929–27942.