Supplemental digital content-1

METHODS

This study was approved by the Institutional Review Board (IRB#2, Pediatric and Pregnant Woman IRB) of Orlando Health, Orlando, FL.

Patients

We limited our study to non-Hispanic white pediatric patients because of the relative scarcity of samples and the lower prevalence of reported pathogenic mutations in other races.

Power analysis with alpha 0.05, beta 0.2/power 0.8 gave a sample size of 109. We enrolled 125 abnormal (low) sucrase cases considering potential data contraction during processing and assay. We selected 500 normal sucrase cases (250 with moderate and 250 with high sucrase activities) to get 4:1 ratio of normal (n=500) vs abnormal (n=125) sucrase cases. All 625 patients had normal duodenal histology and none had underlying diseases such as celiac disease that would have influenced the enzyme activities. The mean age was 12.6 ± 4.5 years in the abnormal sucrase activity group, 10.5 ± 5.3 years in the moderate normal sucrase and 9.7 ± 5.1 years in the high normal sucrase activity groups (Table 1A). The gender distribution was essentially equal in all three groups.

The prevalence of *SI* gene mutations was 1.2% in sequenced non-Hispanic whites in the Exome Variant Server, a relevant proxy database (National Heart, Lung, and Blood Institute, Exome Variant Server, GO Exome Sequencing Project. http://evs.gs.washington.edu/EVS, 2012) and the prevalence of heterozygosity in the *SI* gene in IBS was reported as twice higher (1-3), therefore we expected that at least 5% of cases with abnormal sucrase will have a *SI* gene mutation.

Disaccharidase assays

Disaccharidase activities were measured in the authors' laboratory following a modified Dahlqvist method (4-6). Duodenal biopsy specimens were kept frozen (-80^{o}C) until used for the assays. A modified Dahlqvist method was used. The biopsies were homogenized and using lactose, maltose, sucrose, palatinose (isomaltose), and maltodextrose substrates, the glucose production was measured by glucose oxidase. Total protein concentration was measured using Pierce BCA protein assay kit (Pierce, Thermo Scientific, USA). Specific activities of enzymes were expressed as units (U), defined as micromoles of glucose released per minute per gram of mucosal protein at 37°C. The cut off values of enzymes were based on the analysis of over 9000 assays using standard protocols. Low (abnormal) disaccharidase activity was defined by – 2 SD below the means. Low or abnormal enzyme activities were considered below the following cutoff values: sucrase, <25.8U; lactase, <15.4U; maltase <103.7U; palatinase, <8.6 U; and glucoamylase, <24.6.

Next-Generation Sequencing of *SI* gene exons using FFPE Tissue DNA

Next-Generation Sequencing (NGS; Illumina Inc. San Diego, CA, USA) of the entire coding sequence (48 exons) of the *SI* gene using DNA extracted from formalin fixed paraffin embedded (FFPE) tissue samples was performed to detect known pathogenic *SI* gene or CSID variants (Table S1, Supplemental Digital Content-2). DNA extraction was optimized using QIAamp DNA FFPE Tissue Kit (QIAGEN, USA). DNA samples were purified and checked for quality as per Illumina's guidelines. The NGS method was optimized with FFPE tissue DNA for Illumina's MiSeq platform using custom amplicon design [TruSeq Custom Amplicon (TSCA) and AmpliSeq custom panel gene designs] targeting all the 48 exons of the *SI* gene. A lab optimized TruSeq Custom Amplicon Low Input Library preparation protocol and AmpliSeq protocol were used to generate the sequencing libraries and sequencing was performed using V3 600 cycle

sequencing reagent on the MiSeq platform (Illumina Inc., USA). Run quality parameters were maintained within the limits of good sequencing output. Sequence data was analyzed using MiSeq Reporter software to generate variant call files (VCF). The VCF files were analyzed with Illumina's Variant Studio software to identify the variants using the human genome variant database. For AmpliSeq sequencing protocol the sequencing data were analyzed using the Illumina's cloud based DNA Amplicon BaseSpace App to generate VCF files (https://support.illumina.com/help/BaseSpace_App_DNA_Amplicon_v2_OLH_1000000041403/Content/Source/HomePages/Home_Page_DNA_Amplicon_App.htm). BaseSpace Variant Interpreter, an interpretation and reporting platform, was used to identify the variants using the VCF files (https://support.illumina.com/help/BaseSpace_VariantInterpreter_OLH_001129/Content/Source/HomePages/Home_Page_BaseSpace_Variant_Interpreter.htm).

A list of 41 *SI* gene mutations including known CSID genetic variants was prepared based on various published studies and pathogenic probability analysis available through SNP databases including the Exome Variant Server (EVS) ExAC database (7-9) (Table S1, Supplemental Digital Content-2).

Study design

Study subjects were selected from the laboratory database based on the sucrase activities. Then the electronic medical records were used to collect clinical histories, symptoms and management by the ordering physicians. The FFPE tissue samples of the 625 cases from the previous 12 years (from 2006 to 2018) were used for DNA extraction to assess the prevalence of *SI* gene mutations.

As mentioned above, the study subjects were classified into three groups based on sucrase activities: (a) low or abnormal sucrase activities, (b) moderate normal sucrase activities, and (c) high normal sucrase activities. The selection criteria for the abnormal sucrase group were (i) abnormal sucrase activity (≤ 25.8U), (ii) a primary symptom of functional gastrointestinal disorders (FGIDs). The selection criteria for the moderate sucrase activity group were (i) a moderate level of sucrase activity ≥25.8 U - ≤55 U, (ii) FGIDs may or may not be present. The high sucrase activity group patients had (i) sucrase activity >55 U, (ii) FGIDs may or may not be present. Patterns of disaccharidase deficiencies and clinical management of the low sucrase cases were also added to the database.

We reviewed clinical symptoms to determine FGIDs by using Rome IV classification (10-12).

Statistical analysis of clinical and *SI* gene variant data

We used Pearson's Chi-square to test the cumulative frequency of pathogenic *SI* variants identified compared to the high normal sucrase activity group as seen in Table 1B.

We plotted the average sucrase activities across variants within the 36 low sucrase case group. We also plotted the individual sucrase activities of cases within each variant on top of the bars. We point out that there are very few low sucrase cases and therefore we expect a lot of variance in measurements of sucrase activity split by variant. Some variant groups only have one patient, and show highly varied low sucrase activities.

We used multinomial logistic regression with LASSO (L1 regularization) and cross validation on variant presence data to predict whether a patient had low (abnormal), and moderate or high normal sucrase activity. LASSO (least absolute shrinkage and selection operator) was used to identify the genetic variants most predictive of patient classification with cross validation over regularization

coefficients from 100, 10, 3, 2.5, 2, 1.67, 1.43, 1, 0.67, 0.5, and 0.4 (13, 14). We weighted the three classes to balance differences in frequency across the three classes.

We adopt the LASSO method with regularization coefficient 1.67. Examining the resulting weight matrix from the regression model, we see that the important variants for predicting low (abnormal), moderate normal, or high normal sucrase category are largely consistent with the predicted pathogenic variants. Importance is defined by high absolute value weight for a given variant (taking the maximum weight for a variant across sucrase activity categories).

The purpose of this regression is not to build a good classifier, but rather to perform feature selection on the variants to determine which of the detected pathogenic variants chosen from literature were most important for separating normal (moderate and high) and low or abnormal sucrase activity cases given our collected data.

We also performed more visual analysis by plotting the frequency of low enzyme activity vs. variant. For each variant, we assessed the frequency of low enzyme activity. We ignored variants with sample sizes less than 5 as well as highly prevalent clinically silent variants p.Val15Phe, p.Met1523Ile, and p.Thr231Ala, which are common in public variant datasets such as ExAC as shown in Figure 2. We used the same method to plot symptom frequency vs. variants in Figure 2. Abdominal pain is generally more common than diarrhea and thus has higher frequency.

We used sucrase:lactase ratio to classify low (abnormal) sucrase patients with at least one pathogenic variant. We then plotted an ROC for this classifier. This turns out to be a moderate classifier, giving an AUC (area under curve) of 0.71. The AUC is above chance due to the fact that amongst the cases, the average ratio of sucrase to lactase is lower than the ratio of sucrase to lactase amongst the controls (most cases have only slightly lower lactase activity but much lower sucrase

activity). A cutoff value of 1.43 (meaning ratios below 1.43 are classified as having at least one pathogenic variant) gives an accuracy of 75.2% with sensitivity of 86.5% and specificity of 50.0%%. Note: This disagrees with the recommendation for CSID diagnosis using sucrase:lactase ratio by Treem (Treem, 2012; Ref #9). Using a threshold for sucrase:lactase that is <1 will result in a lower sensitivity (higher false negative rate for CSID diagnosis).

To perform all these analyses, Python 3 was used with the Pandas (15), NumPy (16), Scikit-learn (17), and Matplot libraries (18).

References:

1 Garcia-Etxebarria K, Zheng T, Bonfiglio F, et al. Increased Prevalence of Rare Sucrase-isomaltase Pathogenic Variants in Irritable Bowel Syndrome Patients. Clin Gastroenterol Hepatol 2018;16(10):1673-76.

2 Henstrom M, Diekmann L, Bonfiglio F, et al. Functional variants in the sucrase-isomaltase gene associate with increased risk of irritable bowel syndrome. Gut 2018;67(2):263-70.

3 Husein DM, Wanes D, Marten LM, et al. Heterozygotes Are a Potential New Entity among Homozygotes and Compound Heterozygotes in Congenital Sucrase-Isomaltase Deficiency. Nutrients 2019;11(10).

4 Dahlqvist A Assay of intestinal disaccharidases. Scand J Clin Lab Invest 1984;44(2):169-72.

5 Dahlqvist A, Hammond JB, Crane RK, et al. Assay of Disaccharidase Activities in Peroral Biopsies of the Small-Intestinal Mucosa. Acta Gastroenterol Belg 1964;27(543-55.

6 Dahlqvist A, Nordstrom C The distribution of disaccharidase activities in the villi and crypts of the small-intestinal mucosa. Biochim Biophys Acta 1966;113(3):624-6.

7 Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536(7616):285-91.

8 Uhrich S, Wu Z, Huang JY, et al. Four mutations in the SI gene are responsible for the majority of clinical symptoms of CSID. J Pediatr Gastroenterol Nutr 2012;55 Suppl 2(S34-5.

9 Sander P, Alfalah M, Keiser M, et al. Novel mutations in the human sucrase-isomaltase gene (SI) that cause congenital carbohydrate malabsorption. Hum Mutat 2006;27(1):119.

10 Edwards T, Friesen C, Schurman JV Classification of pediatric functional gastrointestinal disorders related to abdominal pain using Rome III vs. Rome IV criterions. BMC Gastroenterol 2018;18(1):41.

11 Palsson OS, Whitehead WE, van Tilburg MA, et al. Rome IV Diagnostic Questionnaires and Tables for Investigators and Clinicians. Gastroenterology 2016.

12 Lacy BE, Patel NK Rome Criteria and a Diagnostic Approach to Irritable Bowel Syndrome. J Clin Med 2017;6(11).

13 Friedman J, Hastie T, Tibshirani R Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9(3):432-41.
14 Friedman J, Hastie T, Tibshirani R Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33(1):1-22.
15 McKinney W Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference 2010:51-56.
16 Stéfan van der Walt SCCaGV The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering 2011;13(22-30.
17 Fabian Pedregosa GV, Alexandre Gramfort, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12(2825-30.
18 Hunter JD Matplotlib: A 2D Graphics Environment Computing in Science & Engineering 2007;9(90-95.

Table S1. List of 41 *SI* Gene and CSID specific Variants

Chromosome	Position*	RSID	Reference	Alternate	Protein Consequence	Transcript Consequence	Annotation	Grantham Score**	PolyPhen***	Comment
CSID Variants										
3	165046998	rs121912615	A	C	**p.Val577Gly**	c.1730T>G	missense	109	probable	13 variants detected in abnormal sucrase cases are in bold fonts (blue and black)
3	164983015	rs79717168	A	C	**p.Phe1745Cys**	c.5234T>G	missense	205	probable	
3	165021265	rs121912616	C	T	**p.Gly1073Asp**	c.3218G>A	missense	94	probable	
3	165059937	rs138434001	C	T	**p.Val371Met**	c.1111G>A	missense	21	probable	**Variants in Blue fonts are most probable pathogenic variants -Val577Gly, Gly1073Asp, Phe1745Cys, Pro348Leu, and Val371Met as per our and other data, and predictive analysis also supports their pathogenic potentials**
3	165060005	rs77546399	G	A	**p.Pro348Leu**	c.1043C>T	missense	98	probable	
3	165049844	rs144972103	C	A	**p.Gly515Val**	c.1544G>T	missense	109	probable	
3	165009325	rs148831941	A	C	**p.Ile1378Ser**	c.4133T>G	missense	142	Probable	
3	165023746	rs146785675	A	G	**p.Tyr975His**	c.2923T>C	missense	83	probable	
3	164991462	rs142018224	C	G	**p.Val1667Leu**	c.4999G>C	missense	32	benign	
3	165037925	rs200972419	C	A	**p.Glu801Ter**	c.2401G>T	stop gained	NA	NA	
3	165009359	rs143388292	T	C	**p.Arg1367Gly**	c.4099A>G	missense	125	probable	
3	165037931	rs150246328	T	C	**p.Ile799Val**	c.2395A>G	missense	29	benign	
3	165030864	rs199706219	G	T	**p.Leu914Ile**	c.2740C>A	missense	5	benign	
3	165019655	rs200451408	G	A	p.Arg1124Ter	c.3370C>T	stop gained	NA	NA	
3	165038006	rs147207752	T	C	p.Arg774Gly	c.2320A>G	missense	125	probable	
3	165049235	rs376816463	T	A	p.Asp536Val	c.1607A>T	missense	152	Possible	
3	165007927	rs142090504	A	C	p.Tyr1417Ter	c.4251T>G	stop gained	NA	NA	
3	164982379	rs145556619	C	T	p.Gly1760Asp	c.5279G>A	missense	94	benign	
3	164982379	rs145556619	C	A	p.Gly1760Val	c.5279G>T	missense	109	benign	
3	164982274	rs139504152	G	A	p.Ser1795Leu	c.5384C>T	missense	145	benign	
3	165015986	rs375443860	A	G	p.Ile1285Thr	c.3854T>C	missense	89	probable	
3	165032659	rs140230726	A	G	p.Tyr867His	c.2599T>C	missense	83	probable	
3	164992390	rs202225928	C	G	p.Asp1617His	c.4849G>C	missense	81	benign	
3	165067419	rs142447888	A	G	p.Ser186Pro	c.556T>C	missense	74	possible	
3	164982253	rs9917722	G	C	p.Thr1802Ser	c.5405C>G	missense	58	benign	
3	165030815	rs150927256	T	C	p.Gln930Arg	c.2789A>G	missense	43	benign	
3	164998629	rs145246112	C	T	p.Arg1484His	c.4451G>A	missense	29	probable	
3	164992209	rs139876383	C	T	p.Val1651Ile	c.4951G>A	missense	29	benign	
3	165019732	rs121912611	T	G	p.Gln1098Pro	c.3293A>C	missense	76	NA	
3	165046891	rs201055347	C	A	p.Glu613Ter	c.1837G>T	stop gained	NA	NA	
3	164998653	rs758043919	C	G	p.Gly1476Ala	c.4427G>C	missense	60	NA	
3	164996597	rs767701775	G	A	p.Arg1544Cys	c.4630C>T	missense	180	NA	
3	165046948	rs765433197	A	G	p.Ser594Pro	c.1780T>C	missense	74	NA	
3	165041019	rs780664460	T	A	p.Thr694Ser	c.2080A>T	missense	58	NA	
3	164996634	rs779692980	A	C	p.Cys1531Trp	c.4593T>G	missense	215	NA	
3	164994281	rs376062850	G	A	p.Thr1606Ile	c.4817C>T	missense	89	NA	
3	165039910	rs771409581	G	A	p.Leu741Phe	c.2221C>T	missense	22	NA	
3	165038005	rs143885457	C	T	p.Arg774Lys	c.2321G>A	missense	26	benign	
Common Mutations										
3	165075970	rs9290264	C	A	p.Val15Phe	c.43G>T	missense	50	possible	High frequency in all groups
3	165065377	rs9283633	T	C	p.Thr231Ala	c.691A>G	missense	58	benign	High frequency in all groups
3	164996744	rs4855271	C	T	p.Met1523Ile	c.4569G>A	missense	10	benign	High frequency in all groups

***Position:** Reference genome GRCh38 (GRCh38 = Genome Reference Consortium Human Genome Build 38; synonym hg38 - UCSC Genome Browser assembly ID: hg38; UCSC, University of California Santa Cruz).
****Grantham Score:** Predicts the distance between two amino acids and thus indicates the possible impact of amino acid substation caused by genetic variation. Higher score indicates higher chances of having deleterious effects. Ref: Grantham, R. (1974) Amino-acid difference formula to help explain protein evolution. Science, 185, 862–864.
*****PolyPhen (Polymorphism phenotyping):** PolyPhen tool is used to predict possible effect of amino acid substitution in a protein caused by genetic variants. Ref: Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E.,Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010). A method and server for predicting damaging missense mutations.
Nat. Methods, 7, 248–249.
Grantham score and PolyPhen ref: Knecht C1, Mort M2, Junge O1, Cooper DN2, Krawczak M1, Caliebe A1. 2017. IMHOTEP-a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants. Nucleic Acids Res. 2017 Feb 17; 45(3):e13.
NA, not available.

Table S2: Clinical symptoms, FGIDs diagnosis, and *SI* gene variants detected in abnormal sucrase patients

Low/abnormal Sucrase Case	Clinical symptoms				Diagnosis			Type of mutation	Pathogenic variants (n=13)												
	Diarrhea	Constipation	Nausea	Vomit	FGID	LD	PDD		p.Val577Gly	p.Gly1073Asp	p.Phe1745Cys	p.Pro348Leu	p.Val371Met	p.Ile1378Ser	p.Tyr975His	p.Val1667Leu	p.Glu801Ter	p.Arg1367Gly	p.Gly515Val	p.Ile799Val	p.Leu914Ile
1	+				IBS-D			1	**1**												
2					FAP			1			**1**										
3		+	+	+	IBS-C, FV	LD	PDD	1				**1**									
4			+	+	NERD	LD		1		**1**											
5			+	+	NERD	LD		1		**1**											
6			+	+	NERD			1		**1**											
7	+	+	+		IBS-M	LD		1	**1**												
8	+		+		IBS-D			1								**1**					
9	+		+	+	FD	LD		1					**1**								
10			+		NERD			2		**1**								**1**			
11	+		+	+	FAP	LD		1					**1**								
12					FAP	LD	PDD	1	**1**												
13					FAP	LD	PDD	1			**1**										
14			+		FAP			1			**1**										
15	+	+			IBS-M	LD	PDD	1	**1**												
16					FAP-IBD	LD		1											**1**		
17		+	+	+	FAP			**4**					**2**						**2**		
18					FAP	LD	PDD	**3**				**2**									
19	+				IBS-D	LD		1			**1**										
20			+		NERD			1				**1**									
21	+		+		IBS-D, FD	LD	PDD	1			**1**										
22					FAP	LD	PDD	2		**1**					**1**						
23			+	+	FD	LD	PDD	1							**1**						
24					FD	LD		1	**1**												
25					FAP			1				**1**									
26			+	+	FD			1		**1**											
27	+		+		IBS-D	LD	PDD	1												**1**	
28		+	+		IBS-C	LD	PDD	2	**1**				**1**								
29	+		+		IBS-D	LD	PDD	2	**1**					**1**							
30	+				IBS-D	LD	PDD	1	**1**												
31			+		FD	LD	PDD	1		**1**											
32				+	FD			1			**1**										
33					NERD			1			**1**										
34			+	+	NERD	LD		1									**1**				
35		+			IBS-C			2	**1**												**1**
36			+		FD	LD		1	**1**												

FGIDs = Functional gastrointestinal disorders; LD = Lactase deficient; PDD = Pan-disaccharidase deficiency; FD = Functional dyspepsia; IBS-C = Irritable bowel syndrome with constipation; IBS-D = Irritable bowel syndrome with diarrhea; IBS-M = Irritable bowel syndrome mixed; NERD = Non-erosive reflux disease; and FAP = Functional Abdominal Pain.

Clinical symptoms: +, Symptom present and left blank if symptom not reported.

Type of mutations: 1 (heterozygous), 2 (compound heterozygous), 3 (homozygous), 4 (combined homozygous).

Mutation scoring: 0 (no mutation detected), 1 (heterozygous), and 2 (homozygous).

Five most significant *SI* gene pathogenic variants (1. p.Val577Gly, 2. p.Gly1073Asp, 3. p.Phe1745Cys, 4. p.Pro348Leu, 5.p.Val371Met) were detected in 31 of 36 abnormal sucrase cases (31/36; 86%; see Table S4).

Table S3: Different forms and combinations of variants detected in all the sucrase activity groups

Abnormal - Low Sucrase Group (≤25.8U)			Normal - Moderate Sucrase Group (≥25.8U to ≤55U)			Normal - High Sucrase Group (>55U)		
Heterozygous		Comments	Heterozygous		Comments	Heterozygous		Comments
Variant	Patients (n)		Variant	Patients (n)		Variant	Patients (n)	
p.Val577Gly	7	10 variants in heterzygous form detected in 29 cases. One of the top 4 most common mutations in CSID and IBS -reported by -Uhrich et al., 2012, and Henstrom et al, Gut, 2018, et al, respectively.	p.Val577Gly	3	10 variants in heterozygous form were detected in 21 abnormal low sucrase cases of which 5 were not detected in heterozygous form in abnormal sucrase group. Only 3 of top 5 were detected in 7 moderate normal sucrase group. One of the top 4 most common mutations in CSID and IBS -reported by -Uhrich et al., 2012, and Henstrom et al, Gut, 2018, et al, respectively.	p.Val577Gly	0	7 variants in heterozygous form were detected in 11 abnormal sucrase cases of which 5 were not detected in heterozygous form in abnormal cases. Only 1 of top 9 were detected in 9 high normal sucrase group. None of the top 4 most common pathogenic mutaions were detected in the high normal sucrase group.
p.Gly1073Asp	5	One of the top 4 most common mutations in CSID and IBS -reported by -Uhrich et al., 2012, and Henstrom et al, Gut, 2018, et al, respectively.	p.Gly1073Asp	2	One of the top 4 most common mutations in CSID and IBS -reported by -Uhrich et al., 2012, and Henstrom et al, Gut, 2018, et al, respectively.	p.Gly1073Asp	0	
p.Phe1745Cys	7	One of the top 4 most common mutations in CSID and IBS -reported by -Uhrich et al., 2012, and Henstrom et al, Gut, 2018, et al, respectively.	p.Phe1745Cys	0		p.Phe1745Cys	0	
p.Pro348Leu	3	reported as SI-rare variant. - in IBS - Garcia et al, 2018 - D'Amato	p.Pro348Leu	2		p.Pro348Leu	0	
p.Val371Met	2	reported as SI-rare variant -in IBS- Garcia et al, 2018 - D'Amato	p.Val371Met	0		p.Val371Met	0	
p.Gly515Val	1		p.Gly515Val	0		p.Gly515Val	0	
p.Tyr975His	1	reported as SI-rare variant in IBS - Garcia et al, 2018 - D'Amato	p.Tyr975His	2	Below 10%-ile cut off value of 32.8U. p.Arg1484His was reported as SI-rare variant. Garcia et al, 2018 - D'Amato	p.Tyr975His	1	reported as SI-rare variant - in IBS - Garcia et al, 2018 - D'Amato
p.Val1667Leu	1	reported as SI-rare variant - in IBS - Garcia et al, 2018 - D'Amato	p.Val1667Leu	0		p.Val1667Leu	0	
p.Glu801Ter	1		p.Glu801Ter	0		p.Glu801Ter	0	
p.Ile799Val	1		p.Ile799Val	5		p.Ile799Val	4	
Total	29		p.Gly1760Val	2		p.Arg774Gly	1	reported as SI-rare variant. Garcia et al, 2018 - D'Amato
Compound Heterozygous			p.Ser1795Leu	1		p.Gly1760Asp	1	
Variants	# Patients		p.Thr1802Ser	3		p.Gly1760Val	0	
p.Val577Gly+Val371Met	1	7 different variants were detected in compound hetezygouys state in total 5 cases - left column.	p.Val1651Ile	1		p.Ser186Pro	1	
p.Val577Gly+Ile1378Ser	1	p.Ile1378Ser was detected in only one case with p.Val577Gly in compound heterozygous combination	p.Gly1476Ala	1		p.Thr1802Ser	2	
p.Val577Gly+Leu914Ile	1		Compound Heterozygous			p.Val1651Ile	1	
Gly1073Asp+Arg1367Gly	1	Arg1367Gly - reported as SI-rare variant. Garcia et al, 2018 - D'Amato	Variants	Patients (n)		Compound Heterozygous		None
Gly1073Asp+Tyr975His	1		p.Tyr975His + p.Arg1484His	2	Below 10%-ile cut off value of 32.8U. p.Arg1484His was reported as SI-rare variant. Garcia et al, 2018 - D'Amato			
Total Compound Heterozygous	5				One of the13 variants (p.Tyr975His) was detected in 2 patients in compound heterozygous form who had sucrase activity value <10%-ile cut off value of 32.8U			
Homozygous								
Pro348Leu	1	One variants was detected in homozygous state.						
Combined (double) Homozygous								
Val371Met + Gly515Val	1	Two of the 13 variants were deteced in a combined homozygous form in one case						

Most significant top five variants: Val577Gly, Gly1073Asp, Phe1745Cys, Pro348Leu, and Val371Met

Table S4: Frequency of the five most significant *SI* gene pathogenic variants in the three sucrase activity groups and in the abnormal sucrase group with patterns of disaccharidase activities and symptoms.

Patient groups	Sample numbers (n)[$]	Number of cases with mutation* (frequency %)	*p*-value (versus high Controls)**
Normal Sucrase Activity group			
High Sucrase activity (>55U[a])	250	0 (0.0)	
Moderate Sucrase activity (>25.8 U - <55U)	250	7 (2.80)	7.71E-03
Abnormal sucrase activity Case group			
Abnormal Sucrase activity (<25.8U)	125	31 (24.80)	2.02E-16
Abnormal Case Subsets with Disaccharidase activity patterns			
Abnormal sucrase with normal lactase >= 15.4	28	12 (42.86)	3.62E-26
Abnormal sucrase with abnormal lactase < 15.4	97	19 (19.59)	6.13E-13
Abnormal sucrase with PDD	51	11 (21.57)	7.37E-14
Moderate Cases with 10th Percentile Sucrase < 32.8	38	1 (2.63)	1.02E-02
Abnormal sucrase with Diarrhea (D)	10	2 (20.00	1.26E-12
Abnormal sucrase with Abdominal Pain (AP)	87	23 (26.44)	3.70E-17
Abnormal sucrase with D and AP	28	6 (21.43)	1.37E-13

Five most significant pathogenic variants: 1. p.Val577Gly, 2. p.Gly1073Asp, 3. p.Phe1745Cys, 4. p.Pro348Leu, 5.p.Val371Met (see Table 2). The five most significant pathogenic variants (this table) are also part of the 13 pathogenic variants (Table S2). These five variants are the most significant ones compared to the rest of the 8 variants of the 13 variants identified in 36 low or abnormal sucrase cases (Table S2). Through linear regression analysis including the symptom complex and prevalence these 5 variants are the most significant ones to cause CSID symptoms (this table).

n = number;

*Presence of one or more than one of these five mutations in one patient was counted as one case.

[a]U, unit; µM/min/gram protein.

***p* versus high normal sucrase group, meaning significantly higher mutation frequency is present in the low sucrase group.

[$], sample number in different sucrase activity level groups consisting of normal (moderate and high) and abnormal case (low sucrase) groups, and the low sucrase case group subsets with different symptoms.

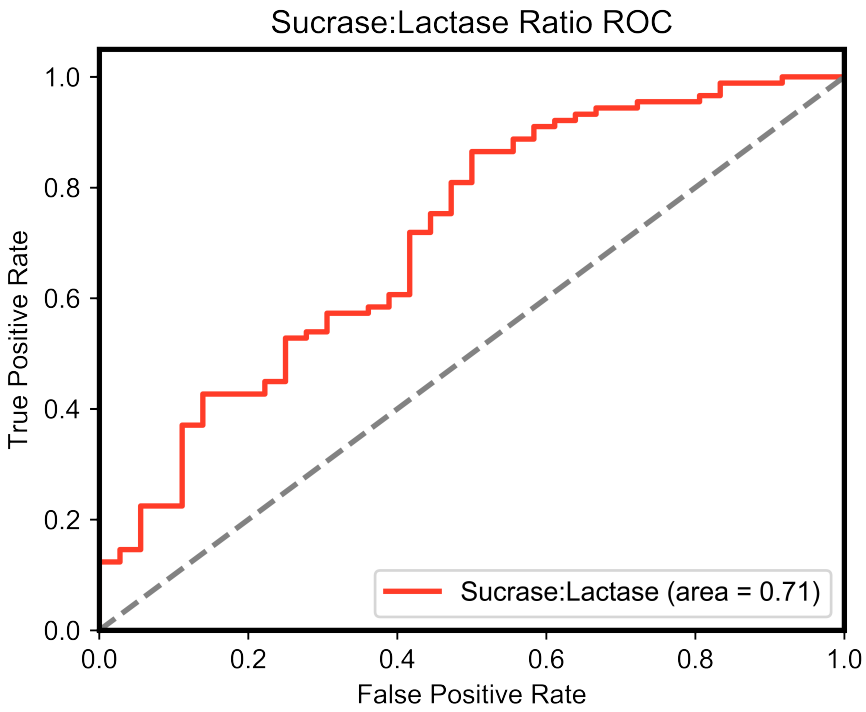

Figure S1. ROC curve for sucrase:lactase ratio to classify low sucrase

We show an ROC (receiver operating characteristic) curve using sucrase:lactase ratio to classify whether a patient has low sucrase or not. We see that this ratio decently classifies patients with an AUC of 0.71. This AUC is better than chance because amongst the cases, the average ratio of sucrase to lactase is lower than amongst the controls. We find a cutoff value of 1.43, (meaning ratios below 1.43 are classified as having at least one pathogenic variant) gives an accuracy of 75.2% with sensitivity of 86.5% and specificity of 50.0%. See Supplementary Digital Content-1 for more details.

Note: This disagrees with the recommendation for CSID diagnosis using sucrase:lactase ratio by Treem (Treem, 2012; Ref #9). Using a threshold for sucrase:lactase that is < 1 will result in a lower sensitivity (higher false negative rate for CSID diagnosis).