Genomics 113 (2021) 1867–1875

Contents lists available at ScienceDirect

Genomics

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

Original Article

Key genetic variants associated with variation of milk oligosaccharides from diverse human populations

Janet E. Williams [a], Michelle K. McGuire [b], Courtney L. Meehan [c], Mark A. McGuire [a], Sarah L. Brooker [a], Elizabeth W. Kamau-Mbuthia [d], Egidioh W. Kamundia [d], Samwel Mbugua [d], Sophie E. Moore [e,f], Andrew M. Prentice [f], Gloria E. Otoo [g], Juan M. Rodríguez [h], Rossina G. Pareja [i], James A. Foster [j], Daniel W. Sellen [k], Debela G. Kita [l], Holly L. Neibergs [m], Brenda M. Murdoch [a,*]

[a] Department of Animal, Veterinary, and Food Sciences, University of Idaho, Moscow, ID 83844, USA
[b] Margaret Ritchie School of Family and Consumer Sciences, University of Idaho, Moscow, ID 83844, USA
[c] Department of Anthropology, Washington State University, Pullman, WA 99164, USA
[d] Department of Human Nutrition, Egerton University, Nakuru, Kenya
[e] Department of Women and Children's Health, King's College London, London, United Kingdom
[f] MRC Unit The Gambia at the London School of Hygiene and Tropical Medicine, Fajara, The Gambia
[g] Department of Nutrition and Food Science, University of Ghana, Accra, Ghana
[h] Dpto. of Nutrition and Food Science, Complutense University of Madrid, Madrid, Spain
[i] Instituto de Investigación Nutricional, Lima, Peru
[j] Department of Biological Sciences and Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA
[k] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[l] Department of Anthropology, Hawassa University, Hawassa, Ethiopia
[m] Department of Animal Sciences, Washington State University, Pullman, WA, USA

ARTICLE INFO

Keywords:
Human milk
Oligosaccharides
GWAS
FUT2
FUT3

ABSTRACT

Human milk oligosaccharides (HMO), the third most abundant component of human milk, are thought to be important contributors to infant health. Studies have provided evidence that geography, stage of lactation, and Lewis and secretor blood groups are associated with HMO profile. However, little is known about how variation across the genome may influence HMO composition among women in various populations. In this study, we performed genome-wide association analyses of 395 women from 8 countries to identify genetic regions associated with 19 different HMO. Our data support *FUT2* as the most significantly associated ($P < 4.23^{-9}$ to $P < 4.5^{-70}$) gene with seven HMO and provide evidence of balancing selection for *FUT2*. Although polymorphisms in *FUT3* were also associated with variation in lacto-N-fucopentaose II and difucosyllacto-N-tetrose, we found little evidence of selection on *FUT3*. To our knowledge, this is the first report of the use of genome-wide association analyses on HMO.

1. Introduction

Milk is highly complex and contains a variety of constituents including carbohydrates, lipids, unconjugated complex carbohydrates (human milk oligosaccharides, HMO), proteins, immune factors (including immune cells), and non-nutritive biologically-active substances (e.g., enzymes, hormones) [1]. In human milk, aside from water, carbohydrates are the predominant component with lactose concentrations ranging from approximately 63–77 g/L and lipids comprising approximately 30–45 g/L [2]. HMO are the third most abundant milk component with combined totals ranging between approximately 5–15 g/L [3]. Although little variation has been reported in human milk lactose levels, concentrations of both lipid and HMO profiles appear to be associated with a number of factors. For example, diet and genetic variation of the fatty acid desaturase genes are thought to influence fatty acid composition of human milk [4]. HMO

* Corresponding author at: University of Idaho, 875 Perimeter Dr MS2330, Moscow, ID 83844-2330, USA.
E-mail address: bmurdoch@uidaho.edu (B.M. Murdoch).

https://doi.org/10.1016/j.ygeno.2021.04.004
Received 18 September 2020; Received in revised form 20 March 2021; Accepted 4 April 2021
Available online 6 April 2021

concentrations have been reported to vary across lactation stage, seasonality, and geographical locations [5–8]. The variability in HMO concentration has also been partially explained by functional properties of enzyme products from two genes: fucosyltransferase 2 (*FUT2*) and fucosyltransferase 3 (*FUT3*). However, aside from these two genes, little is known about other potential genetic drivers of this variation.

There have been up to 200 different HMO structures characterized [9]. Core oligosaccharide structures can be modified with addition of moieties such as fucose, *N*-acetylglucosamine (GlcNAc), and sialic acid (*N*-acetylneuraminic acid, NeuAc). The addition of these components to the core structures is regulated by enzymes such as α-1,2-fucosyltransferase and α-1-3/4-fucosyltransferase, which notably, are the same enzymes involved in determination of blood group characteristics. These enzymes are encoded by *FUT2* and *FUT3*, respectively. For instance, women who produce milk containing higher concentrations of some HMO, including 2′-fucosyllactose (2'FL) and difucosyllactose (DFLac), are referred to as "secretors" and thought to have a functional *FUT2* gene product (α-1,2-fucosyltransferase). Conversely, women who produce milk with negligible amounts of these HMO are classified as "non-secretors" and are thought to lack a functional α-1,2-fucosyltransferase [10–12]. *FUT3*, also known as the Lewis gene, regulates the expression of Le^a and Le^b antigens, and women who lack a functional enzyme from this gene produce milk with little lacto-N-fucopentaose-2 (LNFP II) [13]. Substantial research has been conducted to investigate how HMO can be grouped by secretor and/or Lewis phenotypic status; however, a thorough investigation into the relationships between genetic polymorphisms across these genes and across the human genome and HMO composition has not been attempted. Filling this research gap is important because secretor status is not only associated with HMO composition but also has been linked to increased susceptibility to disease [14] and conversely, resistance to disease [15] in some infants. Elucidating the genetic components that influence secretor status and HMO composition are key to understanding the underlying mechanisms that influence the health status of the mother and her infant.

In the present study, we conducted a genome-wide association study (GWAS) to examine how maternal genetic variation is related to HMO profiles and concentrations. Furthermore, we analyzed the genetic diversity of *FUT2* and *FUT3* to assess the likelihood that evolutionary selective forces have impacted variation in these genes.

2. Methods

2.1. Study populations

Samples used in these analyses were collected as part of a cross-sectional study that has been described previously [8]. Briefly, milk ($n = 410$) and saliva samples ($n = 405$) were obtained from women living in 2 regions of the United States (ethnically heterogeneous residents of southeastern Washington and northwestern Idaho [USW] and self-identified Hispanic residents of southern California [USC]), Peru (peri-urban area of Lima [PE]), Sweden (near Helsingborg and self-identified as having Nordic heritage [SW]), Spain (recruited in Madrid, Zaragoza, Huesca, and Vizcaya with no conditions regarding ethnicity [SP]), and 6 sub-Saharan African groups (rural and urban Ethiopia [ETR and ETU, respectively], rural and urban Gambia [GBR and GBU, respectively], Kenya [KE], and Ghana [GN]). Ethics approvals were obtained for all procedures from each participating institution, and overarching approval was obtained from the Washington State University Institutional Review Board (13264). After being translated from English (when needed), informed written consent or verbal (depending on the locale, subject's literacy level, and when approved by the IRB boards) was acquired from each participating woman.

2.2. Human milk and oligosaccharide analysis

Milk collection and oligosaccharide analysis have been described previously [8,16]. Absolute concentrations were calculated on the basis of standard response curves for each of the following HMO: 2′FL, 3-fucosyllactose (3FL), 3′-sialyllactose (3′SL), 6′-sialyllactose (6′SL), difucosyllactose (DFLac), difucosyllacto-*N*-hexaose (DFLNH), difucosyllacto-*N*-tetrose (DFLNT), disialyllacto-*N*-hexaose (DSLNH), disialyllacto-*N*-tetraose (DSLNT), fucodisialyllacto-*N*-hexaose (FDSLNH), fucosyllacto-*N*-hexaose (FLNH), lacto-*N*-fucopentaose (LNFP) I, LNFP II, LNFP III, lacto-*N*-hexaose (LNH), lacto-*N*-neotetraose (LNnT), lacto-*N*-tetrose (LNT), sialyl-lacto-*N*-tetraose b (LSTb), and sialyl-lacto-*N*-tetraose c (LSTc). "Secretor" milk was defined as having a 2'FL concentration > 200 nmol/mL. HMO data (nmol/mL) were tested for assumptions of normality and were transformed as needed by performing a squared function.

2.3. Maternal saliva collection and genotyping

Saliva was collected from each participant using the DNAgard Saliva kit (USA Biomātrica, San Diego, CA). Genomic DNA was extracted from 1 mL preserved saliva using the Gentra Puregene Blood DNA extraction kit (Qiagen, Valencia, CA) following the manufacturer's instructions and genotypes were obtained with Illumina Multi-Ethnic Global-8 v1.0 arrays (MEGA; https://www.illumina.com/products/by-type/microarray-kits/infinium-multi-ethnic-global.html) following manufacturer's recommended protocol by Neogen Genomics (GeneSeek, Lincoln, NE). The MEGA contains $>1.7 \times 10^6$ single-nucleotide polymorphisms (SNPs) spanning all chromosomes of the human genome and includes markers that include "highly informative SNP for GWAS analyses in European and East Asian descent populations for backwards compatibility with other genotyping arrays" [17]. Additionally, the customized content of markers was chosen to expand the discovery of associations related to metabolic, cardiovascular, renal, inflammatory, anthropometric, and other traits across multiple ethnicities and associations to less frequent (1–5%) and rare (<1%) genetic variants. Genotypes were called using GenomeStudio 2.0 (Illumina Inc., San Diego, CA) and analyzed using SNP & Variation Suite software (SVS; version Win64 8.7.2; Golden Helix, Bozeman, MT). Quality control approaches excluded markers based on the following criteria: call rates <89.5% (10,027 SNPs), minor allele frequencies <0.0095 (640,954 SNPs), and Hardy-Weinberg equilibrium (HWE) of $P < 1 \times 10^{-70}$ (552 SNPs). Two samples were excluded due to call rates <87%. After filtering, data from 395 women and 1,128,286 SNPs were retained for subsequent analyses.

2.4. Statistical tests

Associations between the HMO phenotypes and SNPs were evaluated using the efficient mixed-model association eXpedited (EMMAX) method [18] as implemented in SVS with an additive genetic model. Briefly, this method approximates the variance components, uses the same variance for all probes, and adjusts for pair-wise genetic relatedness among all samples using a kinship matrix. Single-locus models were performed initially, followed by multi-locus models for any HMO that displayed associations across chromosomes [19,20]. All reported gene and SNP locations were reported based on the human genome build, GRCh38 (https://www.ncbi.nlm.nih.gov/genome/guide/human/). Significance for an association was declared at $P < 1 \times 10^{-8}$.

2.5. Haplotype analysis

Phased haplotypes for *FUT2*, consisting of 24 markers spanning from the 1st intron (*rs418821*; 19:48696547) to the 3′ untranslated region (UTR) (*rs503279*; 19:48705753) and *FUT3*, consisting of 8 markers from the 1st intron to the UTR of *FUT3*, were obtained using default parameters in PHASE 2.1 [21,22]. Median joining networks [23] were constructed using phased genotypes and PopART (http://popart.otago.ac.nz) [24]. Haplotypes for *FUT2* were considered to be '*se*' or nonfunctional if the allele at *rs601338* coded for a stop codon and '*Se*' if the allele

was functional.

2.6. Genetic diversity analysis

Two measures of nucleotide diversity, namely the average number of pairwise differences in a given region (π) and Tajima's D [25], were obtained for *FUT2* and *FUT3* (24 and 8 markers, respectively) using PopART and used to compare SNP frequencies in *FUT2* and *FUT3* gene regions. Tajima's D is a test statistic used to determine if the differences in distribution of allele frequencies are greater or less than the expected variation under the neutral theory, i.e. from that evolving randomly. Significant positive values for Tajima's D indicate an excess of intermediate frequency variants and suggest balancing selection; while negative Tajima's D indicate a selective sweep and high frequency of rare variants. Nucleotide diversity (π) provides a summary statistic that is used to detect selection; whereby, low diversity is considered a sign of directional selection.

3. Results

3.1. Genome-wide association analyses of individual HMO

A GWAS analysis was performed to identify genomic regions or genes associated with various HMO. Of the 19 HMO examined, variation in concentrations of 11 were associated with variation in 5 different regions of the human genome. The most significant associations for each HMO are provided in Table 1; all significant associations, defined as those less than $P < 10^{-8}$, are provided in Supplementary Table 1. Seven HMO (2'FL, DFLac, LNFP I, LNFP II, FSDLNH, LNT, and LSTb; Supplementary Fig. S1) were associated with many of the same genetic variants within the *FUT2* gene of chromosome 19. Specifically, *rs679574* and *rs516316* exhibited the strongest association ($P < 4.50 \times 10^{-70}$) with variation in 2'FL concentration (Fig. 1A, B). The majority of women with the GG genotype at these loci produced milk with little to no detectable 2'FL. These two SNPs are in complete linkage and explained 55% of the variation observed in 2'FL concentration of milk. Additionally, *rs492602* (which is also in complete linkage with *rs679574* and *rs516316*) and *rs601338* (which is in very high linkage, D′ = 0.99, with *rs679574*, *rs516316*, and *rs492602*) were strongly associated ($P < 3.91 \times 10^{-69}$ and $P < 4.73 \times 10^{-69}$, respectively) with 2'FL (Fig. 1C). Overall, there were 24 additional SNPs in this region of chromosome 19 associated with 2'FL variation (Fig. 1B). Many of the same SNPs in the region of the *FUT2* gene also exhibited a strong association with concentrations of DFLac and LNFP I; *rs601338* exhibited the strongest association ($P < 8.13 \times 10^{-50}$) with DFLac, whereas *rs492602* was most associated ($P < 4.31 \times 10^{-41}$) with LNFP I. Within the *FUT2* gene, several SNPs, including *rs516246* and *rs485186*, were also found to be associated with concentrations of LNFP II, LNT, and FDSLNH (Fig. 1D). It is noteworthy, however, that the above-mentioned SNPs exhibited an inverse relationship with the concentrations of these HMO, in comparison to that of these SNPs to 2'FL, DFLac, and LNFP I (Table 2).

Associations observed with other HMO include three SNPs associated with varying concentrations of DFLNT: *rs708686* which lies just upstream of *FUT6*, and *rs778986* and *rs2561796* which are both located in the *FUT3* region of chromosome 19, with *rs778986* coding a missense variant and *rs2561796* residing in an intronic region of *FUT3*. Additionally, other associations included *rs75248127* on chromosome 10 and *rs185875168* in the intron of *RAB31L1* on chromosome 11 that were associated with FLNH and LNH, respectively (Supplementary Table S1).

3.2. Multi-locus GWAS

Although most of the HMO were associated with only one region of the genome, a few exhibited associations with more than one location; therefore, multi-locus mixed model analyses were conducted for FLNH, LNFP II, LNFP III and LNH. Using the extended Bayesian information criterion (EBIC) to determine the optimal model for each HMO, two SNPs (one on chromosome 5 and the other on chromosome 10) were associated with FLNH ($P < 3.62 \times 10^{-9}$ and $P < 2.49 \times 10^{-11}$, respectively); two SNPs on chromosome 19 (*rs708686* located between *FUT6* and *FUT3* and *rs516246* located in *FUT2*) were associated with LNFP II ($P < 5.04 \times 10^{-15}$ and $P < 2.97 \times 10^{-25}$, respectively); two SNPs (*rs185875168* on chromosome 11 and *rs778986* on chromosome 19) were associated with LNH ($P < 4.59 \times 10^{-11}$ and $P < 1.52 \times 10^{-10}$, respectively); and two SNPs (*rs35231001* on chromosome 4 and *rs10504422* on chromosome 8) were associated with LNFP III ($P < 2.01 \times 10^{-12}$ and $P < 5.52 \times 10^{-10}$, respectively) (Supplementary Fig. S2).

3.3. Genome-wide association analyses by secretor status

Given that the majority of HMO variation is explained by null mutations found in *FUT2* and to better understand how variation in concentrations of HMO may be regulated differently in secretors and non-secretors, we performed separate analyses with secretor women (n =

Table 1
Most significant single nucleotide polymorphisms associated with variation in oligosaccharide concentrations of milk produced by 11 cohorts of women from 8 countries.

HMO	Significant SNPs (#)	Most significant SNP(s)	Chr[a]	Position	Alternate allele	Reference allele[b]	Assay alleles[c]	Proportion of variance explained	*P* value
2'FL[e]	27	*rs679574*	19	48,702,85	G	C	[G/C]	0.55	4.50×10^{-70}
		rs516316	19	1 48,702,888	C	G	[G/C][d]		
DFLac[e]	22	*rs601338*	19	48,703,417	A	G	[A/G]	0.43	8.13×10^{-50}
LNFP I[e]	24	*rs492602*	19	48,703,160	G	A	[C/T][d]	0.36	4.31×10^{-41}
LNFP II[e]	26	*rs516246*	19	48,702,915	T	C	[A/G][d]	0.22	4.02×10^{-23}
LNFP III[e]	14	*rs35231001*	4	22,968,709	G	A	[G/A]	0.14	1.56×10^{-14}
DFLNT[e]	3	*rs708686*	19	5,840,608	T	C	[T/C]	0.13	6.07×10^{-14}
FDSLNH[e]	13	*rs516246*	19	48,702,915	T	C	[A/G][d]	0.13	7.14×10^{-14}
LNT[e]	13	*rs485186*	19	48,703,94	G	A	[C/T][d]	0.11	3.73×10^{-11}
		rs603985	19	9 48,704,000	C	T	[C/T]		
LSTb[e]	6	*rs2251034*	19	48,704,535	A	G	[A/G]	0.09	1.06×10^{-9}
FLNH[e]	5	*rs75248127*	10	36,616,538	G	A	[G/A]	0.09	1.37×10^{-9}
LNH[e]	3	*rs185875168*	11	61,898,945	A	G	[A/G]	0.09	1.70×10^{-9}

[a] Chromosome.
[b] Reference allele was designated as the allele found in the top strand of the displayed sequence in dbSNP generated from using GRCh38.p12.
[c] Assay alleles are listed with the alternate allele or the allele complementary to the alternate allele listed first and the reference allele or allele complementary to reference listed second.
[d] Opposite strand used in array design.
[e] 2'FL - 2′-fucosyllactose; DFLac - difucosyllactose; LNFP - lacto-N-fucopentaose; DFLNT - difucosyllacto-N-tetrose; FDSLNH - fucodisialyllacto-N-hexaose; LNT - lacto-N-tetrose; LSTb - sialyl-lacto-N-tetraose b; FLNH - fucosyllacto-N-hexaose; LNH - lacto-N-hexaose.

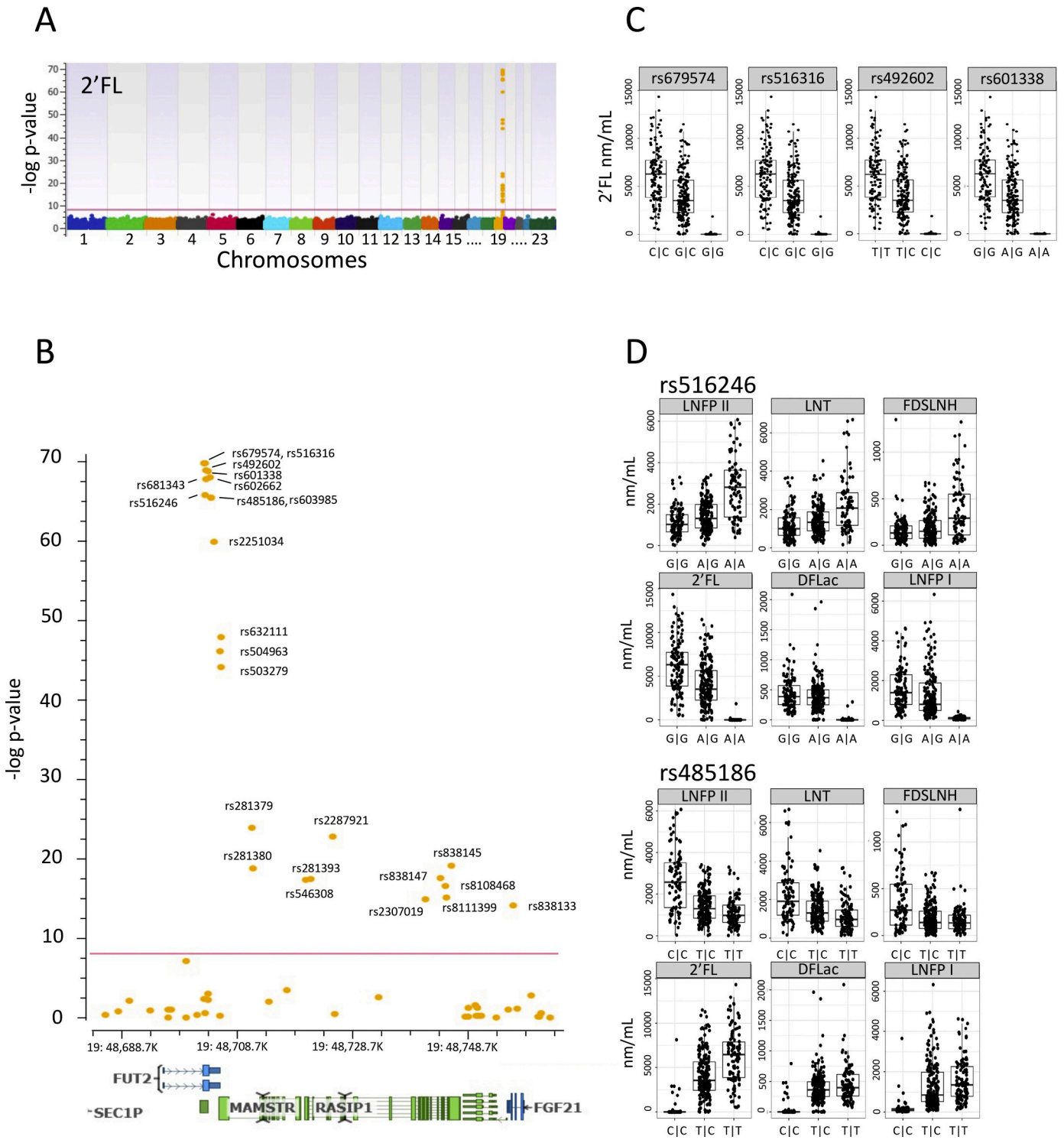


Fig. 1. (A) Manhattan plot of GWAS analyses with 2'FL. The –log10 *P*-values for each SNP from a genome-wide scan are plotted against their position on each of the 23 chromosomes. The red horizontal line indicates the genome-wide significance threshold ($P < 1 \times 10^{-8}$). (B) Zoomed in view of Manhattan plot around *FUT2* gene region. (C) Boxplots of the 4 most significant loci for 2'FL where dots represent individual women's concentrations of HMO by genotype. (D) Boxplots of most significant SNP for LNFP II and FDSLNH (rs516246) and most significant SNP for LNT (rs485186) and their associations with concentrations of LNFP II, LNT, and FDSLNH, and the inverse relationships with concentrations of 2FL, DFLac, and LNFP I. Dots represent individual women's samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

305) and non-secretor women ($n = 90$). When data from secretor only women were evaluated using EMMAX, concentrations of LNFP II and DFLNT were still associated with *rs708686* on chromosome 19 (Supplementary Fig. S3; Table S2). Additionally, LNFP I was significantly ($P < 7.48 \times 10^{-10}$) associated with *rs708686*. DFLNT was also once again associated with *rs2561796* and *rs778986* in the *FUT3* region of chromosome 19. One SNP on chromosome 1 (*rs77634072*) was associated with 3FL concentration ($P < 5.76 \times 10^{-10}$). These analyses were generally consistent with the analyses done with all of the women included. When only data from non-secretor women were evaluated,

Table 2
SNP that display significant associations with HMO but have different directions of the beta coefficient associated with the model fit depending on the HMO it is associated with.

rsID	Chr[a]	Position	Gene: Consequence	HMO Negative Beta[b]	HMO Positive Beta[b]
rs679574	19	48,702,851	FUT2: intron variant; LOC105447645: non-coding transcript variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs516316	19	48,702,888	FUT2: intron variant; LOC105447645: non-coding transcript variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs516246	19	48,703,160	FUT2: intron variant; LOC105447645: non-coding transcript variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs492602	19	48,703,160	FUT2: synonymous variant; LOC105447645: non-coding transcript variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs681343	19	48,703,205	FUT2: stop gained; LOC105447645: non-coding transcript variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs601338	19	48,703,417	FUT2: stop gained; LOC105447645: non-coding transcript variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs602662	19	48,703,728	FUT2: missense variant; LOC105447645: non-coding transcript variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs485186	19	48,703,949	FUT2: synonymous variant; LOC105447645: 2 KB upstream variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH, LSTb
rs603985	19	48,704,000	FUT2: 3' UTR variant; LOC105447645: 2 KB upstream variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH, LSTb
rs2251034	19	48,704,535	FUT2: 3' UTR variant; LOC105447645: 2 KB upstream variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH, LSTb
rs504963	19	48,705,608	FUT2: 3' UTR variant; LOC105447645: 2 KB upstream variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs632111	19	48,705,721	FUT2: 3' UTR variant; LOC105447645: 2 KB upstream variant	2'FL, DFLac, LNFP I	LNFP II, LNT, FDSLNH
rs503279	19	48,705,753	FUT2: 3' UTR variant; LOC105447645: 2 KB upstream variant	2'FL, DFLac, LNFP I	LNFP II, FDSLNH
rs281379	19	48,711,017		2'FL, DFLac, LNFP I	LNFP II
rs281380	19	48,711,213		LNFP II	2'FL, DFLac, LNFP I
rs281393	19	48,721,227	RASIP1: missense variant; MAMSTR: 2 KB upstream variant	LNFP II	2'FL, DFLac, LNFP I
rs2287921	19	48,725,015	RASIP1: intron variant	2'FL, DFLac, LNFP I	LNFP II
rs838147	19	48,743,609	IZUMO1: intron variant	2'FL, DFLac, LNFP I	LNFP II
rs8108468	19	48,744,529	IZUMO1: synonymous variant	LNFP II	2'FL, LNFP I
rs8111399	19	48,744,706	IZUMO1: intron variant	LNFP II	2'FL, LNFP I
rs838145	19	48,745,473	IZUMO1: intron variant	2'FL, DFLac, LNFP I	LNFP II

[a] Chromosome.
[b] 2'FL - 2′-fucosyllactose; DFLac - difucosyllactose; LNFP - lacto-N-fucopentaose; FDSLNH - fucodisialyllacto-*N*-hexaose; LNT - lacto-*N*-tetrose; LSTb - sialyl-lacto-*N*-tetraose b.

one SNP in lncRNA on chromosome 5 (*rs17312027*) was associated with variation in DFLNH concentration; and *rs79318201* on chromosome 10 was associated with LSTc (Supplementary Fig. S3; Table S3).

3.4. Genetic relatedness of study participants

To examine the distribution in the genetic relatedness of the 395 women, we performed a PCA (Fig. 2). Visual inspection of the PCA plot suggested that samples clustered into 4 main groups. Samples from The Gambia, Kenya, and Ghana cohorts clustered together (PC group 4), while samples from the two Ethiopian cohorts formed another cluster (PC group 3). Samples from Spain and Sweden clustered with most of the samples from USW and a few from USC (PC group 2), and most of the samples from USC and Peru formed a broader cluster (PC group 1) (Supplementary Table S4).

3.5. Haplotype block analyses of genetic variation in FUT2 and FUT3 genes

Haplotype block analysis revealed two blocks across the *FUT2* gene. Haplotype blocks are defined as arrays of two or more SNPs with high linkage disequilibrium (LD). Haplotype block 1 consisted of six SNPs that were all in very high LD among each other, (e.g. R^2 ranged from 0.98–1.0 among *rs679574, rs516316, rs516246, rs492602, rs681343,* and *rs601338* (Supplementary Table S5). Haplotype block 2 consisted of seven SNPs (*rs602662, rs485186, rs603985, rs2251034, rs504963, rs632111* and *rs503279*) that although not as high as that observed for Block 1, still exhibited strong LD (e.g. R^2 ranged from 0.76–1.0). The frequency of block 1 and 2 diplotypes across the PC groups are displayed in Fig. 3A. Most of the women in PC Group 1 (86%; 48/56) were homozygous for the Block 1 diplotype CCGTCG|CCGTCG; whereas each of the other 3 PC groups displayed a more varied frequency with presence of up to three or four different diplotypes (Fig. 3A). Block 1 and block 2 diplotypes also exhibited linkage disequilibrium. For example, those with the CCGTCG|CCGTCG diplotype in block 1 more often (85%) had the GTTGCAA|GTTGCAA diplotype in block 2; conversely, those with the GGACTA|GGACTA diplotype were more often (82%) homozygous for the block 2 ACCATGG|ACCATGG diplotype. Interestingly, no haplotype blocks were detected in the *FUT3* region.

3.6. Haplotypes by PC group

After phasing the 24 markers found within *FUT2*, 49 unique haplotypes were inferred (Supplementary Table S5 and S6). To visualize the relationships among the haplotypes and better understand the distribution of haplotypes by individuals with similar genetic backgrounds, a median joining network of the haplotypes was constructed for *FUT2* (Fig. 3B). The haplotype network is divided into two main groups, Se and se, with the null mutation variant *rs601338* described by Kelly et al. 1995 [11] defining the two groups. Within the Se group there were two predominant haplotypes (Se15 and Se21) with a third less common haplotype (Se32) in PC group 1. In the se group, one major haplotype (se12) was observed, along with se3 being another haplotype that was found predominantly in individuals in PC group 4. These two haplotypes only differed by one marker (*rs590017*) which is found in the intron of *FUT2*. Less common haplotypes were often found in PC Groups 3 and 4.

To investigate *FUT3*, we examined how the haplotype structure for *FUT3* might vary across the four PC groups. Fourteen unique *FUT3* haplotypes were inferred, with one (Hap6) containing the previously reported null mutation at *rs59796499* (Supplementary Table S7). However, none of the individuals in this study were homozygous for this variant. Five haplotypes were commonly observed for *FUT3* in PC groups 2, 3, and 4; whereas, in PC groups 3 and 4 some additional less common haplotypes were observed (Supplementary Fig. S4).

3.7. Population genetic statistics

To test if *FUT2* and *FUT3* are under selective pressure, we examined various neutrality indices. Utilizing the markers for *FUT2*, Tajima's D for the entire study population was found to be significant and positive (3.30, $P = 0.003$) suggesting *FUT2* is under balancing selection. Interestingly, when evaluating by PC groups, PC group 2, 3, and 4 were each positive (2.88, 2.38, 3.0; $P < 0.025$). Conversely, although not significant, a negative Tajima's D was observed for PC group 1 (−0.67; $P = 0.727$). Additionally, nucleotide diversity (π) of *FUT2* was examined and found to be similar among PC groups 2, 3, and 4 ($\pi = 0.31, 0.32, 0.34$, respectively), whereas, PC group 1 had a value of 0.12 for π. When evaluating *FUT3,* Tajima's D was not significant (0.34; $P = 0.693$) for the entire study population, but when examined by PC group, PC group 2

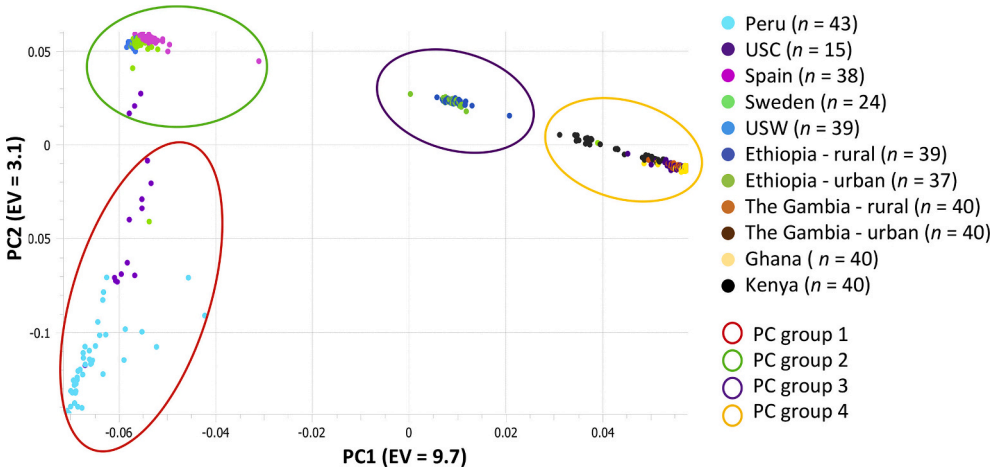


Fig. 2. Principal component analysis (PCA) plot displaying the genetic relatedness of the women participating in the INSPIRE study. Dots represent individual samples and are colored by site of sample collection. Colored ellipses display PC groupings. EV = eigenvalue.

had a significant positive Tajima's D (2.30; $P = 0.033$) and all others were not significant. Values of π for PC groups 1, 2, 3, and 4 were similar for *FUT3* (π = 0.13, 0.16, 0.17, 0.16, respectively).

4. Discussion

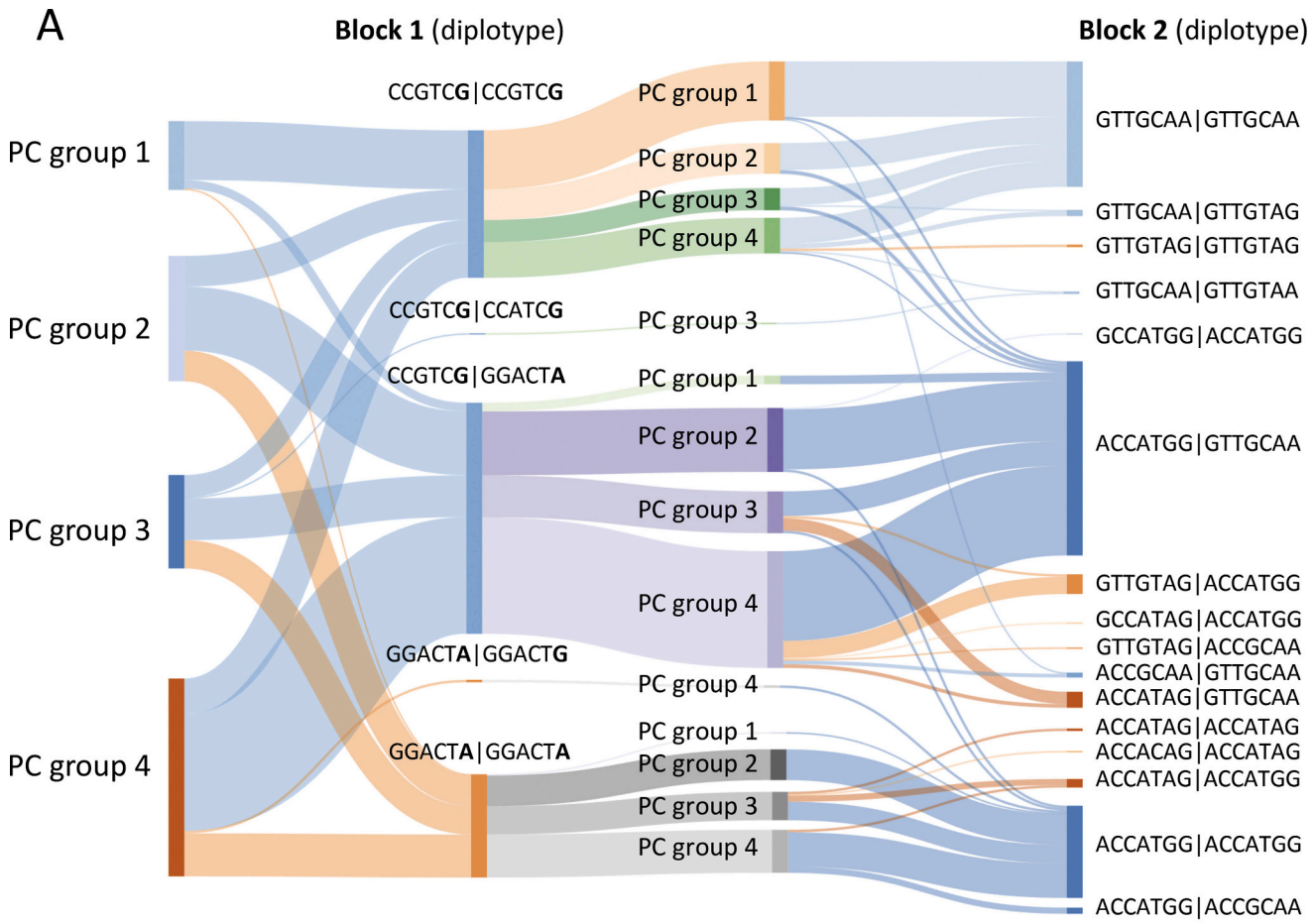
This study represents the first GWAS of variation in HMO, collected from approximately 400 women residing in 8 different countries. We identified genetic loci that are significantly associated with 11 different HMO, the most significant being with 2'FL ($P < 4.50 \times 10^{-70}$). Polymorphisms identified in the *FUT2* and *FUT3* genes were associated with variation in the concentration of eight HMO. These associations are consistent with findings from previous research that have demonstrated a strong correlation between HMO profile and activity of fucosyltransferases that are encoded by the *FUT2* and *FUT3* genes [13,26–28].

The multiple associations among these SNPs with some of the HMO such as 2'FL were due to the high LD found in this region. For example, the genotypes in *rs679574* and *rs516316* are in complete LD and consequently both of these SNPs are in very high linkage disequilibrium ($R^2 = 0.99$) with *rs601338*, a previously described nonsense mutation in the *FUT2* gene [11]. Interestingly, several SNPs located within *FUT2* were significantly associated with a number of different oligosaccharides including 2'FL, DFLac, LNFP I, LNFP II, LNT, FDSLNH, and LSTb; however, the direction of the beta coefficients for the relationships with 2'FL, DFlac, and LNFP I was opposite from the direction of those with LNFP II, LNT, FDSLNH, and LSTb. This is evidence that the negative correlation, previously documented by McGuire et al. (2017) and Azad et al. (2018), can be attributed to specific genetic variants. It also provides additional evidence that women without a missense SNP have a functioning enzyme encoded from *FUT2*, and thus have an HMO profile rich in 2'FL, DFLac, LNFP I, whereas women with genotypes that include the missense mutation at *rs60188* have a nonfunctioning enzyme resulting in a different HMO profile with higher concentrations of HMO such as LNFP II and LNT.

Our data are consistent with results from a previous study that utilized restriction fragment length polymorphism PCR and enzyme digestion to distinguish the genotype at *rs601338* [29]. In this study, the authors found the genotypes from this SNP were correlated with the levels of HMO. The authors observed that all mothers who were determined to be homozygous non-secretors at this allele were also non-secretors in phenotype, i.e. produced milk with little to no 2'FL. However, of the secretors, two individuals produced milk with little 2'FL early in lactation (e.g., day 6); but later in lactation (day 120), they produced levels indicative of a secretor. In our study, we also observed that the majority of women displayed consistency between the secretor/non-secretor genotypes and expected profiles for HMO such as 2'FL. However, there were few instances where the relationships between SNPs and HMO concentrations did not hold true. Notably, there were 6 individuals (5 from GN and 1 from GBR) whose genotypes at the stop-gain marker *rs601338* in *FUT2* were predictive of a functional enzyme but had <200 nmol/mL 2'FL in their milk. It is possible that there are other mutations or indels present but which are not represented on the MEGA and/or other environmental factors which resulted in a non-secretor phenotype. Furthermore, two individuals exhibited a diplotype consistent with a non-secretor diplotype. However, they differed at one specific site, the functional allele at *rs601338*; milk produced by these individuals had >200 nmol/mL 2'FL.

This study also elucidated genomic regions other than *FUT2* associated with HMO composition. Two SNPs located in *FUT3* (*rs778986* and *rs2561796*) and one SNP located near *FUT6* (*rs708686*) were associated with concentrations of DFLNT. Interestingly, *rs708686*, along with other SNPs in *FUT6*, a gene that encodes a fucosyltransferase enzyme that catalyzes the addition of a fucose exclusively onto type 2 precursor chains [30], and SNPs in *FUT2* (including *rs602662* and *rs601338*) are also associated with plasma vitamin B12 status [31]. Using a meta-analysis approach, Nongmaithem and coworkers (2017) observed an association ($P = 5.7 \times 10^{-15}$) between *rs708686* and plasma B12 concentrations in 4419 Indo-European and Dravidian individuals from six cohorts. They noted that the relationships persisted across a wide range of factors including age, gender, and ethnicities, suggesting a strong genetic contribution to variation in plasma vitamin B12 concentrations. The authors also investigated potential differences in the binding of transcription factors across variants in *FUT6*. Although they reported promotor and/or enhancer activity for some of the SNP in *FUT6*, they did not observe any allele-specific differential binding of transcription factors with variation at *rs708686*. The biological implication, however, for alterations in both plasma vitamin B12 and HMO concentrations being induced by genetic variation in *FUT6* warrants further research.

The fucosyltransferase gene family has long been studied both from a biological and evolutionary standpoint as these genes play an integral role in determining cell surface markers and blood group antigens [32,33]. The evolutionary history of these genes, in particular for *FUT2* and *FUT3*, have been described in several populations, many of which overlap with our study populations [34–37]. For example, the primary non-functional allele in *FUT2, rs601338*, observed in our cohorts has also been reported to be the primary nonfunctional allele in Europeans and Africans [33]. Similar to results reported by Soejima and coworkers (2007) who reported evidence of balancing selection for *FUT2* in a Ghanaian cohort, our findings for Tajima's D also indicate that the genetic diversity in *FUT2* in our study populations is not only the result of



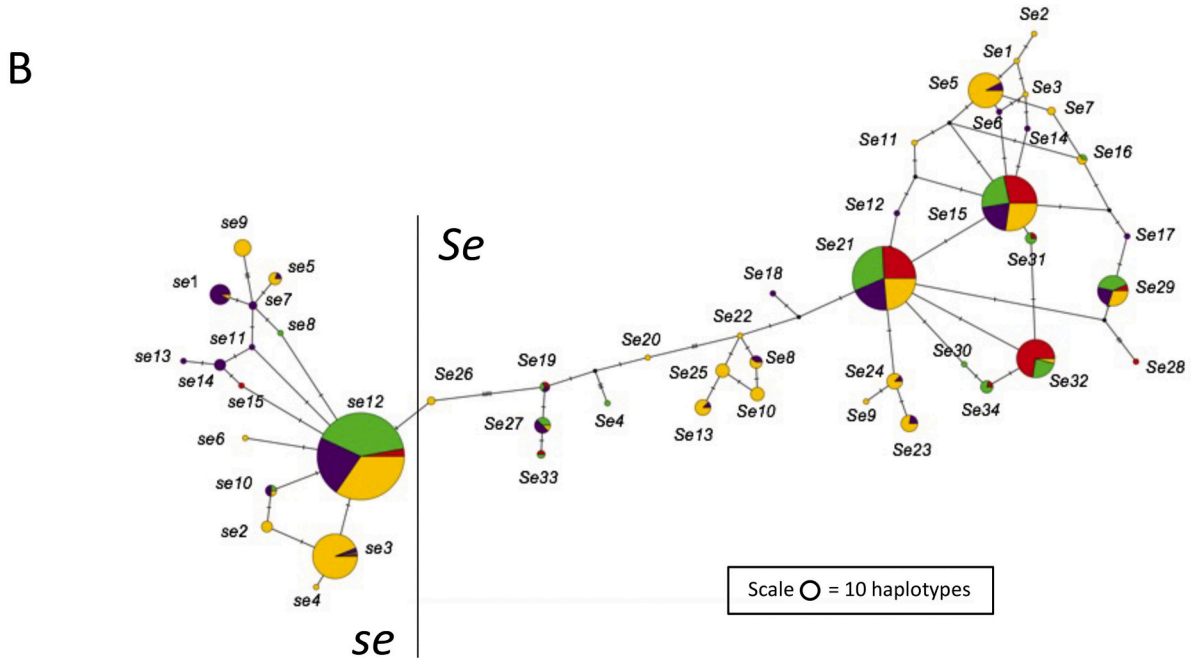


Fig. 3. (A) Distribution of Block 1 and Block 2 *FUT2* diplotypes by PC Group. Block 1 consists of 6 markers, rs679574, rs516316, rs516246, rs492602, rs681343, and rs601338, with the bold letter in the Block 1 diplotype representing the rs601338 marker. Block 2 consists of 7 markers, rs602662, rs485186, rs603985, rs2251034, rs504963, rs632111, rs503279 (B) Median joining network of *FUT2* haplotypes found in the 4 PC groups. Each circle represents a specific haplotype and the size of the circle is proportional to the frequency of that haplotype. The colored areas within each circle are proportional to the frequency of the haplotype for the PC group (PC group 1 - red, PC group 2 - green, PC group 3 - purple, and PC group 4 - yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mutation and genetic drift but also of balancing selection.

Balancing selection is an adaptive force that maintains genetic variation in populations. The evidence of balancing selection observed for *FUT2* in PC groups 2, 3, and 4 suggest that diversity of alleles in *FUT2* may offer some benefit in these populations. Conversely, in the case of PC group 1 (primarily secretors [49/50]) where no evidence of balancing selection was found, there may be benefit to secretor status or alternatively, allelic variation has been reduced by genetic drift which removes polymorphisms from populations. It is noteworthy, however, that as these fucosyltransferase genes are also linked to ABO histo-blood groupings, disentangling the contribution of HMO to the molecular evolution of these pleiotropic genes is difficult. For example, presence or absence of glycoproteins on the surface of epithelial cells and erythrocytes and/or presence or absence of HMO may increase or decrease the risk of disease or infection by various pathogens [3,38].

Work from the 1000 Genomes Project Consortium indicates that individuals with African ancestry have the greatest number of variant sites, which is also observed in our study populations [39]. Additionally, we observed 31 unique haplotypes and 24 segregating sites for *FUT2* in PC group 4 which was comprised mainly of individuals living in Africa. In alignment with what others have reported, we also observed PC group 4 as having the highest Tajima's D value across the PC groups. However, in contrast to the balancing selection observed for *FUT2*, we did not observe evidence of selective pressure for *FUT3*. This could be due to the low number of SNPs in *FUT3* which had passed the filtering step in our analyses and resulting in few SNP being evaluated in the neutrality tests.

Although our study was somewhat limited by having a relatively small number of individuals in each cohort, these results provide strong evidence for the impact of mutations in *FUT2* and *FUT3* on HMO composition. Our data also support previous results that have shown the null mutation at *rs601338* in *FUT2* has a high prevalence among Caucasians as well as African populations. When this nonsense mutation is homozygous in individuals, we observed very little 2'FL and other HMO that stem from this branch-chain base. Surprisingly, we observed little evidence of genes other than *FUT2* and *FUT3* being associated with HMO composition. We also attempted to elucidate additional variants that would help explain the variation in HMO concentrations by examining the HMO profiles in milk from secretor (all HMO) and non-secretor women (excluding 2'FL, DFLac, LNFP I) separately. Because the HMO profiles are distinct between these two groups, i.e. secretor and non-secretor women, we could then potentially identify genetic associations that could explain the variation of specific HMO that might be more prevalent in one group or the other. Once again, we found *rs708686* to be associated with DFLNT and LNFP II in secretor women. We did not however, find these association in non-secretor women. Instead, we found *rs17312027* and *rs79318201* were associated with DFLNH and and LSTc, respectively. It is possible that a large number of individuals in these populations would be necessary to uncover the potential small gene effect on HMO diversity. More research is needed to identify additional mechanisms and factors that result in such wide ranges of HMO concentrations.

In conclusion, our study is the first report to use a GWAS to investigate the contribution of genetic variation to HMO composition. Understanding the influence of genetic polymorphisms on HMO synthesis is important not only as evidence continues to mount that various HMO play a role in the health of the infant, but also in that these same genes may impact the health of the infant through other mechanisms, like vitamin B12 status [40] and risk of virus infection and respiratory and gastrointestinal illnesses [as reviewed [14,41].

Author statement

Contributions - MKM, BMM, HLN, MAM, and JEW designed the current project. MKM and CLM oversaw the parent INSPIRE project which included collection of milk and saliva samples. CLM, EWK-M, EWK, SM, SEM, AMP, GEO, JMR, RGP, DGK collected or processed the samples. JEW, HLN, and BMM oversaw the genotyping. JEW and BMM analyzed the data and wrote the initial manuscript. All authors read, contributed to, and approved the manuscript.

Data availability

The data from this cohort are not available as participants were not consented for data sharing.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

The authors would like to thank Kelcey McBride, Hannah Jaeger, and Jennifer Kiser for their assistance in processing the samples or data. Funding for this project was supported by grants from the National Science Foundation (DBI-0939454, 1344288, 1917476) and Washington State University Office of Research grand challenges nutritional genomics initiative, and the USDA National Institute of Food and Agriculture (Hatch projects IDA01643 and IDA01566).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.04.004.

References

[1] O.T. Oftedal, The evolution of milk secretion and its ancient origins, Animal. 6 (3) (2012) 355–368.
[2] R.G. Jensen, Miscellaneous factors affecting composition and volume of human and bovine milks, in: R.G. Jensen (Ed.), Handbook of Milk Composition, Academic Press, 1995, pp. 237–271.
[3] L. Bode, Human milk oligosaccharides: every baby needs a sugar mama, Glycobiology. 22 (9) (2012 Sep) 1147–1162.
[4] L. Xie, S.M. Innis, Genetic variants of the *FADS1 FADS2* gene cluster are associated with altered (n-6) and (n-3) essential fatty acids in plasma and erythrocyte phospholipids in women during pregnancy and in breast milk during lactation, J. Nutr. 138 (11) (2008) 2222–2228.
[5] R.M. Erney, W.T. Malone, M.B. Skelding, A.A. Marcon, K.M. Kleman-Leyer, M. L. O'Ryan, et al., Variability of human milk neutral oligosaccharides in a diverse population, J. Pediatr. Gastroenterol. Nutr. 30 (2) (2000) 181–192.
[6] P. Chaturvedi, C.D. Warren, M. Altaye, A.L. Morrow, G. Ruiz-Palacios, L. K. Pickering, D.S. Newburg, Fucosylated human milk oligosaccharides vary between individuals and over the course of lactation, Glycobiology. 11 (2001) 365–372.
[7] M.B. Azad, B. Robertson, F. Atakora, A.B. Becker, P. Subbarao, T.J. Moraes, et al., Human milk oligosaccharide concentrations are associated with multiple fixed and modifiable maternal characteristics, environmental factors, and feeding practices, J. Nutr. 148 (11) (2018 Nov 1) 1733–1742.
[8] M.K. McGuire, C.L. Meehan, M.A. McGuire, J.E. Williams, J. Foster, D.W. Sellen, et al., What's normal? Oligosaccharide concentrations and profiles in milk produced by healthy women vary geographically, Am. J. Clin. Nutr. 105 (2017) 1086–1100.
[9] M.R. Ninonuevo, Y. Park, H. Yin, J. Zhang, R.E. Ward, B.H. Clowers, et al., A strategy for annotating the human milk glycome, J. Agric. Food Chem. 54 (20) (2006) 7471–7480.
[10] E.F. Grollman, V. Ginsburg, Correlation between secretor status and the occurrence of 2′-fucosyllactose in human milk, Biochem. Biophys. Res. Commun. 28 (1967) 50–53.
[11] R.J. Kelly, S. Rouquier, D. Giorgi, G.G. Lennon, J.B. Lowe, Sequence and expression of a candidate for the human secretor blood group alpha(1,2)fucosyltransferase gene (*FUT2*). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype, J. Biol. Chem. 270 (1995) 4640–4649.
[12] A. Kobata, Possible application of milk oligosaccharides for drug development, Chang Gung Med. J. 26 (2003) 621–636.
[13] S. Thurl, M. Munzert, J. Henker, G. Boehm, B. Müller-Werner, J. Jelinek, B. Stahl, Variation of human milk oligosaccharides in relation to milk groups and lactational periods, Br. J. Nutr. 104 (9) (2010) 1261–1271.
[14] J.M. Colston, R. Francois, N. Pisanic, P. Peñataro Yori, B.J.J. McCormick, M. P. Olortegui, et al., Effects of child and maternal histo-blood group antigen status on symptomatic and asymptomatic enteric infections in early childhood, J. Infect. Dis. 220 (1) (2019) 151–162.
[15] D.C. Payne, R.L. Currier, M.A. Staat, L.C. Sahni, R. Selvarangan, N.B. Halasa, et al., Epidemiologic association between *FUT2* secretor status and severe rotavirus

gastroenteritis in children in the United States, JAMA Pediatr. 169 (11) (2015) 1040–1045.
[16] L. Bode, L. Kuhn, H.Y. Kim, L. Hsiao, C. Nissan, M. Sinkala, et al., Human milk oligosaccharide concentration and risk of postnatal transmission of HIV through breastfeeding, Am. J. Clin. Nutr. 96 (4) (2012) 831–839.
[17] S.A. Bien, G.L. Wojcik, N. Zubair, C.R. Gignoux, A.R. Martin, J.M. Kocarnik, et al., Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array, PLoS One 11 (12) (2016), e0167758.
[18] H.M. Kang, J.H. Sul, S.K. Service, N.A. Zaitlen, S.Y. Kong, N.B. Freimer, et al., Variance component model to account for sample structure in genome-wide association studies, Nat. Genet. 42 (4) (2010) 348–354.
[19] V. Segura, B.J. Vihjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, M. Nordborg, An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations, Nat. Genet. 44 (2012) 825–830.
[20] B. Vilhjalmsson, mixmogam, https://github.com/bvilhjal/mixmogam. Commit a40f3e2c95, 2012.
[21] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001) 978–989.
[22] M. Stephens, P. Donnelly, A comparison of Bayesian methods for haplotype reconstruction, Am. J. Hum. Genet. 73 (2003) 1162–1169.
[23] H. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies, Mol. Biol. Evol. 16 (1) (1999) 37–48.
[24] J.W. Leigh, D. Bryant, PopART: full-feature software for haplotype network construction, Methods Ecol. Evol. 6 (9) (2015) 1110–1116.
[25] F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics. 123 (1989) 585–595.
[26] S. Thurl, J. Henker, M. Siegel, K. Tovar, G. Sawatzki, Detection of four human milk groups with respect to Lewis blood group dependent oligosaccharides, Glycoconj. J. 14 (7) (1997) 795–799.
[27] B. Stahl, S. Thurl, J. Henker, M. Siegel, B. Finke, G. Sawatzki, Detection of four human milk groups with respect to Lewis-blood-group-dependent oligosaccharides by serologic and chromatographic analysis, Adv. Exp. Med. Biol. 501 (2001) 299–306.
[28] R. Cabrera-Rubio, C. Kunz, S. Rudloff, I. García-Mantrana, E. Crehuá-Gaudiza, C. Martínez-Costa, M.C. Collado, Association of maternal secretor status and human milk oligosaccharides with milk microbiota: an observational pilot study, J. Pediatr. Gastroenterol. Nutr. 68 (2) (2019) 256–263.
[29] Z.T. Lewis, S.M. Totten, J.T. Smilowitz, M. Popovic, E. Parker, D.G. Lemay, et al., Maternal fucosyltransferase 2 status affects the gut bifidobacterial communities of breastfed infants, Microbiome. 3 (2015) 13.
[30] F. Dupuy, A. Germot, R. Julien, A. Maftah, Structure/function study of Lewis α3- and α3/4-fucosyltransferases: the α1,4 fucosylation requires an aromatic residue in the acceptor-binding domain, Glycobiology. 14 (4) (2004) 347–356.
[31] S.S. Nongmaithem, C.V. Joglekar, G.V. Krishnaveni, S.A. Sahariah, M. Ahmad, S. Ramachandran, et al., GWAS identifies population-specific new regulatory variants in FUT6 associated with plasma B12 concentrations in Indians, Hum. Mol. Genet. 26 (13) (2017) 2551–2564. Erratum in: Hum Mol Genet. 2017;26(13):2589.
[32] M. Costache, A. Cailleau, P. Fernandez-Mateos, R. Oriol, R. Mollicone, Advances in molecular genetics of α-2- and α-3/4-fucosyltransferases, Transfus. Clin. Biol. 4 (1997) 367–382.
[33] Y. Koda, M. Soejima, H. Kimura, The polymorphisms of fucosyltransferases, Legal Med. 3 (2001) 2–14.
[34] M. Soejima, H. Pang, Y. Koda, Genetic variation of *FUT2* in a Ghanaian population: identification of four novel mutations and inference of balancing selection, Ann. Hematol. 86 (2007) 199–204.
[35] M. Soejima, L. Munkhtulga, S. Iwamoto, Y. Koda, Genetic variation of *FUT3* in Ghanaians, Caucasians, and Mongolians, Transfusion. 49 (5) (2009) 959–966.
[36] A. Ferrer-Admetlla, M. Sikora, H. Laayouni, A. Esteve, F. Roubinet, A. Blancher, et al., A natural history of *FUT2* polymorphism in humans, Mol. Biol. Evol. 26 (9) (2009) 1993–2003.
[37] M. Soejima, Y. Koda, Genetic variation of *FUT2* in a Peruvian population: identification of a novel LTR-mediated deletion and characterization of 4 nonsynonymous single-nucleotide polymorphisms, Transfusion. 59 (7) (2019) 2415–2421.
[38] J. Le Pendu, Histo-blood group antigen and human milk oligosaccharides: genetic polymorphism and risk of infectious diseases, Adv. Exp. Med. Biol. 554 (2004) 135–143.
[39] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E. P. Garrison, H.M. Kang, J.O. Korbel, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.
[40] A. Velkova, J.E.L. Diaz, F. Pangilinan, A.M. Molloy, J.L. Mills, B. Shane, et al., The *FUT2* secretor variant p.Trp154Ter influences serum vitamin B12 concentration via holo-haptocorrin, but not holo-transcobalamin, and is associated with haptocorrin glycosylation, Hum. Mol. Genet. 26 (2017) 4975–4988.
[41] S.J. Barton, R. Murray, K.A. Lillycrop, H.M. Inskip, N.C. Harvey, C. Cooper, et al., FUT2 genetic variants and reported respiratory and gastrointestinal illnesses during infancy, J. Infect. Dis. 219 (5) (2019) 836–843.